Acta Numerica (2025), pp. 491–577
doi:10.1017/S0962492925000029

Optimization problems governed by systems of PDEs with uncertainties

Matthias Heinkenschloss
Department of Computational Applied Mathematics and Operations Research, Rice University, 6100 Main St, Houston, TX 77005, USA
E-mail: heinken@rice.edu

Drew P. Kouri
Optimization and Uncertainty Quantification, Sandia National Laboratories, PO Box 5800, Albuquerque, NM 87125, USA
E-mail: dpkouri@sandia.gov

This paper reviews current theoretical and numerical approaches to optimization problems governed by partial differential equations (PDEs) that depend on random variables or random fields. Such problems arise in many engineering, science, economics and societal decision-making tasks. This paper focuses on problems in which the governing PDEs are parametrized by the random variables/fields, and the decisions are made at the beginning and are not revised once uncertainty is revealed. Examples of such problems are presented to motivate the topic of this paper, and to illustrate the impact of different ways to model uncertainty in the formulations of the optimization problem and their impact on the solution. A linear–quadratic elliptic optimal control problem is used to provide a detailed discussion of the set-up for the risk-neutral optimization problem formulation, study the existence and characterization of its solution, and survey numerical methods for computing it. Different ways to model uncertainty in the PDE-constrained optimization problem are surveyed in an abstract setting, including risk measures, distributionally robust optimization formulations, probabilistic functions and chance constraints, and stochastic orders. Furthermore, approximation-based optimization approaches and stochastic methods for the solution of the large-scale PDE-constrained optimization problems under uncertainty are described. Some possible future research directions are outlined.

2020 Mathematics Subject Classification: Primary 49K20, 49K45, 49M41, 90C15
Secondary 49J50, 65K05, 65N30

CONTENTS

1. Introduction

Optimization problems constrained by ordinary differential equations (ODEs) or partial differential equations (PDEs) arise in many decision-making tasks in engineering, science and economics. Examples include flow control and shape optimization (e.g. Gunzburger 2003, Conti *et al.* 2011, Dambrine, Dapogny and Harbrecht 2015, Geihe, Lenz, Rumpf and Schultz 2013, Mohammadi and Pironneau 2004, Schulz and Schillings 2013, Royset, Bonfiglio, Vernengo and Brizzolara 2017), topological design for additively manufactured systems (e.g. Bendsøe and Sigmund 2003, Lazarov, Schevenels and Sigmund 2012*a*,*b*, Zhou, Lazarov and Sigmund 2014) and optimal resource allocation in oil field and gas pipeline operations (e.g. Carter and Rachford Jr 2003, Brouwer and Jansen 2004, Zandvliet, Van Essen and Brouwer 2008, Bangerth *et al.* 2006). In these applications, uncertainty is pervasive, arising from environmental variability, unknown system inputs and coefficients, variability in the execution of the decision, and unverifiable modelling assumptions. Often a decision needs to be made before the uncertainty is revealed, leading to deterministic decision variables, such as the system control or design, that do not *anticipate* the unobserved realization. This class of problems can be broadly modelled as stochastic programs (albeit infinite-dimensional) with underlying differential equation constraints. This paper focuses on the current theoretical and numerical treatment of optimization problems governed by PDEs depending on random variables.[1]

In these optimization problems, the underlying system is described by PDEs that depend on uncertain inputs/coefficients and on deterministic optimization variables that model the user's decision. The PDE is often called the *state equation*, and its solution is called the *state*. Given a realization of the random variables and given an instance of the optimization variables, the underlying PDE can be solved by standard methods for deterministic PDEs. In addition to the governing PDE, we also have a scalar-valued 'cost' or 'loss' function, which depends on the PDE solution and, perhaps, on the optimization variables and additional random variables. Here, cost

[1] More generally, the governing PDE could depend on random fields, but to focus our initial discussion, we assume dependence on random variables.

or loss does not necessarily correspond to a monetary cost or loss, but is a scalar quantification of the under-performance of the system. For example, if the goal is to steer the state to a desired state, the distance between the actual and the desired state would be the cost or loss. Given optimization variables, each realization of random variables gives a realized cost of the system. The goal of the optimization is to determine optimization variables that in some sense minimize the cost over all possible realizations of the random variables. For example, one could determine optimization variables that minimize the expected cost. More generally, one may also have constraints on the states and/or controls. If the constraints depend on the states, they correspondingly depend on the random variables. Consequently, one must determine optimization variables that, in an appropriate sense, satisfy the random constraints. The basic mathematical formulation of such problems, along with their analysis and numerical solution, is the subject of this paper.

The optimization research that is the focus of this paper lies at the intersection of several active research areas: stochastic programming/optimization under uncertainty (see e.g. Birge and Louveaux 2011, Ruszczynski and Shapiro 2003, Shapiro, Dentcheva and Ruszczyński 2014), (deterministic) PDE-constrained optimization (see e.g. Hinze, Pinnau, Ulbrich and Ulbrich 2009) and numerical solution of PDEs with random parameters and uncertainty quantification of PDEs (see e.g. Gunzburger, Webster and Zhang 2014). However, the formulation and efficient solution of the optimization problems in this paper requires techniques that are typically not considered in any of the above areas alone. For example, the existence of solutions of the state equation, its dependence on the random variable and its differentiability properties with respect to the optimization variable are not covered in stochastic programming. The numerical solution of these problems requires discretization of the PDE in space (or space and time) and often also optimization variables, which in control and design applications are functions in infinite-dimensional spaces. For example, by varying the spatial discretization, we can construct discretizations with different levels of fidelity and computational cost that can be integrated with discretization/sampling of the random parameter to achieve substantial speed-ups in the solution of the optimization problems. Such approaches are not considered in stochastic programming. While many such techniques originated in the solution of PDEs with random parameters and uncertainty quantification of PDEs, their use in PDE-constrained optimization under uncertainty requires important modifications to successfully integrate them into the optimization beyond their straightforward use as PDE solvers.

In the optimization problems studied in this paper, the governing PDEs are parametrized by random variables, that is, given a realization of the random variables, the PDE is deterministic. Later, we will use optimization problems governed by stationary PDEs as examples, but there are many optimization problems of this type that are governed by time-dependent PDEs, for example optimization of instationary heating processes, where the uncertainty enters through the time-independent material properties. In the time-dependent case, the PDEs we consider are also

called *random PDEs*; see Smith (2014, Section 4.7). These random PDEs are fundamentally different from stochastic PDEs, which are crucial for modelling other important problems, but require completely different analysis and solution approaches; see e.g. Smith (2014, Section 4.7) and Lord, Powell and Shardlow (2014, Chapters 8 and 9). Moreover, in the optimization problems studied in this paper, the decisions are made at the beginning and are not revised. This is different from multistage stochastic programs studied, for example, in Birge and Louveaux (2011), Pflug and Pichler (2014) and Shapiro *et al.* (2014), or stochastic optimal control problems studied in Fabbri, Gozzi and Świech (2017) and Kushner and Dupuis (2001), where decisions are made in stages or over time, incorporating the information about uncertainties revealed in previous stages or times. Note, however, that the problem class we consider includes optimal control problems governed by time-dependent PDEs and with time-dependent controls, such as the control of an instationary heating process where the uncertainty enters through the time-independent material properties, and once the process is started, no information about the uncertainty is collected to update the controls. Multistage problems are interesting and important, but because of their increased computational cost, they have not yet been considered for more general PDE-constrained problems.

This paper is organized as follows. To motivate the topic of this paper, we present three examples in Section 2 to illustrate the impact of different ways to model the uncertainty in the formulations of the optimization problem and its impact on the solution. In Section 3 we use a linear–quadratic elliptic optimal control problem to discuss in detail the set-up of the risk-neutral optimization problem formulation, study the existence and characterization of its solution, and survey numerical methods for computing this solution. Different ways to model uncertainty in the PDE-constrained optimization problem are surveyed in an abstract setting in Section 4, including risk measures, distributionally robust optimization formulations, probabilistic functions and chance constraints, and stochastic orders. Section 5 describes approximation-based optimization approaches and stochastic methods used to solve PDE-constrained optimization problems. The research on the topic of this paper is rapidly evolving, and some possible extensions and future directions are outlined in Section 6.

Notation

In the following, we summarize basic notation used in this paper.

General vector spaces. For a normed vector space $\mathcal{V}$, we denote the norm on $\mathcal{V}$ by $\|\cdot\|_{\mathcal{V}}$. The dual of a normed vector space $\mathcal{V}$ is denoted by $\mathcal{V}^*$, and $\langle\cdot,\cdot\rangle_{\mathcal{V}^*,\mathcal{V}}$ denotes the duality product between $\mathcal{V}^*$ and $\mathcal{V}$. Specifically, for a bounded linear functional $\ell \in \mathcal{V}^*$, we have $\langle \ell, v\rangle_{\mathcal{V}^*,\mathcal{V}} = \ell(v)$. When $\mathcal{V}$ is a Hilbert space, we denote the associated inner product by $\langle\cdot,\cdot\rangle_{\mathcal{V}}$. Given a Hilbert space $\mathcal{V}$ and a non-empty, closed and convex subset $\mathcal{S} \subset \mathcal{V}$, the projection of $v \in \mathcal{V}$ onto the subset $\mathcal{S}$ is

uniquely defined and is given by

$$\operatorname{proj}_{\mathcal{S}}(v) := \operatorname*{arg\,min}_{v' \in \mathcal{S}} \|v - v'\|_{\mathcal{V}}.$$

See e.g. Bauschke and Combettes (2017, Theorem 3.16). In addition, when $\mathcal{V}$ is a Hilbert space, the normal cone to a subset $\mathcal{S} \subset \mathcal{V}$ at a point $v \in \mathcal{V}$ is

$$N_{\mathcal{S}}(v) := \begin{cases} \{\eta \in \mathcal{V} \mid \langle \eta, v' - v \rangle_{\mathcal{V}} \leq 0 \ \ \forall\, v' \in \mathcal{S}\} & \text{if } v \in \mathcal{S}, \\ \emptyset & \text{if } v \notin \mathcal{S}. \end{cases} \tag{1.1}$$

Function spaces. Given a domain $D \subset \mathbb{R}^n$ with boundary ∂D, we use $L^p(D)$, $p \geq 1$ or $p = \infty$ to denote the usual Lebesgue spaces, and $W^{k,p}(D)$, $p \geq 1$ or $p = \infty$ and $k \in \mathbb{N}$ to denote Sobolev spaces of k-times weakly differentiable functions. In the case $p = 2$, we set $H^k(D) = W^{k,2}(D)$. These spaces are equipped with their usual norms, which are denoted by $\|\cdot\|_{L^p(D)}$, $\|\cdot\|_{W^{k,p}(D)}$ or $\|\cdot\|_{H^k(D)}$. The space $H_0^1(D)$ is the space of all functions in $H^1(D)$ that are zero (in the trace sense) on the boundary ∂D. For more details of these spaces, see e.g. Adams and Fournier (2003) or Brenner and Scott (2008).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, where Ω is the set of outcomes, $\mathcal{F} \subset 2^{\Omega}$ is a σ-algebra of events, and $\mathbb{P}\colon \mathcal{F} \to [0,1]$ is a probability measure. Given a Banach space $\mathcal{Y}$, we define the Bochner spaces

$$L^q_{\mathbb{P}}(\Omega, \mathcal{Y}) := \Big\{ v\colon \Omega \to \mathcal{Y} \;\Big|\; v \text{ is strongly measurable and } \int_{\Omega} \|y(\omega)\|_{\mathcal{Y}}^q \, \mathrm{d}\mathbb{P}(\omega) < \infty \Big\}$$

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

$$L^{\infty}_{\mathbb{P}}(\Omega, \mathcal{Y}) := \Big\{ v\colon \Omega \to \mathcal{Y} \;\Big|\; v \text{ is strongly measurable and } \mathbb{P}\text{-}\operatorname*{ess\,sup}_{\omega \in \Omega} \|y(\omega)\|_{\mathcal{Y}} < \infty \Big\}.$$

When $\mathcal{Y} = \mathbb{R}$, we simplify notation to $L^p_{\mathbb{P}}(\Omega, \mathbb{R}) = L^p_{\mathbb{P}}(\Omega)$.

Probability and statistics. Given a random variable X defined on $(\Omega, \mathcal{F}, \mathbb{P})$, we denote the cumulative distribution function (CDF) associated with X by

$$\Psi_X(t) := \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \leq t\}) = \mathbb{P}(X \leq t).$$

We further denote the associated quantile function by $q_X(\beta) := \Psi_X^{-1}(\beta)$. In addition, we denote the expectation of an integrable random variable $X \in L^1_{\mathbb{P}}(\Omega)$ by

$$\mathbb{E}[X] := \int_{\Omega} X(\omega) \, \mathrm{d}\mathbb{P}(\omega)$$

and its variance, if $X \in L^2_{\mathbb{P}}(\Omega)$, by

$$\mathbb{V}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Often we consider a vector $\boldsymbol{\xi}\colon (\Omega, \mathcal{F}) \to (\Xi, \mathcal{B}_\Xi)$ of random variables, where $\Xi \subset \mathbb{R}^M$ is a non-empty set and $\mathcal{B}_\Xi$ is the Borel σ-algebra on Ξ. The distribution (or law) of $\boldsymbol{\xi}$ on the σ-algebra $\sigma(\boldsymbol{\xi}) = \{\boldsymbol{\xi}^{-1}(B) \mid B \in \mathcal{B}_\Xi\}$ is $\mathbb{P}^{\boldsymbol{\xi}} = \mathbb{P} \circ \boldsymbol{\xi}^{-1}$. We use bold font $\boldsymbol{\xi}$ to denote the vector of random variables and corresponding normal font ξ to denote a realization of $\boldsymbol{\xi}$, i.e. $\xi = \boldsymbol{\xi}(\omega)$ for some $\omega \in \Omega$. Furthermore, we use $\boldsymbol{\xi}^{(m)}$ to denote the mth component of $\boldsymbol{\xi}$, and ξ_i to denote the ith sample of $\boldsymbol{\xi}$.

Shapiro *et al.* (2014, Chapter 7) summarize background material on probability and other topics relevant to stochastic programming.

2. Example optimization problems

This section describes three example optimization problems to illustrate the impact of random parameters in the PDE constraints on the computed solution, and the impact of different ways to model the uncertainty in the formulations of the optimization problem. The deterministic versions of the example problems in this section or related problems are studied in the books by Bendsøe and Sigmund (2003), Gunzburger (2003), Hinze *et al.* (2009), Litvinov (2000), Lions (1971) and Tröltzsch (2010).

2.1. Elliptic optimal control problem

Our first example is an elliptic optimal control problem. The spatial domain is $D = (0,1) \times (0,1)$ with control boundary $\partial D_c = \{0\} \times [0,1]$ and Neumann boundary $\partial D_n = \partial D \setminus \partial D_c$. Given parametrized coefficient functions

$$\kappa(x,\xi) = \begin{cases} \xi^{(1)}, & x \in [0,1] \times [0,0.6), \\ \xi^{(2)}, & x \in [0,1] \times [0.6,1], \end{cases} \qquad c(x) = \begin{pmatrix} 1 \\ \xi^{(3)} \end{pmatrix},$$

$$f(x,\xi) = 20 \exp\left(-\frac{(x_1 - \xi^{(4)})^2}{0.1}\right) \exp\left(-\frac{(x_2 - \xi^{(5)})^2}{0.1}\right),$$

with $\xi = \left(\xi^{(1)}, \xi^{(2)}, \xi^{(3)}, \xi^{(4)}, \xi^{(5)}\right)^\top$ and a function $u \in H^1(\partial D_c)$, we consider the linear elliptic PDE

$$-\nabla \cdot (\kappa(x,\xi)\nabla y(x,\xi)) + c(x,\xi) \cdot \nabla y(x,\xi) = f(x,\xi), \quad x \in D, \tag{2.1a}$$

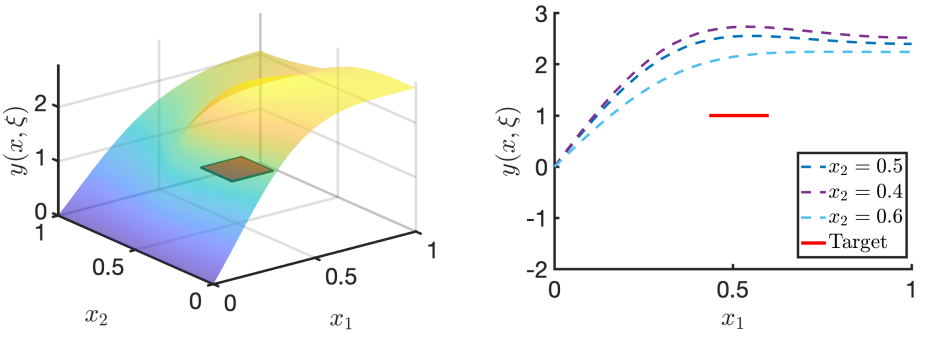
$$y(x,\xi) = u(x), \quad x \in \partial D_c, \tag{2.1b}$$

$$(\kappa(x,\xi)\nabla y(x,\xi)) \cdot n(x) = 0, \quad x \in \partial D_n. \tag{2.1c}$$

For demonstration, we set the components of ξ to be

$$\xi^{(1)} = 0.3, \quad \xi^{(2)} = 0.8, \quad \xi^{(3)} = 0.1, \quad \xi^{(4)} = 0.25, \quad \xi^{(5)} = 0.45. \tag{2.2}$$

Figure 2.1 depicts a finite-element approximation of the solution of (2.1) with $u \equiv 0$ and parameter ξ in (2.2). All results shown in this section are computed using



(a) (b)

Figure 2.1. (a) Solution y of (2.1) with $u \equiv 0$ and parameter ξ in (2.2). (b) Slices $y(x_1, x_2)$ of the solution at $x_2 = 0.4, 0.5, 0.6$.

piecewise linear finite elements on a uniform triangulation obtained by dividing the domain D into squares of size $h \times h$, $h = 1/30$, and then dividing each square into two triangles.

Now assume that we want to compute $u \in H^1(\partial D_c)$ such that the resulting solution $y(u; \cdot, \xi)$ of (2.1) is ideally below a target, here chosen to be 1, in the observation region $D_o = [0.4, 0.6] \times [0.4, 0.6]$ (red square in Figure 2.1(a)). We quantify how much $y(u; \cdot, \xi)$ exceeds the target in D_o using

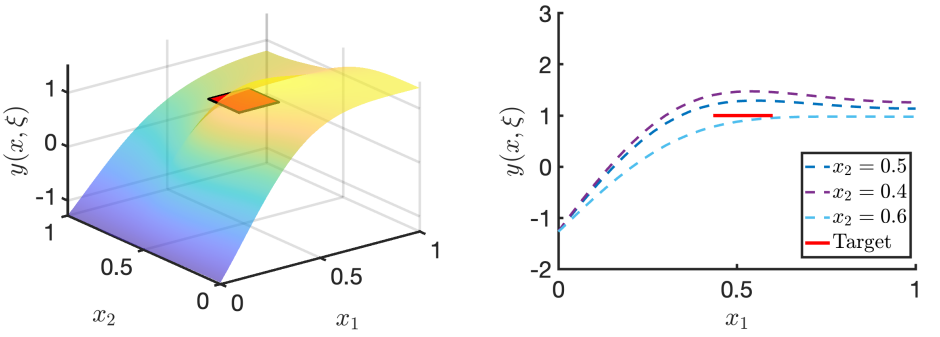
$$\frac{1}{2} \int_{D_o} (y(u; x, \xi) - 1)_+^2 \, \mathrm{d}x,$$

where $z_+ = \max\{z, 0\}$. We add a term $(\alpha/2)\|u\|^2_{H^1(\partial D_c)}$ that penalizes the use of large controls. Here, $\alpha > 0$ is the penalty parameter, and we set $\alpha = 10^{-2}$. This leads to the optimization problem

$$\min_{u \in H^1(\partial D_c)} \frac{1}{2} \int_{D_o} (y(u; x, \xi) - 1)_+^2 \, \mathrm{d}x + \frac{10^{-2}}{2} \|u\|^2_{H^1(\partial D_c)}, \tag{2.3}$$

where $y(u; \cdot, \xi)$ is the solution of (2.1) given u and ξ in (2.2). The function u is referred to as the control, the corresponding solution $y(u; \cdot, \xi)$ of (2.1) is called the state, and the PDE (2.1) is called the state equation. Optimal control problems of the type (2.3) with given parameter ξ and numerical methods for their solution are analysed in the books by Hinze *et al.* (2009) and Tröltzsch (2010), for example.

The deterministic optimal control problem (2.3) has a unique solution u_* and the optimal state $y(u_*; \cdot, \xi)$ is shown in Figure 2.2. The optimal control is shown in Figure 2.6 below. The optimal control moved the state closer to the target in the observation region $D_o = [0.4, 0.6] \times [0.4, 0.6]$ (red square in Figure 2.2(a)). However, this is only true if the control is applied to (2.1) with ξ in (2.2). In this survey we are interested in the case where the parameter is not deterministic but is



(a) (b)

Figure 2.2. (a) The optimal state $y(u_*;\cdot,\xi)$ with parameter ξ in (2.2). (b) Slices $y(u_*;x_1,x_2,\xi)$ of the optimal state at $x_2 = 0.4, 0.5, 0.6$.

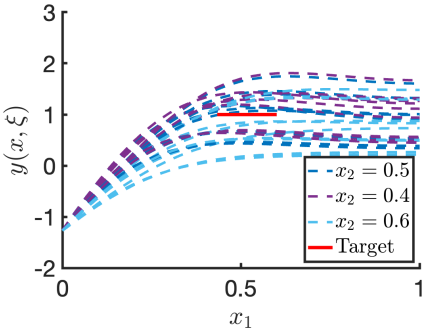


Figure 2.3. Slices at $x_2 = 0.4, 0.5, 0.6$ of the solution $y(u_*;x_1,x_2,\xi)$ of (2.1) with the optimal control u_* computed as the solution of the deterministic problem (2.3) and 10 samples of $\boldsymbol{\xi}$.

a realization of a random variable $\boldsymbol{\xi} = \left(\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \boldsymbol{\xi}^{(3)}, \boldsymbol{\xi}^{(4)}, \boldsymbol{\xi}^{(5)}\right)^\top$, with

$$\begin{aligned} &\boldsymbol{\xi}^{(1)} \sim U(0.2, 0.4), \quad \boldsymbol{\xi}^{(2)} \sim U(0.7, 0.9), \quad \boldsymbol{\xi}^{(3)} \sim U(0.0, 0.2), \\ &\boldsymbol{\xi}^{(4)} \sim U(0.1, 0.4), \quad \boldsymbol{\xi}^{(5)} \sim U(0.3, 0.6). \end{aligned} \tag{2.4}$$

If we compute the optimal control as the solution of the deterministic problem (2.3) with a fixed parameter ξ in (2.2), and then apply this optimal control to the state equation (2.1) with samples of $\boldsymbol{\xi}$, we obtain the results in Figure 2.3. The optimal control computed with fixed ξ can perform poorly if applied with different values of the parameter. Thus the randomness in the parameter must be incorporated into the computation of the control. Note that we are interested in one control that is applied to all possible outcomes of the random variable $\boldsymbol{\xi}$. This is needed when we must decide on the control before the uncertainty is revealed.

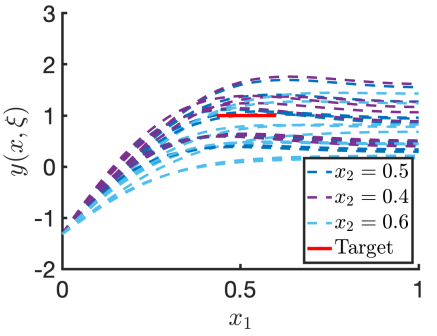


Figure 2.4. Slices at $x_2 = 0.4, 0.5, 0.6$ of the solution $y(u_*; x_1, x_2, \xi)$ of (2.1) with the optimal control u_* computed as the solution of the risk-neutral problem (2.6) and 10 samples of $\boldsymbol{\xi}$.

We consider $\boldsymbol{\xi}$ as in (2.4), and we set $\Xi = [0.2, 0.4] \times [0.7, 0.9] \times [0, 0.2] \times [0.1, 0.4] \times [0.3, 0.6]$ and density $\rho(\xi) \equiv 10^5/72$. Given a control $u \in H^1(\partial D_c)$, we again consider the objective function

$$\frac{1}{2}\int_{D_o} (y(u; x, \xi) - 1)_+^2 \, \mathrm{d}x + \frac{10^{-2}}{2}\|u\|^2_{H^1(\partial D_c)}, \tag{2.5}$$

where $y(u; \cdot, \xi)$ solves (2.1), but now we consider (2.5) for all $\xi \in \Xi$. We want to compute a control u that makes (2.5) small in some sense for all $\xi \in \Xi$. Since now, for given $u \in H^1(\partial D_c)$, (2.5) is a function in $\xi \in \Xi$, we need to quantify its size. One possibility is to take its expected value. This leads to the optimal control problem

$$\min_{u \in H^1(\partial D_c)} \int_\Xi \rho(\xi) \left[\frac{1}{2}\int_{D_o} (y(u; x, \xi) - 1)_+^2 \, \mathrm{d}x\right] \mathrm{d}\xi + \frac{10^{-2}}{2}\|u\|^2_{H^1(\partial D_c)}, \tag{2.6}$$

where $y(u; \cdot, \cdot)$ solves (2.1). The formulation (2.6) using the expected value is also known as the risk-neutral formulation of the optimal control problem. The problem (2.6) has a unique solution. We use a sample average approximation with $N = 100$ Monte Carlo samples to compute an approximation of the optimal control. We will discuss this and other solution methods in later sections. The computed control is shown in Figure 2.6 below. We compute the solution of the state equation (2.1) with u given by the solution u_* of (2.6) and with the same 10 samples of $\boldsymbol{\xi}$ used to generate Figure 2.3. Cross-sections of the solution $y(u_*; x_1, x_2, \xi)$ at $x_2 = 0.4, 0.5, 0.6$ are shown in Figure 2.4.

In this example and for the 10 samples of $\boldsymbol{\xi}$ used, there is little difference between Figures 2.3 and 2.4. In particular, if the optimal control u_* is computed as the solution of the risk-neutral problem (2.6), samples $y(u_*; \cdot, \xi)$ of the corresponding state can be significantly larger than the target.

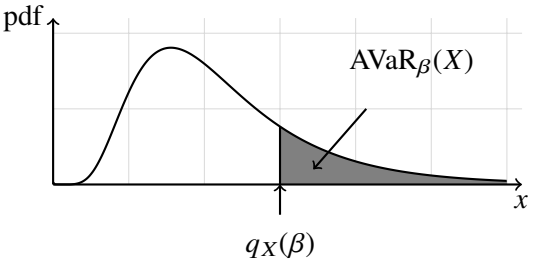


Figure 2.5. Schematic of the average value-at-risk. The label $q_X(\beta)$ denotes the β-quantile or value-at-risk of the random variable X, while the average of the shaded region is AVaR$_\beta(X)$.

In some applications, for example if $y(u;\cdot,\xi)$ represents the concentration of a substance, it may be more harmful if $y(u;\cdot,\xi)$ greatly exceeds the target than if $y(u;\cdot,\xi)$ barely exceeds the target. In these cases it is beneficial to compute a control u_* so that the $(1-\beta)\times 100\%$ of the worst cases are minimized in some sense. This can be quantified, for example, using the *average value-at-risk* (AVaR). For a random variable $X \in L^1_{\mathbb{P}}(\Omega)$ and a confidence level $\beta \in (0,1)$, the average value-at-risk is defined by

$$\text{AVaR}_\beta(X) := \min_{t\in\mathbb{R}} \left\{ t + \frac{1}{1-\beta}\mathbb{E}[\max\{0, X-t\}] \right\}. \tag{2.7}$$

For continuous random variables X, AVaR$_\beta(X)$ is the average of the $(1-\beta)\times 100\%$ largest outcomes of X, thus providing a measure of the distribution tail weight as depicted in Figure 2.5. AVaR and other risk measures will be discussed in Section 4.1.

We apply AVaR to

$$X = \frac{1}{2}\int_{D_o} (y(u;x,\boldsymbol{\xi}) - 1)_+^2 \, \mathrm{d}x.$$

This leads to the optimization problem

$$\min_{u\in H^1(\partial D_c)} \text{AVaR}_\beta\left[\frac{1}{2}\int_{D_o} (y(u;x,\boldsymbol{\xi}) - 1)_+^2 \, \mathrm{d}x\right] + \frac{10^{-2}}{2}\|u\|^2_{H^1(\partial D_c)}. \tag{2.8}$$

We choose $\beta = 0.95$, reflecting that we are concerned with mitigating the highest 5% of outcomes. The optimal control is shown in Figure 2.6. Again, we compute the solution of the state equation (2.1) with u given by the solution u_* of (2.8) and using the same 10 samples of $\boldsymbol{\xi}$ used to generate Figure 2.3. Figure 2.7 depicts cross-sections of the solution $y(u_*;x_1,x_2,\xi)$ at $x_2 = 0.4, 0.5, 0.6$. Comparing Figures 2.4 and 2.7 shows that when AVaR is used to compute the optimal control, most samples $y(u_*;\cdot,\xi)$ of the corresponding state are below or close to the target.

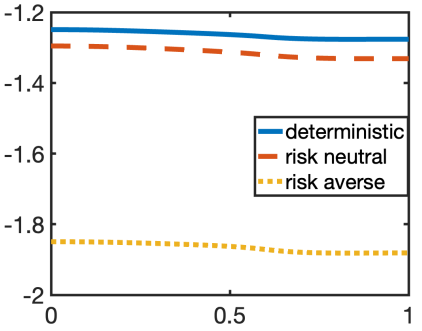


Figure 2.6. Optimal controls computed using the deterministic problem (2.3), the risk-neutral problem (2.6) and the risk-averse problem (2.8).

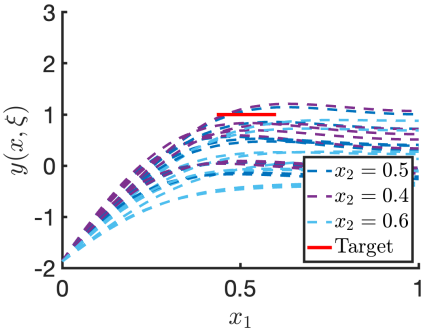


Figure 2.7. Slices at $x_2 = 0.4, 0.5, 0.6$ of the solution $y(u_*; x_1, x_2, \xi)$ of (2.1) with the optimal control u_* computed as the solution of the AVaR problem (2.8) and 10 samples of $\boldsymbol{\xi}$.

2.2. *Optimal control of thermally convected flow*

The example in this section is motivated by the transport process in high-pressure chemical vapour deposition (CVD) reactors, which can be modelled using the Boussinesq flow equations; see e.g. Ito and Ravindran (1998, Section 5.2). In particular, hot wall CVD reactors heat the walls of the reaction chamber, producing more uniform temperature and deposition profiles. In this example, we describe an optimal control problem for the Boussinesq flow equations, drawing inspiration from the control of hot wall CVD reactors.

Consider the domain depicted in Figure 2.8, where $D = (0, 1)^3$, inflow boundary $\Gamma_i = [1/3, 2/3] \times \{(1, 1)\}$, outflow boundary $\Gamma_o = ([0, 1/3] \cup [2/3, 1]) \times \{(1, 1)\}$, reactor bottom $\Gamma_b = [0, 1] \times \{0\} \times [0, 1]$ and side walls (i.e. control boundaries) $\Gamma_c = (\{0, 1\} \times [0, 1]^2) \cup ([0, 1]^2 \times \{0, 1\})$. The Boussinesq flow equations on this

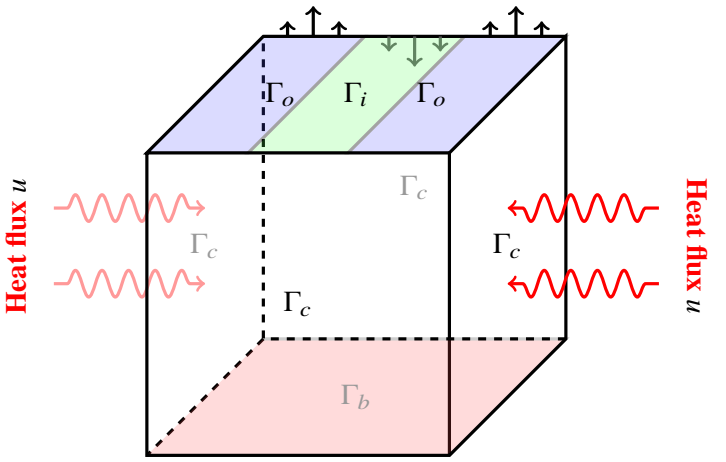


Figure 2.8. The physical domain for the three-dimensional CVD control problem. The blue faces denote the outflow boundaries, the green face denotes the inflow boundary and the red face denotes the substrate boundary. The remaining faces are the control boundaries.

domain are

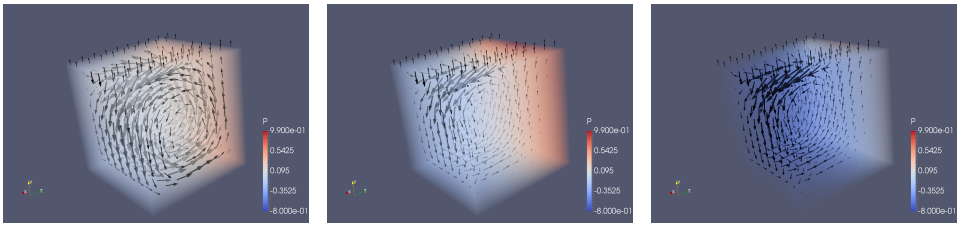
$$-\mu\Delta v + (v\cdot\nabla)v + \nabla p + \eta T g = 0 \quad \text{in } D, \tag{2.9a}$$
$$\nabla\cdot v = 0 \quad \text{in } D, \tag{2.9b}$$
$$-\kappa\Delta T + v\cdot\nabla T = 0 \quad \text{in } D, \tag{2.9c}$$
$$v - v_i = 0, \qquad T = 0 \quad \text{on } \Gamma_i, \tag{2.9d}$$
$$v - v_o = 0, \qquad \kappa\nabla T\cdot n = 0 \quad \text{on } \Gamma_o, \tag{2.9e}$$
$$v = 0, \qquad T - T_b = 0 \quad \text{on } \Gamma_b, \tag{2.9f}$$
$$v = 0, \qquad \kappa\nabla T\cdot n + h(u - T) = 0 \quad \text{on } \Gamma_c, \tag{2.9g}$$

Here, g denotes the acceleration due to gravity, μ is the kinematic viscosity, η is the coefficient of thermal expansion, κ is the thermal conductivity, T_b is the bottom wall temperature, and h is the convection coefficient. Aside from g, these coefficients are often uncertain. The coefficients μ, η and κ satisfy the following relationships:

$$\mu = \frac{1}{\mathrm{Re}}, \quad \eta = \mu^2\mathrm{Gr} \quad \text{and} \quad \kappa = \frac{\mu}{\mathrm{Pr}},$$

where Re is the Reynolds number, Gr is the Grashof number and Pr is the Prandtl number. For this demonstration, v_i and v_o are deterministic, Re, Gr and Pr are random variables, and we model the uncertainty in h and T_b using products of truncated Karhunen–Loève (KL) expansions of Brownian bridge random processes, each with 10 terms, resulting in the 53-dimensional random vector $\boldsymbol{\xi}$.

One possible goal in operating CVD reactors is to promote uniform deposition by minimizing the vorticity within the reactor. To achieve this, we will control the



(a) uncontrolled (b) risk-neutral (c) risk-averse

Figure 2.9. Velocity and pressure fields for the uncontrolled system (a), the risk-neutral controlled system (b), and the risk-averse controlled system (c).

thermal flux on the side walls. Given a risk measure $\mathcal{R}(\cdot)$ such as $\mathcal{R}(\cdot) = \mathbb{E}[\cdot]$ or $\mathcal{R}(\cdot) = \mathrm{AVaR}_\beta(\cdot)$, we formulate the control problem as

$$\min_{u \in \mathcal{U}} \mathcal{R}\left(\frac{1}{2}\int_D (\nabla \times v(u;\cdot,\boldsymbol{\xi}))(x)\ \mathrm{d}x\right) + \frac{\alpha}{2}\int_{\Gamma_c} |u(x)|^2\,\mathrm{d}x,$$

where

$$\begin{aligned} y(u) &= (v(u), p(u), T(u)) \\ &= (v, p, T) \in L^2_{\mathbb{P}}(\Omega, H^1(D))^3 \times L^2_{\mathbb{P}}(\Omega, L^2(D)) \times L^2_{\mathbb{P}}(\Omega, H^1(D)) \end{aligned}$$

solves the weak form of the Boussinesq flow equations (2.9) for almost all realizations of the random vector $\boldsymbol{\xi}$.

We discretized this problem using second-order finite elements for the velocity and temperature fields, and first-order finite elements for the pressure field on a uniform hexahedral mesh, resulting in a discretized system with $O(10^4)$ state variables per sample. We used $N = 80$ Monte Carlo samples. Figure 2.9 depicts the velocity (arrows) and pressure (colours) fields when no control (i.e. $u = 0$) is applied, when the optimal risk neutral (i.e. $\mathcal{R}(\cdot) = \mathbb{E}[\cdot]$) control is applied, and when the optimal risk-averse control is applied. For the risk-averse solution, we employed the entropic risk measure

$$\mathcal{R}(X) = \sigma^{-1} \log \mathbb{E}[\exp(\sigma X)], \quad \sigma > 0,$$

with $\sigma = 2$, which arises from considerations in expected utility theory. Both the risk-neutral and risk-averse controls reduce the variability in the system by a factor of about 2.5 when compared with the uncontrolled system. Furthermore, both risk-neutral and entropic risk significantly reduce the magnitude of the vorticity by approximately 2.3-fold. This fact can also be qualitatively seen in Figure 2.9, noting that the magnitude of the vorticity is significantly smaller in the controlled systems.

2.3. Topology optimization

Recent advances in additive manufacturing have drastically increased design possibilities, giving topology optimization a central role in engineering design. See Bendsøe and Sigmund (2003) for an overview of topology optimization. The archetypal topology optimization problem is to place material in some domain $D \subset \mathbb{R}^d$, $d = 2, 3$, to minimize the compliance, or equivalently maximize the stiffness, of the resulting component whose displacements satisfy the linear elasticity equations

$$-\nabla \cdot (\mathbf{E}(u) : \varepsilon) = f \quad \text{in } D, \tag{2.10a}$$

$$\varepsilon = \tfrac{1}{2}(\nabla y + \nabla y^\top) \quad \text{in } D, \tag{2.10b}$$

$$\varepsilon n = T \quad \text{on } \Gamma_t, \tag{2.10c}$$

$$y = 0 \quad \text{on } \Gamma_d. \tag{2.10d}$$

Here, $u \in L^2(D)$ is the optimization variable, which represents the material distribution (i.e. $u(x) = 0$ signifies no material and $u(x) = 1$ signifies material at x), $f : D \to \mathbb{R}^d$ is a volumetric load, $\Gamma_d \subset \partial D$ denotes the segment of the boundary where the elastic body is fixed, $\Gamma_t = \partial D \setminus \Gamma_d$ denotes the traction boundary, and $T : \Gamma_t \to \mathbb{R}^d$ is a traction load. To obtain a material distribution, u, that is nearly binary and that respects a minimal length scale, it is common to employ a material model like the *solid isotropic material with penalization* (SIMP) model and a density filter such as the volume-preserving Helmholtz filter described in Lazarov and Sigmund (2011). In this setting, the material tensor $\mathbf{E}(u)$ takes the form

$$\mathbf{E}(u) = (\rho_{\min} + (\rho_{\max} - \rho_{\min})\mathbf{F}(u)^3)\mathbf{E}_0,$$

where $0 < \rho_{\min} < \rho_{\max}$, $\mathbf{E}_0$ is a nominal material tensor, and $z = \mathbf{F}(u) \in H^1(D)$ solves the weak form of the elliptic PDE

$$\begin{aligned} -r^2 \Delta z + z &= u \quad \text{in } D, \\ \nabla z \cdot n &= 0 \quad \text{on } \partial D. \end{aligned} \tag{2.11}$$

Here, $r > 0$ dictates the length scale of the optimal design.

Given a volume fraction $v_0 \in (0, 1)$, the compliance minimization problem is formulated as

$$\min_{u \in L^2(D)} \int_D f(x) \cdot y(u; x)\, \mathrm{d}x + \int_{\Gamma_t} T(x) \cdot y(u; x)\, \mathrm{d}x \tag{2.12a}$$

$$\text{subject to} \quad \int_D u(x)\, \mathrm{d}x \le v_0 |D|, \quad 0 \le u \le 1 \ \text{ a.e.} \tag{2.12b}$$

where the displacements $y = y(u) \in H^1_{\Gamma_d}(D)^d$ solve the weak form of the linear elasticity equations (2.10). The objective function in (2.12a) is the compliance of the structure defined by u, and the first constraint in (2.12b) ensures that its volume is no more than $(v_0 \times 100)\%$ of the total volume of the domain D.

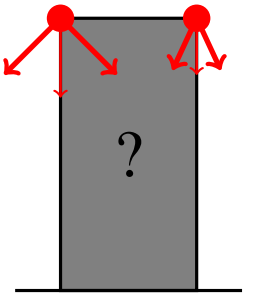


Figure 2.10. Topology optimization problem set-up. The random loads are depicted by the red arrows.

To ensure a reliable design, one must account for uncertainties in (2.12), which may arise from manufacturing variabilities, uncertain external loading scenarios, and unknown internal loads such as residual stresses introduced by the additive manufacturing process. For this example we consider (2.12) with $D = (0,1)\times(0,2)$, $\Gamma_d = [0,1]\times\{0\}$ and $\Gamma_t = [0,1/8]\cup[7/8,1]\times\{2\}$, where the internal load is zero (i.e. $f \equiv 0$) and the traction force is given by

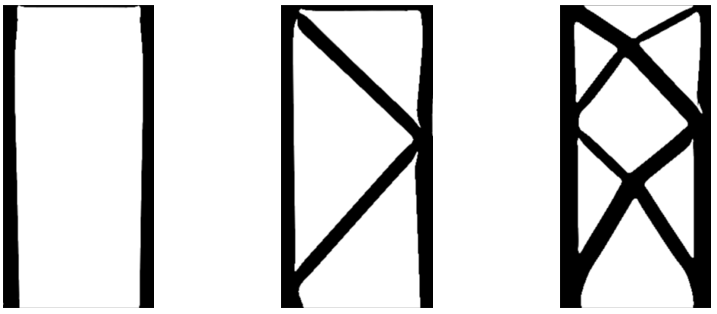
$$T(x,\xi) = \begin{cases} \left(\xi^{(1)}\cos\left(\xi^{(2)}\right), \xi^{(1)}\sin\left(\xi^{(2)}\right)\right)^\top & \text{if } x_1 \in [0,1/8] \text{ and } x_2 = 2, \\ \left(\xi^{(3)}\cos\left(\xi^{(4)}\right), \xi^{(3)}\sin\left(\xi^{(4)}\right)\right)^\top & \text{if } x_1 \in [7/8,1] \text{ and } x_2 = 2, \\ (0,0)^\top & \text{otherwise.} \end{cases}$$

Here, $\boldsymbol{\xi} = \left(\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \boldsymbol{\xi}^{(3)}, \boldsymbol{\xi}^{(4)}\right)$ is a uniformly distributed random vector. The magnitude and angle of the left traction load $\left(\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}\right)$ are uniformly distributed on $[0.25, 1.75]\times[225°, 315°]$, while the magnitude and angle of the right load $\left(\boldsymbol{\xi}^{(3)}, \boldsymbol{\xi}^{(4)}\right)$ are uniformly distributed on $[0.75, 1.25]\times[245°, 295°]$. Figure 2.10 depicts a schematic of the physical domain D and the traction load T. The thick arrows correspond to the upper and lower angle bounds whereas the lengths of the arrows correspond to the maximum magnitudes. As in Sections 2.1 and 2.2, we often ensure that the computed density u produces a reliable design by minimizing a conservative measure of risk such as the AVaR (2.7). Employing AVaR within our topology optimization problem yields the risk-averse optimization problem

$$\min_{u\in L^2(D)} \operatorname{AVaR}_\beta\left(\int_{\Gamma_t} T(x,\boldsymbol{\xi})\cdot y(u;x,\boldsymbol{\xi})\,\mathrm{d}x\right) \tag{2.13a}$$

$$\text{subject to} \quad \int_D u(x)\,\mathrm{d}x \le v_0|D|, \quad 0 \le u \le 1 \ \text{ a.e.} \tag{2.13b}$$

where the displacements $y = y(u) \in L^2_{\mathbb{P}}(\Omega; (H^1_{\Gamma_d}(D))^d)$ solve the weak form of the linear elasticity equations (2.10) for almost all realizations of the uncertain traction load $T(\cdot,\boldsymbol{\xi})$.



(a) deterministic (b) risk-neutral (c) risk-averse

Figure 2.11. Optimal designs for deterministic, risk-neutral and risk-averse topology optimization.

We discretized the elastic displacements y in (2.10) and the filtered density variables $\mathbf{F}(u)$ in (2.11) using continuous piecewise bilinear finite elements defined on a uniform quadrilateral mesh. We further discretize the density variable u using a piecewise constant ansatz on the same mesh. This results in a discretized problem with $O(10^4)$ discretized density variables and $O(10^4)$ state variables per sample. We set the material tensor $\mathbf{E}_0$ to the plain stress tensor with Young's modulus 200 gigapascals and Poisson ratio 0.29, which are common values for steel at room temperature. We further set the volume fraction $v_0 = 0.5$, the filter radius $r = 0.1$, the minimum and maximum densities $\rho_{\min} = 10^{-4}$ and $\rho_{\max} = 1$, respectively, and the AVaR confidence level $\beta = 0.95$. We consider three formulations: the deterministic mean-value formulation in which the random vector $\boldsymbol{\xi}$ is replaced by its mean value, i.e. $\mathbb{E}[\boldsymbol{\xi}] = (1, 270^\circ, 1, 270^\circ)$; the risk-neutral formulation in which AVaR_β in (2.13) is replaced by the expectation $\mathbb{E}$; and the risk-averse formulation (2.13). We approximate the expectation in the risk-neutral and risk-averse formulations using sample average approximation with 1000 Monte Carlo samples. Figure 2.11 depicts the computed densities for the three problem formulations. Intuitively, the deterministic design places material in the direction of the single deterministic load given by the mean value $\mathbb{E}[\boldsymbol{\xi}] = (1, 270^\circ, 1, 270^\circ)$ of the uncertain parameters $\boldsymbol{\xi}$. The risk-neutral and risk-averse designs differ considerably from the deterministic mean-value design, accounting for the various loading scenarios modelled by $T(\cdot, \boldsymbol{\xi})$.

3. Model problem

In this section we use a linear–quadratic elliptic optimal control problem to discuss the risk-neutral optimization problem formulation, study the existence and characterization of its solution, and survey numerical methods for its solution. This model problem class has been studied extensively in the literature. The deterministic versions of this model problem and related problems are studied in the books by

Hinze *et al.* (2009), Lions (1971), Quarteroni (2009) and Tröltzsch (2010). The purpose of this section is to provide an introduction to PDE-constrained optimization under uncertainty, to highlight some issues that need to be addressed when moving from deterministic problems to ones governed by PDEs with random parameters, and to describe solution approaches using a simple model problem and the – arguably easiest to tackle – risk-neutral formulation of the optimization problem.

3.1. Problem with deterministic parameters

3.1.1. State equation

The control and state spaces are the Hilbert spaces $\mathcal{U}$ and $\mathcal{V}$, respectively. Given a control $u \in \mathcal{U}$, the state $y \in \mathcal{V}$ satisfies the variational equation

$$a(y, \varphi) + b(u, \varphi) = \ell(\varphi) \quad \text{for all } \varphi \in \mathcal{V}, \tag{3.1}$$

where

$$a : \mathcal{V} \times \mathcal{V} \to \mathbb{R} \tag{3.2a}$$

is a $\mathcal{V}$-coercive and continuous bilinear form. In other words, there exist $0 < a_{\min} \leq a_{\max}$ such that

$$a_{\min}\|y\|_{\mathcal{V}}^2 \leq a(y, y) \quad \text{and} \quad |a(y, \varphi)| \leq a_{\max}\,\|y\|_{\mathcal{V}}\|\varphi\|_{\mathcal{V}} \quad \text{for all } y, \varphi \in \mathcal{V}. \tag{3.2b}$$

Further,

$$b : \mathcal{U} \times \mathcal{V} \to \mathbb{R} \tag{3.2c}$$

is a continuous bilinear form, that is, there exists $0 < b_{\max}$ such that

$$|b(u, \varphi)| \leq b_{\max}\,\|u\|_{\mathcal{U}}\|\varphi\|_{\mathcal{V}} \quad \text{for all } u \in \mathcal{U}, \varphi \in \mathcal{V}, \tag{3.2d}$$

and $\ell \in \mathcal{V}^*$ is a bounded linear form on $\mathcal{V}$. The variational equation (3.1) can be equivalently written as a linear operator equation,

$$Ay + Bu = \ell \quad \text{in } \mathcal{V}^*, \tag{3.3}$$

where $A \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$ and $B \in \mathcal{L}(\mathcal{U}, \mathcal{V}^*)$ are defined by

$$\langle Ay, \varphi\rangle_{\mathcal{V}^*,\mathcal{V}} = a(y, \varphi), \quad \langle Bu, \varphi\rangle_{\mathcal{V}^*,\mathcal{V}} = b(u, \varphi) \quad \text{for all } y, \varphi \in \mathcal{V}, u \in \mathcal{U}.$$

Because a is $\mathcal{V}$-coercive, the linear operator A is continuously invertible:

$$A^{-1} \in \mathcal{L}(\mathcal{V}^*, \mathcal{V}).$$

The Lax–Milgram theorem (see e.g. Brenner and Scott 2008, Theorem 2.7.7) gives the following existence and uniqueness result.

Theorem 3.1. If $\mathcal{V}$ and $\mathcal{U}$ are Hilbert spaces, $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, $b : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ are bilinear forms satisfying (3.2), and $\ell \in \mathcal{V}^*$, then for every $u \in \mathcal{U}$ the variational equation (3.1) has a unique solution $y(u) \in \mathcal{V}$, and this solution satisfies

$$\|y(u)\|_{\mathcal{V}} \leq a_{\min}^{-1}(\|\ell\|_{\mathcal{V}^*} + b_{\max}\|u\|_{\mathcal{U}}) \quad \text{for all } u \in \mathcal{U}.$$

An example for (3.1) is given by the following elliptic diffusion equation.

Example 3.2. Given a bounded domain $D \subset \mathbb{R}^n$ with boundary ∂D, given functions

$$f \in L^2(D), \quad \kappa \in L^\infty(D), \tag{3.4a}$$

such that

$$\kappa_{\max} \geq \kappa(x) \geq \kappa_{\min} > 0 \quad \text{a.e. in } D, \tag{3.4b}$$

and given $u \in L^2(D)$, consider the elliptic PDE

$$-\nabla \cdot (\kappa(x)\nabla y(x)) = f(x) + u(x), \quad x \in D, \tag{3.5a}$$

$$y(x) = 0, \quad x \in \partial D. \tag{3.5b}$$

The weak form of the problem (3.5) is as follows: find $y \in H_0^1(D)$ such that

$$\int_D \kappa(x)\nabla y(x)\nabla\varphi(x)\,\mathrm{d}x = \int_D (f(x) + u(x))\varphi(x)\,\mathrm{d}x \quad \text{for all } \varphi \in H_0^1(D). \tag{3.6}$$

If we define $\mathcal{V} = H_0^1(D)$ with norm $\|\varphi\|_{\mathcal{V}} = \|\nabla\varphi\|_{L^2(D)}$, $\mathcal{U} = L^2(D)$, the bilinear forms

$$a\colon H_0^1(D) \times H_0^1(D) \to \mathbb{R},$$

$$(y, \varphi) \mapsto a(y, \varphi) = \int_D \kappa(x)\nabla y(x) \cdot \nabla\varphi(x)\,\mathrm{d}x, \tag{3.7a}$$

$$b\colon L^2(D) \times H_0^1(D) \to \mathbb{R},$$

$$(u, \varphi) \mapsto b(u, \varphi) = -\int_D u(x)\varphi(x)\,\mathrm{d}x, \tag{3.7b}$$

and the linear functional

$$\ell\colon H_0^1(D) \to \mathbb{R}, \quad \ell(\varphi) = \int_D f(x)\varphi(x)\,\mathrm{d}x, \tag{3.7c}$$

then (3.6) can be written as (3.1). Under the conditions (3.4), the bilinear form a in (3.7) is $H_0^1(D)$-coercive and continuous on $H_0^1(D) \times H_0^1(D)$,

$$\kappa_{\min}\|y\|^2_{H_0^1(D)} \leq a(y, y), \quad |a(y, \varphi)| \leq \kappa_{\max}\|y\|_{H_0^1(D)}\|\varphi\|_{H_0^1(D)}$$

for all $y, \varphi \in H_0^1(D)$. Moreover, the bilinear form b in (3.7) is continuous on $L^2(D) \times H_0^1(D)$, that is,

$$|b(u, \varphi)| \leq c_D\,\|u\|_{L^2(D)}\|\varphi\|_{H_0^1(D)} \quad \text{for all } u \in L^2(D), \varphi \in H_0^1(D),$$

where c_D is the constant in the Poincaré inequality, and ℓ is a bounded linear functional on $H_0^1(D)$:

$$|\ell(\varphi)| \leq c_D\,\|f\|_{L^2(D)}\|\varphi\|_{H_0^1(D)} \quad \text{for all } \varphi \in H_0^1(D).$$

3.1.2. Optimal control problem

The equations (3.1) are called the state equations, and the solution $y(u)$ is referred to as the state. In the optimal control setting, we want to find a control $u \in \mathcal{U}$ such that a 'cost' or 'loss' of the system modelled by the state equation is minimized. As described in Section 1, cost or loss do not necessarily correspond to a monetary cost or loss, but are a scalar quantification of the under-performance of the system. Note that we write optimization problems as minimization problems. If we want to maximize the performance of the system, then we can achieve this by minimizing the negative of the quantification of performance.

The cost is a scalar quantity that depends on the solution $y(u) \in \mathcal{V}$ of (3.1) and possibly also directly on $u \in \mathcal{U}$, and it is the objective functional (sometimes also called the cost functional) in the optimal control problem. We consider a quadratic objective functional

$$\frac{1}{2}q(y,y) + c(y) + \frac{1}{2}r(u,u), \tag{3.8}$$

where

$$q : \mathcal{V} \times \mathcal{V} \to \mathbb{R} \tag{3.9a}$$

is a symmetric, non-negative, continuous bilinear form. In other words, $q(y,\varphi) = q(\varphi,y)$ for all $y, \varphi \in \mathcal{V}$,

$$q(y,y) \geq 0 \quad \text{for all } y \in \mathcal{V}, \tag{3.9b}$$

and there exists $0 < q_{\max}$ such that

$$|q(y,\varphi)| \leq q_{\max} \|y\|_{\mathcal{V}} \|\varphi\|_{\mathcal{V}} \quad \text{for all } y, \varphi \in \mathcal{V}. \tag{3.9c}$$

Further,

$$r : \mathcal{U} \times \mathcal{U} \to \mathbb{R} \tag{3.9d}$$

is a symmetric, $\mathcal{U}$-coercive and continuous bilinear form, i.e. $r(u,\psi) = r(\psi,u)$ for all $u, \psi \in \mathcal{U}$, and there exists $0 < r_{\min} \leq r_{\max}$ such that

$$r_{\min}\|u\|_{\mathcal{U}}^2 \leq r(u,u) \quad \text{and} \quad |r(u,\psi)| \leq r_{\max} \|u\|_{\mathcal{U}} \|\psi\|_{\mathcal{U}} \quad \text{for all } u, \psi \in \mathcal{U}, \tag{3.9e}$$

and c is a bounded linear form on $\mathcal{V}$:

$$c \in \mathcal{V}^*. \tag{3.9f}$$

Our optimal control problem is given by

$$\min_{u \in \mathcal{U}} \frac{1}{2}q(y(u), y(u)) + c(y(u)) + \frac{1}{2}r(u,u), \tag{3.10}$$

where $y(u) \in \mathcal{V}$ is the solution of the state equation (3.1) given $u \in \mathcal{U}$.

The problem (3.10) is a convex, elliptic, linear–quadratic optimal control problem, and such problems are analysed in the books by Lions (1971), Hinze *et al.*

(2009), Quarteroni (2009) and Tröltzsch (2010), for example. The objective function in (3.10) is Fréchet-differentiable, and the Fréchet derivative can be computed via the adjoint equation approach, as detailed in the following result.

Theorem 3.3. If $\mathcal{V}$ and $\mathcal{U}$ are Hilbert spaces, $a: \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, $b: \mathcal{U} \times \mathcal{V} \to \mathbb{R}$, $q: \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, $r: \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ are bilinear forms satisfying (3.2) and (3.9), and $\ell, c \in \mathcal{V}^*$, then the objective function in (3.10),

$$f: \mathcal{U} \to \mathbb{R}, \quad u \mapsto f(u) := \frac{1}{2} q(y(u), y(u)) + c(y(u)) + \frac{1}{2} r(u, u), \tag{3.11}$$

is Fréchet-differentiable and the Fréchet derivative applied to δu is

$$f'(u)\delta u = r(u, \delta u) + b(\delta u, \lambda), \tag{3.12a}$$

where $\lambda \in \mathcal{V}$ solves

$$a(\varphi, \lambda) + q(u, \varphi) = -c(\varphi) \quad \text{for all } \varphi \in \mathcal{V}. \tag{3.12b}$$

Proof. See e.g. Section 1.6.3 in Hinze *et al.* (2009). □

The following result addresses existence, uniqueness and characterization of the solution of (3.10).

Theorem 3.4. If $\mathcal{V}$ and $\mathcal{U}$ are Hilbert spaces, $a: \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, $b: \mathcal{U} \times \mathcal{V} \to \mathbb{R}$, $q: \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, $r: \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ are bilinear forms satisfying (3.2) and (3.9), and $\ell, c \in \mathcal{V}^*$, then the optimal control problem (3.10) has a unique solution $u \in \mathcal{U}$. Furthermore, $u \in \mathcal{U}$ solves (3.8) if and only if there exist $y \in \mathcal{V}$ and $\lambda \in \mathcal{V}$ such that y, u, λ solve

$$a(\varphi, \lambda) + q(y, \varphi) = -c(\varphi) \quad \text{for all } \varphi \in \mathcal{V}, \tag{3.13a}$$

$$r(u, \psi) + b(\psi, \lambda) = 0 \quad \text{for all } \psi \in \mathcal{U}, \tag{3.13b}$$

$$a(y, \varphi) + b(u, \varphi) = \ell(\varphi) \quad \text{for all } \varphi \in \mathcal{V}. \tag{3.13c}$$

Proof. See e.g. Section 1.51 in Hinze *et al.* (2009). □

Because (3.10) is a convex optimization problem, the optimality conditions (3.13) are necessary and sufficient, and because (3.10) is a linear–quadratic optimal control problem, the optimality conditions (3.13) are a system of linear variational equations. The equation (3.13a) is called the adjoint equation, and its solution λ is called the adjoint. The equation (3.13c) is just the state equation and its solution y is the state.

The following is an example of an optimal control problem governed by the elliptic diffusion equation in Example 3.2.

Example 3.5. Given a domain $D_o \subset D$ and a desired state $\widehat{y} \in L^2(D_0)$, we want to find $u \in L^2(D)$ such that $y(u; \cdot)$ is close to $\widehat{y}$ in the L^2 sense, that is, for our example the cost is $\frac{1}{2} \int_{D_o} (y(u; x) - \widehat{y}(x))^2 \, \mathrm{d}x$. Rather than just minimizing the deviation of the state from the desired state, we add a penalty term with parameter

$\alpha > 0$ for the control. Thus our optimal control problem is given by

$$\min_{u\in L^2(D)} \frac{1}{2}\int_{D_o} (y(u;x) - \widehat{y}(x))^2 \,\mathrm{d}x + \frac{\alpha}{2}\int_D u(x)^2 \,\mathrm{d}x, \tag{3.14}$$

where $y(u;\cdot) \in H_0^1(D)$ is the solution of (3.6) given $u \in L^2(D)$.

If we drop the constant

$$\frac{1}{2}\int_{D_o} \widehat{y}(x)^2 \,\mathrm{d}x,$$

the objective functional in (3.14) is a special case of (3.8) with

$$q(y,\varphi) = \int_{D_o} y(x)\varphi(x)\,\mathrm{d}x,$$
$$c(y) = -\int_{D_o} y(x)\widehat{y}(x)\,\mathrm{d}x,$$
$$r(u,\psi) = \int_D u(x)\psi(x)\,\mathrm{d}x.$$

Application of Theorem 3.4 gives the following result on the existence, uniqueness and characterization of the solution of (3.14).

Corollary 3.6. If equations (3.7) are satisfied, the optimal control problem (3.14) has a unique solution $u \in L^2(D)$. Furthermore, $u \in L^2(D)$ solves (3.14) if and only if there exist $y \in H_0^1(D)$ and $\lambda \in H_0^1(D)$ such that y, u, λ solve

$$-\nabla\cdot(\kappa(x)\nabla\lambda(x)) = -(y(x) - \widehat{y}(x)), \quad x \in D, \tag{3.15a}$$
$$\lambda(x) = 0, \quad x \in \partial D, \tag{3.15b}$$
$$\alpha u(x) - \lambda(x) = 0, \quad x \in D, \tag{3.15c}$$
$$-\nabla\cdot(\kappa(x)\nabla y(x)) = f(x) + u(x), \quad x \in D, \tag{3.15d}$$
$$y(x) = 0, \quad x \in \partial D. \tag{3.15e}$$

3.2. Problem with random parameters

In applications, material parameters or forces may not be known, but are subject to random variations. In this case, we must compute a control that, in some sense, minimizes the cost of the system over all variations in material parameters or forces. To find a suitable formulation for such an optimal control problem, we must first extend the state equation (3.1) to allow for randomness in the parameters.

3.2.1. State equation

Let $\mathcal{V}$ and $\mathcal{U}$ be Hilbert spaces and let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. Furthermore, let $\mathcal{B}$ be the Borel σ-algebra on $\mathbb{R}$. Given a realization $\omega \in \Omega$, we consider

$$a(y,\varphi,\omega) + b(u,\varphi,\omega) = \ell(\varphi,\omega) \quad \text{for all } \varphi \in \mathcal{V}. \tag{3.16}$$

In (3.16),

$$a(\cdot,\cdot,\omega)\colon \mathcal{V}\times\mathcal{V}\to\mathbb{R},\quad b(\cdot,\cdot,\omega)\colon \mathcal{U}\times\mathcal{V}\to\mathbb{R} \tag{3.17a}$$

are bilinear forms and

$$\ell(\cdot,\omega)\in\mathcal{V}^*. \tag{3.17b}$$

We assume that

$$\begin{aligned}&\text{for each } y,\varphi\in\mathcal{V},\ u\in\mathcal{U},\text{ the functions}\\ &a(y,\varphi,\cdot),\ b(u,\varphi,\cdot)\colon\Omega\to\mathbb{R}\text{ are }(\mathcal{F},\mathcal{B})\text{-measurable.}\end{aligned} \tag{3.17c}$$

Moreover, we assume that there exist measurable functions

$$0<a_{\min}(\omega)\le a_{\max}(\omega)\quad\text{and}\quad 0<b_{\max}(\omega) \tag{3.17d}$$

such that for a.a. $\omega\in\Omega$,

$$a(y,y,\omega)\ge a_{\min}(\omega)\|y\|_{\mathcal{V}}^2 \qquad \text{for all } y\in\mathcal{V}, \tag{3.17e}$$

$$|a(y,\varphi,\omega)|\le a_{\max}(\omega)\,\|y\|_{\mathcal{V}}\|\varphi\|_{\mathcal{V}} \qquad \text{for all } y,\varphi\in\mathcal{V}, \tag{3.17f}$$

$$|b(u,\varphi,\omega)|\le b_{\max}(\omega)\,\|u\|_{\mathcal{U}}\|\varphi\|_{\mathcal{V}} \qquad \text{for all } u\in\mathcal{U},\varphi\in\mathcal{V}. \tag{3.17g}$$

Because of their measurability properties (3.17c), and their continuity properties (3.17f) and (3.17g), the function $a\colon\mathcal{V}\times\mathcal{V}\times\Omega\to\mathbb{R}$ is a Carathéodory function, and for every $u\in\mathcal{U}$ the function $b(u,\cdot,\cdot)\colon\mathcal{V}\times\Omega\to\mathbb{R}$ is a Carathéodory function.

Analogously to the deterministic case, the parametrized variational equation (3.16) can be equivalently written as a parametrized linear operator equation

$$A(\omega)y(\omega)+B(\omega)u=\ell(\omega)\quad\text{in }\mathcal{V}^*, \tag{3.18}$$

where for a.a. $\omega\in\Omega$ the operators $A(\omega)\in\mathcal{L}(\mathcal{V},\mathcal{V}^*)$ and $B(\omega)\in\mathcal{L}(\mathcal{U},\mathcal{V}^*)$ are defined by

$$\langle A(\omega)y,\varphi\rangle_{\mathcal{V}^*,\mathcal{V}}=a(y,\varphi,\omega),\quad \langle B(\omega)u,\varphi\rangle_{\mathcal{V}^*,\mathcal{V}}=b(u,\varphi,\omega)$$

for all $y,\varphi\in\mathcal{V},u\in\mathcal{U}$. Because of (3.17d),

$$A(\omega)^{-1}\in\mathcal{L}(\mathcal{V}^*,\mathcal{V})\quad\text{for a.a. }\omega\in\Omega.$$

If (3.17) is satisfied, Theorem 3.1 guarantees that for every $u\in\mathcal{U}$ and for a.a. $\omega\in\Omega$ the parametrized variational equation (3.16) has a unique solution $y(u;\omega)\in\mathcal{V}$. However, we need measurability and integrability properties of the function $y(u;\cdot)\colon\Omega\to\mathcal{V}$, which will require additional conditions. There are two possible avenues: we can consider variational equations in the Hilbert space $L^2_{\mathbb{P}}(\Omega,\mathcal{V})$, or we can consider the parametrized equation (3.16). We will consider both approaches below. In either case, we assume that $\mathcal{V}$ is a separable Hilbert space. In this case, the Carathéodory functions $a\colon\mathcal{V}\times\mathcal{V}\times\Omega\to\mathbb{R}$ and $b(u,\cdot,\cdot)\colon\mathcal{V}\times\Omega\to\mathbb{R}$, $u\in\mathcal{U}$, are jointly measurable; see e.g. Aliprantis and Border (2006, Lemma 4.51).

To establish a variational equation in the Hilbert space $L^2_{\mathbb{P}}(\Omega, \mathcal{V})$, we assume that

$$\operatorname*{ess\,inf}_{\omega\in\Omega} a_{\min}(\omega) > 0, \tag{3.19a}$$

$$a_{\max} \in L^\infty_{\mathbb{P}}(\Omega), \quad b_{\max} \in L^2_{\mathbb{P}}(\Omega), \tag{3.19b}$$

and

$$\ell \in L^2_{\mathbb{P}}(\Omega, \mathcal{V}^*). \tag{3.19c}$$

Given a control $u \in \mathcal{U}$, we seek the solution $y \in L^2_{\mathbb{P}}(\Omega, \mathcal{V})$ of the variational equation

$$\begin{aligned}\int_\Omega a(y(\omega), \varphi(\omega), \omega)\,\mathrm{d}\mathbb{P}(\omega) + \int_\Omega b(u, \varphi(\omega), \omega)\,\mathrm{d}\mathbb{P}(\omega)\\ = \int_\Omega \ell(\varphi(\omega), \omega)\,\mathrm{d}\mathbb{P}(\omega) \quad \text{for all } \varphi \in L^2_{\mathbb{P}}(\Omega, \mathcal{V}).\end{aligned} \tag{3.20}$$

Under the conditions (3.17) and (3.19),

$$L^2_{\mathbb{P}}(\Omega, \mathcal{V}) \times L^2_{\mathbb{P}}(\Omega, \mathcal{V}) \ni (y, \varphi) \mapsto \int_\Omega a(y(\omega), \varphi(\omega), \omega)\,\mathrm{d}\mathbb{P}(\omega)$$

is a bounded, $L^2_{\mathbb{P}}(\Omega, \mathcal{V})$-coercive bilinear form, and

$$\mathcal{U} \times L^2_{\mathbb{P}}(\Omega, \mathcal{V}) \ni (u, \varphi) \mapsto \int_\Omega b(u, \varphi(\omega), \omega)\,\mathrm{d}\mathbb{P}(\omega)$$

is a bounded bilinear form.

We can again apply the Lax–Milgram theorem (see e.g. Brenner and Scott 2008, Theorem 2.7.7), now applied with the Hilbert space $L^2_{\mathbb{P}}(\Omega, \mathcal{V})$, to establish the following existence and uniqueness result, which is analogous to Theorem 3.1.

Theorem 3.7. If $\mathcal{V}$ is a separable Hilbert space, if $\mathcal{U}$ is a Hilbert space, and if the bilinear forms a, b and linear form ℓ satisfy (3.17) and (3.19), then for every $u \in \mathcal{U}$ the variational equation (3.20) has a unique solution $y(u) \in L^2_{\mathbb{P}}(\Omega, \mathcal{V})$, and this solution satisfies

$$\|y(u)\|_{L^2_{\mathbb{P}}(\Omega,\mathcal{V})} \le \frac{1}{\operatorname{ess\,inf}_{\omega\in\Omega} a_{\min}} \left(\|\ell\|_{L^2_{\mathbb{P}}(\Omega,\mathcal{V}^*)} + \|b_{\max}\|_{L^\infty_{\mathbb{P}}(\Omega)}\|u\|_{\mathcal{U}}\right) \quad \text{for all } u \in \mathcal{U}.$$

Under the assumptions of Theorem 3.7, the solution $y(u) \in L^2_{\mathbb{P}}(\Omega, \mathcal{V})$ of (3.20) also satisfies (3.16) for a.a. $\omega \in \Omega$, that is,

$$a(y(u;\omega), \varphi, \omega) + b(u, \varphi, \omega) = \ell(\varphi, \omega) \quad \text{for all } \varphi \in \mathcal{V}, \text{ a.a. } \omega \in \Omega. \tag{3.21}$$

This can be seen as follows. Selecting test functions $\varphi(\omega) = \phi\,\chi_E(\omega)$, where $\phi \in \mathcal{V}$ and χ_E is the indicator function for the set $E \in \mathcal{F}$, the variational equality (3.20)

becomes

$$\int_E a(y(u;\omega),\phi,\omega)\,\mathrm{d}\mathbb{P}(\omega) + \int_E b(u,\phi,\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int_E \ell(\phi,\omega)\,\mathrm{d}\mathbb{P}(\omega) \quad \text{for all } \phi \in \mathcal{V}, E \in \mathcal{F}. \tag{3.22}$$

If (3.17) and (3.19) hold, the integrands in (3.22) are in $L^1_{\mathbb{P}}(\Omega)$ for any given $\phi \in \mathcal{V}$, and therefore (3.22) implies that $y(u;\omega)$ satisfies (3.21) (see e.g. Folland 1999, Proposition 2.23).

The existence result in Theorem 3.7 requires that $a_{\min}$ is uniformly bounded away from zero in the sense of (3.19a). This assumption is too strong in some applications. For example, for a parametrized system (3.16) resulting from an elliptic PDE with log-normally distributed diffusion coefficient, this assumption is not valid. See e.g. Babuška, Nobile and Tempone (2007) or Charrier, Scheichl and Teckentrup (2013), and Example 3.9 below.

As mentioned earlier, if (3.17) is satisfied, Theorem 3.1 guarantees that for every $u \in \mathcal{U}$ and for a.a. $\omega \in \Omega$, the parametrized variational equation (3.16) has a unique solution $y(u;\omega) \in \mathcal{V}$, and that this solution satisfies

$$\|y(u,\omega)\|_{\mathcal{V}} \le a_{\min}(\omega)^{-1}(\|\ell(\omega)\|_{\mathcal{V}^*} + b_{\max}(\omega)\|u\|_{\mathcal{U}}) \quad \text{for all } u \in \mathcal{U}. \tag{3.23}$$

To prove that $\omega \mapsto y(u,\omega)$ is measurable, we show that it is the limit of measurable functions. Let $u \in \mathcal{U}$ be given. To simplify the notation, we temporarily drop the dependence of the solution on u and let $y(\omega) \in \mathcal{V}$ denote the solution of (3.16).

Assume that (3.17), (3.19b) and (3.19c) hold. We consider (3.20) with the bilinear form a replaced by

$$a_n(\cdot,\cdot,\omega) \coloneqq a(\cdot,\cdot,\omega) + \frac{1}{n}\langle\cdot,\cdot\rangle_{\mathcal{V}} \colon \mathcal{V}\times\mathcal{V} \to \mathbb{R}.$$

For all $\omega \in \Omega$, this bilinear form satisfies

$$a_n(y,y,\omega) \ge (a_{\min}(\omega) + 1/n)\|y\|^2_{\mathcal{V}} \qquad \text{for all } y \in \mathcal{V}, \tag{3.24a}$$

$$|a_n(y,\varphi,\omega)| \le (a_{\max}(\omega) + 1/n)\,\|y\|_{\mathcal{V}}\|\varphi\|_{\mathcal{V}} \quad \text{for all } y,\varphi \in \mathcal{V}. \tag{3.24b}$$

Because $\operatorname{ess\,inf}_{\omega\in\Omega}(a_{\min}(\omega) + 1/n) \ge 1/n > 0$, Theorem 3.7 guarantees the existence of a unique solution $y_n \in L^2_{\mathbb{P}}(\Omega,\mathcal{V})$ of the variational equation (3.20) with a replaced by a_n. Moreover, this solution satisfies

$$a_n(y_n(\omega),\varphi,\omega) + b(u,\varphi,\omega) = \ell(\varphi,\omega) \quad \text{for all } \varphi \in \mathcal{V}, \text{ a.a. } \omega \in \Omega. \tag{3.25}$$

For a.a. ω the solution $y_n(\omega)$ of (3.25) and the solution $y(\omega)$ of (3.16) satisfy

$$\begin{aligned} a_n(y_n(\omega) - y(\omega),\varphi,\omega) &= a_n(y_n(\omega),\varphi,\omega) - a_n(y(\omega),\varphi,\omega) \\ &= \ell(\varphi,\omega) - b(u,\varphi,\omega) - a_n(y(\omega),\varphi,\omega) \\ &= a(y(\omega),\varphi,\omega) - a_n(y(\omega),\varphi,\omega) \\ &= n^{-1}\langle y(\omega),\varphi\rangle_{\mathcal{V}}. \end{aligned}$$

Inserting $\varphi = y_n(\omega) - y(\omega)$ and using (3.24a) implies

$$\|y_n(\omega) - y(\omega)\|_{\mathcal{V}} \le \frac{1}{n}(a_{\min}(\omega) + 1/n)^{-1}\|y(\omega)\|_{\mathcal{V}}.$$

Hence $\lim_{n\to\infty} y_n(\omega) = y(\omega)$ for a.a. ω, which implies that y is measurable (see e.g. Aliprantis and Border 2006, Lemma 4.29).

As in Babuška *et al.* (2007, Lemma 1.2), we can now use the measurability of the solution of $y(u;\omega)$ (3.16) and the bound (3.23) to establish a moment estimate.

Theorem 3.8. If $\mathcal{V}$ is a separable Hilbert space, if $\mathcal{U}$ is a Hilbert space, and if the bilinear forms a, b and linear form ℓ satisfy (3.17) (3.19b) and (3.19c), then for every $u \in \mathcal{U}$ the pointwise variational equation (3.16) has a unique solution $y(u;\omega)$, and $\omega \mapsto y(u;\omega)$ is measurable.

Let $p, q \ge 1$ with $1/p + 1/q = 1$ and $k \in \mathbb{N}$. If in addition $\ell \in L^{kp}_{\mathbb{P}}(\Omega, \mathcal{V}^*)$, $b_{\max} \in L^{kp}_{\mathbb{P}}(\Omega)$ and $a_{\min}^{-1} \in L^{kq}_{\mathbb{P}}(\Omega)$, then the unique solution of (3.16) satisfies $y(u;\cdot) \in L^k_{\mathbb{P}}(\Omega, \mathcal{V})$, and

$$\int_\Omega \|y(u;\omega)\|^k_{\mathcal{V}}\,\mathrm{d}\mathbb{P}(\omega) \le \left(\int_\Omega \left(\frac{1}{a_{\min}(\omega)}\right)^{kq} \mathrm{d}\mathbb{P}(\omega)\right)^{1/q} \times \left(\int_\Omega (\|\ell(\omega)\|_{\mathcal{V}^*} + b_{\max}(\omega)\|u\|_{\mathcal{U}})^{kp}\,\mathrm{d}\mathbb{P}(\omega)\right)^{1/p}.$$

Theorem 3.7 is a special case of Theorem 3.8 if (3.19a) holds and we choose $p = 1$, $q = \infty$, $k = 2$.

Example 3.9. An example of (3.16) is the elliptic PDE

$$-\nabla\cdot(\kappa(x,\omega)\nabla y(x,\omega)) = f(x,\omega) + u(x), \quad x \in D,\ \omega \in \Omega, \tag{3.26a}$$

$$y(x,\omega) = 0, \quad x \in \partial D,\ \omega \in \Omega, \tag{3.26b}$$

where

$$u \in L^2(D), \quad f \in L^2_{\mathbb{P}}(\Omega, L^2(D)), \quad \kappa \in L^\infty_{\mathbb{P}}(\Omega, L^\infty(D)), \tag{3.27a}$$

and there exist measurable functions $\kappa_{\min}$ and $\kappa_{\max}$ such that

$$\kappa_{\max}(\omega) \ge \kappa(x,\omega) \ge \kappa_{\min}(\omega) > 0, \quad x \in D, \text{ a.a. } \omega \in \Omega. \tag{3.27b}$$

Given a realization $\omega \in \Omega$, the weak form of the problem (3.26) is as follows: find $y(\cdot,\omega) \in H^1_0(D)$ such that

$$\int_D \kappa(x,\omega)\nabla y(x,\omega)\nabla\varphi(x)\,\mathrm{d}x \tag{3.28}$$

$$= \int_D (f(x,\omega) + u(x))\varphi(x)\,\mathrm{d}x \quad \text{for all } \varphi \in H^1_0(D). \tag{3.29}$$

The parametrized weak form (3.28) is a special case of (3.16) with $\mathcal{V} = H_0^1(D)$, $\mathcal{U} = L^2(D)$,

$$a(\cdot,\cdot,\omega)\colon H_0^1(D)\times H_0^1(D)\to\mathbb{R},$$

$$(y,\varphi)\mapsto a(y,\varphi,\omega)=\int_D \kappa(x,\omega)\nabla y(x)\cdot\nabla\varphi(x)\,\mathrm{d}x, \quad (3.30a)$$

$$b(\cdot,\cdot,\omega)\colon L^2(D)\times H_0^1(D)\to\mathbb{R},$$

$$(u,\varphi)\mapsto b(u,\varphi,\omega)=-\int_D u(x)\varphi(x)\,\mathrm{d}x \quad (3.30b)$$

and

$$\ell(\cdot,\omega)\colon H_0^1(D)\to\mathbb{R},\quad \ell(\varphi,\omega)=\int_D f(x,\omega)\varphi(x)\,\mathrm{d}x,\quad \varphi\in H_0^1(D). \quad (3.30c)$$

In this example, the bilinear form (3.30) does not depend on ω, but in general it can.

Under the conditions (3.27), the bilinear forms $a(\cdot,\cdot,\omega)$, $b(\cdot,\cdot,\omega)$ in (3.30) satisfy (3.17e), (3.17f) and (3.17g) with $a_{\min}=\kappa_{\min}$, $a_{\max}=\kappa_{\max}$ and $b_{\max}=c_D$, where c_D is the constant in the Poincaré inequality, and $\ell(\cdot,\omega)\in H^{-1}(D)$, $\ell\in L^2_{\mathbb{P}}(\Omega,H^{-1}(D))$.

Now consider a lognormally distributed κ,

$$\kappa(x,\omega)=\exp\left(\sum_{m=1}^{M} b^{(m)}(x)\,\boldsymbol{\xi}^{(m)}(\omega)\right), \quad (3.31)$$

where $b^{(m)}\in L^\infty(D)$, $m=1,\dots,M$ and $\boldsymbol{\xi}^{(m)}\sim N(0,1)$, $m=1,\dots,M$, are independent and identically distributed (i.i.d.). With this diffusivity, (3.27) holds, but the bilinear form (3.30a) does not satisfy (3.19a). However, as in Example 1 in Babuška *et al.* (2007), one can select the lower bound

$$\kappa_{\min}(\omega)=\exp\left(-\sum_{m=1}^{M}\|b^{(m)}\|_{L^\infty(D)}\,|\boldsymbol{\xi}^{(m)}(\omega)|\right),$$

which for all $k\in\mathbb{N}$ and $1<q<\infty$ satisfies $\kappa_{\min}^{-1}\in L^{kq}_{\mathbb{P}}(\Omega)$. Hence, if $f\in L^{k(1+\epsilon)}_{\mathbb{P}}(\Omega,L^2(D))$ for some $\epsilon>0$, Theorem 3.8 guarantees the existence of a unique solution $y(u;\cdot)\in L^k_{\mathbb{P}}(\Omega,H_0^1(\Omega))$ of (3.28).

3.2.2. *Optimal control problem*

Now we turn to the extension of the optimal control problem (3.10). Let r again be given as in (3.9), but now q and c can depend on ω. For $\omega\in\Omega$, let

$$q(\cdot,\cdot,\omega)\colon \mathcal{V}\times\mathcal{V}\to\mathbb{R}\quad\text{and}\quad c(\cdot,\omega)\in\mathcal{V}^* \quad (3.32a)$$

be such that $q(y,\varphi,\omega)=q(\varphi,y,\omega)$ for all $y,\varphi\in\mathcal{V}$. Moreover, we assume that

$$q(y,\varphi,\cdot)\colon\Omega\to\mathbb{R}\text{ is }(\mathcal{F},\mathcal{B})\text{-measurable for all } y,\varphi\in\mathcal{V}, \quad (3.32b)$$

and that there exists a measurable function $q_{\max}(\omega) \geq 0$ such that for a.a. $\omega \in \Omega$,

$$q(y, y, \omega) \geq 0 \qquad \text{for all } y \in \mathcal{V}, \tag{3.32c}$$

$$|q(y, \varphi, \omega)| \leq q_{\max}(\omega)\, \|y\|_{\mathcal{V}} \|\varphi\|_{\mathcal{V}} \qquad \text{for all } y, \varphi \in \mathcal{V}. \tag{3.32d}$$

Finally, we also assume

$$q_{\max} \in L^{\infty}_{\mathbb{P}}(\Omega), \quad c \in L^{2}_{\mathbb{P}}(\Omega, \mathcal{V}^*). \tag{3.33}$$

Given a realization $\omega \in \Omega$, let $y(u;\omega) \in \mathcal{V}$ be the solution of (3.16). The cost functional is now $\frac{1}{2}q(y(u;\omega), y(u;\omega), \omega) + c(y(u;\omega), \omega) + \frac{1}{2}r(u,u)$ instead of (3.8) in the deterministic case. Again, we sometimes call this functional the cost or the loss, but, as in Section 3.1.2, cost or loss do not necessarily correspond to monetary cost or loss, but are a scalar quantification of the under-performance of the system given a realization $\omega \in \Omega$. Since we must decide on our control before the realization $\omega \in \Omega$ is known, we must consider the cost for all $\omega \in \Omega$, that is, we must consider

$$\omega \mapsto \frac{1}{2}q(y(u;\omega), y(u;\omega), \omega) + c(y(u;\omega), \omega) + \frac{1}{2}r(u,u). \tag{3.34}$$

Under the assumptions (3.32), $q\colon \mathcal{V} \times \mathcal{V} \times \Omega \to \mathbb{R}$ and $c\colon \mathcal{V} \times \Omega \to \mathbb{R}$ are Carathéodory functions and are jointly measurable; see e.g. Aliprantis and Border (2006, Lemma 4.51). Thus (3.34) is measurable, i.e. a random variable in ω on $(\Omega, \mathcal{F}, \mathbb{P})$. To construct an objective function from this random variable, we need to 'scalarize' it.

If (3.32), (3.33) hold and the solution of (3.16) satisfies $y(u;\cdot) \in L^2_{\mathbb{P}}(\Omega, \mathcal{V})$, then (3.34) is integrable and we can use the expected value of (3.34) as our objective function, that is, we can minimize the expected cost. This leads to the problem

$$\min_{u \in \mathcal{U}} \int_{\Omega} \frac{1}{2}q(y(u;\omega), y(u;\omega), \omega) + c(y(u;\omega), \omega)\, \mathrm{d}\mathbb{P}(\omega) + \frac{1}{2}r(u,u), \tag{3.35}$$

where $y(u;\omega)$ is the solution of (3.16). The problem (3.35) is a linear–quadratic optimal control problem with controls in $u \in \mathcal{U}$ and states $y \in L^2_{\mathbb{P}}(\Omega, \mathcal{V})$.

The control $u \in \mathcal{U}$ is deterministic because we need to decide on the control before the uncertainty is revealed, i.e. before the outcome $\omega \in \Omega$ is known. In (3.35) we compute the control by minimizing the expected cost but, as we have seen in Section 2 and will further discuss in Section 4, scalarizations of (3.34) other than the expected value are often preferable.

The first term in the objective function in (3.35) involves the following maps. The first map is the control-to-state map

$$S\colon \mathcal{U} \to L^2_{\mathbb{P}}(\Omega, \mathcal{V}), \tag{3.36a}$$

$$u \mapsto y(u;\cdot), \tag{3.36b}$$

where $y(u;\omega)$ solves (3.16). The second map is

$$J\colon \mathcal{V}\times\Omega\to\mathbb{R}, \tag{3.37a}$$

$$(y,\omega)\mapsto \frac{1}{2}q(y,y,\omega)+c(y,\omega), \tag{3.37b}$$

and the associated map

$$\mathcal{J}\colon L^2_{\mathbb{P}}(\Omega,\mathcal{V})\to L^1_{\mathbb{P}}(\Omega), \tag{3.38a}$$

$$y\mapsto [\mathcal{J}(y)](\cdot)=J(y(\cdot),\cdot), \tag{3.38b}$$

which defines the cost function. The final map is the expected value

$$\mathcal{R}\colon L^1_{\mathbb{P}}(\Omega)\to\mathbb{R} \tag{3.39a}$$

$$\xi\mapsto \mathbb{E}[\xi]=\int_\Omega \xi(\omega)\,\mathrm{d}\mathbb{P}(\omega), \tag{3.39b}$$

which is used to scalarize the cost function. With these maps, the objective function (3.35) can be written as

$$\min_{u\in\mathcal{U}}\ \mathcal{R}(\mathcal{J}(S(u)))+\frac{1}{2}r(u,u). \tag{3.40}$$

As we will see in Section 4, this problem structure arises often, although with more involved versions of the maps S, $\mathcal{J}$ and $\mathcal{R}$.

The Fréchet derivatives of the maps S and $\mathcal{J}$ play an important role in the solution of (3.40), and we will discuss Fréchet differentiability of (3.36), (3.37) and (3.38) next. Because (3.16) is affine linear in y and u, and (3.37) is quadratic in y, establishing Fréchet differentiability is fairly straightforward. However, there are differences between Fréchet differentiability of the underlying maps at a given $\omega\in\Omega$ and the maps into/on $L^2_{\mathbb{P}}(\Omega,\mathcal{V})$. Establishing Fréchet differentiability of the latter maps requires additional assumptions.

Lemma 3.10. For $\omega\in\Omega$ and $u\in\mathcal{U}$, let $y(u;\omega)\in\mathcal{V}$ denote the solution of (3.16).

If (3.17) holds, then for a.a. $\omega\in\Omega$ the map $\mathcal{U}\ni u\mapsto y(u;\omega)\in\mathcal{V}$ is Fréchet-differentiable on $\mathcal{U}$ and the Fréchet derivative $w(\omega)=y_u(u;\omega)\delta u$ is the solution of

$$a(w(\omega),\varphi,\omega)+b(\delta u,\varphi,\omega)=0\quad\text{for all }\varphi\in\mathcal{V}. \tag{3.41}$$

If the assumptions of Theorem 3.8 are satisfied with $k=2$, then the control-to-state map (3.36) is Fréchet-differentiable on $\mathcal{U}$ and the Fréchet derivative $S'(u)\delta u=y_u(u;\cdot)\delta u$ is computed pointwise for $\omega\in\Omega$ as the solution of (3.41).

Proof. Fréchet differentiability of $\mathcal{U}\ni u\mapsto y(u;\omega)\in\mathcal{V}$, for given $\omega\in\Omega$, follows because $e(\omega)=y(u+\delta u,\omega)-y(u;\omega)-y_u(u;\omega)\delta u$ satisfies $a(e(\omega),\varphi,\omega)=0$ for all $\varphi\in\mathcal{V}$, which implies $e(\omega)=0$ in $\mathcal{V}$, and because the solution of (3.41) satisfies $\|y_u(u,\omega)\delta u\|_{\mathcal{V}}\le a_{\min}(\omega)^{-1}b_{\max}(\omega)\|\delta u\|_{\mathcal{U}}$ for all $\delta u\in\mathcal{U}$ (cf. (3.23)), which implies that $\mathcal{U}\ni\delta u\mapsto y_u(u;\omega)\delta u\in\mathcal{V}$ is a bounded linear operator.

If the assumptions of Theorem 3.8 are satisfied with $k = 2$, then one can use the same arguments used to prove Theorem 3.8 to show that, for all $\delta u \in \mathcal{U}$, the unique solution of (3.41) satisfies $w(\cdot) = y_u(u;\cdot)\delta u \in L^k_{\mathbb{P}}(\Omega, \mathcal{V})$ and

$$\int_\Omega \|y_u(u;\cdot)\,\delta u\|^2_{\mathcal{V}}\,\mathrm{d}\mathbb{P}(\omega) \le \left(\int_\Omega \left(\frac{1}{a_{\min}(\omega)}\right)^{2q} \mathrm{d}\mathbb{P}(\omega)\right)^{1/q} \times \left(\int_\Omega b_{\max}(\omega)^{2p}\,\mathrm{d}\mathbb{P}(\omega)\right)^{1/p} \|\delta u\|^2_{\mathcal{U}},$$

that is, $S'(u) \in \mathcal{L}(\mathcal{U}, L^2_{\mathbb{P}}(\Omega, \mathcal{V}))$. Since (3.16) is affine linear in y and u, $S(u+\delta u) = S(u) + S'(u)\delta u$. □

Lemma 3.11. If (3.32) holds, then (3.37) is Fréchet-differentiable in y on $\mathcal{V} \times \Omega$, and the partial Fréchet derivative with respect to y is

$$J_y(y,\omega)\delta y = q(y, \delta y, \omega) + c(\delta y, \omega).$$

If (3.32) and (3.33) hold, then (3.38) is Fréchet-differentiable and the Fréchet derivative is

$$[\mathcal{J}_y(y)\,\delta y](\cdot) = J_y(y(\cdot),\cdot)\,\delta y(\cdot).$$

Proof. The Fréchet differentiability of (3.37) with respect to y follows immediately from

$$J(y+\delta y,\omega) - J(y,\omega) - J_y(y,\omega)\delta y = \frac{1}{2}q(\delta y, \delta y, \omega) \le \frac{q_{\max}(\omega)}{2}\|\delta y\|^2_{\mathcal{V}}$$

and the fact that $J_y(y,\omega) \in \mathcal{V}^*$ if (3.32) holds. Fréchet differentiability of (3.38) follows from

$$\begin{aligned}&\int_\Omega \left|[\mathcal{J}(y+\delta y)](\omega) - [\mathcal{J}(y)](\omega) - [\mathcal{J}_y(y)\,\delta y](\omega)\right| \mathrm{d}\mathbb{P}(\omega)\\ &\quad= \int_\Omega \left|\frac{1}{2}q((\delta y)(\omega), (\delta y)(\omega), \omega)\right| \mathrm{d}\mathbb{P}(\omega) \le \frac{\|q_{\max}\|_{L^\infty_{\mathbb{P}}(\Omega)}}{2}\int_\Omega \|(\delta y)(\omega)\|^2_{\mathcal{V}}\,\mathrm{d}\mathbb{P}(\omega)\end{aligned}$$

and the fact that $\mathcal{J}_y(y) \in \mathcal{L}(L^2_{\mathbb{P}}(\Omega, \mathcal{V}), L^1_{\mathbb{P}}(\Omega))$ if (3.32) and (3.33) hold. □

Lemma 3.12. If the assumptions of Theorem 3.8 are satisfied with $k = 2$, and if (3.32), (3.33) hold, then the composition $\mathcal{J} \circ S$ of the maps (3.36) and (3.38) is Fréchet-differentiable and the Fréchet derivative is

$$\begin{aligned}[\mathcal{J}_y(S(u))\,S'(u)\delta u](\cdot) &= q(y(u;\cdot), y_u(u;\cdot)\delta u, \cdot) + c(y_u(u;\cdot)\delta u, \cdot) &\text{(3.42a)}\\ &= b(\delta u, \lambda(\cdot), \cdot), &\text{(3.42b)}\end{aligned}$$

where $\lambda \in L^1_{\mathbb{P}}(\Omega, \mathcal{V})$ is computed pointwise as the solution of

$$a(\varphi, \lambda(\omega), \omega) + q(y(u;\omega), \varphi, \omega) = -c(\varphi, \omega) \quad \text{for all } \varphi \in \mathcal{V},\ \text{a.a. } \omega \in \Omega. \tag{3.43}$$

If in addition $a^{-1}_{\min} \in L^\infty_{\mathbb{P}}(\Omega)$, then the solution λ of (3.43) satisfies $\lambda \in L^2_{\mathbb{P}}(\Omega, \mathcal{V})$.

Proof. The Fréchet differentiability of $\mathcal{J} \circ S$ and (3.42a) follow from Lemmas 3.10 and 3.11.

To prove (3.42b) we first show that λ defined as the pointwise solution of (3.43) satisfies $\lambda \in L^1_{\mathbb{P}}(\Omega, \mathcal{V})$. Note that $a_{\min}^{-1} \in L^{2q}_{\mathbb{P}}(\Omega) \subset L^2_{\mathbb{P}}(\Omega)$. Because $c \in L^2_{\mathbb{P}}(\Omega, \mathcal{V}^*)$ and $y(u;\cdot) \in L^2_{\mathbb{P}}(\Omega, \mathcal{V})$, one can apply the same arguments used to prove Theorem 3.8 to (3.43) (with $k = 1$, $p = q = 2$) to show that the pointwise variational equation (3.43) has a unique solution $\lambda(\omega)$, that $\omega \mapsto \lambda(\omega)$ is measurable, and that

$$\int_\Omega \|\lambda(\omega)\|_{\mathcal{V}} \, \mathrm{d}\mathbb{P}(\omega) \leq \left(\int_\Omega \left(\frac{1}{a_{\min}(\omega)} \right)^2 \mathrm{d}\mathbb{P}(\omega) \right)^{1/2} \times \left(\int_\Omega (\|c(\cdot,\omega)\|_{\mathcal{V}^*} + q_{\max}(\omega)\|y(u;\omega)\|_{\mathcal{V}})^2 \, \mathrm{d}\mathbb{P}(\omega) \right)^{1/2}.$$

Inserting $\varphi = y_u(u;\omega)\delta u$ in (3.43) and $\varphi = \lambda(\omega)$ in (3.41) implies

$$q(y(u;\omega), y_u(u;\omega)\delta u, \omega) + c(y_u(u;\omega)\delta u, \omega) = b(\delta u, \lambda(\omega), \omega),$$

which yields (3.42b).

If, in addition, $a_{\min}^{-1} \in L^\infty_{\mathbb{P}}(\Omega)$, then we can apply the arguments used to prove Theorem 3.8 to (3.43) with $k = 2$, $p = 1$, $q = \infty$ to show that $\lambda \in L^2_{\mathbb{P}}(\Omega, \mathcal{V})$. In fact, in this case we can also apply the techniques used to prove Theorem 3.7. □

Weak lower semicontinuity of the objective function in (3.35) (or the equivalent representation in (3.40)) is important for the existence of solutions of (3.35). If the assumptions of Theorem 3.8 are satisfied with $k = 2$ and if (3.32) and (3.33) hold, then the function $\mathcal{U} \ni u \mapsto \mathcal{R}(\mathcal{J}(S(u)))$ with S, $\mathcal{J}$ and $\mathcal{R}$ given by (3.36), (3.38) and (3.39) is convex and continuous (even Fréchet-differentiable). Moreover, if r satisfies the conditions (3.9d) and (3.9e), then $\mathcal{U} \ni u \mapsto r(u,u)$ is convex and continuous. Since a convex and continuous functional on a normed space is weakly lower semicontinuous (see e.g. Ekeland and Temam 1999, Section 2.2, or Jahn 2007, Section 2.2), the objective function in (3.35) (or its equivalent representation in (3.40)) is weakly lower semicontinuous.

Lemma 3.13. If the assumptions of Theorem 3.8 are satisfied with $k = 2$ and if (3.9d), (3.9e), (3.32) and (3.33) hold, then the function

$$f : \mathcal{U} \to \mathbb{R}, \quad u \mapsto f(u) := \mathcal{R}(\mathcal{J}(S(u))) + \frac{1}{2} r(u,u)$$

is weakly lower semicontinuous.

One can prove the following existence and uniqueness result.

Theorem 3.14. If the assumptions of Theorem 3.8 are satisfied with $k = 2$, and if (3.9d), (3.9e), (3.32) and (3.33) hold, then the optimal control problem (3.35) has a unique solution $u \in \mathcal{U}$. Moreover, $u \in \mathcal{U}$ solves (3.35) if and only if there

exist $y \in L^2_{\mathbb{P}}(\Omega, \mathcal{V})$ and $\lambda \in L^1_{\mathbb{P}}(\Omega, \mathcal{V})$ such that y, u, λ solve

$$a(\varphi, \lambda(\omega), \omega) + q(y(\omega), \varphi, \omega) = -c(\varphi, \omega) \quad \text{for all } \varphi \in \mathcal{V}, \text{ a.a. } \omega \in \Omega, \tag{3.44a}$$

$$\int_\Omega b(\psi, \lambda(\omega), \omega)\, \mathrm{d}\mathbb{P}(\omega) + r(u, \psi) = 0 \quad \text{for all } \psi \in \mathcal{U}, \tag{3.44b}$$

$$a(y(\omega), \varphi, \omega) + b(u, \varphi, \omega) = \ell(\varphi, \omega) \quad \text{for all } \varphi \in \mathcal{V}, \text{ a.a. } \omega \in \Omega. \tag{3.44c}$$

Proof. The conditions (3.9d) and (3.9e) imply $\mathcal{R}(\mathcal{J}(S(u))) + \frac{1}{2} r(u, u) \to \infty$ as $\|u\|_{\mathcal{U}} \to \infty$. Together with the weak lower semicontinuity, it implies the existence of a solution. Strict convexity of the objective function implies uniqueness. Since the function is convex and Fréchet-differentiable, $u \in \mathcal{U}$ solves (3.35) if and only if the Fréchet derivative of the objective function at $u \in \mathcal{U}$ is zero, which by Lemma 3.12 and linearity of the expected value is equivalent to (3.44). □

The following example is an extension of Example 3.5.

Example 3.15. Consider the state equation in Example 3.9, and assume that for $u \in L^2(D)$ the state equation (3.28) has a unique solution $y(u; \cdot) \in L^2_{\mathbb{P}}(\Omega, H^1_0(D))$. Given a function $\eta \in L^\infty_{\mathbb{P}}(\Omega, L^\infty(D))$ with $\eta \geq 0$ a.e. in $D \times \Omega$, we want to solve

$$\min_{u \in L^2(D)} \int_\Omega \frac{1}{2} \int_D \eta(x, \omega)(y(u; x, \omega) - \widehat{y}(x))^2 \, \mathrm{d}x \, \mathrm{d}\mathbb{P}(\omega) + \frac{\alpha}{2} \int_D u(x)^2 \, \mathrm{d}x, \tag{3.45}$$

where $y(u; \cdot) \in L^2_{\mathbb{P}}(\Omega, H^1_0(D))$ is the unique solution of (3.28) given $u \in L^2(D)$.

If we drop the constant

$$\frac{1}{2} \int_\Omega \int_D \eta(x, \omega) \widehat{y}(x)^2 \, \mathrm{d}x \, \mathrm{d}\mathbb{P}(\omega),$$

the objective functional in (3.14) is a special case of (3.8) with

$$q(y, \varphi, \omega) = \int_\Omega \int_D \eta(x, \omega) y(x, \omega) \varphi(x, \omega) \, \mathrm{d}x \, \mathrm{d}\mathbb{P}(\omega),$$

$$c(y, \omega) = -\int_\Omega \int_D \eta(x, \omega) y(x, \omega) \widehat{y}(x) \, \mathrm{d}x \, \mathrm{d}\mathbb{P}(\omega),$$

and $r(u, \psi) = \int_D u(x)\psi(x) \, \mathrm{d}x$. Application of Theorem 3.14 gives the following result on the existence, uniqueness and characterization of the solution of (3.45).

Corollary 3.16. If the assumptions in Example 3.9 are satisfied with $k = 2$, the optimal control problem (3.14) has a unique solution $u \in L^2(D)$. Furthermore, $u \in L^2(D)$ solves (3.14) if and only if there exist $y \in L^2_{\mathbb{P}}(\Omega, H^1_0(D))$ and $\lambda \in L^1_{\mathbb{P}}(\Omega, H^1_0(D))$ such that y, u, λ solve

$$-\nabla \cdot (\kappa(x, \omega) \nabla \lambda(x, \omega)) = -\eta(x, \omega)(y(x, \omega) - \widehat{y}(x)), \quad x \in D, \text{ a.a. } \omega \in \Omega, \tag{3.46a}$$

$$\lambda(x, \omega) = 0, \quad x \in \partial D, \text{ a.a. } \omega \in \Omega, \tag{3.46b}$$

$$\int_\Omega \lambda(x, \omega) \, \mathrm{d}\mathbb{P}(\omega) = \alpha u(x), \quad x \in D, \tag{3.46c}$$

$$-\nabla \cdot (\kappa(x,\omega)\nabla y(x,\omega)) = f(x,\omega) + u(x), \qquad x \in D, \text{ a.a. } \omega \in \Omega, \quad (3.46d)$$

$$y(x,\omega) = 0, \qquad x \in \partial D, \text{ a.a. } \omega \in \Omega. \quad (3.46e)$$

3.3. Problems with finite noise

Often the linear and bilinear forms in (3.48) do not depend on ω directly but through a finite-dimensional vector of random variables $\boldsymbol{\xi}$. This is the case for the example problems in Section 2 and also for the example (3.26) with (3.31). The results from the previous Section 3.2 immediately translate to this setting. We collect the main results here for easier referencing.

As in the previous section, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. Furthermore, let $\Xi \subset \mathbb{R}^M$ be a non-empty set, let $\mathcal{B}_\Xi$ be the Borel σ-algebra on Ξ, and let $\boldsymbol{\xi}\colon (\Omega, \mathcal{F}) \to (\Xi, \mathcal{B}_\Xi)$ be a vector of random variables with distribution $\mathbb{P}^{\boldsymbol{\xi}} = \mathbb{P} \circ \boldsymbol{\xi}^{-1}$ on the sigma algebra $\sigma(\boldsymbol{\xi}) = \{\boldsymbol{\xi}^{-1}(B) \mid B \in \mathcal{B}_\Xi\}$.

Because $\boldsymbol{\xi}$ is a vector of M random variables, this set-up is also referred to as the finite-noise case, or we say that the problem satisfies the finite-noise assumption. Sometimes we can approximate problems of the type considered in Section 3.3 by a problem with finite noise. For example, a series representation of the diffusivity coefficient,

$$\kappa(x,\omega) = \bar{\kappa}(x) + \sum_{m\geq 1} \kappa^{(m)}(x)\, \boldsymbol{\xi}^{(m)}(\omega), \quad (3.47)$$

and of the right-hand side f, e.g. via the Karhunen–Loève expansion, can allow us to reformulate elliptic PDE (3.26) as a problem in parameters $\{\xi^{(m)}\}_{m\geq 1}$. Then, under suitable assumptions on the coefficient functions, a truncation of these series representations for κ and f allow us to approximate the elliptic PDE (3.26) by one with finite noise. Such approaches are analysed in the papers by Cohen, DeVore and Schwab (2010) and Cohen and DeVore (2015), among others, for the approximation of PDEs. However, we note that in the optimization context, error estimates for an approximation of the state equation are not enough. In addition, one must also derive estimates for the corresponding approximation of the adjoint equation. In the case of the Example 3.15, one must analyse finite-noise approximations of the state equation (3.46d), (3.46e) and of the adjoint equation (3.46a), (3.46b).

The set-up of the finite-noise version of the optimal control problem (3.35) is fairly straightforward, and the existence, differentiability and optimality condition results from Section 3.2.2 hold with $(\Omega, \mathcal{F}, \mathbb{P})$ replaced by $(\Xi, \mathcal{B}_\Xi, \mathbb{P}^{\boldsymbol{\xi}})$. We summarize the set-up and the main results for easier referencing later.

In the finite-noise case, the set-up of the state equation and the optimal control problem is as follows. For $\xi \in \Xi$, let $\ell(\cdot,\xi) \in \mathcal{V}^*$, and let $a(\cdot,\cdot,\xi)\colon \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ and $b(\cdot,\cdot,\xi)\colon \mathcal{U}\times\mathcal{V} \to \mathbb{R}$ be bilinear forms such that

$$\text{for each } y, \varphi \in \mathcal{V},\ u \in \mathcal{U}, \text{ the functions}$$

$$a(y,\varphi,\cdot), b(u,\varphi,\cdot)\colon \xi \to \mathbb{R} \text{ are } (\mathcal{B}_\Xi, \mathcal{B})\text{-measurable}, \quad (3.48a)$$

and such that there exist measurable functions $0 < a_{\min}(\xi) \leq a_{\max}(\xi)$ and $0 < b_{\max}(\xi)$ with

$$a(y, y, \xi) \geq a_{\min}(\xi)\|y\|_{\mathcal{V}}^2 \qquad \text{for all } y \in \mathcal{V}, \text{ a.a. } \xi \in \Xi, \tag{3.48b}$$

$$|a(y, \varphi, \xi)| \leq a_{\max}(\xi)\,\|y\|_{\mathcal{V}}\|\varphi\|_{\mathcal{V}} \qquad \text{for all } y, \varphi \in \mathcal{V}, \text{ a.a. } \xi \in \Xi, \tag{3.48c}$$

$$|b(u, \varphi, \xi)| \leq b_{\max}(\xi)\,\|u\|_{\mathcal{U}}\|\varphi\|_{\mathcal{V}} \qquad \text{for all } u \in \mathcal{U}, \varphi \in \mathcal{V}, \text{ a.a. } \xi \in \Xi. \tag{3.48d}$$

We now consider the state equation

$$a(y, \varphi, \xi) + b(u, \varphi, \xi) = \ell(\varphi, \xi) \quad \text{for all } \varphi \in \mathcal{V}, \text{ a.a. } \xi \in \Xi. \tag{3.49}$$

There is an existence result analogous to Theorem 3.8. We state it for the case $k = 2$.

Theorem 3.17. Let $p, q \geq 1$ with $1/p + 1/q = 1$. If $\mathcal{V}$ is a separable Hilbert space, if $\mathcal{U}$ is a Hilbert space, and if the bilinear forms a, b and linear form ℓ satisfy (3.48) and $a_{\max} \in L^{\infty}_{\mathbb{P}^\xi}(\Xi)$, $a_{\min}^{-1} \in L^{2q}_{\mathbb{P}^\xi}(\Xi)$, $b_{\max} \in L^{2p}_{\mathbb{P}^\xi}(\Xi)$ and $\ell \in L^{2p}_{\mathbb{P}^\xi}(\Xi, \mathcal{V}^*)$, then for every $u \in \mathcal{U}$, the pointwise variational equation (3.49) has a unique solution $y(u; \xi)$, and this solution satisfies $y(u; \cdot) \in L^2_{\mathbb{P}^\xi}(\Xi, \mathcal{V})$.

The objective function in the optimal control problem now involves the following bilinear and linear forms. Let r again be given as in (3.9), $c \in L^2_{\mathbb{P}^\xi}(\Xi, \mathcal{V}^*)$, and for $\xi \in \Xi$, let $q(\cdot, \cdot, \xi)\colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be such that $q(y, \varphi, \xi) = q(\varphi, y, \xi)$ for all $y, \varphi \in \mathcal{V}$ and

$$q(y, \varphi, \cdot)\colon \xi \to \mathbb{R} \text{ is } (\mathcal{B}_\Xi, \mathcal{B})\text{-measurable for all } y, \varphi \in \mathcal{V}. \tag{3.50a}$$

Moreover, assume that there exists $q_{\max} \in L^{\infty}_{\mathbb{P}^\xi}(\Xi)$ such that for all $\xi \in \Xi$,

$$0 \leq q(y, y, \xi) \quad \text{and} \quad |q(y, \varphi, \xi)| \leq q_{\max}(\xi)\,\|y\|_{\mathcal{V}}\|\varphi\|_{\mathcal{V}} \quad \text{for all } y, \varphi \in \mathcal{V}. \tag{3.50b}$$

This leads to the optimal control problem

$$\min_{u \in \mathcal{U}} \int_\Xi \frac{1}{2} q(y(u; \xi), y(u; \xi), \xi) + c(y(u; \xi), \xi)\,\mathrm{d}\mathbb{P}^\xi(\xi) + \frac{1}{2} r(u, u), \tag{3.51}$$

where $y(u; \xi)$ is the solution of (3.49). If Theorem 3.17 and (3.50) hold, the objective function in (3.51) is well-defined for each $u \in \mathcal{U}$.

Analogously to Theorem 3.14, one can prove the following existence and uniqueness result.

Theorem 3.18. If the assumptions of Theorem 3.17 are satisfied and if (3.9d), (3.9e) and (3.50) hold, then the optimal control problem (3.51) has a unique solution

$u \in \mathcal{U}$. Moreover, $u \in \mathcal{U}$ solves (3.51) if and only if there exist $y \in L^2_{\mathbb{P}^\xi}(\Xi, \mathcal{V})$ and $\lambda \in L^1_{\mathbb{P}^\xi}(\Xi, \mathcal{V})$ such that y, u, λ solve

$$a(\varphi, \lambda(\xi), \xi) + q(y(\xi), \varphi, \xi) = -c(\varphi, \xi) \quad \text{for all } \varphi \in \mathcal{V}, \text{ a.a. } \xi \in \Xi, \tag{3.52a}$$

$$\int_\Xi b(\psi, \lambda(\xi), \xi)\, \mathrm{d}\mathbb{P}^\xi(\xi) + r(u, \psi) = 0 \quad \text{for all } \psi \in \mathcal{U}, \tag{3.52b}$$

$$a(y(\xi), \varphi, \xi) + b(u, \varphi, \xi) = \ell(\varphi, \xi) \quad \text{for all } \varphi \in \mathcal{V}, \text{ a.a. } \xi \in \Xi. \tag{3.52c}$$

We note that under the assumptions of Theorem 3.18, the objective function

$$f(u) = \int_\Xi \frac{1}{2} q(y(u;\xi), y(u;\xi), \xi) + c(y(u;\xi), \xi)\, \mathrm{d}\mathbb{P}^\xi(\xi) + \frac{1}{2} r(u, u)$$

in (3.51) is Fréchet-differentiable and

$$f'(u)\delta u = \int_\Xi b(\delta u, \lambda(\xi), \xi)\, \mathrm{d}\mathbb{P}^\xi(\xi) + r(u, \delta u) \quad \text{for all } \delta u \in \mathcal{U},$$

where λ is the solution of the adjoint equation (3.52a).

3.4. Problems with random control

We have focused thus far on PDE-constrained optimization under uncertainty, where the controls or, more generally, the decisions must be determined before the uncertainty is revealed. In this case the control $u \in \mathcal{U}$ is deterministic, and only the state depends on the random parameter $y(u;\cdot) \in L^p_{\mathbb{P}}(\Omega, \mathcal{V})$ for some $p \geq 1$.

Some authors have considered the case where the control is also allowed to depend on the random parameter. In this setting, controls parametrized by random parameters are computed, and once the uncertainty is revealed, the control instance corresponding to the realized uncertainty is applied. This problem is easier to solve and eliminates several of the problem formulation issues and resulting numerical challenges that are the focus of this paper.

In the context of the model problem in Section 3.2, the problem set-up with random controls is as follows. Let $\mathcal{V}$ be a separable Hilbert space, let $\mathcal{U}$ be a Hilbert space, the bilinear forms a, b and linear form ℓ satisfy (3.17), (3.19), the bilinear form q and linear form c satisfy (3.32), (3.33), and let r be given as follows. For $\omega \in \Omega$, let

$$r(\cdot, \cdot, \omega) \colon \mathcal{U} \times \mathcal{U} \to \mathbb{R} \tag{3.53a}$$

be such that $r(u, \psi, \omega) = r(\psi, u, \omega)$ for all $u, \psi \in \mathcal{U}$. Moreover, assume that

$$r(y, \varphi, \cdot) \colon \Omega \to \mathbb{R} \text{ is } (\mathcal{F}, \mathcal{B})\text{-measurable for all } u, \psi \in \mathcal{U} \tag{3.53b}$$

and that there exist functions $r_{\min}, r_{\max}$ that are positive a.e., that satisfy

$$r_{\min}^{-1}, r_{\max} \in L^\infty_{\mathbb{P}}(\Omega), \tag{3.53c}$$

and, for a.a. $\omega \in \Omega$,

$$r(u,u,\omega) \geq r_{\min}(\omega)\,\|u\|_{\mathcal{U}}^2 \qquad \text{for all } u \in \mathcal{U}, \tag{3.53d}$$

$$|r(u,\psi,\omega)| \leq r_{\max}(\omega)\,\|u\|_{\mathcal{U}}\|\psi\|_{\mathcal{U}} \qquad \text{for all } u,\psi \in \mathcal{U}. \tag{3.53e}$$

Given a control $u \in L^2_{\mathbb{P}}(\Omega,\mathcal{U})$, the state equation is given as follows. Compute the solution $y \in L^2_{\mathbb{P}}(\Omega,\mathcal{V})$ of the variational equation

$$\int_\Omega a(y(\omega),\varphi(\omega),\omega)\,\mathrm{d}\mathbb{P}(\omega) + \int_\Omega b(u(\omega),\varphi(\omega),\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int_\Omega \ell(\varphi(\omega),\omega)\,\mathrm{d}\mathbb{P}(\omega) \quad \text{for all } \varphi \in L^2_{\mathbb{P}}(\Omega,\mathcal{V}). \tag{3.54}$$

If (3.17), (3.19) hold, then, by the Lax–Milgram theorem, for every $u \in L^2_{\mathbb{P}}(\Omega,\mathcal{U})$, the state equation (3.54) has a unique solution $y(u) \in L^2_{\mathbb{P}}(\Omega,\mathcal{V})$. Similar to our discussion in Section 3.4, this solution also satisfies

$$a(y(u;\omega),\varphi,\omega) + b(u(\omega),\varphi,\omega) = \ell(\varphi,\omega) \quad \text{for all } \varphi \in \mathcal{V},\ \text{a.a. } \omega \in \Omega. \tag{3.55}$$

Instead of (3.20) we now consider

$$\omega \mapsto \frac{1}{2}q(y(u;\omega),y(u;\omega),\omega) + c(y(u;\omega),\omega) + \frac{1}{2}r(u(\omega),u(\omega),\omega), \tag{3.56}$$

where $y(u)$ solves (3.54). Since $u \in L^2_{\mathbb{P}}(\Omega,\mathcal{U})$ and $y(u) \in L^2_{\mathbb{P}}(\Omega,\mathcal{V})$, the random variable (3.56) is integrable, and we can minimize the expectation of the cost (3.56). This leads to the problem

$$\min_{u\in L^2_{\mathbb{P}}(\Omega,\mathcal{U})} \int_\Omega \frac{1}{2}q(y(u;\omega),y(u;\omega),\omega) + c(y(u;\omega),\omega)\,\mathrm{d}\mathbb{P}(\omega) + \int_\Omega \frac{1}{2}r(u(\omega),u(\omega),\omega)\,\mathrm{d}\mathbb{P}(\omega), \tag{3.57}$$

where $y(u)$ solves (3.54). Alternatively, one could also consider a family of optimal control problems parametrized by $\omega \in \Omega$,

$$\min_{u(\omega)\in\mathcal{U}} \frac{1}{2}q(y(u;\omega),y(u;\omega),\omega) + c(y(u;\omega),\omega) + \frac{1}{2}r(u(\omega),u(\omega),\omega). \tag{3.58}$$

The equivalence between (3.57) and (3.58) follows because the objective function is Carathéodory (hence, a normal integrand), $\mathcal{U}$ is a separable Hilbert space and $L^2_{\mathbb{P}}(\Omega,\mathcal{V})$ is decomposable. See e.g. the proof of Theorem 2 in Rockafellar (1971) for general details on interchanging the integral and minimization operations.

The necessary and sufficient optimality conditions for (3.57) or (3.58) now lead to

$$a(\varphi,\lambda(\omega),\omega) + q(y(\omega),\varphi,\omega) = -c(\varphi,\omega) \quad \text{for all } \varphi \in \mathcal{V},\ \text{a.a. } \omega \in \Omega, \tag{3.59a}$$

$$b(\psi,\lambda(\omega),\omega) + r(u(\omega),\psi,\omega) = 0 \quad \text{for all } \psi \in \mathcal{U},\ \text{a.a. } \omega \in \Omega, \tag{3.59b}$$

$$a(y(\omega),\varphi,\omega) + b(u(\omega),\varphi,\omega) = \ell(\varphi,\omega) \quad \text{for all } \varphi \in \mathcal{V},\ \text{a.a. } \omega \in \Omega. \tag{3.59c}$$

We now have to compute a control $u \in L^2_{\mathbb{P}}(\Omega, \mathcal{U})$, but in contrast to (3.44) the optimality conditions (3.59) decouple across ω. This makes the problem (3.35) computationally easier to solve.

As mentioned before, in this problem setting we compute a control $u(\omega)$ for all possible realizations $\omega \in \Omega$. Then, once the realization $\omega \in \Omega$ is observed, we apply the corresponding control $u(\omega)$. This eliminates many of the problem formulation issues and the need for the efficient solution of the chosen optimization problem formulation we have to address when a deterministic control must be computed before the uncertainty is revealed. These problem formulations and corresponding solution methods are the focus of this paper.

We will not consider problems where the control depends on $\omega \in \Omega$, and we refer to the literature for further discussions of numerical methods for this problem class; see e.g. the papers by Borzì (2010), Tiesler, Kirby, Xiu and Preusser (2012), Rosseel and Wells (2012), Chen and Quarteroni (2014), Chen, Quarteroni and Rozza (2016) Benner, Onwunta and Stoll (2016) and Ahmad Ali, Ullmann and Hinze (2017).

3.5. Discretization

3.5.1. Discretization of the problem with deterministic parameters

To solve the optimal control problem (3.10), we need to discretize it. Discretizations of (3.10) typically build on discretizations of the underlying state equation (3.1), but additional requirements are needed. There are two fundamental approaches to the discretization of optimal control problems: discretize-then-optimize and optimize-then-discretize. See also Chapter 3 of Hinze *et al.* (2009) or Section 16.12 of Quarteroni (2009). Next we describe both approaches in the context of the optimal control model problem (3.10).

In the discretize-then-optimize approach, we first discretize the optimal control problem (3.10) to approximate it by a finite-dimensional optimization problem, and then we solve this finite-dimensional optimization problem. Specifically, the control space is replaced by a finite-dimensional space. Next, given a control from this finite-dimensional control space, the state equation (3.1) is approximated, and finally the objective function (3.8) is approximated based on the choices for the control and state discretizations. For PDE-constrained optimal control problems, the discretization of optimal control problem (3.10) typically builds on a discretization of the underlying state equation (3.1). We mention finite-element and discontinuous Galerkin discretizations, but other methods have also been used.

Discretization of the control space $\mathcal{U}$ and of the state equation (3.1) leads to a linear system

$$\mathbf{A}\mathbf{y} + \mathbf{B}\mathbf{u} = \mathbf{b}, \tag{3.60}$$

where $\mathbf{y} \in \mathbb{R}^{n_y}$ represents the discretized state, $\mathbf{u} \in \mathbb{R}^{n_u}$ represents the discretized control, and $\mathbf{A} \in \mathbb{R}^{n_y \times n_y}$, $\mathbf{B} \in \mathbb{R}^{n_y \times n_u}$, $\mathbf{b} \in \mathbb{R}^{n_y}$. For example, in a conforming

Galerkin finite-element approximation, we construct

$$\mathcal{U}_h = \operatorname{span}\{\psi_1, \dots, \psi_{n_u}\} \subset \mathcal{U}, \quad \mathcal{V}_h = \operatorname{span}\{\varphi_1, \dots, \varphi_{n_y}\} \subset \mathcal{V}, \tag{3.61a}$$

approximate the control and the state by

$$u_h = \sum_{i=1}^{n_u} \mathbf{u}_i \psi_i, \quad y_h = \sum_{i=1}^{n_y} \mathbf{y}_i \varphi_i, \tag{3.61b}$$

and replace (3.1) with

$$a(y_h, \varphi) + b(u_h, \varphi) = \ell(\varphi) \quad \text{for all } \varphi \in \mathcal{V}_h. \tag{3.62a}$$

This leads to (3.60) with

$$\begin{aligned} \mathbf{A}_{ij} &= a(\varphi_j, \varphi_i), \quad && i, j = 1, \dots, n_y, \\ \mathbf{B}_{ij} &= b(\psi_j, \varphi_i), \quad && i = 1, \dots, n_y, \, j = 1, \dots, n_u, \\ \mathbf{b}_i &= \ell(\varphi_i), && i = 1, \dots, n_y. \end{aligned} \tag{3.62b}$$

The assumptions (3.2) imply invertibility of $\mathbf{A} \in \mathbb{R}^{n_y \times n_y}$.

The state and control space discretizations lead to the following discretization of the objective function (3.8),

$$\frac{1}{2}\mathbf{y}^\top \mathbf{Q}\,\mathbf{y} + \mathbf{c}^\top \mathbf{y} + \frac{1}{2}\mathbf{u}^\top \mathbf{R}\,\mathbf{u}, \tag{3.63}$$

where $\mathbf{Q} = \mathbf{Q}^\top \in \mathbb{R}^{n_y \times n_y}$, $\mathbf{R} = \mathbf{R}^\top \in \mathbb{R}^{n_u \times n_u}$, $\mathbf{c} \in \mathbb{R}^{n_y}$, and, as before, $\mathbf{y} \in \mathbb{R}^{n_y}$, $\mathbf{u} \in \mathbb{R}^{n_u}$ represent the discretized state and control, respectively. For example, inserting the Galerkin discretization (3.61) into (3.8) leads to (3.63) with

$$\mathbf{Q}_{ij} = q(\varphi_i, \varphi_j), \quad i, j = 1, \dots, n_y, \quad \mathbf{R}_{ij} = r(\psi_i, \psi_j), \quad i, j = 1, \dots, n_u,$$

and $\mathbf{c}_i = c(\varphi_i)$, $i = 1, \dots, n_y$. The assumptions (3.9) on the bilinear forms q and r imply that $\mathbf{Q} \in \mathbb{R}^{n_y \times n_y}$ is symmetric positive semidefinite and $\mathbf{R} \in \mathbb{R}^{n_u \times n_u}$ is symmetric positive definite.

For a general discretization, if the matrix $\mathbf{A} \in \mathbb{R}^{n_y \times n_y}$ in (3.60) is invertible, then the control-to-state map is

$$\mathbf{y}(\mathbf{u}) = \mathbf{A}^{-1}(\mathbf{b} - \mathbf{B}\mathbf{u}).$$

The discretization of the optimal control problem (3.10) is given as

$$\begin{aligned} \min_{\mathbf{u} \in \mathbb{R}^{n_u}} \; & \frac{1}{2}\mathbf{u}^\top (\mathbf{B}^\top \mathbf{A}^{-\top} \mathbf{Q}\, \mathbf{A}^{-1} \mathbf{B} + \mathbf{R})\, \mathbf{u} - (\mathbf{B}^\top \mathbf{A}^{-\top} (\mathbf{c} + \mathbf{Q}\, \mathbf{A}^{-1} \mathbf{b}))^\top \mathbf{u} \\ & + \frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-\top} \mathbf{Q}\, \mathbf{A}^{-1} \mathbf{b} + \mathbf{c}^\top \mathbf{A}^{-1} \mathbf{b}. \end{aligned} \tag{3.64}$$

If $\mathbf{A} \in \mathbb{R}^{n_y \times n_y}$ is invertible, $\mathbf{Q} \in \mathbb{R}^{n_y \times n_y}$ is symmetric positive semidefinite and $\mathbf{R} \in \mathbb{R}^{n_y \times n_u}$ is symmetric positive definite, then (3.64) is a strongly convex quadratic optimization problem. Standard results from quadratic optimization,

found in Section 16 of Nocedal and Wright (2006), for example, give the following theorem.

Theorem 3.19. Let $\mathbf{A} \in \mathbb{R}^{n_y \times n_y}$ be invertible, let $\mathbf{Q} \in \mathbb{R}^{n_y \times n_y}$ be symmetric positive semidefinite and let $\mathbf{R} \in \mathbb{R}^{n_y \times n_u}$ be symmetric positive definite. The problem (3.64) has a unique solution $\mathbf{u} \in \mathbb{R}^{n_u}$. Moreover, $\mathbf{u} \in \mathbb{R}^{n_u}$ solves (3.64) if and only if

$$(\mathbf{B}^\top \mathbf{A}^{-\top} \mathbf{Q}\, \mathbf{A}^{-1} \mathbf{B} + \mathbf{R})\, \mathbf{u} = \mathbf{B}^\top \mathbf{A}^{-\top} (\mathbf{c} + \mathbf{Q}\, \mathbf{A}^{-1} \mathbf{b}), \tag{3.65}$$

and if and only if there exists $\mathbf{y} \in \mathbb{R}^{n_y}$ and $\boldsymbol{\lambda} \in \mathbb{R}^{n_y}$ such that

$$\begin{pmatrix} \mathbf{Q} & \mathbf{0} & \mathbf{A}^\top \\ \mathbf{0} & \mathbf{R} & \mathbf{B}^\top \\ \mathbf{A} & \mathbf{B} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{u} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} -\mathbf{c} \\ \mathbf{0} \\ \mathbf{b} \end{pmatrix}. \tag{3.66}$$

The system (3.65) is the Schur complement of (3.66).

An advantage of the discretize-then-optimize approach is that it always leads to a symmetric matrix in the optimality system (3.66), which is beneficial for its solution.

The quadratic problem (3.64) can be solved using, for example, the (preconditioned) conjugate gradient method. Alternatively, one can also solve the system of necessary and sufficient optimality conditions (3.66). The system (3.66) is also known as a Karush–Kuhn–Tucker (KKT) system, a special case of a saddle point problem that arises in optimization. The iterative solution of KKT systems such as (3.66) has been extensively researched, and many efficient methods exist; see e.g. Benzi, Golub and Liesen (2005), Borzì and Schulz (2009, 2012), Pearson and Pestana (2020) and Wathen (2015).

For a discretization error analysis of the discretize-then-optimize approach, we need to interpret the first equation in (3.66), that is,

$$\mathbf{A}^\top \boldsymbol{\lambda} + \mathbf{Q}\mathbf{y} = -\mathbf{c}, \tag{3.67}$$

as a discretization of the adjoint equation (3.13a), and $\boldsymbol{\lambda} \in \mathbb{R}^{n_y}$ as a representation of a discretization of the adjoint variable $\lambda \in \mathcal{V}$. For example, if the Galerkin discretization (3.61), (3.62) is used, (3.67) is equivalent to

$$a(\varphi, \lambda_h) + q(y_h, \varphi) = -c(\varphi) \quad \text{for all } \varphi \in \mathcal{V}_h, \tag{3.68a}$$

where

$$\lambda_h = \sum_{i=1}^{n_y} \boldsymbol{\lambda}_i \varphi_i \in \mathcal{V}_h. \tag{3.68b}$$

Equation (3.68) is the Galerkin discretization of the adjoint equation (3.13a). In general, however, the interpretation of (3.67) as a discretization of the adjoint equation (3.13a) may be less straightforward. In particular, (3.67) may result in a non-standard adjoint equation discretization that is of lower accuracy than the underlying discretization of the state equation. This is, for example, the case for

some stabilized finite-element discretizations applied to advection-dominated problems (see e.g. Collis and Heinkenschloss 2002) or for some discontinuous Galerkin methods (see Leykekhman 2012). If the adjoint equation (3.13a) discretization corresponding to (3.67) is of lower accuracy than the underlying discretization of the state equation, the discretization error is typically limited by this lower accuracy.

Discretization error analysis for Galerkin finite-element discretizations (3.61), (3.62), (3.68) of the optimal control model problem (3.10) are given, for example, in the book by Hinze *et al.* (2009, Chapter 3). For example, for the discretization of the optimal control problem (3.14) with $D_o = D$ using piecewise linear finite elements, which satisfy Assumption 3.1 of Hinze *et al.* (2009) for the state and the control, Hinze *et al.* (2009, Theorem 3.5) prove the following estimate. Suppose $u \in L^2(D)$ solves the optimal control problem (3.14), $y(u) \in H_0^1(D)$ is the corresponding state, $u_h \in \mathcal{U}_h \subset L^2(D)$ solves the finite-element discretization of the optimal control problem, and $y_h(u_h) \in \mathcal{V}_h \subset H_0^1(D)$ is the corresponding solution of the finite-element discretization of the state equation. Then there exists $c > 0$ such that

$$\|u - u_h\|_{L^2(D)} + h\|y(u) - y_h(u_h)\|_{H_0^1(D)} \leq ch^2(\|y(u)\|_{L^2(D)} + \|u\|_{L^2(D)}). \quad (3.69)$$

In the optimize-then-discretize approach applied to the optimal control model problem (3.10), we approximate each equation in the optimality system (3.13) individually. This leads to a system

$$\begin{pmatrix} \widetilde{\mathbf{Q}} & \mathbf{0} & \widetilde{\mathbf{A}}^\top \\ \mathbf{0} & \mathbf{R} & \mathbf{B}^\top \\ \mathbf{A} & \mathbf{B} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{u} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} -\mathbf{c} \\ \mathbf{0} \\ \mathbf{b} \end{pmatrix}, \quad (3.70)$$

where, in general, $\mathbf{A}^\top \neq \widetilde{\mathbf{A}}^\top$ and $\widetilde{\mathbf{Q}}$ may not be symmetric. This is, for example, the case when we apply streamline upwind/Petrov Galerkin (SUPG) stabilized finite elements to an optimal control problem (3.10) governed by an advection–diffusion equation (see e.g. Collis and Heinkenschloss 2002, Heinkenschloss and Leykekhman 2010, or Quarteroni 2009, Section 16.13), or we apply some discontinuous Galerkin methods (see e.g. Leykekhman 2012 and Leykekhman and Heinkenschloss 2012). The system matrix in (3.70) is invertible, provided the conditions (3.2) and (3.9) are satisfied for the optimal control problem and the discretization is sufficiently fine. However, since the system matrix in (3.70) is not symmetric, the iterative solution of (3.70) may be more expensive than the solution of (3.66) given the same discretization sizes. Because each equation in the optimality system (3.13) is discretized individually, these individual discretizations determine the overall discretization error. In particular, the optimize-then-discretize approach tends to have better approximation properties. However, the price is the lack of symmetry of (3.70).

The optimize-then-discretize approach also makes a difference when control constraints are present. In this case the equation corresponding to (3.13b) suggests

a discretization of the controls u that is derived from the discretization of the adjoint λ and the control constraints. Hinze (2005) shows that this implied control discretization, termed variational discretization, leads to a control discretization error of higher order. See also Hinze *et al.* (2009, Chapter 3).

For the linear–quadratic model problem (3.10), the optimality conditions (3.13) are necessary and sufficient and are a system of linear equations. This is no longer true for general nonlinear problems, and the numerical solution of optimal control problems using optimize-then-discretize becomes more involved. Examples are discussed by Gunzburger (2003, Chapter 4).

Ideally, we choose a discretization for which the discretize-then-optimize and optimize-then-discretize approaches commute. As shown above, this is the case for a standard Galerkin discretization (3.61), (3.62), (3.68). It is also true for several discontinuous Galerkin methods; see e.g. Leykekhman (2012).

Finally, adaptive approximations have been investigated where the problem discretization is not fixed, but is refined based on estimates of the discretization error between the solution of the current discretized problem and the true solution. Such adaptive discretizations are discussed in Becker *et al.* (2007), Becker, Kapp and Rannacher (2000), Dedé and Quarteroni (2005) and Hintermüller and Hoppe (2008), for example.

3.5.2. Discretization of the problem with random parameters

A numerical solution of the optimization problems (3.35) and (3.51) requires discretization in $\mathcal{V}, \mathcal{U}$, but also in Ω or Ξ. We focus on the problem (3.51) and on the discretization in Ξ. Some of the approaches discussed in this section do not require a finite-noise assumption and could be applied to (3.35). In other cases, optimization problems can be approximated by problems with finite noise; see the discussion at the beginning of Section 3.3.

Given a control $u \in \mathcal{U}$, discretization methods for the state equation (3.49) are discussed in the overview paper by Gunzburger *et al.* (2014) and in the books by Xiu (2010) and Lord *et al.* (2014, Chapter 9). Specifically, multilevel and multi-fidelity Monte Carlo methods (see e.g. Giles 2015, Krumscheid and Nobile 2018, Teckentrup, Scheichl, Giles and Ullmann 2013, Peherstorfer, Willcox and Gunzburger 2016, 2018, Peherstorfer 2019), multilevel and multi-fidelity quasi-Monte Carlo methods (see e.g. Dick, Kuo and Sloan 2013, Kuo and Nuyens 2016, Kuo *et al.* 2017), sparse-grid methods (see e.g. Bungartz and Griebel 2004, Griebel 2006, Garcke and Griebel 2013) or polynomial expansions (see e.g. Xiu 2010, Eldred 2011) can be used to approximate the objective function in (3.51) for a given $u \in \mathcal{U}$. However, as in the case of deterministic optimal control problems, it is not sufficient for a discretization of (3.51) to only approximate the state equation (3.49) and the objective function: it also needs to approximate the adjoint equation (3.52a) and the equation (3.52b), which defines the gradient of the objective function in (3.51).

We also note that the expectation

$$\mathbb{E}[b(\psi, \lambda(\boldsymbol{\xi}), \boldsymbol{\xi})] = \int_{\Xi} b(\psi, \lambda(\xi), \xi) \, \mathrm{d}\mathbb{P}^{\xi}(\xi)$$

in (3.52b) couples the optimality system across $\xi \in \Xi$. This coupling arises because a deterministic control $u \in \mathcal{U}$ has to be computed as the solution of (3.51) before the uncertainty is revealed.

For a given control $u \in \mathcal{U}$, the objective function in (3.51) involves the expectation of the function

$$F = \mathcal{J} \circ S \colon \mathcal{U} \to L^1_{\mathbb{P}^{\xi}}(\Xi), \tag{3.71a}$$

$$u \mapsto [F(u)](\cdot) = \frac{1}{2} q(y(u;\cdot), y(u;\cdot), \cdot) + c(y(u;\cdot), \cdot), \tag{3.71b}$$

where $y(u;\xi)$ solves (3.49). Monte Carlo, quasi-Monte Carlo, multilevel/multi-fidelity Monte Carlo and sparse-grid methods approximate

$$\begin{aligned}
\mathbb{E}[F(u)] &= \int_{\Xi} [F(u)](\xi) \, \mathrm{d}\mathbb{P}^{\xi}(\xi) \\
&\approx \sum_{i=1}^{N} \zeta_i [F_{h_i}(u)](\xi_i) \\
&\coloneqq \sum_{i=1}^{N} \zeta_i \left(\frac{1}{2} q(y_{h_i}(u;\xi_i), y_{h_i}(u;\xi_i), \xi_i) + c(y_{h_i}(u;\xi_i), \xi_i) \right),
\end{aligned} \tag{3.72}$$

where $\zeta_i \in \mathbb{R}$ and $y_{h_i}(u;\xi_i)$ is an approximation of the solution of (3.49) at $\xi = \xi_i$. For example, if the Galerkin discretization (3.61) with $\mathcal{V}_h$ replaced by $\mathcal{V}_{h_i} = \operatorname{span}\{\varphi_1^{(i)}, \ldots, \varphi_{n_{y,i}}^{(i)}\} \subset \mathcal{V}$ is used for the state space, then $y_{h_i}(u;\xi_i) \in \mathcal{V}_{h_i}$ solves (3.62) with $u_h = u$. In general, the subspace $\mathcal{V}_{h_i} \subset \mathcal{V}$ can depend on the sample. If, in addition, the Galerkin discretization (3.61) is used for the control, then the algebraic representation of $y_{h_i}(u_h;\xi_i)$ at a control $u_h \in \mathcal{U}_h$ is given by

$$\mathbf{y}_i(\mathbf{u}) = \mathbf{A}_i^{-1}(\mathbf{b}_i - \mathbf{B}_i \mathbf{u}),$$

and the algebraic representation of (3.72) at a control $u_h \in \mathcal{U}_h$ is

$$\begin{aligned}
&\sum_{i=1}^{N} \zeta_i \left(\frac{1}{2} \mathbf{u}^{\top} \mathbf{B}_i^{\top} \mathbf{A}_i^{-\top} \mathbf{Q}_i \mathbf{A}_i^{-1} \mathbf{B}_i \, \mathbf{u} - (\mathbf{B}_i^{\top} \mathbf{A}_i^{-\top} (\mathbf{c}_i + \mathbf{Q}_i \mathbf{A}_i^{-1} \mathbf{b}_i))^{\top} \mathbf{u} \right) \\
&\quad + \sum_{i=1}^{N} \zeta_i \left(\frac{1}{2} \mathbf{b}_i^{\top} \mathbf{A}_i^{-\top} \mathbf{Q}_i \mathbf{A}_i^{-1} \mathbf{b}_i + \mathbf{c}_i^{\top} \mathbf{A}_i^{-1} \mathbf{b}_i \right).
\end{aligned} \tag{3.73}$$

Given a discretization $\mathcal{U}_h \subset \mathcal{U}$ of the control space, the discretization of the optimal control problem (3.51) corresponding to (3.72) is given by

$$\min_{u_h \in \mathcal{U}_h} \sum_{i=1}^{N} \zeta_i \left(\frac{1}{2} q(y_{h_i}(u_h;\xi_i), y_{h_i}(u_h;\xi_i), \xi_i) + c(y_{h_i}(u_h;\xi_i), \xi_i) \right) + \frac{1}{2} r(u_h, u_h). \tag{3.74}$$

If we use the Galerkin discretization (3.61), (3.62) with $\mathcal{V}_h$ replaced by

$$\mathcal{V}_{h_i} = \operatorname{span}\{\varphi_1^{(i)}, \ldots, \varphi_{n_{y,i}}^{(i)}\} \subset \mathcal{V},$$

the algebraic representation of (3.74) is given by

$$\begin{aligned} \min_{\mathbf{u} \in \mathbb{R}^{n_u}} & \sum_{i=1}^{N} \zeta_i \left(\frac{1}{2} \mathbf{u}^\top \mathbf{B}_i^\top \mathbf{A}_i^{-\top} \mathbf{Q}_i \mathbf{A}_i^{-1} \mathbf{B}_i \, \mathbf{u} - (\mathbf{B}_i^\top \mathbf{A}_i^{-\top} (\mathbf{c}_i + \mathbf{Q}_i \mathbf{A}_i^{-1} \mathbf{b}_i))^\top \mathbf{u} \right) \\ & + \frac{1}{2} \mathbf{u}^\top \mathbf{R} \, \mathbf{u} + \sum_{i=1}^{N} \zeta_i \left(\frac{1}{2} \mathbf{b}_i^\top \mathbf{A}_i^{-\top} \mathbf{Q}_i \mathbf{A}_i^{-1} \mathbf{b}_i + \mathbf{c}_i^\top \mathbf{A}_i^{-1} \mathbf{b}_i \right). \end{aligned} \tag{3.75}$$

The approximations (3.74), (3.75) are sample average approximations (SAA) of the optimal control problem (3.51). The difference between SAA for finite-dimensional stochastic optimization problems (see e.g. Shapiro *et al.* 2014, Chapter 5) is that an SAA of (3.51) also involves a discretization of the underlying state equation (3.49) for a given sample $\xi = \xi_i$. Typically, hierarchies of discretizations with different approximation properties and computational costs are available. More generally, different computational models of the state equation (3.49) may be available that yield different approximation fidelities at different computational costs, for example by approximating the physics or by applying reduced-order models. This is strategically used in the overall approximation (3.72) and typically leads to large gains in the computational efficiency of the solution of (3.51).

Monte Carlo and quasi-Monte Carlo methods. Monte Carlo and quasi-Monte Carlo methods are equal-weight quadrature methods, i.e. $\zeta_i = 1/N$. In the execution of the Monte Carlo method, we draw N i.i.d. sample points $\xi_1, \ldots, \xi_N \in \Xi$ and approximate

$$\begin{aligned} \mathbb{E}[F(u)] &\approx \mathbb{E}[F_h(u)] \\ &\approx \frac{1}{N} \sum_{i=1}^{N} [F_h(u)](\xi_i) \\ &:= \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{2} q(y_h(u;\xi_i), y_h(u;\xi_i), \xi_i) + c(y_h(u;\xi_i), \xi_i) \right), \end{aligned} \tag{3.76}$$

where $y_h(u;\xi_i)$ is an approximation of the solution of (3.49) at $\xi = \xi_i$. Here the same approximation is used for all samples. For example, if the Galerkin discretization (3.61) is used for the state space, then $y_h(u;\xi_i) \in \mathcal{V}_h$ solves (3.62)

with $u_h = u$. To analyse the Monte Carlo error, we consider the random variable $N^{-1}\sum_{i=1}^{N}[F_h(u)](\boldsymbol{\xi}_i)$, where $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N$ are i.i.d. random variables. The right-hand side in (3.76) is a realization of this random variable. The expected value of $N^{-1}\sum_{i=1}^{N}[F_h(u)](\boldsymbol{\xi}_i)$ is $\mathbb{E}[F_h(u)]$ and its variance is $\mathbb{V}[F_h(u)]/N$. The standard error estimates for the Monte Carlo method (Dick *et al.* 2013, Section 2.2, Gunzburger *et al.* 2014, Section 3.3.2) applied to (3.76) for a fixed $u \in \mathcal{U}$ give

$$\mathbb{E}\left[\left(\mathbb{E}[F_h(u)] - \frac{1}{N}\sum_{i=1}^{N}[F_h(u)](\boldsymbol{\xi}_i)\right)^2\right] \leq \frac{1}{N}\mathbb{V}[F_h(u)], \tag{3.77a}$$

and, by the Chebyshev inequality,

$$\mathbb{P}\left(\left|\mathbb{E}[F_h(u)] - \frac{1}{N}\sum_{i=1}^{N}[F_h(u)](\boldsymbol{\xi}_i)\right| \geq \epsilon\right) \leq \frac{\mathbb{V}[F_h(u)]}{\epsilon^2 N}, \tag{3.77b}$$

where $\mathbb{V}[F_h(u)]$ is the variance of $F_h(u)$. Estimates for the error between $F(u)$ and $F_h(u)$ can be used to derive analogous error bounds between $\mathbb{E}[F(u)]$ and

$$\frac{1}{N}\sum_{i=1}^{N}[F_h(u)](\boldsymbol{\xi}_i).$$

Unfortunately, error estimates of the type (3.77) for fixed $u \in \mathcal{U}$ are not sufficient to analyse the error between the solutions $u \in \mathcal{U}$ and $u_{h,N} \in \mathcal{U}_h$ of the optimal control problems (3.51) and (3.74). For finite-dimensional problems, the error between the solution of a stochastic optimization problem and its Monte Carlo SAA approximation is analysed in Shapiro *et al.* (2014, Chapter 5). Recent work by Milz (2023*a*,*b*,*c*), Milz and Ulbrich (2024) and Römisch and Surowiec (2024) provides analyses for classes of PDE-constrained optimization problems. For example, for a linear quadratic optimal control problem similar to the one in Example 3.15, and controls approximated by piecewise constant finite elements and states approximated by piecewise linear finite elements, Milz (2023*b*) proves that

$$\mathbb{E}\left[\|u_{h,N} - u\|^2_{L^2(D)}\right] \leq c_1 h^2 + c_2/N, \tag{3.78a}$$

and that for each $\delta \in (0,1)$, with probability $1-\delta$,

$$\|u_{h,N} - u\|_{L^2(D)} \leq \tilde{c}_1 h + \tilde{c}_2\sqrt{2\ln(2/\delta)/N}, \tag{3.78b}$$

where $u \in \mathcal{U}$ is the solution of (3.51) and $u_{h,N} \in \mathcal{U}_h$ is the solution of (3.74), and $c_1, c_2, \tilde{c}_1, \tilde{c}_2$ are deterministic problem-dependent parameters. We note that the finite-noise assumption is not needed for this analysis, and the Monte Carlo discretization can be applied to (3.35).

To improve on the $O(1/\sqrt{N})$ convergence rate of Monte Carlo methods, quasi-Monte Carlo methods generate quadrature points deterministically but still use equal weights $\zeta_i = 1/N$. Guth *et al.* (2021) use a finite-element discretization in

space and quasi-Monte Carlo rule in the random variables to obtain an SAA discretization (3.74) of a linear–quadratic optimal control problem similar to the one in Example 3.15. The quasi-Monte Carlo rule used is based on shifted rank-one lattice rules. It is assumed that the diffusivity (κ in Example 3.15) has a Karhunen–Loève-type expansion, and assumptions on the coefficient functions in this expansion are made that allow a finite-noise approximation by truncation. The quasi-Monte Carlo method is then applied to discretize the finite-noise approximation in the random variables and a finite element approximation is used to discretize in space. Error estimates between the solution $u \in \mathcal{U}$ of (3.51) and the solution $u_{h,N,M} \in \mathcal{U}_h$ of (3.74) (with a truncation of the expansion of the diffusion coefficient after M terms) are provided. The numerical results show a nearly $O(1/N)$ convergence with respect to the number of quadrature points.

Multilevel/multi-fidelity Monte Carlo. Monte Carlo error estimates of the type (3.77) depend on the variance $\mathbb{V}[F_h(u)]/N$ of the estimator $N^{-1}\sum_{i=1}^{N}[F_h(u)](\boldsymbol{\xi}_i)$. Multilevel/multi-fidelity Monte Carlo methods use models $F_{h_\ell}(u)$, $\ell = 0, \ldots, L$, of different levels/fidelities and of different computational costs to compute an unbiased estimator with lower variance than the Monte Carlo estimator using only the high-fidelity model $F_{h_0}(u)$, and with computational cost equal to the computational cost of the Monte Carlo estimator using only the high-fidelity model. We assume that $F_{h_0}(u)$ is the high fidelity model and $F_{h_L}(u)$ is the lowest fidelity model.

Next, we describe the multilevel Monte Carlo method and refer to the literature for multi-fidelity Monte Carlo methods. By linearity of the expected value,

$$\mathbb{E}[F_{h_0}(u)] = \mathbb{E}[F_{h_0}(u)] + \sum_{\ell=1}^{L} \mathbb{E}[F_{h_\ell}(u) - F_{h_{\ell-1}}(u)].$$

Estimators for the quantities on the right-hand side are

$$\frac{1}{N_0}\sum_{i=1}^{N_0}[F_{h_0}(u)](\boldsymbol{\xi}_{i,0}), \quad \frac{1}{N_\ell}\sum_{i=1}^{N_\ell}[F_{h_\ell}(u)](\boldsymbol{\xi}_{i,\ell}) - [F_{h_{\ell-1}}(u)](\boldsymbol{\xi}_{i,\ell}), \quad \ell = 1, \ldots, L,$$

where $\boldsymbol{\xi}_{1,0}, \ldots, \boldsymbol{\xi}_{N_0,0}, \ldots, \boldsymbol{\xi}_{1,L}, \ldots, \boldsymbol{\xi}_{N_L,L}$ are i.i.d. random variables. The estimator of $\mathbb{E}[F_{h_0}(u)]$ is

$$Q_{\mathbf{N}}[F_{h_0}(u)] \coloneqq \frac{1}{N_0}\sum_{i=1}^{N_0}[F_{h_0}(u)](\boldsymbol{\xi}_{i,0}) + \sum_{\ell=1}^{L}\frac{1}{N_\ell}\sum_{i=1}^{N_\ell}[F_{h_\ell}(u)](\boldsymbol{\xi}_{i,\ell}) - [F_{h_{\ell-1}}(u)](\boldsymbol{\xi}_{i,\ell}), \tag{3.79}$$

where $\mathbf{N} = (N_0, \ldots, N_L)$. The estimator (3.79) is unbiased, $\mathbb{E}[Q_{\mathbf{N}}[F_{h_0}(u)]] = \mathbb{E}[F_{h_0}(u)]$, and has variance

$$\mathbb{V}[Q_{\mathbf{N}}[F_{h_0}(u)]] = \frac{1}{N_0}\mathbb{V}[F_{h_0}(u)] + \sum_{\ell=1}^{L}\frac{1}{N_\ell}\mathbb{V}[F_{h_\ell}(u) - F_{h_{\ell-1}}(u)]. \tag{3.80}$$

From estimates of the variances $\mathbb{V}[F_{h_0}(u)]$, $\mathbb{V}[F_{h_\ell}(u) - F_{h_{\ell-1}}(u)]$, $\ell = 1, \dots, L$, which in the multilevel case can be obtained from estimates of the state solution $y_{h_\ell}(u, \xi)$ error at different discretization levels, and from costs of sampling the objective $F_{h_\ell}(u)$ at different discretization levels, one can determine sample sizes $\mathbf{N} = (N_0, \dots, N_L)$ so that the variance (3.80) is below a certain tolerance, and one can compute the corresponding cost of applying the estimator (3.79); for details, see e.g. the papers by Giles (2015), Gunzburger *et al.* (2014), Krumscheid and Nobile (2018) and Teckentrup *et al.* (2013).

The estimator (3.79) is of the form (3.72), but it involves positive and negative weights ζ_i. The same is true when multi-fidelity Monte Carlo methods are used. As a consequence, the discretized optimal control problem (3.74) obtained with a multilevel/multi-fidelity Monte Carlo method using samples chosen *a priori* and kept fixed for all u_h may not be convex. For these discretizations, there are currently no error estimates like (3.78) for single-level Monte Carlo.

Sparse grids. We assume that $\boldsymbol{\xi} = (\boldsymbol{\xi}^{(1)}, \dots, \boldsymbol{\xi}^{(M)})$ with $\boldsymbol{\xi}^{(m)} \colon \Omega \to \Xi_m \subseteq \mathbb{R}$ and $\Xi = \Xi_1 \times \cdots \times \Xi_M$, and that $\boldsymbol{\xi}$ has the joint Lebesgue density $\rho = \rho_1 \otimes \cdots \otimes \rho_M$ with $\rho_m \colon \Xi_m \to [0, \infty)$. Under this assumption, the expectation of $F(u)$ can be written as

$$\begin{aligned}
&\int_\Xi [F(u)](\xi) \, \mathrm{d}\mathbb{P}^{\boldsymbol{\xi}}(\xi) \\
&\quad = \int_\Xi \rho(\xi) [F(u)](\xi) \, \mathrm{d}\xi \\
&\quad = \int_{\Xi_1} \rho_1\big(\xi^{(1)}\big) \dots \int_{\Xi_M} \rho_M\big(\xi^{(M)}\big) [F(u)]\big(\xi^{(1)}, \dots, \xi^{(M)}\big) \, \mathrm{d}\xi^{(M)} \dots \, \mathrm{d}\xi^{(1)}.
\end{aligned} \tag{3.81}$$

In principle, the integrals over $\Xi = \Xi_1 \times \cdots \times \Xi_M$ in (3.81) can be approximated by tensor product quadrature rules based on one-dimensional quadrature. However, even for small $M > 1$, this naive approach leads to a huge number of quadrature points in Ξ.

Sparse-grid quadrature operators are constructed from one-dimensional quadrature operators. For $m = 1, \dots, M$, let $\{\mathbb{E}_m^i\}_{i \geq 1}$ denote a sequence of one-dimensional quadrature operators built on the quadrature points $\mathcal{N}_m^i \subset \Xi_m$ such that $\mathbb{E}_m^i$ is exact for polynomials of degree up to $d_m^i - 1$, where $\{d_m^i\}_{i=1}^\infty \subset \mathbb{N}$ is an increasing sequence, and

$$\mathbb{E}_m^i[X] \to \mathbb{E}_m[X] := \int_{\Xi_m} \rho_m(\xi^{(m)}) \, X(\xi^{(m)}) \, \mathrm{d}\xi^{(m)} \quad \text{as } i \to \infty$$

for sufficiently regular $X \in C^0_{\rho_m}(\Xi_m)$. Define the one-dimensional difference quadrature operators

$$\Delta_m^1 := \mathbb{E}_m^1 \quad \text{and} \quad \Delta_m^i := \mathbb{E}_m^i - \mathbb{E}_m^{i-1} \quad \text{for } i \geq 2.$$

The idea behind sparse grids is that for smooth functions, some of the increments Δ_m^i are small and can be dropped. To define the M-dimensional quadrature rule on $\Xi = \Xi_1 \times \cdots \times \Xi_M$, let $\mathbf{i} = (i_1, \ldots, i_M)$ be a multi-index and let $\mathcal{I} \subset \mathbb{N}_+^M$ be a finite multi-index set, where $\mathbb{N}_+ = \{1, 2, \ldots\}$. The general sparse-grid quadrature operator is defined as

$$\mathbb{E}_{\mathcal{I}} := \sum_{\mathbf{i} \in \mathcal{I}} \left(\Delta_1^{i_1} \otimes \cdots \otimes \Delta_M^{i_M} \right). \tag{3.82}$$

If

$$\mathcal{I} = \{(i_1, \ldots, i_M) \in \mathbb{N}_+ \mid i_m \leq L, m = 1, \ldots, M\},$$

then $\mathbb{E}_{\mathcal{I}} = \left(\mathbb{E}_1^L \otimes \cdots \otimes \mathbb{E}_M^L \right)$ is just a standard tensor product quadrature rule. However, for smooth functions, some of the increments Δ_m^i are small and can be dropped without substantially increasing the quadrature error. For example, a popular choice due to Smolyak (1963) is

$$\mathcal{I} = \left\{ (i_1, \ldots, i_M) \in \mathbb{N}_+ \;\middle|\; \sum_{m=1}^{M} i_m \leq L + M - 1 \right\}.$$

Gerstner and Griebel (2003) compute $\mathcal{I}$ adaptively.

Applied to (3.81), the sparse-grid approximation is

$$\int_{\Xi} \rho(\xi)[F(u)](\xi)\, \mathrm{d}\xi \approx \mathbb{E}_{\mathcal{I}}[F(u)] = \sum_{\mathbf{i} \in \mathcal{I}} \left(\Delta_1^{i_1} \otimes \cdots \otimes \Delta_M^{i_M} \right)[F(u)]. \tag{3.83}$$

Gerstner and Griebel (1998, Section 4.4) collect estimates for the error

$$\left| \int_{\Xi} \rho(\xi)[F(u)](\xi)\, \mathrm{d}\xi - \mathbb{E}_{\mathcal{I}}[F(u)] \right|$$

depending on the smoothness of the function $\xi \mapsto [F(u)](\xi)$ and on the one-dimensional quadrature used.

The quadrature rule in (3.83) is expressed via one-dimensional difference quadrature operators Δ_m^i. If the index set $\mathcal{I} \subset \mathbb{N}_+^M$ is admissible in the sense that for all $\mathbf{i} = (i_1, \ldots, i_M) \in \mathcal{I}$,

$$\mathbf{j} = (j_1, \ldots, j_M) \in \mathbb{N}_+^M \quad \text{and} \quad j_m \leq i_m \quad \text{for all } m = 1, \ldots, M \quad \Longrightarrow \quad \mathbf{j} \in \mathcal{I},$$

then we can use the so-called combination technique to write (3.83) as

$$\mathbb{E}_{\mathcal{I}}[F(u)] = \sum_{\mathbf{i} \in \mathcal{I}} \left(\Delta_1^{i_1} \otimes \cdots \otimes \Delta_M^{i_M} \right)[F(u)] = \sum_{i=1}^{N} \zeta_i [F_i(u)](\xi_i). \tag{3.84}$$

The sparse-grid quadrature points $\xi_1, \ldots, \xi_N \in \Xi \subset \mathbb{R}^M$ required to evaluate $\mathbb{E}_{\mathcal{I}}$ (the sparse grid associated with $\mathcal{I}$) and the sparse-grid quadrature weights $\zeta_1, \ldots, \zeta_N \in \mathbb{R}$ are computed from the original one-dimensional quadrature formulas; see e.g. Gerstner and Griebel (1998).

Some of the weights ζ_i in the sparse-grid approximation (3.84) are negative. As a consequence, the discretized optimal control problem (3.74) obtained with a sparse-grid approximation (3.84) and sparse grid chosen *a priori* and kept fixed for all u_h may not be convex. However, numerical experiments suggest that for sufficiently fine sparse grids, the discretized problem is convex. If the sparse-grid discretized problem has a solution, then Kouri (2012, Section 3.4.1, 2014, page 62) proves a bound for the error $\|u - u_{h,\mathcal{I}}\|_{L^2(D)}$ between the solution u of the infinite-dimensional optimal control problem (3.51) and the solution $u_{h,\mathcal{I}}$ of the sparse-grid discretized optimal control problem (3.74).

Optimal control combination technique. If the expectation is of the form (3.81) with small M, one can use standard tensor quadrature rules with positive weights. This leads to a discretized optimal control problem (3.74) with positive weights, and this problem has a unique solution. One can then consider different discretization levels in different dimensions, e.g. for the integrals $\int_{\Xi_m} \rho_m(\xi^{(m)}) \dots \, \mathrm{d}\xi^{(m)}$ in the different components of the vector of random variables. Each level requires the solution of an optimal control problem. The solution at the finest level can be written using telescoping sums of differences of controls. Some of these differences will be small and can be eliminated. Nobile and Vanzan (2024) use error and computational work estimates to determine which differences of controls to eliminate and write the sum of the remaining differences as weighted sums of controls, each obtained as the solution of discretized optimal control problem (3.74) with select tensor quadrature rules and spatial grids. It is important to note that here the combination technique is applied to sequences of optimal controls. This is different from sparse-grid approximations where sparse grids and recombinations are applied to a single optimal control problem.

Discretized optimal control problem. Once the samples $\xi_i \in \Xi$, the weights $\zeta_i \in \mathbb{R}$ and the state approximations $y_{h_i}(u_h; \xi_i)$ are chosen, the discretized SAA problem (3.74), a large-scale finite-dimensional quadratic optimization problem (3.75), has to be solved. Analogously to Theorem 3.19, we have the following result on the existence, uniqueness and characterization of the solution of (3.75). Note that unless all ζ_i are non-negative, symmetric positive semidefiniteness of the $\mathbf{Q}_i \in \mathbb{R}^{n_{y,i} \times n_{y,i}}$, $i = 1, \dots, N$, and symmetric positive definiteness of $\mathbf{R} \in \mathbb{R}^{n_u \times n_u}$ are not enough to ensure the positive definiteness of the Hessian $\mathbf{R} + \sum_{i=1}^N \zeta_i \mathbf{B}_i^\top \mathbf{A}_i^{-\top} \mathbf{Q}_i \mathbf{A}_i^{-1} \mathbf{B}_i$ in (3.75). Therefore positive definiteness of the Hessian needs to be assumed.

Theorem 3.20. Let the matrices $\mathbf{A}_i \in \mathbb{R}^{n_{y,i} \times n_{y,i}}$, $i = 1, \dots, N$, be invertible, let $\mathbf{R} \in \mathbb{R}^{n_u \times n_u}$ be symmetric positive definite, and let $\mathbf{R} + \sum_{i=1}^N \zeta_i \mathbf{B}_i^\top \mathbf{A}_i^{-\top} \mathbf{Q}_i \mathbf{A}_i^{-1} \mathbf{B}_i$ be symmetric positive definite. The problem (3.75) has a unique solution $\mathbf{u} \in \mathbb{R}^{n_u}$. Moreover, $\mathbf{u} \in \mathbb{R}^{n_u}$ solves (3.75) if and only if

$$\left(\sum_{i=1}^N \zeta_i \mathbf{B}_i^\top \mathbf{A}_i^{-\top} \mathbf{Q}_i \mathbf{A}_i^{-1} \mathbf{B}_i + \mathbf{R} \right) \mathbf{u} = \sum_{i=1}^N \zeta_i \mathbf{B}_i^\top \mathbf{A}_i^{-\top} \left(\mathbf{c}_i + \mathbf{Q}_i \mathbf{A}_i^{-1} \mathbf{b}_i \right) \tag{3.85}$$

and if and only if there exist

$$\vec{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, \quad \vec{\boldsymbol{\lambda}} = \begin{pmatrix} \boldsymbol{\lambda}_1 \\ \vdots \\ \boldsymbol{\lambda}_N \end{pmatrix},$$

such that

$$\begin{pmatrix} \vec{\mathbf{Q}} & \mathbf{0} & \vec{\mathbf{A}}^\top \\ \mathbf{0} & \mathbf{R} & \vec{\mathbf{B}}^\top \\ \vec{\mathbf{A}} & \vec{\mathbf{B}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \vec{\mathbf{y}} \\ \mathbf{u} \\ \vec{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} -\vec{\mathbf{c}} \\ \mathbf{0} \\ \vec{\mathbf{b}} \end{pmatrix}, \tag{3.86}$$

where

$$\vec{\mathbf{A}} = \begin{pmatrix} \zeta_1 \mathbf{A}_1 & & \\ & \ddots & \\ & & \zeta_N \mathbf{A}_N \end{pmatrix}, \quad \vec{\mathbf{Q}} = \begin{pmatrix} \zeta_1 \mathbf{Q}_1 & & \\ & \ddots & \\ & & \zeta_N \mathbf{Q}_N \end{pmatrix}, \quad \vec{\mathbf{B}} = \begin{pmatrix} \zeta_1 \mathbf{B}_1 \\ \vdots \\ \zeta_N \mathbf{B}_N \end{pmatrix}$$

and

$$\vec{\mathbf{b}} = \left(\zeta_1 \mathbf{b}_1^\top, \ldots, \zeta_N \mathbf{b}_N^\top\right)^\top, \quad \vec{\mathbf{c}} = \left(\zeta_1 \mathbf{c}_1^\top, \ldots, \zeta_N \mathbf{c}_N^\top\right)^\top.$$

The system (3.85) is the Schur complement of (3.86). Given $\mathbf{u} \in \mathbb{R}^{n_u}$, the $\mathbf{y}_i, \boldsymbol{\lambda}_i, \in \mathbb{R}^{n_{y,i}}$ in (3.86) represent Galerkin approximations of the solutions $y(u_h, \xi_i)$ and $\lambda(u_h, \xi_i)$ of (3.52a), (3.52c) at $\xi = \xi_i$, $u = u_h$, respectively.

Once a discretization is determined, the quadratic problem (3.74) or its equivalent representation (3.75) can be solved using the (preconditioned) conjugate gradient (CG) method, for example. If the Hessian $\mathbf{R} + \sum_{i=1}^N \zeta_i \mathbf{B}_i^\top \mathbf{A}_i^{-\top} \mathbf{Q}_i \mathbf{A}_i^{-1} \mathbf{B}_i$ has negative eigenvalues, the objective function in (3.75) is unbounded from below and (3.75) does not have a solution. In this case, the CG method will detect a direction $\mathbf{v}$ of negative curvature, $\mathbf{v}^\top \left(\mathbf{R} + \sum_{i=1}^N \zeta_i \mathbf{B}_i^\top \mathbf{A}_i^{-\top} \mathbf{Q}_i \mathbf{A}_i^{-1} \mathbf{B}_i\right) \mathbf{v} < 0$, and terminate, provided the CG stopping tolerance is sufficiently small.

If the assumptions of Theorem 3.20 hold, one can also solve the system (3.86) of the necessary and sufficient optimality conditions. The matrix in (3.86) has a similar structure to the matrix in (3.66), except that $\vec{\mathbf{Q}}$ is only positive semidefinite if all weights ζ_i are non-negative and all matrices $\mathbf{Q}_i \in \mathbb{R}^{n_{y,i} \times n_{y,i}}$, $i = 1, \ldots, N$, are symmetric positive semidefinite. Block preconditioned Krylov subspace methods developed for (3.66) can be extended to (3.86). Examples of such a preconditioned Krylov subspace approach are given in Kouri and Ridzal (2018, Section 5.2) and Nobile and Vanzan (2023). Ciaramella, Nobile and Vanzan (2024) analyse a multigrid algorithm to solve (3.86), where the number of samples N is fixed and a sequence of spatial meshes are used to discretize the problem in $\mathcal{U}, \mathcal{V}$.

The discretization approaches discussed in this section fall into the discretize-then-optimize category. Optimize-then-discretize approaches have been used as well, and we will discuss these together with optimization methods in the next section.

3.6. *Optimization*

We describe optimization methods for the solution of the optimal control problem (3.51). However, many of these methods could also be applied to (3.35). The optimal control problem (3.51) is of the form

$$\min_{u\in\mathcal{U}} \mathbb{E}[F(u)] + \wp(u), \tag{3.87}$$

where $F = \mathcal{J} \circ S$ is the composition of the control-to-state map $S\colon \mathcal{U} \to L^2_{\mathbb{P}^\xi}(\Xi, \mathcal{V})$ and the objective function map $\mathcal{J}\colon L^2_{\mathbb{P}^\xi}(\Xi, \mathcal{V}) \to L^1_{\mathbb{P}^\xi}(\Xi)$ defined analogously to (3.36) and (3.38), respectively, and $\wp(u) = \frac{1}{2}r(u, u)$. Under the assumptions of Theorem 3.18, the objective function $f(\cdot) := \mathbb{E}[F(\cdot)] + \wp(\cdot)\colon \mathcal{U} \to \mathbb{R}$ in (3.87) is Fréchet-differentiable.

Optimization methods for the solution of the optimal control problem (3.51) can be roughly classified into two classes: approximation methods and stochastic approximation methods. Approximation methods are based on discretizations of the optimal control problem, and range from discretize-then-optimize approaches to optimize-then-discretize approaches. Typically, different discretizations, both for $\mathcal{V}$ and $\mathcal{U}$ and for the expectation, are integrated for computational efficiency. Stochastic approximation methods use the fact that computationally inexpensive estimators for the gradient of the objective function are available, and use them in a stochastic gradient method. The gradient of the objective function in (3.51) involves the computation of an expectation, and in a stochastic gradient method, a sample-based approximation (often using only one sample) is used. A numerical realization of the stochastic gradient method also involves a discretization in $\mathcal{V}$, which leads to a bias in the final gradient estimator – a bias that can be controlled by adjusting the discretization level.

Approximation methods. In the discretize-then-optimize approach, the optimal control problem (3.51) is discretized and approximated by (3.74), and the discretized problem is solved. Often, however, the discretization level needed to ensure that the solution of (3.74) is a good enough approximation of the solution of (3.51) is not known *a priori*. Once an approximate solution of (3.74) is computed, its quality is assessed. If needed, the discretization is refined to obtain a better approximation of the solution of (3.51). Even if an appropriate approximation can be determined *a priori*, different levels of coarser discretizations can be used to compute approximate solutions at lower computational cost or to approximately solve subproblems in the optimization algorithm applied to the fine discretization. Approaches that use different spatial (or spatial and temporal) discretization levels have been used successfully in deterministic PDE-constrained optimization. See e.g. the papers by Nash (2000), Lewis and Nash (2005), Gratton, Sartenaer and Toint (2008), Ziems and Ulbrich (2011) and Borzì and Schulz (2012), which have been extended to the solution of optimal control problems such as (3.51). Specifically, Kouri (2014) uses sparse-grid stochastic collocation with different levels to

extend the MG/OPT approach from Nash (2000) and Lewis and Nash (2005). Here levels are defined in terms of the sparse-grid quadrature, while the discretization for $\mathcal{V}$ and $\mathcal{U}$ is constant. Van Barel and Vandewalle (2021) use an MG/OPT-type algorithm based on the multilevel Monte Carlo estimator (3.79) of $\mathbb{E}[F_{h_0}(u)]$. Specifically, the spatial discretizations and samples used in the multilevel Monte Carlo estimator (3.79) of the fine-level objective $\mathbb{E}[F_{h_0}(u)]$ is computed so that the gradient has a desired root-mean-square error. These spatial discretizations and samples are then used as the fine-level discretization of the optimal control problem The coarser-level approximations of this fine-level discretization are also of the type (3.79), but are constructed by limiting the highest level, i.e. using levels $\ell = k, \dots, L$ (recall that in (3.79) $\ell = 0$ is the finest level and $\ell = L$ the coarsest), and using only a fraction $q^k N_\ell$, $\ell = k, \dots, L$, $q \in (0, 1)$, of the samples used in fine-level discretization (3.79).

The previous approaches solve a discretization or a sequence of discretizations of the optimal control problem and are therefore discretize-then-optimize approaches. Alternatively, we can also use optimize-then-discretize approaches. One way to do this is to use trust-region or line-search optimization algorithms, such as those described in Kouri, Heinkenschloss, Ridzal and van Bloemen Waanders (2013, 2014) and Grundvig and Heinkenschloss (2024), which employ objective function models with tunable accuracy. Given an iterate u_k, these optimization algorithms generate a model m_k of the objective function (3.87). The model fidelity is adjusted based on the progress of the optimization algorithm. These trust-region and line-search optimization algorithms only use the objective function models m_k, but generate iterates so that $\liminf_{k\to\infty} \|\nabla f(u_k)\|_{\mathcal{U}} = 0$ or $\lim_{k\to\infty} \|\nabla f(u_k)\|_{\mathcal{U}} = 0$. We will provide details of a trust-region algorithm in Section 5.1.1. In the context of solving the risk-neutral PDE-constrained optimization problem under uncertainty (3.87), the model m_k is generated by a discretization of the objective function at u_k, as discussed in Section 3.5.2.

Stochastic approximation. Under the assumptions of Theorem 3.18, the objective function in (3.87) is Fréchet-differentiable, and the derivative applied to a direction $\delta u \in \mathcal{U}$ is $\int_\Xi b(\delta u, \lambda(\xi), \xi)\, \mathrm{d}\mathbb{P}^\xi(\xi) + r(u, \delta u)$. The gradient of the objective function in (3.87), $\nabla\mathbb{E}[F(u)] + \nabla\wp(u)$, is the Riesz representation of the derivative, and is defined by

$$\int_\Xi b(\delta u, \lambda(\xi), \xi)\, \mathrm{d}\mathbb{P}^\xi(\xi) + r(u, \delta u) = \langle \nabla\mathbb{E}[F(u)] + \nabla\wp(u), \delta u\rangle_{\mathcal{U}} \quad \text{for all } \delta u \in \mathcal{U}. \tag{3.88}$$

A gradient method applied to (3.87) generates a sequence of iterates

$$u_{k+1} = u_k - \gamma_k \nabla f(u_k) \tag{3.89}$$

with a suitable step size $\gamma_k > 0$, typically chosen such that $f(u_{k+1})$ is sufficiently smaller than $f(u_k)$. Specifically,

$$f(u_{k+1}) \le f(u_k) + c\gamma_k \|\nabla f(u_k)\|_{\mathcal{U}}^2 \tag{3.90}$$

for some $c \in (0, 1)$, typically $c = 10^{-4}$; see e.g. the books by Dennis Jr and Schnabel (1996, Chapter 6) and Nocedal and Wright (2006, Chapter 3).

One issue when using a gradient method for the solution of (3.87) is that each objective function evaluation and each gradient evaluation involves an expectation, that is, it requires integration over all ξ, which is expensive. In their basic form, stochastic gradient methods avoid this expense by replacing the gradient $\nabla f(u_k)$ with a stochastic function $G(u_k, \xi_k)$ such that $\mathbb{E}[G(u_k, \xi_k)] \approx \nabla \mathbb{E}[F(u)]$ and then generate a sequence of iterates

$$u_{k+1} = u_k - \gamma_k (G(u_k, \xi_k) + \nabla \wp(u_k)) \tag{3.91}$$

with step size $\gamma_k > 0$. Since $G(u_k, \xi_k)$ is a stochastic function, the iterates u_k are now random variables. The stochastic gradient method does not access the function f. In particular, there is no monotonicity in function values like (3.90). We will provide details in Section 5.2. In practice, $G(u_k, \xi_k)$ is computed by replacing the expected value in (3.88) by just using one randomly selected sample, that is,

$$b(\delta u, \lambda(\xi_k), \xi_k) + r(u_k, \delta u) = \langle G(u_k, \xi_k) + \nabla \wp(u_k), \delta u \rangle_{\mathcal{U}} \quad \text{for all } \delta u \in \mathcal{U}, \tag{3.92}$$

or by using the mean of a small number of randomly selected samples.

Geiersbach and Wollner (2020) analysed and applied a stochastic gradient method to solve problems like the one in Example 3.15 (with control constraints) using an adaptive finite-element discretization to solve state and adjoint equations. For implementation, the stochastic gradient method (3.91) with (3.92) still requires discretizations in $\mathcal{V}$ and $\mathcal{U}$, and this discretization introduces a bias in the discretized stochastic gradient (additional bias is introduced if the discretized state and adjoint equations needed to compute the discrete version of $G(u_k, \xi_k)$ are solved iteratively and therefore inexactly). This bias is controlled by adapting the finite-element discretization. Martin, Krumscheid and Nobile (2021) provide convergence and complexity results of the stochastic gradient method with minibatch and constant as well as variable mesh-size finite-element discretizations applied to the problem essentially equal to the one in Example 3.15. Van Barel and Vandewalle (2019) use multilevel Monte Carlo methods to compute estimates of the gradient, and use them in a nonlinear conjugate gradient method to solve a problem similar to the one in Example 3.15 but with an additional variance term in the objective. This is not a stochastic gradient method and no convergence analysis exists, but this method can be seen as a stochastic approximation method if samples used in the multilevel Monte Carlo methods are varied between iterations.

We will discuss stochastic methods more broadly in Section 5.2.

3.7. Extensions

We have used a linear–quadratic elliptic optimal control problem to discuss the risk-neutral optimization problem formulation, study the existence and characterization of its solution, and survey numerical methods for its solution. The solution of

PDE-constrained optimization problems under uncertainty quickly becomes even more involved when aspects of the problem are changed.

If the state equation is not linear in the states and controls, a theory for existence and uniqueness of the state equation has to be developed, together with regularity and integrability of the control-to-state map S that allows the application of the expected value or of risk measures. Moreover, the Fréchet differentiability of the control-to-state map has to be proved. These investigations are guided by the theory of the corresponding deterministic problem with state space $\mathcal{V}$, but, as we have already seen in the case of the model problem, additional assumptions are needed for the analysis in a state space $L^q_{\mathbb{P}}(\Omega, \mathcal{V})$ for some $q \geq 1$. Moreover, if the state equation is not linear in states and controls, the control-to-state map S is nonlinear. As a consequence, the function $F = \mathcal{J} \circ S$ arising in the objective function is no longer convex, and therefore the optimization problem is no longer convex. This impacts optimization algorithms and discretization error estimates, even in the deterministic case.

In this section we have only considered the risk-neutral formulation of the model problem, in which the expected value of $F = \mathcal{J} \circ S$ is minimized together with a penalty of the control (see (3.40)). As seen in the examples in Section 2, other risk measures such as AVaR lead to better results. However, many such risk measures are non-differentiable, which impacts solutions of these risk-averse problems. Instead of using risk measures, one can consider other formulations to include uncertainty in the optimization formulation, such as chance constraints or dominance constraints. However, these lead to additional challenges in the analysis of the resulting problem and its numerical solution.

If the underlying deterministic optimal control problem has state-dependent constraints, then in the extension to PDEs with uncertain parameters, these state-dependent constraints will be parametrized by the random parameter. We have to model how the resulting constraints will be incorporated into the optimization formulation. Should the state constraints hold in expectation, should they hold almost surely, or should some risk measure be applied? Again, the formulation impacts the analysis of the resulting problem formulation and its solution.

In the following sections we will discuss some of these extensions.

4. Problem formulation

Let $\mathcal{U}_{\text{ad}} \subseteq \mathcal{U}$ encapsulate deterministic constraints on the optimization variables. The problems described in Section 2 can be formulated as the optimization problem

$$\min \quad \mathcal{R}(F(u)) + \wp(u), \tag{4.1a}$$

$$\text{subject to } G(u) \leq B, \tag{4.1b}$$

$$u \in \mathcal{U}_{\text{ad}}, \tag{4.1c}$$

where $F\colon \mathcal{U} \to L^p_{\mathbb{P}}(\Omega)$ is an uncertain cost function, $\wp\colon \mathcal{U} \to \mathbb{R}$ is a deterministic cost function, $G\colon \mathcal{U} \to L^r_{\mathbb{P}}(\Omega)$ is an uncertain constraint function, $B \in L^r_{\mathbb{P}}(\Omega)$

is a benchmark random variable, and $\mathcal{R}\colon L^p_{\mathbb{P}}(\Omega) \to \mathbb{R}$ is a risk measure. As in Section 3.2.2, in PDE-constrained optimization, the function F is given as the composition of a cost function $\mathcal{J}$ with the control-to-state map S, $F = \mathcal{J} \circ S$. The constraint function G may have a similar representation.

In order to make sense of (4.1), $\mathcal{R}$ must scalarize the random function F for example using risk measures that model the stakeholder's aversion to risk. In contrast, the random constraint $G(u) \leq B$ can be understood in various ways. For example, we can require that $G(u) \leq B$ holds almost surely (i.e. with probability one) or with some probability less than one. We could also employ stochastic orders to define this constraint.

4.1. Risk measures

When uncertainties are present in applications, the resulting random objective function no longer quantifies a loss but rather the possibility of loss, which is often referred to as *risk*. In this setting, a critical question arises: How do we reformulate our objective function to (perhaps, conservatively) quantify this risk? If the uncertainties are well characterized, we can employ functionals called risk measures. Commonly used in financial mathematics, risk measures are statistics that inherit the units of the underlying random objective function and are used to model risk preference. We will discuss risk measures applied to a generic random variable $X \in L^p_{\mathbb{P}}(\Omega)$. We sometimes refer to X as a cost or a loss, but, as in earlier sections, X can be a scalar quantification of the under-performance of a system and may not represent a monetary cost or loss.

Coherent risk measures, introduced by Artzner, Delbaen, Eber and Heath (1999), have garnered much recent attention in PDE-constrained optimization. A function $\mathcal{R}\colon L^p_{\mathbb{P}}(\Omega) \to \mathbb{R}$ is a coherent risk measure if it satisfies the following properties. For all X, $X' \in L^p_{\mathbb{P}}(\Omega)$ and $t \in \mathbb{R}$:

(R1) *Subadditivity.* $\mathcal{R}(X + X') \leq \mathcal{R}(X) + \mathcal{R}(X')$.

(R2) *Monotonicity.* If $X \leq X'$ a.s., then $\mathcal{R}(X) \leq \mathcal{R}(X')$.

(R3) *Translation equivariance.* $\mathcal{R}(X + t) = \mathcal{R}(X) + t$.

(R4) *Positive homogeneity.* If $t \geq 0$, then $\mathcal{R}(tX) = t\mathcal{R}(X)$.

Conditions (R1) and (R4) are equivalent to $\mathcal{R}$ being convex and positively homogeneous. Consequently, (R1) is often replaced by:

(R1′) *Convexity.* If $t \in [0, 1]$, then $\mathcal{R}(tX + (1 - t)X') \leq t\mathcal{R}(X) + (1 - t)\mathcal{R}(X')$.

In financial applications, subadditivity (R1) incentivizes diversification or decentralization of risk. For general applications, monotonicity (R2) is a natural condition ensuring that risk is decreased if the overall loss is decreased. When combined with (R4), translation equivariance (R3) ensures that deterministic quantities are *risk-free* (i.e. $\mathcal{R}(t) = t$ for all $t \in \mathbb{R}$). Finally, positive homogeneity (R4) enables consistent change of units.

Föllmer and Schied (2010) refer to $\mathcal{R}\colon L^p_{\mathbb{P}}(\Omega) \to \mathbb{R}$ that satisfies (R2) and (R3) as a *monetary* risk measure, and a monetary risk measure that satisfies (R1′) as a *convex* risk measure. A basic approach to constructing risk measures is through *acceptance sets*, as in Föllmer and Schied (2002). Let $\mathcal{A} \subseteq L^p_{\mathbb{P}}(\Omega)$ denote a set of acceptable random outcomes. For a random variable $X \in L^p_{\mathbb{P}}(\Omega)$, the risk measure $\mathcal{R}_{\mathcal{A}}$ associated with the acceptance set $\mathcal{A}$ is defined as the smallest value m such that $X - m$ is acceptable, that is,

$$\mathcal{R}_{\mathcal{A}}(X) = \inf\{m \in \mathbb{R} \mid X - m \in \mathcal{A}\}.$$

If the acceptance set $\mathcal{A}$ satisfies

$$\mathcal{A} \cap \mathbb{R} \neq \emptyset, \tag{4.2a}$$

$$\inf\{m \in \mathbb{R} \mid X - m \in \mathcal{A}\} > -\infty \quad \text{for all } X \in L^p_{\mathbb{P}}(\Omega), \tag{4.2b}$$

$$X \in \mathcal{A},\ X' \in L^p_{\mathbb{P}}(\Omega),\ X' \geq X \text{ a.s.} \quad \Longrightarrow \quad X' \in \mathcal{A}, \tag{4.2c}$$

then $\mathcal{R}_{\mathcal{A}}$ is a monetary risk measure. In addition, if $\mathcal{A}$ is convex, then $\mathcal{R}_{\mathcal{A}}$ is a convex risk measure and if $\mathcal{A}$ is a convex cone, then $\mathcal{R}_{\mathcal{A}}$ is coherent. Furthermore, a risk measure $\mathcal{R}$ is coherent if and only if there exists $\mathcal{R} = \mathcal{R}_{\mathcal{A}}$, where $\mathcal{A}$ is a convex cone satisfying (4.2). In this case,

$$\mathcal{A} = \{X \in L^p_{\mathbb{P}}(\Omega) \mid \mathcal{R}(X) \leq 0\}.$$

Ben-Tal and Teboulle (2007) introduce another useful approach for constructing risk measures based on expected utility theory called *optimized certainty equivalents*. Optimized certainty equivalent risk measures have the form

$$\mathcal{R}(X) = \inf_{t \in \mathbb{R}} \{t - \mathbb{U}(t - X)\}, \tag{4.3}$$

where $\mathbb{U}\colon L^p_{\mathbb{P}}(\Omega) \to \mathbb{R}$ is an *expected utility function*; see von Neumann and Morgenstern (2007). Commonly, $\mathbb{U}$ is concave, monotonic and satisfies $\mathbb{U}(0) = 0$.

Coherent risk measures satisfy numerous desirable properties. For example, (R2) and (R1′) ensure that $\mathcal{R}$ is continuous (see Shapiro *et al.* 2014, Proposition 6.6), and consequently the Fenchel–Moreau theorem (see e.g. Ekeland and Temam 1999, Proposition 4) ensures that

$$\mathcal{R}(X) = \sup_{\theta \in L^q_{\mathbb{P}}(\Omega)} \{\mathbb{E}[\theta X] - \mathcal{R}^*(\theta)\}, \tag{4.4}$$

where $1/p + 1/q = 1$ and $\mathcal{R}^*\colon L^q_{\mathbb{P}}(\Omega) \to (-\infty, +\infty]$ is the Legendre–Fenchel transformation of $\mathcal{R}$, that is,

$$\mathcal{R}^*(\theta) = \sup_{X \in L^p_{\mathbb{P}}(\Omega)} \{\mathbb{E}[\theta X] - \mathcal{R}(X)\}.$$

Recall that $\mathcal{R}^*$ is proper, lower semicontinuous and convex (see the discussion in Ekeland and Temam 1999, page 18), and so its effective domain

$$\operatorname{dom} \mathcal{R}^* = \{\theta \in L^q_{\mathbb{P}}(\Omega) \mid \mathcal{R}^*(\theta) < +\infty\}$$

is non-empty, closed and convex. In addition, if (R2) holds, then $\theta \in \operatorname{dom} \mathcal{R}^*$ if and only if $\theta \geq 0$ a.s., and if (R3) holds, then $\theta \in \operatorname{dom} \mathcal{R}^*$ if and only if $\mathbb{E}[\theta] = 1$. In particular, if $\mathcal{R}$ is a convex risk measure, then $\operatorname{dom} \mathcal{R}^*$ consists of probability density functions in $L^q_{\mathbb{P}}(\Omega)$. Finally, when (R4) holds, $\mathcal{R}^*(\theta) = 0$ for all $\theta \in \operatorname{dom} \mathcal{R}^*$. In fact, $\mathcal{R}$ is a coherent risk measure if and only if there exists a non-empty, closed and convex subset $\mathfrak{A} \subseteq L^q_{\mathbb{P}}(\Omega)$, sometimes called the *risk envelope*, satisfying

$$\theta \in \mathfrak{A} \quad \Longrightarrow \quad \theta \geq 0 \ \text{ a.s.} \quad \text{and} \quad \mathbb{E}[\theta] = 1$$

such that

$$\mathcal{R}(X) = \sup_{\theta \in \mathfrak{A}} \mathbb{E}[\theta X], \tag{4.5}$$

in which case $\mathfrak{A} = \operatorname{dom} \mathcal{R}^*$; see Theorems 6.5 and 6.7 in Shapiro *et al.* (2014) for more details.

Perhaps a more fundamental property for risk measures is *law invariance*. Two random variables are *distributionally equivalent*, denoted $X \overset{D}{\sim} X'$, if their CDFs are equal, i.e. $\Psi_X(t) = \Psi_{X'}(t)$ for all $t \in \mathbb{R}$. Using this definition, a risk measure $\mathcal{R}$ is *law-invariant* if

$$X \overset{D}{\sim} X' \quad \Longrightarrow \quad \mathcal{R}(X) = \mathcal{R}(X') \tag{4.6}$$

for any two random variables X, $X' \in L^p_{\mathbb{P}}(\Omega)$. In particular, a law-invariant risk measure is a function of the distribution, not the values, of a random variable. A fundamental law-invariant, coherent risk measure is the average value-at-risk[2] (AVaR) (see e.g. Rockafellar and Uryasev 2002, Uryasev and Rockafellar 2001), which is defined for a random variable $X \in L^p_{\mathbb{P}}(\Omega)$ as

$$\operatorname{AVaR}_\beta(X) := \frac{1}{1-\beta} \int_\beta^1 q_X(s)\, \mathrm{d}s \tag{4.7}$$

for fixed confidence level $\beta \in [0, 1)$, where $q_X(\cdot)$ is the quantile function of the random variable X.[3] For continuous random variables X, $\operatorname{AVaR}_\beta(X)$ is the average of the $(1-\beta) \times 100\%$ largest outcomes of X. See also Figure 2.5. Moreover, AVaR is a fundamental building block of law-invariant, coherent risk measures. In particular, if $(\Omega, \mathcal{F}, \mathbb{P})$ is non-atomic and $\mathcal{R}$ is law-invariant and coherent, then there exists a set of probability measures $\mathfrak{M}$ on $[0, 1)$ for which

$$\mathcal{R}(X) = \sup_{\mu \in \mathfrak{M}} \int_0^1 \operatorname{AVaR}_{1-s}(X)\, \mathrm{d}\mu(s). \tag{4.8}$$

Equation (4.8) is referred to as the Kusuoka representation of $\mathcal{R}$.

[2] Also called conditional value-at-risk, expected shortfall, expected tail loss and superquantile.
[3] The quantile function is sometimes called the *value-at-risk*.

Example 4.1 (average value-at-risk). As we have already discussed, AVaR (4.7) is a law-invariant, coherent risk measure. As shown in Rockafellar and Uryasev (2002), AVaR can be equivalently written as

$$\mathcal{R}(X) = \text{AVaR}_\beta(X) = \inf_{t\in\mathbb{R}}\left\{t + \frac{1}{1-\beta}\mathbb{E}[\max\{0, X-t\}]\right\} \tag{4.9}$$

for $\beta \in (0,1)$. The representation (4.9) is particularly useful in optimization, and this representation of AVaR was used in optimization problems discussed in Sections 2.1 and 2.3. Specifically, if $X = F(u)$, then $\min_{u\in\mathcal{U}_{\text{ad}}} \text{AVaR}_\beta(F(u)) + \wp(u)$ is equivalent to

$$\min_{t\in\mathbb{R},\, u\in\mathcal{U}_{\text{ad}}}\left\{t + \frac{1}{1-\beta}\mathbb{E}[\max\{0, F(u)-t\}] + \wp(u)\right\}. \tag{4.10}$$

It is worth noting that AVaR at $\beta = 0$ is $\text{AVaR}_0(X) = \mathbb{E}[X]$, and its limit as $\beta \to 1$ is

$$\lim_{\beta\uparrow 1} \text{AVaR}_\beta(X) = \operatorname{ess\,sup} X.$$

For a continuously distributed random variable X, $\text{AVaR}_\beta(X)$ is the expectation of X conditioned on the event

$$\{\omega \in \Omega \mid X(\omega) > q_X(\beta)\}.$$

Informally, $\text{AVaR}_\beta(X)$ is the β-tail average of X. Since AVaR is coherent, it has the biconjugate representation (4.5) with risk envelope

$$\operatorname{dom}\mathcal{R}^* = \{\theta \in L^\infty_{\mathbb{P}}(\Omega) \mid \mathbb{E}[\theta] = 1,\ 0 \le \theta \le (1-\alpha)^{-1} \text{ a.s.}\}.$$

Krokhmal (2007) extended AVaR to account for higher-order tail moments, introducing the higher-moment coherent risk measure (HMCR),

$$\mathcal{R}(X) = \inf_{t\in\mathbb{R}}\left\{t + \frac{1}{1-\beta}\mathbb{E}[\max\{0, X-t\}^p]^{1/p}\right\},$$

with $p \in (1,\infty)$. Like AVaR, HMCR is a law-invariant coherent risk measure and has the biconjugate representation (4.5) with risk envelope

$$\operatorname{dom}\mathcal{R}^* = \left\{\theta \in L^q_{\mathbb{P}}(\Omega) \mid \mathbb{E}[\theta] = 1,\ \theta \ge 0 \text{ a.s.},\ \|\theta\|_{L^q_{\mathbb{P}}(\Omega)} \le \frac{1}{1-\alpha}\right\},$$

where $1/p + 1/q = 1$. See Section 5.3.1 in Cheridito and Li (2009) for more information on HMCR.

Example 4.2 (mean-plus-deviation). The mean-plus-deviation risk measure, as used in the pioneering portfolio optimization work of Markowitz (1952), is commonly used in engineering applications because of its simple and intuitive nature. This risk measure, given by

$$\mathcal{R}(X) = \mathbb{E}[X] + c\mathbb{E}[|X - \mathbb{E}[X]|^p]^{1/p}, \quad c > 0$$

for $p \in [1, \infty)$, is law-invariant and satisfies (R1), (R3) and (R4), but is not monotonic. This can lead to paradoxical scenarios in which one value of the objective function is always smaller than the other but has larger risk. Example 6.62 in Shapiro *et al.* (2014) is a simple example of this situation. The lack of monotonicity results from the equal penalization of deviations above and below the expected value, which can be fixed using the mean-plus-semideviation risk measure, for example

$$\mathcal{R}(X) = \mathbb{E}[X] + c\mathbb{E}[\max\{0, X - \mathbb{E}[X]\}^p]^{1/p}, \quad c \in [0, 1].$$

The mean-plus-semideviation is monotonic and hence coherent, yielding the representation (4.5) with risk envelope

$$\operatorname{dom} \mathcal{R}^* = \{\theta \in L^q_{\mathbb{P}}(\Omega) \mid \theta = 1 + \theta' - \mathbb{E}[\theta'],\ \|\theta'\|_{L^q_{\mathbb{P}}(\Omega)} \leq c,\ \theta' \geq 0 \text{ a.s.}\},$$

where $1/p + 1/q = 1$.

Example 4.3 (entropic risk). The entropic risk measure is the *certainty equivalent* associated with the exponential utility function $v(t) = \exp(\sigma t)$, that is,

$$\mathcal{R}(X) = \sigma^{-1} \log(\mathbb{E}[\exp(\sigma X)]), \quad \sigma > 0. \tag{4.11}$$

In particular, if a stakeholder with utility function v is presented a random outcome X, then $\mathcal{R}(X) = v^{-1}(\mathbb{E}[v(X)])$ is the certain amount for which they remain unconcerned with the outcome. See Ben-Tal and Teboulle (2007) for more discussion of the various types of certainty equivalents. Interestingly, the entropic risk is also the optimized certainty equivalent (4.3) associated with the exponential utility function. The entropic risk is convex, monotonic and translation-equivariant, but is not positively homogeneous and therefore not coherent. The natural domain for the entropic risk measure is $L^\infty_{\mathbb{P}}(\Omega)$, which significantly complicates the analysis of $\mathcal{R}$ since its topological dual space is a space of measures. To overcome this challenge, we typically consider $L^\infty_{\mathbb{P}}(\Omega)$ equipped with the weak* topology paired with $L^1_{\mathbb{P}}(\Omega)$ equipped with the norm topology. See Section 6.3 of Shapiro *et al.* (2014) for a discussion of the theoretical challenges associated with essentially bounded random variables. The entropic risk measure was used in the optimal control of a thermally convected flow in Section 2.2.

We conclude this section with a discussion of risk-averse optimization problems. Let $\mathcal{U}$ be a reflexive Banach space, let $\mathcal{U}_{\text{ad}} \subseteq \mathcal{U}$, let $\mathcal{V}$ be a reflexive Banach space, and consider the problem

$$\min \quad \mathcal{R}(F(u)) + \wp(u), \tag{4.12a}$$

$$\text{subject to } u \in \mathcal{U}_{\text{ad}}, \tag{4.12b}$$

where, similar to (3.36)–(3.40), $F = \mathcal{J} \circ S$ is the composition of a control-to-state map

$$S \colon \mathcal{U} \to L^2_{\mathbb{P}}(\Omega, \mathcal{V}) \tag{4.13}$$

and an objective function map

$$\mathcal{J}\colon L^2_{\mathbb{P}}(\Omega, \mathcal{V}) \to L^p_{\mathbb{P}}(\Omega), \tag{4.14}$$

and where $\mathcal{R}\colon L^p_{\mathbb{P}}(\Omega) \to \mathbb{R}$ is a proper, closed, convex and monotonic risk measure, and $\wp\colon \mathcal{U} \to \mathbb{R}$ is proper, closed and convex. Kouri and Surowiec (2018) prove the existence of a solution of (4.12) and optimality conditions for (4.12) under fairly general assumptions on S, $\mathcal{J}$, $\mathcal{R}$ and $\wp$. Convexity of $\mathcal{R}$ and the relationship (4.5) play an important role in the optimality conditions. Additional results for PDE-constrained optimization problems governed by a semilinear elliptic PDE, i.e. where the control-to-state map (4.13) corresponds to the solution of a semilinear elliptic PDE, are given in Kouri and Surowiec (2020*b*). These results generalize those in Section 3.2 to semilinear elliptic PDE and proper, closed, convex and monotonic risk measures $\mathcal{R}$. Subdifferentials of $\mathcal{R}$ are used in the optimality conditions in Kouri and Surowiec (2018, 2020*b*), because, in contrast to the risk-neutral formulation discussed in Section 3, coherent risk measures are generally not Fréchet-differentiable. In fact, Kouri and Surowiec (2020*a*, Theorem 1) prove the following result.

Theorem 4.4. Let $\mathcal{R}\colon L^p_{\mathbb{P}}(\Omega) \to \mathbb{R}$ be a coherent risk measure with $p \in [1, \infty)$. Then $\mathcal{R}$ is Fréchet-differentiable if and only if there exists $\theta \in L^q_{\mathbb{P}}(\Omega)$ with $1/p + 1/q = 1$ such that $\theta \geq 0$, $\mathbb{E}[\theta] = 1$, and $\mathcal{R}(X) = \mathbb{E}[\theta X]$ for all $X \in L^p_{\mathbb{P}}(\Omega)$.

Consequently, only linear coherent risk measures are Fréchet-differentiable. The lack of Fréchet differentiability also complicates the numerical solution of (4.12) and (4.1). We will return to this issue in Section 5.1.3.

4.2. Distributionally robust

Distributionally robust optimization (DRO) models the situation when the underlying probabilistic characterization of the uncertainty – encapsulated by the probability measure $\mathbb{P}$ – is uncertain. For example, this is the case when certain coefficients of the PDE are estimated from data using robust Bayesian inference; see Berger (1994). To produce solutions that are resilient to this uncertainty, DRO minimizes the worst-case expectation over a set of probability measures $\mathfrak{P}$ defined on the measurable space $(\Omega, \mathcal{F})$. Relating this to (4.1), DRO employs functionals $\mathcal{R}$ given by

$$\mathcal{R}(X) = \sup_{P \in \mathfrak{P}} \int_{\Omega} X(\omega)\, \mathrm{d}P(\omega) = \sup_{P \in \mathfrak{P}} \mathbb{E}_P[X]. \tag{4.15}$$

In view of (4.5), DRO is a generalization of coherent risk-averse optimization. In particular, given a coherent risk measure and associated risk envelope $\mathfrak{A}$, the associated *uncertainty set* is

$$\mathfrak{P} = \{P\colon \mathcal{F} \to [0, 1] \mid \exists\, \theta \in \mathfrak{A} \text{ such that } \mathrm{d}P = \theta\, \mathrm{d}\mathbb{P}\}.$$

On the other hand, if $\mathfrak{P}$ consists of probability measures that are absolutely continuous with respect to $\mathbb{P}$, then the associated DRO $\mathcal{R}$ in (4.15) is a coherent risk measure. For example, given a ϕ-divergence $D(\cdot\|\cdot)$ (such as the Kullback–Leibler, Hellinger, or χ^2, α divergence) and a tolerance $\tau > 0$, we can define $\mathfrak{P}$ to consist of all probability measures that are absolutely continuous with respect to $\mathbb{P}$ (denoted $P \ll \mathbb{P}$) that satisfy

$$D(P\|\mathbb{P}) \leq \tau.$$

For a given scalar function $\phi\colon [0,+\infty) \to (-\infty,+\infty]$ satisfying $\phi(t) < +\infty$ for all $t > 0$, $\phi(1) = 0$ and $\phi(0) = \lim_{t\to 0^+} \phi(t)$, the associated ϕ-divergences are given by

$$D(P\|\mathbb{P}) = \mathbb{E}[\phi(\mathrm{d}P/\mathrm{d}\mathbb{P})].$$

In this setting we can rewrite $\mathcal{R}$ as an optimization problem over two scalar variables

$$\mathcal{R}(X) = \inf_{\lambda\in\mathbb{R},\,\mu\geq 0} \{\tau\mu + \lambda + \mathbb{E}[(\mu\phi)^*(X-\lambda)]\}, \tag{4.16}$$

where $(\mu\phi)^*(t) = \sup_{s\in\mathbb{R}}\{st - (\mu\phi)(s)\}$ is the Fenchel conjugate of $(\mu\phi)$ (see Shapiro 2017).

Perhaps a more fundamental approach to defining the uncertainty set $\mathfrak{P}$ – first used by Scarf (1958) – is through *moment matching*. Given $K \in \mathbb{N}$ real-valued measurable functions $\psi_k\colon \Omega \to \mathbb{R}$ and scalars $m_k \in \mathbb{R}$ for $k = 1,\dots,K$, the moment-matching uncertainty set is defined as

$$\mathfrak{P} = \{P\colon \mathcal{F} \to [0,1] \mid P(\Omega) = 1,\ \mathbb{E}_P[\psi_k] \leq m_k,\ k = 1,\dots,K\}.$$

Owing to classical results on the *problem of moments* by Rogosinski (1958), there exists a maximizing probability measure in (4.15) with a moment-matching uncertainty set that is supported on at most $K+1$ points (see Propositions 6.66 and Theorem 7.37 in Shapiro *et al.* 2014), which enables the reformulation

$$\mathcal{R}(X) = \begin{cases} \displaystyle\max_{\substack{\omega_1,\dots,\omega_{K+1}\in\Omega \\ p\in\mathbb{R}_+^{K+1}}} \sum_{\ell=1}^{K+1} p_\ell X(\omega_\ell) \\ \text{subject to } \displaystyle\sum_{\ell=1}^{K+1} p_\ell \psi_k(\omega_\ell) \leq m_k,\quad k = 1,\dots,K,\quad \sum_{\ell=1}^{K+1} p_\ell = 1. \end{cases} \tag{4.17}$$

DRO formulations of PDE-constrained optimization under uncertainty have been considered by various authors; see e.g. Kolvenbach, Lass and Ulbrich (2018), Kouri (2017) and Kouri and Shapiro (2018). As discussed, when the uncertainty set $\mathfrak{P}$ consists of probability measures that are absolutely continuous with respect to $\mathbb{P}$, the associated function $\mathcal{R}$ is a coherent risk measure and the methods employed for risk-averse optimization apply. For example, we can employ (4.16) to rewrite the min-max problem as a minimization problem over an augmented optimization space, analogously to the AVaR case (4.10). Moreover, we can apply the discussion at the end of Section 4.1 concerning the existence of solutions of risk-averse

optimization problems and optimality conditions. However, for more general $\mathfrak{P}$, such as the moment-matching uncertainty set, solving a min-max problem involving a generally non-concave inner maximization problem, as in (4.17), is inevitable. In an attempt to make general DRO tractable for PDE-constrained optimization, Kouri (2017) introduces a convergent approximation for $\mathfrak{P}$ that simplifies the maximization problem defining $\mathcal{R}$, requiring the solution of a linear program to evaluate $\mathcal{R}$ in the case of (4.17).

4.3. Probabilistic functions and chance constraints

An analogous approach to risk measures is to use probabilistic functions of the form

$$p_F(u) = \mathbb{P}(F(u) > F(u_0)) \quad \text{and} \quad p_G(u) = \mathbb{P}(G(u) > B),$$

where $u_0 \in \mathcal{U}$ is a benchmark optimization variable. The probabilistic objective function p_F and chance constraints of the form

$$p_G(u) \leq 1 - \alpha$$

for fixed $\alpha \in (0, 1)$ are popular because they are intuitive. However, multiple complications exist for probabilistic functions. For example, p_F and p_G are generally not differentiable or convex even if F and G are. Even worse, p_F and p_G need not be continuous; see e.g. Uryasev (1994, 1995, 2000).

An approach that overcomes these challenges is the *buffered probability* introduced by Rockafellar and Royset (2010). Roughly speaking, the buffered probability that a random variable X exceeds a threshold x is the probability that the tail weight of X exceeds x. Mathematically, this is defined as

$$\text{bPOE}_x(X) = 1 - \beta, \quad \text{where } \beta \in [0, 1) \text{ solves } x = \text{AVaR}_\beta(X).$$

In some applications, the buffered probability is preferred over p_F or p_G because it not only quantifies the probability of exceeding the prescribed threshold but also encodes the magnitudes of the events in excess of the threshold. Furthermore, the buffered probability has many desirable properties. Mafusalov and Uryasev (2018) show that it is monotonic, quasi-convex and lower semicontinuous. In fact, the buffered probability is the smallest upper bound for the probability over all such functions. In particular,

$$p_F(u) \leq \text{bPOE}_0(F(u) - F(u_0)) \quad \text{and} \quad p_G(u) \leq \text{bPOE}_0(G(u) - B).$$

Mafusalov and Uryasev (2018, Proposition 2.2) show that, analogous to the average value-at-risk, evaluating the buffered probability requires the solution of a one-dimensional minimization problem:

$$\text{bPOE}_x(X) = \min_{t \geq 0} \mathbb{E}[\max\{0, t(X - x) + 1\}]. \tag{4.18}$$

Additionally, the relationship between the buffered probability and the average

value-at-risk enables us to equivalently reformulate a constraint on the buffered probability to a constraint on the average value-at-risk, that is,

$$\mathrm{bPOE}_x(G(u) - B) \leq 1 - \alpha \quad \Longleftrightarrow \quad \mathrm{AVaR}_\alpha(G(u) - B) \leq 0.$$

The latter constraint has the benefit that if G is convex, then $u \mapsto \mathrm{AVaR}_\alpha(G(u) - B)$ is also convex. Although the buffered probability has many convenient mathematical properties, it is often not differentiable. When derivatives are required, one can instead use the higher-moment buffered probability introduced in Kouri (2019).

Although intuitive, the use of probability and buffered probability constraints in PDE-constrained optimization is limited. This has to do with the theoretical and numerical intricacies associated with these constraints. For initial results on these fronts; see Chaudhuri *et al.* (2022), Kouri (2019) and Kouri, Staudigl and Surowiec (2023).

The constraints (4.1b) involve scalar-valued functions. In PDE-constrained optimization, one could also consider pointwise constraints on the PDE solution. Probabilistic or almost sure formulations of such pointwise state constraints have recently been discussed in Geiersbach and Henrion (2024), Geiersbach, Henrion and Pérez-Aros (2025) and Geiersbach and Hintermüller (2022).

4.4. Stochastic orders

As discussed, the inequality $G(x) \leq B$ can be understood in various ways including almost surely or employing a preference order using an expected utility function $\mathbb{U}$, i.e. $\mathbb{U}(G(x)) \leq \mathbb{U}(B)$, or a risk measure $\mathcal{R}$, i.e. $\mathcal{R}(G(x)) \leq \mathcal{R}(B)$.

Related to preference orders is the notion of *stochastic orders*; see e.g. Levy (1992) for a survey of stochastic orders and dominance. A random variable X dominates another random variable X' with respect to the *first stochastic order*, denoted $X \succeq_{(1)} X'$, if the CDFs Ψ_X and $\Psi_{X'}$ associated with X and X' satisfy

$$\Psi_X(t) \leq \Psi_{X'}(t) \quad \text{for all } t \in \mathbb{R}. \tag{4.19}$$

Relating first-order stochastic dominance to preference orders using utility functions, (4.19) is equivalent to

$$\mathbb{E}[v(X)] \leq \mathbb{E}[v(X')] \tag{4.20}$$

for all non-decreasing functions $v : \mathbb{R} \to \mathbb{R}$ for which the expectation exists. On the other hand, X dominates X' with respect to the *second stochastic order*, denoted $X \succeq_{(2)} X'$, if

$$\int_{-\infty}^{t} \Psi_X(s)\,\mathrm{d}s \leq \int_{-\infty}^{t} \Psi_{X'}(s)\,\mathrm{d}s \quad \text{for all } t \in \mathbb{R}, \tag{4.21}$$

which is equivalent to the condition

$$\mathbb{E}[(t - X)_+] \leq \mathbb{E}[(t - X')_+] \quad \text{for all } t \in \mathbb{R}. \tag{4.22}$$

As gleaned from the second-order stochastic dominance condition (4.22), the second stochastic order requires the left tail of X to be *smaller* than that of X'. In many applications we are instead interested in comparing the right tails of X and X', which is achieved using the *increasing convex order*. In particular, X dominates X' in the increasing convex order, denoted $X \succeq_{\text{icx}} X'$, if

$$\mathbb{E}[v(X)] \geq \mathbb{E}[v(X')], \tag{4.23}$$

for all non-decreasing convex functions $v : \mathbb{R} \to \mathbb{R}$ for which the expectation exists. Relating this back to the second stochastic order, we have that $X \succeq_{\text{icx}} X'$ if and only if $-X' \succeq_{(2)} -X$.

Stochastic orders and risk measures are closely related. Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is non-atomic and $\mathcal{R}$ is law-invariant. Theorem 6.50 in Shapiro *et al.* (2014) shows that if $X \succeq_{(1)} X'$, then $\mathcal{R}(X) \geq \mathcal{R}(X')$ if and only if $\mathcal{R}$ satisfies the monotonicity condition (R2). Similarly, Theorem 6.51 in Shapiro *et al.* (2014) shows that if $\mathcal{R}$ is a convex risk measure (i.e. it satisfies (R1′), (R2) and (R3)), then $X \succeq_{\text{icx}} X'$ implies $\mathcal{R}(X) \geq \mathcal{R}(X')$. These relationships indicate that stochastic orders impart additional risk aversion when compared with preference orders based on monetary risk measures. For more information on stochastic dominance constraints, see Dentcheva and Ruszczyński (2003).

Recently, Conti, Rumpf, Schultz and Tölkes (2018) employed stochastic dominance constraints on compliance for elastic topology optimization to compute minimum volume designs. Stochastic dominance constraints were also used by Jakeman, Kouri and Huerta (2022) to build reliable surrogate models for engineering applications.

5. Optimization methods

There are two fundamentally different approaches to solving the optimization problem (4.1): approximation-based approaches and stochastic methods. Approximation-based approaches either first discretize the stochastic objective function and constraints using e.g. polynomial approximations or sample average approximation. They then solve the resulting deterministic problem, or integrate discretization-based models with optimization algorithms applied to the infinite-dimensional setting, or some combination. In contrast, stochastic methods, such as stochastic approximation, use stochastic estimators of the gradient to converge to a solution of the original problem.

5.1. Approximation methods

Much of the recent research on approximation methods centres around sample average approximation (SAA) and its asymptotic properties. Although many authors have investigated other various approximation techniques, including sparse-grid quadrature (Kouri *et al.* 2013, 2014, Kouri 2014, Zahr, Carlberg and Kouri 2019), polynomial chaos (Rosseel and Wells 2012) and Voronoi-based piecewise constant

approximations (Zou, Kouri and Aquino 2018, 2022), SAA appears to be the only truly dimension-independent approach. Moreover, SAA does not suffer from the additional regularity requirements of polynomial-based approximations, enabling the solution of coherent risk-averse problems. As mentioned in Section 3.5.2, Römisch and Surowiec (2024), Milz (2023*a*,*b*,*c*) and Milz and Ulbrich (2024) establish bounds for the error between the solution of a risk-neutral formulation of an optimal control problem similar to the one in Example 3.15 and its SAA approximation. Phelps, Royset and Gong (2016) provide consistency results for an SAA approximation of a risk-neutral formulation of an optimal control problem governed by ordinary differential equations.

Given an approximate problem, one can employ any off-the-shelf optimization algorithm to solve the resulting problem. Although practical, this approach does not account for the infinite-dimensional nature of PDE-constrained optimization problems. In particular, there is typically no mechanism for leveraging different discretization fidelities – a critical and distinguishing feature of PDE-constrained optimization under uncertainty when compared to traditional stochastic programming. As we have already mentioned in Section 3.6, even if an appropriate approximation can be determined *a priori*, different levels of coarser discretizations can be used to compute approximate solutions at lower computational cost or to approximately solve subproblems in the optimization algorithm applied to the fine discretization. We have reviewed approaches in Section 3.6.

In the remainder of this subsection, we discuss trust-region methods with tunable objective function and gradient approximations and progressive hedging. Lastly, we discuss the primal–dual risk minimization method that generates a sequence of smoothed risk measures to solve risk-averse PDE-constrained optimization problems.

5.1.1. Trust-region methods

Given a Hilbert space $\mathcal{U}$, a closed convex set $\mathcal{U}_{\text{ad}} \subset \mathcal{U}$, $F\colon \mathcal{U} \to L^p_{\mathbb{P}}(\Omega)$, and $\wp\colon \mathcal{U} \to \mathbb{R}$, we consider

$$\min_{u \in \mathcal{U}_{\text{ad}}} \mathcal{R}[F(u)] + \wp(u) \tag{5.1}$$

with a smooth risk measure $\mathcal{R}\colon L^p_{\mathbb{P}}(\Omega) \to \mathbb{R}$, e.g. $\mathcal{R}[\cdot] = \mathbb{E}[\cdot]$ or $\mathcal{R}$ given by the entropic risk (4.11). For the numerical solution of (5.1), $\mathcal{R}[F(u)]$ must be approximated. For example, $\mathcal{R}[\cdot] = \mathbb{E}[\cdot]$ or (4.11) require an approximation of the expectation as well as the state equations that enter the evaluation of $F(u)$. Thus, in practice, one cannot work directly with the objective function

$$f(u) = \mathcal{R}[F(u)] + \wp(u), \tag{5.2}$$

but rather with models m_k of f built around the optimization iterates u_k. Trust-region methods provide a powerful framework for model management in optimization. Unfortunately this framework is currently only available for a limited class

of problems, and therefore we restrict our attention to problems of the type

$$\min_{u \in \mathcal{U}_{\text{ad}}} \; f(u), \tag{5.3}$$

when the objective function f is bounded below on $\mathcal{U}_{\text{ad}}$ and is Lipschitz-continuously differentiable on an open set containing $\mathcal{U}_{\text{ad}}$. Problem (5.1) is in this class, provided $\mathcal{R}[F(\cdot)] + \wp(\cdot)$ has the desirable smoothness properties. In addition, approximations of risk-averse optimization problems with smoothed risk measures are also in this problem class. The latter arise as subproblems in Garreis, Surowiec and Ulbrich (2021) and Kouri and Surowiec (2016, 2020*a*, 2022), and in the subproblem (5.22) discussed later.

Conn, Gould and Toint (2000) provide a comprehensive discussion of trust-region methods. Our presentation follows Kouri *et al.* (2014) because they introduce relaxations on model and gradient errors that are important for the practical implementation of the trust-region algorithm. These relaxations build on earlier work by Carter (1991, 1993), Heinkenschloss and Vicente (2002) and Ziems (2013). We assume Lipschitz-continuous differentiability of f, but one could consider nonsmooth, but proper, closed and convex $\wp(\cdot)$, as in Baraldi and Kouri (2023, 2024).

At the kth iteration of the trust-region method, the algorithm requires the construction of a smooth model m_k of the objective function f around the current iterate u_k. In the context of (5.1), the model m_k could be constructed using quadrature approximations of $\mathcal{R}$, and discretizations of the state equations that enter in the evaluation of $F(u_k)$. Often, m_k is a quadratic function of the form

$$m_k(u) - m_k(u_k) = \langle g_k, u - u_k \rangle_{\mathcal{U}} + \frac{1}{2} \langle B_k(u - u_k), u - u_k \rangle_{\mathcal{U}}, \tag{5.4}$$

where B_k is a self-adjoint bounded linear operator that encapsulates the curvature information of f at u_k, and g_k is an approximation of the gradient $\nabla f(u_k)$. Again, in the context of (5.1), the gradient approximation g_k could be constructed using quadrature approximations of the expected value arising in $\nabla f(u_k) = \nabla \mathcal{R}[F(u_k)] + \nabla \wp(u_k)$, and discretizations of the state and adjoint equations that enter the quadrature.

The model m_k must be computed so that its gradient at u_k sufficiently agrees with $\nabla f(u_k)$. To specify the gradient tolerance, define the norm of the projected model gradient

$$h_k := \frac{1}{r} \| \text{proj}_{\mathcal{U}_{\text{ad}}}(u_k - r \nabla m_k(u_k)) - u_k \|_{\mathcal{U}}, \tag{5.5}$$

where $r > 0$ is fixed, and $\text{proj}_{\mathcal{U}_{\text{ad}}}(\cdot)$ denotes the metric projection onto $\mathcal{U}_{\text{ad}}$. Given a *trust-region radius* $\Delta_k > 0$, which the trust-region algorithm will automatically adapt, we require that the model satisfies

$$\| \nabla f(u_k) - \nabla m_k(u_k) \|_{\mathcal{U}} \leq \kappa_{\text{grad}} \min\{h_k, \Delta_k\}, \tag{5.6}$$

for a given positive constant κ_{grad} independent of k. If m_k is given by (5.4), $\nabla m_k(u_k) = g_k$. The condition in (5.6) is motivated by the classical inexact gradient condition introduced in Carter (1991), and stems from Heinkenschloss and Vicente (2002) and Kouri *et al.* (2013). In particular, (5.6) can be implemented in practice without the need to compute/estimate problem-dependent quantities, as shown in Kouri *et al.* (2013), for example.

Given the trust-region radius $\Delta_k > 0$ and the model m_k satisfying (5.6), the algorithm then computes a trial iterate u_k^+ that approximately solves the trust-region subproblem

$$\min_{u \in \mathcal{U}_{\text{ad}}} m_k(u) \quad \text{subject to} \quad \|u - u_k\|_{\mathcal{U}} \leq \Delta_k. \tag{5.7}$$

The approximate solution u_k^+ of (5.7) must satisfy the so-called *fraction of Cauchy decrease condition*, that is, given constants κ_{rad} and κ_{fcd}, independent of k, $u_k^+ \mathcal{U}_{\text{ad}}$ must satisfy

$$\begin{aligned} \|u_k^+ - u_k\|_{\mathcal{U}} &\leq \kappa_{\text{rad}} \Delta_k, \\ m_k(u_k) - m_k(u_k^+) &\geq \kappa_{\text{fcd}} h_k \min\left\{ \frac{h_k}{1+\omega_k}, \Delta_k \right\}, \end{aligned} \tag{5.8}$$

where h_k is given by (5.5) and ω_k is the maximum Fréchet curvature of m_k over the trust region:

$$\omega_k := \sup\left\{ \frac{2}{\|s\|_{\mathcal{U}}} |m_k(u_k + s) - m_k(u_k) - \langle \nabla m_k(u_k), s \rangle_{\mathcal{U}}| \mid 0 < \|s\|_{\mathcal{U}} \leq \kappa_{\text{rad}} \Delta_k \right\}.$$

When m_k is the quadratic model (5.4), then $\omega_k = \|B_k\|$. Conn *et al.* (2000, Chapter 7) discuss several algorithms that compute such a u_k^+ in the unconstrained case $\mathcal{U}_{\text{ad}} = \mathcal{U}$, and some of these can be generalized to convex constraints.

Given the trial iterate u_k^+, the original trust-region algorithm decides whether to accept the iterate u_k^+ and how to update the trust-region radius Δ_k based on the ratio between the actual decrease in the model,

$$\text{ared}_k = f(u_k) - f(u_k^+),$$

and the *predicted decrease* by the model,

$$\text{pred}_k := m_k(u_k) - m_k(u_k^+).$$

Because we want to avoid evaluation of f, we replace the actual decrease with a *computed decrease*,

$$\text{cred}_k \approx f(u_k) - f(u_k^+),$$

where cred_k is computed using model evaluations. The error between the computed decrease and the actual decrease, $|\text{cred}_k - (f(u_k) - f(u_k^+))|$, needs to be sufficiently small, and we will specify this in (5.9) below.

Given parameters $0 < \eta_1 < \eta_2$ and $0 < \gamma_1 < 1 < \gamma_2$, we decide whether or not to accept the trial iterate and to increase or decrease the trust-region radius as follows. If

$$\text{cred}_k \leq \eta_1 \, \text{pred}_k,$$

we reject the trial iterate and decrease the trust-region radius, setting $u_{k+1} = u_k$ and $\Delta_{k+1} = \gamma_1 \Delta_k$. If

$$\text{cred}_k > \eta_1 \text{pred}_k,$$

we accept the trial iterate, $u_{k+1} = u_k^+$, and update the trust-region radius according to

$$\Delta_{k+1} = \begin{cases} \Delta_k & \text{if } \text{cred}_k \leq \eta_2 \, \text{pred}_k, \\ \gamma_2 \Delta_k & \text{if } \text{cred}_k > \eta_2 \, \text{pred}_k. \end{cases}$$

As noted before, the computed decrease cred_k is computed using model evaluations. We assume there exist positive constants κ_{obj}, $\eta < \min\{\eta_1, 1 - \eta_2\}$ and $\zeta > 1$, independent of k, as well as a positive sequence $\{\theta_k\}$ with $\lim_{k\to\infty} \theta_k = 0$ such that

$$|(f(u_k) - f(u_k^+)) - \text{cred}_k| \leq \kappa_{\text{obj}} [\min\{\eta \, \text{pred}_k, \theta_k\}]^\zeta. \tag{5.9}$$

The condition (5.9) is due to Kouri *et al.* (2014). The conditions (5.6) and (5.9) are designed so that they can be implemented without knowledge of problem parameters such as Lipschitz constants. We will comment on applications below.

The trust-region algorithm with function and gradient model is summarized in Algorithm 1. The following first-order convergence result for Algorithm 1 implies that for arbitrary initial guess $u_1 \in \mathcal{U}_{\text{ad}}$, the limit of every convergence subsequence of the iterates $\{u_k\} \subset \mathcal{U}_{\text{ad}}$ generated by Algorithm 1 is a stationary point of (5.3).

Theorem 5.1. Let $\mathcal{U}$ be a Hilbert space and let $\mathcal{U}_{\text{ad}}$ be non-empty, closed and convex. If $f : \mathcal{U} \to \mathbb{R}$ is bounded below and Lipschitz-continuous Fréchet-differentiable on an open set containing $\mathcal{U}_{\text{ad}}$, and if the Fréchet curvature ω_k of the models m_k satisfies

$$\sum_{k=1}^{\infty} (1 + \max\{\omega_1, \ldots, \omega_k\})^{-1} = +\infty,$$

then the sequence of iterates $\{u_k\} \subset \mathcal{U}_{\text{ad}}$ generated by Algorithm 1 satisfies

$$\liminf_{k\to\infty} \frac{1}{r} \|\text{proj}_{\mathcal{U}_{\text{ad}}}(u_k - r\nabla f(u_k)) - u_k\|_{\mathcal{U}} = 0.$$

The proof of this theorem is a slight modification of the convergence result in Kouri *et al.* (2014, Theorem 4.4). Baraldi and Kouri (2023, Theorem 3) prove a slightly more general version of Theorem 5.1, where their proof is based on the techniques used in Toint (1988).

Algorithm 1 Trust-region algorithm with objective function and gradient approximations.

Require: Initial guess $u_1 \in \mathcal{U}_{\text{ad}}$, initial radius $\Delta_1 > 0$, $0 < \eta_1 < \eta_2 < 1$, and $0 < \gamma_1 < 1 < \gamma_2$.

for $k = 1, 2, \ldots$ **do**
  Model selection: Choose a model m_k that satisfies (5.6).
  Step computation: Compute $u_k^+ \in \mathcal{U}_{\text{ad}}$ that satisfies (5.8).
  Computed reduction: Compute cred_k that satisfies (5.9).
  Step acceptance and radius update:
  if $\text{cred}_k \leq \eta_1 \text{pred}_k$ **then**
    $u_{k+1} \leftarrow u_k$
    $\Delta_{k+1} \leftarrow \gamma_1 \Delta_k$
  else
    $u_{k+1} \leftarrow u_k^+$
    if $\text{cred}_k \leq \eta_2 \, \text{pred}_k$ **then**
      $\Delta_{k+1} \leftarrow \Delta_k$
    else
      $\Delta_{k+1} \leftarrow \gamma_2 \Delta_k$
    end if
  end if
end for

In the context of PDE-constrained optimization under uncertainty, the model m_k is generated by a discretization of the objective function at u_k, as discussed for the model problem in Section 3.5.2. Since both the gradient gradient error $\|\nabla f(u_k) - \nabla m_k(u_k)\|_{\mathcal{U}}$ and the error in the computed decrease cred_k must be controlled as in (5.6) and (5.9), the fidelity of the discretization must be chosen based on the objective function and its gradient. Algorithm 1 is deterministic, that is, error bounds (5.6) and (5.9) are deterministic. This limits the approximation methods used for the model generation. For example, multilevel/multi-fidelity Monte Carlo methods cannot be used, because they provide approximations in expectation or in probability. However, Algorithm 1 could be applied to an SAA of the problem with a large sample size, where the large sample problem is viewed as the underlying problem and small(er) sample models are used to compute iterates.

In the context of PDE-constrained optimization under uncertainty, the conditions (5.6) and (5.9) were used to refine dimension-adaptive sparse-grid discretizations in Kouri *et al.* (2013, 2014), Kouri (2014) and Zahr *et al.* (2019), adaptive reduced-order models and Voronoi approximations in Zou *et al.* (2018, 2022) and adaptive tensor-train approximations in Garreis and Ulbrich (2017).

Bastin, Cirillo and Toint (2006) use a trust-region algorithm to solve a fixed SAA approximation in stochastic programming and mixed logit regression. They compute models by dynamically subsampling and employ model error conditions from Carter (1991) and Conn *et al.* (2000, Sections 8.4, 10.6). For problem (5.1),

Bastin *et al.* (2006) employ the sample variance estimate of the objective function value, that is,

$$\frac{1}{N_k - 1} \sum_{n=1}^{N_k} \left(F_n(u_k) - \frac{1}{N_k} \sum_{m=1}^{N_k} F_m(u_k) \right)^2,$$

and similarly for the gradient, as error estimates to verify their inexactness conditions. Here N_k is the number of samples used at the kth iteration and $F_n(u_k)$ are i.i.d. realizations of $F(u_k)$.

There are trust-region-type and line-search-based algorithms for finite-dimensional optimization problems that use stochastic models; see e.g. Bandeira, Scheinberg and Vicente (2014), Berahas, Cao and Scheinberg (2021) and Curtis, Scheinberg and Shi (2019). However, they have not yet been applied to PDE-constrained optimization.

5.1.2. *Progressive hedging*

Given a Hilbert space $\mathcal{U}$ and a closed convex set $\mathcal{U}_{\text{ad}} \subset \mathcal{U}$, we again consider

$$\min_{u \in \mathcal{U}_{\text{ad}}} \mathbb{E}[F(u)] + \wp(u). \tag{5.10}$$

An alternative method for solving (5.10) is the *progressive hedging algorithm* introduced by Rockafellar and Wets (1991).

Employing sampling or quadrature to approximate $\mathbb{E}[F(u)]$, (5.10) is approximated as

$$\min_{u \in \mathcal{U}_{\text{ad}}} \sum_{n=1}^{N} \zeta_n F_n(u) + \wp(u), \tag{5.11}$$

where $F_n(\cdot)$, $n = 1, \ldots, N$, is an evaluation or i.i.d. realization of $F(\cdot)$, and ζ_n, $n = 1, \ldots, N$, is the associated probability or weight satisfying

$$\zeta_n > 0, \ n = 1, \ldots, N, \quad \zeta_1 + \ldots + \zeta_N = 1.$$

We assume that F_n, $n = 1, \ldots, N$, and $\wp$ are Fréchet-differentiable, but this assumption can be relaxed.

The evaluations of $F_n(u)$ and its derivative can be performed concurrently for each n. However, the computation of the average value and gradient leads to a potential communication bottleneck. Progressive hedging overcomes this using techniques akin to the *alternating direction method of multipliers* (ADMM). By introducing new optimizations u_n that *anticipate* the realization $(F_n + \wp)(\cdot)$, we can equivalently reformulate (5.11) as

$$\min \quad \sum_{n=1}^{N} \zeta_n (F_n(u_n) + \wp(u_n)) \tag{5.12a}$$

$$\text{subject to } u_n = u, \qquad n = 1, \ldots, N, \tag{5.12b}$$

$$u_n \in \mathcal{U}_{\text{ad}}, \quad n = 1, \ldots, N. \tag{5.12c}$$

The Lagrange multipliers λ_n, $n = 1, \ldots, N$, associated with (5.12b) satisfy

$$\sum_{n=1}^{N} \zeta_n \lambda_n = 0. \tag{5.13}$$

The augmented Lagrangian associated with (5.12) is given by

$$\sum_{n=1}^{N} \zeta_n L_n(u_n, u, \lambda_n), \tag{5.14a}$$

where L_n is the augmented Lagrangian associated with the nth realization, that is,

$$L_n(u_n, u, \lambda_n) \coloneqq F_n(u_n) + \wp(u_n) + \langle \lambda_n, u_n - u \rangle_{\mathcal{U}} + \frac{r}{2} \|u_n - u\|_{\mathcal{U}}^2, \tag{5.14b}$$

and $r > 0$ is a penalty parameter.

Given λ_n^{k-1}, $n = 1, \ldots, N$, the augmented Lagrangian method solves

$$\min_{u_n,\, u \in \mathcal{U}_{\mathrm{ad}}} \sum_{n=1}^{N} \zeta_n L_n\big(u_n, u, \lambda_n^{k-1}\big), \tag{5.15}$$

for $u_1^k, \ldots, u_N^k \in \mathcal{U}_{\mathrm{ad}}$ and $u^k \in \mathcal{U}$, and then updates the Lagrange multipliers using

$$\lambda_n^k = \lambda_n^{k-1} + r\big(u_n^k - u^k\big), \quad n = 1, \ldots, N. \tag{5.16}$$

The subproblem (5.15) still couples across all $u_n^k \in \mathcal{U}_{\mathrm{ad}}$, $n = 1, \ldots, N$, $u^k \in \mathcal{U}$. Progressive hedging overcomes this by first minimizing

$$\sum_{n=1}^{N} \zeta_n L_n\big(u_n, u^{k-1}, \lambda_n^{k-1}\big)$$

over $u_n \in \mathcal{U}_{\mathrm{ad}}$, $n = 1, \ldots, N$, given $u^{k-1} \in \mathcal{U}$ and λ_n^{k-1}, $n = 1, \ldots, N$, satisfying (5.13), then minimizing

$$\sum_{n=1}^{N} \zeta_n L_n\big(u_n^k, u, \lambda_n^{k-1}\big)$$

over $u \in \mathcal{U}$ to compute u^k, and finally updating the Lagrange multipliers using (5.16). The function

$$\sum_{n=1}^{N} \zeta_n L_n\big(u_n, u^{k-1}, \lambda_n^{k-1}\big)$$

is separable in u_n, $n = 1, \ldots, N$. Therefore, minimizing this function is equivalent to the solution of N minimization problems

$$\min_{u_n \in \mathcal{U}_{\mathrm{ad}}} L_n\big(u_n, u^{k-1}, \lambda_n^{k-1}\big). \tag{5.17}$$

Algorithm 2 Progressive hedging algorithm.

Require: Initial guess $u^0 \in \mathcal{U}$, penalty parameter $r > 0$, multiplier estimates $\lambda_n^0 \in \mathcal{U}$, $n = 1, \ldots, N$, satisfying $\sum_{n=1}^N \zeta_n \lambda_n^0 = 0$, and a sequence of tolerances $\{\epsilon_k\}$ with $\epsilon_k > 0$ and $\sum_{k=1}^\infty \epsilon_k < \infty$.

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

2: For $n = 1, \ldots, N$, compute an approximate solution $u_n^k \in \mathcal{U}_{\text{ad}}$ of (5.17) satisfying (5.19).

3: Aggregate u_n^k to compute the implementable solution estimate

$$u^k \leftarrow \sum_{n=1}^N \zeta_n u_n^k.$$

4: Update the multiplier estimates as

$$\lambda_n^k \leftarrow \lambda_n^{k-1} + r(u_n^k - u^k), \quad n = 1, \ldots, N.$$

5: **end for**

Given $u_n^k \in \mathcal{U}_{\text{ad}}$, $n = 1, \ldots, N$, and λ_n^{k-1}, $n = 1, \ldots, N$, satisfying (5.13), the solution of

$$\min_{u \in \mathcal{U}} \sum_{n=1}^N \zeta_n L_n\left(u_n^k, u, \lambda_n^{k-1}\right)$$

is

$$u^k = \sum_{n=1}^N \zeta_n u_n^k. \tag{5.18}$$

Because $u_n^k \in \mathcal{U}_{\text{ad}}$, for $n = 1, \ldots, N$, their convex combination must also be in $\mathcal{U}_{\text{ad}}$, i.e. $u^k \in \mathcal{U}_{\text{ad}}$. Finally, note that if λ_n^{k-1}, $n = 1, \ldots, N$, satisfy (5.13), $u_n^k \in \mathcal{U}_{\text{ad}}$, $n = 1, \ldots, N$, and u^k is given by (5.18), the new Lagrange multipliers (5.16) also satisfy (5.13). In practice, the minimization problems (5.17) do not need to be solved exactly, but we have

$$\inf_{\eta \in N_{\mathcal{U}_{\text{ad}}}(u_n^k)} \|\nabla(F_n + \wp)(u_n^k) + \lambda_n^{k-1} + r(u_n^k - u^{k-1}) + \eta\|_{\mathcal{U}} \leq \frac{\epsilon_k}{r} \tag{5.19}$$

for a suitable ϵ_k. Here $N_{\mathcal{U}_{\text{ad}}}(u_n^k)$ is the normal cone to $\mathcal{U}_{\text{ad}}$ at the point u_n^k (cf. (1.1)), and the left-hand side of (5.19) is zero at a solution of (5.17).

We summarize the progressive hedging algorithm in Algorithm 2. When $(F + \wp)(\cdot)$ is convex, Algorithm 2 is a special case of the method of partial inverses from Spingarn (1983) and more generally Douglas–Rachford splitting, which are provably convergent in a general Hilbert space. See e.g. Eckstein and Bertsekas (1992) for the detailed convergence of Douglas–Rachford splitting. We state

Algorithm 2 in the next theorem. Note that this result is a corollary of Theorem 2 in Spingarn (1985).

Theorem 5.2. Suppose $(F_n + \wp)(\cdot)$, $n = 1, \dots, N$, are convex a.s. and Fréchet-differentiable, and let $\{(u_n^k, u^k, \lambda_n^k)\}$ be the sequence of iterates generated by Algorithm 2 so that

$$\inf_{\eta \in N_{\mathcal{U}_{\mathrm{ad}}}(u_n^k)} \left\| \nabla(F_n + \wp)\left(u_n^k\right) + \lambda_n^{k-1} + r\left(u_n^k - u^{k-1}\right) + \eta \right\|_{\mathcal{U}} \leq \frac{\epsilon_k}{r} \text{ with } \sum_{k=1}^{\infty} \epsilon_k < \infty.$$

If the sequence $\{v_k\} \subset [0, +\infty)$ consisting of entries

$$v_k := \left(\sum_{n=1}^{N} \zeta_n \| u^k + \lambda_n^k \|_{\mathcal{U}}^2 \right)^{1/2}$$

is bounded, then there exists a minimizer $\bar{u} \in \mathcal{U}_{\mathrm{ad}}$ to (5.10) for which $u^k \rightharpoonup \bar{u}$.

In finite dimensions, Rockafellar and Wets (1991, Theorem 6.1) extend the convergence theory to non-convex problems, as long as u_n^k are computed as δ-locally optimal solutions. In particular, Rockafellar and Wets (1991) prove that $\{u^k\}$ converges to a stationary point of (5.11). Recently, Rockafellar (2018) extended Algorithm 2 to solve risk-averse problems using optimized certainty equivalent risk measures (4.3), and Rockafellar and Uryasev (2020) extended it to minimize the buffered probability (4.18).

5.1.3. Primal–dual risk minimization

The trust-region method described in Section 5.1.1 requires differentiability of the objective function f, which is often not the case for risk-averse problems; see Theorem 4.4. To overcome this issue, authors such as Kouri and Surowiec (2016, 2020*a*) have investigated the use of smoothed risk measures. Although practical, smoothing techniques lead to approximations of the risk measures and errors in the optimal solutions, and therefore they require the user to tune parameters in the smoothed risk measures. To control these errors, Kouri and Surowiec (2022) introduced the primal–dual risk minimization algorithm. To describe this method, we consider the risk-averse optimization problem

$$\min_{u \in \mathcal{U}_{\mathrm{ad}}} \{ f(u) := \mathcal{R}[F(u)] + \wp(u) \}, \tag{5.20}$$

where $\mathcal{R} \colon L^2_{\mathbb{P}}(\Omega) \to \mathbb{R}$ is a coherent risk measure. Motivated by augmented Lagrangian methods, the primal–dual risk minimization algorithm regularizes the dual form of the risk measure (4.5) with a quadratic penalty, that is,

$$\mathcal{R}_k[X] := \max_{\theta \in \mathfrak{A}} \left\{ \mathbb{E}[\theta X] - \frac{r_k}{2} \| \theta - \theta_k \|^2_{L^2_{\mathbb{P}}(\Omega)} \right\}, \tag{5.21}$$

where $\mathfrak{A} \subset L^2_{\mathbb{P}}(\Omega)$, $r_k > 0$ is the penalty parameter and $\theta_k \in L^2_{\mathbb{P}}(\Omega)$ is an estimate of a so-called *risk identifier*, i.e. a maximizer of (4.5), at the kth iteration. Note that

Algorithm 3 Primal–dual risk minimization algorithm.

Require: Initial primal guess $u_1 \in \mathcal{U}_{\text{ad}}$, initial dual guess $\theta_1 \in \mathfrak{A}$, initial penalty parameter $r_1 > 0$, and positive constants $\tau_1 > 0$, $\delta_r \in (1, +\infty)$ and $\delta_\tau \in (0, 1)$

for $k = 1, 2, \ldots$ **do**
  Step computation: Compute $u_{k+1} \in \mathcal{U}_{\text{ad}}$ that approximately solves (5.22)
  Risk identifier update: Set $\theta_{k+1} \leftarrow \Theta_k(F(u_{k+1}))$
  Penalty parameter update:
  if $\|\theta_{k+1} - \theta_k\|_{L^2_{\mathbb{P}}(\Omega)} > r_k \tau_k$ **then**
    Set $r_{k+1} \leftarrow \delta_r r_k$
    Set $\tau_{k+1} \leftarrow \tau_k$
  else
    Set $r_{k+1} \leftarrow r_k$
    Set $\tau_{k+1} \leftarrow \delta_\tau \tau_k$
  end if
end for

(5.21) is analogous to the derivation of the augmented Lagrangian in Rockafellar (1973), for example. The regularized risk measure $\mathcal{R}_k[\cdot]$ is Lipschitz-continuously differentiable and the maximization problem (5.21) defining $\mathcal{R}_k[X]$ has the unique maximizer

$$\Theta_k(X) := \text{proj}_{\mathfrak{A}}(r_k X + \theta_k).$$

Moreover, the gradient of $\mathcal{R}_k[\cdot]$ is $\Theta_k(\cdot)$ and so the risk-averse objective function at the kth iteration is continuously differentiable with gradient

$$\nabla[\mathcal{R}_k \circ F + \wp](u) = \mathbb{E}[\Theta_k(F(u))F'(u)] + \nabla\wp(u)$$

as long as $F(\cdot)$ and $\wp(\cdot)$ are.

At the kth iteration of the primal–dual risk minimization method, we compute an approximate minimizer for the regularized problem

$$\min_{u \in \mathcal{U}_{\text{ad}}} \{f_k(u) := \mathcal{R}_k(F(u)) + \wp(u)\}, \tag{5.22}$$

and then employ augmented Lagrangian techniques to update the penalty parameter r_{k+1} and the dual estimates θ_{k+1}. One such realization of these updates is given in Algorithm 3. Kouri and Surowiec (2022) prove convergence of the sequence of iterates $\{u_k\}$ generated by Algorithm 3 when u_{k+1} is an ϵ_k-minimizer of (5.22). In particular, if $r_k \to r^\star > 0$ and $\epsilon_k \to \epsilon^\star \geq 0$, then Theorem 1 in Kouri and Surowiec (2022) ensures that any weak accumulation point of $\{u_k\}$ is a $(K^2/r^\star + \epsilon^\star)$-minimizer of (5.20). Here, $K > 0$ is a bound for the risk envelope $\mathfrak{A}$. This result is generally not practical unless (5.20) is convex. For non-convex problems, ensuring ϵ_k-minimizers is impossible. Instead, one can only guarantee that u_{k+1} is an ϵ_k-stationary point of the form

$$\langle \mathbb{E}[\Theta_k(F(u_{k+1}))F'(u_{k+1})] + \nabla\wp(u_{k+1}), u - u_{k+1} \rangle \geq -\epsilon_k \|u - u_{k+1}\| \tag{5.23}$$

for all $u \in \mathcal{U}_{\text{ad}}$, using, for example, Algorithm 1 or Algorithm 2. In this setting, Kouri and Surowiec (2022, Theorem 3) prove convergence to a stationary point of (5.20).

Theorem 5.3. Suppose $\wp\colon \mathcal{U} \to \mathbb{R}$ is weakly lower semicontinous, $F\colon \mathcal{U} \to L^2_{\mathbb{P}}(\Omega)$ is completely continuous, that is,

$$v_k \rightharpoonup v \quad \Longrightarrow \quad F(v_k) \to F(v),$$

$F'(\cdot)$ is completely continuous, $\wp'(\cdot)$ is weakly continuous, and there exists $\gamma \in \mathbb{R}$ such that the γ-level set

$$\{u \in \mathcal{U}_{\text{ad}} \mid \mathcal{R}[F(u)] + \wp(u) \leq \gamma\}$$

is non-empty and bounded. If the sequence of iterates $\{u_k\}$ generated by Algorithm 3 satisfy (5.23) with $\epsilon_k \to 0$ and $r_k \to \infty$, then any weak accumulation point of $\{u_k\}$ is a stationary point of (5.20).

As mentioned earlier, the objective function in (5.22) is differentiable if $F(\cdot)$ and $\wp(\cdot)$ are differentiable. Thus the optimization subproblems (5.22) can be solved using Newton-type optimization methods, such as the trust-region method described in Section 5.1.1. Additional efficiencies can be gained by exploiting the structure of the smoothed risk measure $\mathcal{R}_k$. For example, if the underlying risk measure is $\mathcal{R} = \text{AVaR}_\beta$ defined in (4.9) and an SAA approximation with N samples is used, then Markowski (2022) demonstrates that gradient and Hessian-times-vector information of the SAA approximation of the objective function in (5.22) can be well approximated using only a fraction of the N samples – in later iterations of Algorithm 3 only approximately $(1-\beta)N \ll N$ samples. Moreover, when applying smoothing as in Kouri and Surowiec (2016, 2020*a*) to (4.10), Markowski (2022) proposes a substep in an inexact Newton method to deal with the near-rank deficiency of the Hessian.

Related to Algorithm 3 is the interior point method developed in Garreis *et al.* (2021). Their interior point method considers a restricted class of coherent risk measures of the form (4.3), with the expected utility function

$$\mathbb{U}(X) = \mathbb{E}[\min\{a_1 X, a_2 X\}]$$

for $a_1 \in [0, 1)$ and $a_2 > 1$. This class of risk measures enables the reformulation of (5.20) as

$$\begin{aligned} &\min_{(u,t,W)\in \mathcal{U}_{\text{ad}}\times\mathbb{R}\times L^2_{\mathbb{P}}(\Omega)} \quad t + \mathbb{E}[W] + \wp(u) \\ &\text{subject to} \quad W \geq a_1(F(u) - t) \text{ a.s.}, \quad W \geq a_2(F(u) - t) \text{ a.s.} \end{aligned} \tag{5.24}$$

Garreis *et al.* (2021) apply an interior point method to solve (5.24).

5.2. *Stochastic methods*

The *stochastic approximation* (SA) method, originally introduced by Robbins and Monro (1951), has recently received increased attention for PDE-constrained applications of the type

$$\min_{u\in\mathcal{U}_{\text{ad}}} \ \mathbb{E}[F(u)] + \wp(u), \tag{5.25}$$

where $\mathcal{U}$ is a Hilbert space, $\mathcal{U}_{\text{ad}} \subset \mathcal{U}$ is a closed convex set, $F : \mathcal{U} \to L^1_{\mathbb{P}}(\Omega)$ and $\wp : \mathcal{U} \to \mathbb{R}$. We assume that F and $\wp$ are Fréchet-differentiable. We have discussed some SA methods in Section 3.6 for the model problem without control constraints, i.e. $\mathcal{U}_{\text{ad}} = \mathcal{U}$. More generally, Barty, Roy and Strugarek (2007) and Geiersbach and Pflug (2019) studied SA for convex problems posed in Hilbert space. Building on these works, Geiersbach and Wollner (2020) and Martin *et al.* (2021) combine SA and uniform finite element mesh refinement for convex PDE-constrained optimization problems. In addition, Geiersbach and Scarinci (2023) study SA methods for nonlinear, potentially non-convex, PDE-constrained optimization problems. Relating SA and the methods described in Section 5.1, Martin and Nobile (2021) apply a stochastic gradient method called SAGA (stochastic averaged gradient accelerated method) to a quadrature approximation of (5.25) for the particular case of the model problem in Section 3.

The basic SA method generates the sequence of iterations

$$u_{k+1} = \text{proj}_{\mathcal{U}_{\text{ad}}}(u_k - \gamma_k(G(u_k, \xi_k) + \nabla\wp(u_k))), \tag{5.26}$$

where $G_k(u_k)$ are independent random realizations of the gradient of $F(\cdot)$, and $\gamma_k > 0$ are predetermined step lengths. For example,

$$G_k(u_k) = \frac{1}{N_k}\sum_{n=1}^{N_k} \nabla F_n(u_k), \tag{5.27}$$

where $N_k \in \mathbb{N}$ and $\nabla F_n(u_k)$ are independent realizations of $\nabla F(u_k)$. In this setting, each iteration of SA defined in (5.26) requires N_k state and adjoint solves. Traditionally, N_k is small, like $N_k = 1$, resulting in a low per-iteration cost. However, the convergence of $\{u_k\}$ is probabilistic, often requires convexity of $F(\cdot) + \wp(\cdot)$, and is dependent on the choice of γ_k. The basic SA method (5.26), (5.27) is written in Hilbert space, and in this case one may assume that the gradient estimators $G_k(u_k)$ are unbiased. The numerical realization, however, requires discretizations of state and adjoint PDEs and possibly the iterative solution of these discretized equations, which generates biases. The size of the bias can be controlled by refining the discretization or iterative method stopping criteria and must be integrated into the convergence analysis of the SA method. See e.g. Geiersbach and Pflug (2019), Geiersbach and Wollner (2020), Martin *et al.* (2021) for such analyses applied to the model problem in Section 3.

When there is no bias, a basic analysis of (5.26) can be found, for example, in Shapiro *et al.* (2014, Section 5.9.1). It employs the step sizes $\gamma_k := \kappa/k$ for prescribed constant $\kappa > 0$ and the following assumptions.

(i) There exists a constant $M > 0$ such that

$$\mathbb{E}\big[\|G_k(u)\|_{\mathcal{U}}^2\big] \le M^2, \quad u \in \mathcal{U}_{\mathrm{ad}}. \tag{5.28}$$

(ii) The function $f(\cdot) = \mathbb{E}[F(\cdot)] + \wp(\cdot)$ is Fréchet-differentiable and strongly convex, that is, there exists $c > 0$ such that

$$f(u') \ge f(u) + \langle \nabla f(u), u' - u\rangle_{\mathcal{U}} + \frac{1}{2}c\|u' - u\|_{\mathcal{U}}^2 \quad \text{for all } u,\, u' \in \mathcal{U}. \tag{5.29}$$

For $\kappa > 1/(2c)$, these assumptions ensure that

$$\mathbb{E}\big[\|u_k - \bar{u}\|_{\mathcal{U}}^2\big] = O(k^{-1}), \tag{5.30}$$

where $\bar{u} \in \mathcal{U}_{\mathrm{ad}}$ is the unique solution to (5.25). In words, the expected error at the kth iteration is $O(k^{-1/2})$. Additionally, if $\nabla f(\cdot)$ is Lipschitz-continuous and $\bar{u} \in \mathcal{U}_{\mathrm{ad}}$ satisfies $\nabla f(\bar{u}) = 0$, then

$$\mathbb{E}[f(u_k) - f(\bar{u})] = O(k^{-1}). \tag{5.31}$$

Although convergent, SA is highly sensitive to the choice of κ in the step size $\gamma_k = \kappa/k$. Nemirovski, Juditsky, Lan and Shapiro (2009) demonstrate extremely slow convergence of SA for a poorly chosen κ, when minimizing a simple one-dimensional deterministic quadratic function, reducing the error in (5.30) by only 0.015 after one billion iterations. Convergence can be even slower without strong convexity (5.29) of f. Moreover, this choice of γ_k can be too small to achieve a reasonable convergence rate, while longer steps may not result in a convergent algorithm. For finite-dimensional convex problems, these issues can be overcome using the robust SA and mirror descent methods described in Nemirovski *et al.* (2009), which are based on the earlier work of Nemirovski and Yudin (1978). By employing the averages of the SA iterates u_k given by

$$\hat{u}_{j:k} := \sum_{i=j}^{k} \nu_i u_i \quad \text{for } \nu_\ell := \frac{\gamma_\ell}{\sum_{i=j}^{k} \nu_i},$$

Nemirovski *et al.* (2009) show that

$$\mathbb{E}[f(\hat{u}_{k:N}) - f(\bar{u})] = O(N^{-1/2}) \tag{5.32}$$

after N iterations, even when γ_k is chosen to be a constant, i.e. $\gamma_k = O(N^{-1/2})$, and when $\gamma_k = O(k^{-1/2})$. See equations (2.22) and (2.27) in Nemirovski *et al.* (2009). Although robust SA results in the slower convergence rate (5.32) compared to (5.31), it does not require strong convexity, and the dependence of the convergence rate on the choice of step size is significantly diminished. In particular, scaling γ_k by some positive constant θ results in the right-hand side of (5.32) being scaled by the constant $\max\{\theta, \theta^{-1}\}$.

SA methods are actively researched, primarily in the context of machine learning problems. In principle, these SA methods can be adapted to solve PDE-constrained optimization problems under uncertainty. However, as mentioned before, in this problem class, biases that result from discretization errors or iterative solutions of discretized PDEs at given samples, for example, are inevitable and must be incorporated into the design and analysis of the SA method. Moreover, except for special cases such as our model problem in Section 3.2.2, PDE-constrained optimization problems under uncertainty generally lead to non-convex problems, whereas many SA analyses in the literature assume convexity.

6. Conclusions and outlook

Research on PDE-constrained optimization under uncertainty is rapidly evolving and presents many challenging research issues related to problem formulation, mathematical analysis, numerical methods, and applications.

Much of the current literature has focused on unconstrained or deterministic control-constrained problems (e.g. (4.1c)). However, state-dependent constraints of the type (4.1b) and pointwise (in space and/or time) state constraints play an important role in applications. Even for deterministic PDE-constrained optimization problems, pointwise state constraints are challenging to handle, both theoretically and numerically. The treatment of such constraints in the presence of uncertainty is just beginning. See the references at the end of Section 4.3.

Likewise, the current literature has focused on PDE-constrained optimization problems in which the decisions are made before observing the uncertainty and are not revised after the uncertainty is revealed. However, multistage problems in which the decisions are made in stages, with later decisions incorporating the uncertainty revealed in the present and previous stages, are important in many applications. As already noted in, for example, Birge and Louveaux (2011) and Pflug and Pichler (2014), there is a connection with stochastic control problems, but also with nonlinear model predictive control (see e.g. Grüne and Pannek 2011 and Rawlings, Mayne and Diehl 2024) and reinforcement learning (see e.g. Bertsekas 2019, 2024, Meyn 2022 and Recht 2019).

A major computational bottleneck when trying to extend existing stochastic programming tools to the PDE-constrained case is the computational cost of the evaluation of the control-to-state map, i.e. the solution of the state PDE, that enters the objective function and constraint functions, e.g. F and G in (4.1), and the associated adjoint PDEs required for derivative computations. In the case of the model problem (3.51), these are the PDEs (3.52c) and (3.52c). Thus, methods to reduce the overall number of PDE solutions or the cost of an approximate PDE solution have an immediate impact. For example, for deterministic PDE-constrained optimization problems, projection-based reduced-order models (PROMs) of the state and adjoint PDEs have been integrated into a trust-region algorithm, which was proposed and analysed in Kouri *et al.* (2013, 2014) and which is Algorithm 1 with

$\mathcal{U}_{\text{ad}} = \mathcal{U}$. This integrated approach guarantees convergence to points u_* that satisfy the first-order necessary optimality conditions of the original problem (see also Theorem 5.1) at a computational cost that is a fraction of the cost incurred had only a high-fidelity PDE discretization been used; see e.g. Zahr and Farhat (2015) and Wen and Zahr (2023). The extension to PDE-constrained optimization problems with uncertainty has been limited to the smooth risk-averse formulations and applications with a relatively small number of random variables where the expectation can be approximated by sparse-grid quadrature or Voronoi-based piecewise constant approximation; see e.g. Zahr *et al.* (2019) and Zou *et al.* (2022).

Similarly, for the evaluation of expectations of quantities of interest involving a PDE solution at a fixed control, multilevel and multi-fidelity Monte Carlo methods, which were reviewed in Section 3.5.2, are successfully used to reduce the computational cost. Krumscheid and Nobile (2018) and Giles and Haji-Ali (2019) extend this to functions other than the expected value. Alternatively, Heinkenschloss, Kramer, Takhtaganov and Willcox (2018) and Heinkenschloss, Kramer and Takhtaganov (2020) developed an importance sampling-based approach that uses PROMs and their error estimates for AVaR (4.9) estimation, while Zou, Kouri and Aquino (2017, 2019) developed a locally adaptive PROM approximation for estimating general risk measures. These developments are for the evaluation of some risk measures at a fixed control. Extension of such approaches to the optimization context is an open area of research.

Acknowledgements

The research of MH has been funded in part by AFOSR grant FA9550-22-1-0004 and the research of DPK has been funded in part by AFOSR grant FA9550-22-1-0248 and the US Department of Energy Office of Science, Advanced Scientific Computing Research through the Early Career Research Program. The authors' research on this topic was also funded by previous grants from AFOSR, DARPA, DOE and NSF, which are detailed in the cited papers by the authors.

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell International Inc., for the US Department of Energy's National Nuclear Security Administration (DOE/NNSA) under contract DE-NA0003525. This written work is authored by an employee of NTESS. The employee, not NTESS, owns the right, title and interest in and to the written work and is responsible for its contents. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the US Government. The publisher acknowledges that the US Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for US Government purposes. The DOE will provide public access to results of federally sponsored research in accordance with the DOE Public Access Plan.

References

R. A. Adams and J. J. F. Fournier (2003), *Sobolev Spaces*, second edition, Elsevier Science.

A. Ahmad Ali, E. Ullmann and M. Hinze (2017), Multilevel Monte Carlo analysis for optimal control of elliptic PDEs with random coefficients, *SIAM/ASA J. Uncertain. Quantif.* **5**, 466–492.

C. D. Aliprantis and K. C. Border (2006), *Infinite Dimensional Analysis: A Hitchhiker's Guide*, third edition, Springer.

P. Artzner, F. Delbaen, J.-M. Eber and D. Heath (1999), Coherent measures of risk, *Math. Finance* **9**, 203–228.

I. Babuška, F. Nobile and R. Tempone (2007), A stochastic collocation method for elliptic partial differential equations with random input data, *SIAM J. Numer. Anal.* **45**, 1005–1034.

A. S. Bandeira, K. Scheinberg and L. N. Vicente (2014), Convergence of trust-region methods based on probabilistic models, *SIAM J. Optim.* **24**, 1238–1264.

W. Bangerth, H. Klie, M. Wheeler, P. Stoffa and M. Sen (2006), On optimization algorithms for the reservoir oil well placement problem, *Comput. Geosci.* **10**, 303–319.

R. J. Baraldi and D. P. Kouri (2023), A proximal trust-region method for nonsmooth optimization with inexact function and gradient evaluations, *Math. Program.* **201**, 559–598.

R. J. Baraldi and D. P. Kouri (2024), Local convergence analysis of an inexact trust-region method for nonsmooth optimization, *Optim. Lett.* **18**, 663–680.

K. Barty, J.-S. Roy and C. Strugarek (2007), Hilbert-valued perturbed subgradient algorithms, *Math. Oper. Res.* **32**, 551–562.

F. Bastin, C. Cirillo and P. L. Toint (2006), An adaptive Monte Carlo algorithm for computing mixed logit estimators, *Comput. Manag. Sci.* **3**, 55–79.

H. H. Bauschke and P. L. Combettes (2017), *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, second edition, Springer.

R. Becker, M. Braack, D. Meidner, R. Rannacher and B. Vexler (2007), Adaptive finite element methods for PDE-constrained optimal control problems, in *Reactive Flows, Diffusion and Transport* (W. Jäger, R. Rannacher and J. Warnatz, eds), Springer, pp. 177–205.

R. Becker, H. Kapp and R. Rannacher (2000), Adaptive finite element methods for optimal control of partial differential equations: Basic concepts, *SIAM J. Control Optim.* **39**, 113–132.

A. Ben-Tal and M. Teboulle (2007), An old-new concept of convex risk measures: The optimized certainty equivalent, *Math. Finance* **17**, 449–476.

M. P. Bendsøe and O. Sigmund (2003), *Topology Optimization: Theory, Methods and Applications*, Springer.

P. Benner, A. Onwunta and M. Stoll (2016), Block-diagonal preconditioning for optimal control problems constrained by PDEs with uncertain inputs, *SIAM J. Matrix Anal. Appl.* **37**, 491–518.

M. Benzi, G. H. Golub and J. Liesen (2005), Numerical solution of saddle point problems, *Acta Numer.* **14**, 1–137.

A. S. Berahas, L. Cao and K. Scheinberg (2021), Global convergence rate analysis of a generic line search algorithm with noise, *SIAM J. Optim.* **31**, 1489–1518.

J. O. Berger (1994), An overview of robust Bayesian analysis, *Test* **3**, 5–124.
D. P. Bertsekas (2019), *Reinforcement Learning and Optimal Control*, Athena Scientific Optimization and Computation Series, Athena Scientific.
D. P. Bertsekas (2024), Model predictive control and reinforcement learning: A unified framework based on dynamic programming, *IFAC-PapersOnLine* **58**, 363–383.
J. R. Birge and F. Louveaux (2011), *Introduction to Stochastic Programming*, Springer Series in Operations Research and Financial Engineering, second edition, Springer.
A. Borzì (2010), Multigrid and sparse-grid schemes for elliptic control problems with random coefficients, *Comput. Vis. Sci.* **13**, 153–160.
A. Borzì and V. Schulz (2009), Multigrid methods for PDE optimization, *SIAM Rev.* **51**, 361–395.
A. Borzì and V. Schulz (2012), *Computational Optimization of Systems Governed by Partial Differential Equations*, Vol. 8 of Computational Science & Engineering, SIAM.
S. C. Brenner and L. R. Scott (2008), *The Mathematical Theory of Finite Element Methods*, Vol. 15 of Texts in Applied Mathematics, third edition, Springer.
D. R. Brouwer and J. D. Jansen (2004), Dynamic optimization of waterflooding with smart wells using optimal control theory, *SPE J.* **9**, 391–402.
H.-J. Bungartz and M. Griebel (2004), Sparse grids, *Acta Numer.* **13**, 147–269.
R. G. Carter (1991), On the global convergence of trust region algorithms using inexact gradient information, *SIAM J. Numer. Anal.* **28**, 251–265.
R. G. Carter (1993), Numerical experience with a class of algorithms for nonlinear optimization using inexact function and gradient information, *SIAM J. Sci. Comput.* **14**, 368–388.
R. G. Carter and H. H. Rachford Jr (2003), Optimizing line-pack management to hedge against future load uncertainty, in *PSIG Annual Meeting*, PSIG, art. PSIG-0306.
J. Charrier, R. Scheichl and A. L. Teckentrup (2013), Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods, *SIAM J. Numer. Anal.* **51**, 322–352.
A. Chaudhuri, B. Kramer, M. Norton, J. O. Royset and K. E. Willcox (2022), Certifiable risk-based engineering design optimization, *AIAA J.* **60**, 551–565.
P. Chen and A. Quarteroni (2014), Weighted reduced basis method for stochastic optimal control problems with elliptic PDE constraint, *SIAM/ASA J. Uncertain. Quantif.* **2**, 364–396.
P. Chen, A. Quarteroni and G. Rozza (2016), Multilevel and weighted reduced basis method for stochastic optimal control problems constrained by Stokes equations, *Numer. Math.* **133**, 67–102.
P. Cheridito and T. Li (2009), Risk measures on Orlicz hearts, *Math. Finance* **19**, 189–214.
G. Ciaramella, F. Nobile and T. Vanzan (2024), A multigrid solver for PDE-constrained optimization with uncertain inputs, *J. Sci. Comput.* **101**, art. 13.
A. Cohen and R. DeVore (2015), Approximation of high-dimensional parametric PDEs, *Acta Numer.* **24**, 1–159.
A. Cohen, R. DeVore and C. Schwab (2010), Convergence rates of best N-term Galerkin approximations for a class of elliptic PDEs, *Found. Comput. Math.* **10**, 615–646.
S. S. Collis and M. Heinkenschloss (2002), Analysis of the streamline upwind/Petrov Galerkin method applied to the solution of optimal control problems. Technical report TR02–01, Department of Computational and Applied Mathematics, Rice University, Houston, TX. Also available at arXiv:2411.09828.

A. R. Conn, N. I. M. Gould and P. L. Toint (2000), *Trust-Region Methods*, MPS/SIAM Series on Optimization, SIAM.
S. Conti, H. Held, M. Pach, M. Rumpf and R. Schultz (2011), Risk averse shape optimization, *SIAM J. Control Optim.* **49**, 927–947.
S. Conti, M. Rumpf, R. Schultz and S. Tölkes (2018), Stochastic dominance constraints in elastic shape optimization, *SIAM J. Control Optim.* **56**, 3021–3034.
F. E. Curtis, K. Scheinberg and R. Shi (2019), A stochastic trust region algorithm based on careful step normalization, *INFORMS J. Optim.* **1**, 200–220.
M. Dambrine, C. Dapogny and H. Harbrecht (2015), Shape optimization for quadratic functionals and states with random right-hand sides, *SIAM J. Control Optim.* **53**, 3081–3103.
L. Dedé and A. Quarteroni (2005), Optimal control and numerical adaptivity for advection-diffusion equations, *M2AN Math. Model. Numer. Anal.* **39**, 1019–1040.
J. E. Dennis Jr and R. B. Schnabel (1996), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Vol. 16 of Classics in Applied Mathematics, SIAM.
D. Dentcheva and A. Ruszczyński (2003), Optimization with stochastic dominance constraints, *SIAM J. Optim.* **14**, 548–566.
J. Dick, F. Y. Kuo and I. H. Sloan (2013), High-dimensional integration: The quasi-Monte Carlo way, *Acta Numer.* **22**, 133–288.
J. Eckstein and D. P. Bertsekas (1992), On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators, *Math. Program.* **55**, 293–318.
I. Ekeland and R. Temam (1999), *Convex Analysis and Variational Problems*, Vol. 28 of Classics in Applied Mathematics, SIAM.
M. S. Eldred (2011), Design under uncertainty employing stochastic expansion methods, *Int. J. Uncertain. Quantif.* **1**, 119–146.
G. Fabbri, F. Gozzi and A. Świech (2017), *Stochastic Optimal Control in Infinite Dimension: Dynamic Programming and HJB Equations*, Vol. 82 of Probability Theory and Stochastic Modelling, Springer.
G. B. Folland (1999), *Real Analysis: Modern Techniques and Their Applications*, Pure and Applied Mathematics, second edition, Wiley.
H. Föllmer and A. Schied (2002), Convex measures of risk and trading constraints, *Finance Stoch.* **6**, 429–447.
H. Föllmer and A. Schied (2010), Convex and coherent risk measures, in *Encyclopedia of Quantitative Finance*, Wiley, pp. 355–363.
J. Garcke and M. Griebel, eds (2013), *Sparse Grids and Applications*, Vol. 88 of Lecture Notes in Computational Science and Engineering, Springer.
S. Garreis and M. Ulbrich (2017), Constrained optimization with low-rank tensors and applications to parametric problems with PDEs, *SIAM J. Sci. Comput.* **39**, A25–A54.
S. Garreis, T. Surowiec and M. Ulbrich (2021), An interior-point approach for solving risk-averse PDE-constrained optimization problems with coherent risk measures, *SIAM J. Optim.* **31**, 1–29.
C. Geiersbach and R. Henrion (2024), Optimality conditions in control problems with random state constraints in probabilistic or almost-sure form, *Math. Oper. Res*. Available at doi:10.1287/moor.2023.0177.
C. Geiersbach and M. Hintermüller (2022), Optimality conditions and Moreau–Yosida regularization for almost sure state constraints, *ESAIM Control Optim. Calc. Var.* **28**, art. 80.

C. Geiersbach and G. C. Pflug (2019), Projected stochastic gradients for convex constrained problems in Hilbert spaces, *SIAM J. Optim.* **29**, 2079–2099.
C. Geiersbach and T. Scarinci (2023), A stochastic gradient method for a class of nonlinear PDE-constrained optimal control problems under uncertainty, *J. Differential Equations* **364**, 635–666.
C. Geiersbach and W. Wollner (2020), A stochastic gradient method with mesh refinement for PDE-constrained optimization under uncertainty, *SIAM J. Sci. Comput.* **42**, A2750–A2772.
C. Geiersbach, R. Henrion and P. Pérez-Aros (2025), Numerical solution of an optimal control problem with probabilistic and almost sure state constraints, *J. Optim. Theory Appl.* **204**, art. 7.
B. Geihe, M. Lenz, M. Rumpf and R. Schultz (2013), Risk averse elastic shape optimization with parametrized fine scale geometry, *Math. Program.* **141**, 383–403.
T. Gerstner and M. Griebel (1998), Numerical integration using sparse grids, *Numer. Algorithms* **18**, 209–232.
T. Gerstner and M. Griebel (2003), Dimension-adaptive tensor-product quadrature, *Computing* **71**, 65–87.
M. B. Giles (2015), Multilevel Monte Carlo methods, *Acta Numer.* **24**, 259–328.
M. B. Giles and A.-L. Haji-Ali (2019), Multilevel nested simulation for efficient risk estimation, *SIAM/ASA J. Uncertain. Quantif.* **7**, 497–525.
S. Gratton, A. Sartenaer and P. L. Toint (2008), Recursive trust-region methods for multiscale nonlinear optimization, *SIAM J. Optim.* **19**, 414–444.
M. Griebel (2006), Sparse grids and related approximation schemes for higher dimensional problems, in *Foundations of Computational Mathematics, Santander 2005* (L. M. Pardo *et al.*, eds), Vol. 331 of London Mathematical Society Lecture Note Series, Cambridge University Press, pp. 106–161.
D. S. Grundvig and M. Heinkenschloss (2024), Line-search based optimization using function approximations with tunable accuracy, *Optim. Methods Softw*. Available at doi:10.1080/10556788.2024.2436192.
L. Grüne and J. Pannek (2011), *Nonlinear Model Predictive Control: Theory and Algorithms*, Communications and Control Engineering Series, Springer.
M. D. Gunzburger (2003), *Perspectives in Flow Control and Optimization*, SIAM.
M. D. Gunzburger, C. G. Webster and G. Zhang (2014), Stochastic finite element methods for partial differential equations with random input data, *Acta Numer.* **23**, 521–650.
P. A. Guth, V. Kaarnioja, F. Y. Kuo, C. Schillings and I. H. Sloan (2021), A quasi-Monte Carlo method for optimal control under uncertainty, *SIAM/ASA J. Uncertain. Quantif.* **9**, 354–383.
M. Heinkenschloss and D. Leykekhman (2010), Local error estimates for SUPG solutions of advection-dominated elliptic linear–quadratic optimal control problems, *SIAM J. Numer. Anal.* **47**, 4607–4638.
M. Heinkenschloss and L. N. Vicente (2002), Analysis of inexact trust-region SQP algorithms, *SIAM J. Optim.* **12**, 283–302.
M. Heinkenschloss, B. Kramer and T. Takhtaganov (2020), Adaptive reduced-order model construction for conditional value-at-risk estimation, *SIAM/ASA J. Uncertain. Quantif.* **8**, 668–692.
M. Heinkenschloss, B. Kramer, T. Takhtaganov and K. Willcox (2018), Conditional-value-at-risk estimation via reduced-order models, *SIAM/ASA J. Uncertain. Quantif.* **6**, 1395–1423.

M. Hintermüller and R. H. W. Hoppe (2008), Goal-oriented adaptivity in control constrained optimal control of partial differential equations, *SIAM J. Control Optim.* **47**, 1721–1743.

M. Hinze (2005), A variational discretization concept in control constrained optimization: The linear–quadratic case, *Comput. Optim. Appl.* **30**, 45–63.

M. Hinze, R. Pinnau, M. Ulbrich and S. Ulbrich (2009), *Optimization with PDE Constraints*, Vol. 23 of Mathematical Modelling, Theory and Applications, Springer.

K. Ito and S. S. Ravindran (1998), Optimal control of thermally convected fluid flows, *SIAM J. Sci. Comput.* **19**, 1847–1869.

J. Jahn (2007), *Introduction to the Theory of Nonlinear Optimization*, third edition, Springer.

J. D. Jakeman, D. P. Kouri and J. G. Huerta (2022), Surrogate modeling for efficiently, accurately and conservatively estimating measures of risk, *Reliab. Engrg Syst. Saf.* **221**, art. 108280.

P. Kolvenbach, O. Lass and S. Ulbrich (2018), An approach for robust PDE-constrained optimization with application to shape optimization of electrical engines and of dynamic elastic structures under uncertainty, *Optim. Eng.* **19**, 697–731.

D. P. Kouri (2012), An approach for the adaptive solution of optimization problems governed by partial differential equations with uncertain coefficients. PhD thesis, Department of Computational and Applied Mathematics, Rice University, Houston, TX.

D. P. Kouri (2014), A multilevel stochastic collocation algorithm for optimization of PDEs with uncertain coefficients, *SIAM/ASA J. Uncertain. Quantif.* **2**, 55–81.

D. P. Kouri (2017), A measure approximation for distributionally robust PDE-constrained optimization problems, *SIAM J. Numer. Anal.* **55**, 3147–3172.

D. P. Kouri (2019), Higher-moment buffered probability, *Optim. Lett.* **13**, 1223–1237.

D. P. Kouri and D. Ridzal (2018), Inexact trust-region methods for PDE-constrained optimization, in *Frontiers in PDE-Constrained Optimization* (H. Antil *et al.*, eds), Vol. 163 of IMA Volumes in Mathematics and its Applications, Springer, pp. 83–121.

D. P. Kouri and A. Shapiro (2018), Optimization of PDEs with uncertain inputs, in *Frontiers in PDE-Constrained Optimization* (H. Antil *et al.*, eds), Vol. 163 of IMA Volumes in Mathematics and its Applications, Springer, pp. 41–81.

D. P. Kouri and T. M. Surowiec (2016), Risk-averse PDE-constrained optimization using the conditional value-at-risk, *SIAM J. Optim.* **26**, 365–396.

D. P. Kouri and T. M. Surowiec (2018), Existence and optimality conditions for risk-averse PDE-constrained optimization, *SIAM/ASA J. Uncertain. Quantif.* **6**, 787–815.

D. P. Kouri and T. M. Surowiec (2020*a*), Epi-regularization of risk measures, *Math. Oper. Res.* **45**, 774–795.

D. P. Kouri and T. M. Surowiec (2020*b*), Risk-averse optimal control of semilinear elliptic PDEs, *ESAIM Control Optim. Calc. Var.* **26**, art. 53.

D. P. Kouri and T. M. Surowiec (2022), A primal–dual algorithm for risk minimization, *Math. Program.* **193**, 337–363.

D. P. Kouri, M. Heinkenschloss, D. Ridzal and B. G. van Bloemen Waanders (2013), A trust-region algorithm with adaptive stochastic collocation for PDE optimization under uncertainty, *SIAM J. Sci. Comput.* **35**, A1847–A1879.

D. P. Kouri, M. Heinkenschloss, D. Ridzal and B. G. van Bloemen Waanders (2014), Inexact objective function evaluations in a trust-region algorithm for PDE-constrained optimization under uncertainty, *SIAM J. Sci. Comput.* **36**, A3011–A3029.

D. P. Kouri, M. Staudigl and T. M. Surowiec (2023), A relaxation-based probabilistic approach for PDE-constrained optimization under uncertainty with pointwise state constraints, *Comput. Optim. Appl.* **85**, 441–478.

P. A. Krokhmal (2007), Higher moment coherent risk measures, *Quant. Finance* **7**, 373–387.

S. Krumscheid and F. Nobile (2018), Multilevel Monte Carlo approximation of functions, *SIAM/ASA J. Uncertain. Quantif.* **6**, 1256–1293.

F. Y. Kuo and D. Nuyens (2016), Application of quasi-Monte Carlo methods to elliptic PDEs with random diffusion coefficients: A survey of analysis and implementation, *Found. Comput. Math.* **16**, 1631–1696.

F. Y. Kuo, R. Scheichl, C. Schwab, I. H. Sloan and E. Ullmann (2017), Multilevel quasi-Monte Carlo methods for lognormal diffusion problems, *Math. Comp.* **86**, 2827–2860.

H. J. Kushner and P. Dupuis (2001), *Numerical Methods for Stochastic Control Problems in Continuous Time*, Vol. 24 of Stochastic Modelling and Applied Probability, second edition, Springer.

B. S. Lazarov and O. Sigmund (2011), Filters in topology optimization based on Helmholtz-type differential equations, *Int. J. Numer. Methods Engrg* **86**, 765–781.

B. S. Lazarov, M. Schevenels and O. Sigmund (2012*a*), Topology optimization considering material and geometric uncertainties using stochastic collocation methods, *Struct. Multidiscip. Optim.* **46**, 597–612.

B. S. Lazarov, M. Schevenels and O. Sigmund (2012*b*), Topology optimization with geometric uncertainties by perturbation techniques, *Int. J. Numer. Methods Engrg* **90**, 1321–1336.

H. Levy (1992), Stochastic dominance and expected utility: Survey and analysis, *Manag. Sci.* **38**, 555–593.

R. M. Lewis and S. G. Nash (2005), Model problems for the multigrid optimization of systems governed by differential equations, *SIAM J. Sci. Comput.* **26**, 1811–1837.

D. Leykekhman (2012), Investigation of commutative properties of discontinuous Galerkin methods in PDE constrained optimal control problems, *J. Sci. Comput.* **53**, 483–511.

D. Leykekhman and M. Heinkenschloss (2012), Local error analysis of discontinuous Galerkin methods for advection-dominated elliptic linear–quadratic optimal control problems, *SIAM J. Numer. Anal.* **50**, 2012–2038.

J.-L. Lions (1971), *Optimal Control of Systems Governed by Partial Differential Equations*, Vol. 170 of Grundlehren der mathematischen Wissenschaften, Springer.

W. G. Litvinov (2000), *Optimization in Elliptic Problems with Applications to Mechanics of Deformable Bodies and Fluid Mechanics*, Vol. 119 of Operator Theory: Advances and Applications, Birkhäuser.

G. J. Lord, C. E. Powell and T. Shardlow (2014), *An Introduction to Computational Stochastic PDEs*, Cambridge Texts in Applied Mathematics, Cambridge University Press.

A. Mafusalov and S. Uryasev (2018), Buffered probability of exceedance: Mathematical properties and optimization, *SIAM J. Optim.* **28**, 1077–1103.

H. Markowitz (1952), Portfolio selection, *J. Finance* **7**, 77–91.

M. Markowski (2022), Efficient solution of smoothed risk-averse PDE-constrained optimization problems. PhD thesis, Department of Computational and Applied Mathematics, Rice University, Houston, TX.

M. Martin and F. Nobile (2021), PDE-constrained optimal control problems with uncertain parameters using SAGA, *SIAM/ASA J. Uncertain. Quantif.* **9**, 979–1012.
M. Martin, S. Krumscheid and F. Nobile (2021), Complexity analysis of stochastic gradient methods for PDE-constrained optimal control problems with uncertain parameters, *ESAIM Math. Model. Numer. Anal.* **55**, 1599–1633.
S. Meyn (2022), *Control Systems and Reinforcement Learning*, Cambridge University Press.
J. Milz (2023*a*), Consistency of Monte Carlo estimators for risk-neutral PDE-constrained optimization, *Appl. Math. Optim.* **87**, art. 57.
J. Milz (2023*b*), Reliable error estimates for optimal control of linear elliptic PDEs with random inputs, *SIAM/ASA J. Uncertain. Quantif.* **11**, 1139–1163.
J. Milz (2023*c*), Sample average approximations of strongly convex stochastic programs in Hilbert spaces, *Optim. Lett.* **17**, 471–492.
J. Milz and M. Ulbrich (2024), Sample size estimates for risk-neutral semilinear PDE-constrained optimization, *SIAM J. Optim.* **34**, 844–869.
B. Mohammadi and O. Pironneau (2004), Shape optimization in fluid mechanics, *Annu. Rev. Fluid Mech.* **36**, 255–279.
S. G. Nash (2000), A multigrid approach to discretized optimization problems, *Optim. Methods Softw.* **14**, 99–116.
A. Nemirovski and D. Yudin (1978), On Cezari's convergence of the steepest descent method for approximating saddle point of convex–concave functions, *Soviet Math. Dokl.* **19**, 258–269.
A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro (2009), Robust stochastic approximation approach to stochastic programming, *SIAM J. Optim.* **19**, 1574–1609.
F. Nobile and T. Vanzan (2023), Preconditioners for robust optimal control problems under uncertainty, *Numer. Linear Algebra Appl.* **30**, art. e2472.
F. Nobile and T. Vanzan (2024), A combination technique for optimal control problems constrained by random PDEs, *SIAM/ASA J. Uncertain. Quantif.* **12**, 693–721.
J. Nocedal and S. J. Wright (2006), *Numerical Optimization*, second edition, Springer.
J. W. Pearson and J. Pestana (2020), Preconditioners for Krylov subspace methods: an overview, *GAMM-Mitt.* **43**, art. e202000015.
B. Peherstorfer (2019), Multifidelity Monte Carlo estimation with adaptive low-fidelity models, *SIAM/ASA J. Uncertain. Quantif.* **7**, 579–603.
B. Peherstorfer, K. Willcox and M. D. Gunzburger (2016), Optimal model management for multifidelity Monte Carlo estimation, *SIAM J. Sci. Comput.* **38**, A3163–A3194.
B. Peherstorfer, K. Willcox and M. D. Gunzburger (2018), Survey of multifidelity methods in uncertainty propagation, inference, and optimization, *SIAM Rev.* **60**, 550–591.
G. C. Pflug and A. Pichler (2014), *Multistage Stochastic Optimization*, Springer Series in Operations Research and Financial Engineering, Springer.
C. Phelps, J. O. Royset and Q. Gong (2016), Optimal control of uncertain systems using sample average approximations, *SIAM J. Control Optim.* **54**, 1–29.
A. Quarteroni (2009), *Numerical Models for Differential Problems*, Vol. 2 of MS&A: Modeling, Simulation and Applications, Springer.
J. B. Rawlings, D. Q. Mayne and M. M. Diehl (2024), *Model Predictive Control: Theory, Computation, and Design*, second edition, Nob Hill Publishing.
B. Recht (2019), A tour of reinforcement learning: The view from continuous control, *Annu. Rev.* **2**, 253–279.

H. Robbins and S. Monro (1951), A stochastic approximation method, *Ann. Math. Statist.* **22**, 400–407.
R. T. Rockafellar (1971), Convex integral functionals and duality, in *Contributions to Nonlinear Functional Analysis* (E. H. Zarantonello, ed.), Elsevier, pp. 215–236.
R. T. Rockafellar (1973), Penalty methods and augmented Lagrangians in nonlinear programming, in *Fifth Conference on Optimization Techniques, Part I* (R. Conti and A. Ruberti, eds), Vol. 3 of Lecture Notes in Computer Science, Springer, pp. 418–425.
R. T. Rockafellar (2018), Solving stochastic programming problems with risk measures by progressive hedging, *Set-Valued Var. Anal.* **26**, 759–768.
R. T. Rockafellar and J. O. Royset (2010), On buffered failure probability in design and optimization of structures, *Reliab. Engrg Syst. Saf.* **95**, 499–510.
R. T. Rockafellar and S. Uryasev (2002), Conditional value-at-risk for general loss distributions, *J. Banking Finance* **26**, 1443–1471.
R. T. Rockafellar and S. Uryasev (2020), Minimizing buffered probability of exceedance by progressive hedging, *Math. Program.* **181**, 453–472.
R. T. Rockafellar and R. J.-B. Wets (1991), Scenarios and policy aggregation in optimization under uncertainty, *Math. Oper. Res.* **16**, 119–147.
W. W. Rogosinski (1958), Moments of non-negative mass, *Proc. R. Soc. A* **245**(1240), 1–27.
W. Römisch and T. M. Surowiec (2024), Asymptotic properties of Monte Carlo methods in elliptic PDE-constrained optimization under uncertainty, *Numer. Math.* **156**, 1887–1914.
E. Rosseel and G. N. Wells (2012), Optimal control with stochastic PDE constraints and uncertain controls, *Comput. Methods Appl. Mech. Engrg* **213–216**, 152–167.
J. O. Royset, L. Bonfiglio, G. Vernengo and S. Brizzolara (2017), Risk-adaptive set-based design and applications to shaping a hydrofoil, *J. Mech. Design* **139**, art. 101403.
A. Ruszczynski and A. Shapiro, eds (2003), *Stochastic Programming*, Vol. 10 of Handbooks in Operations Research and Management Science, Elsevier.
H. Scarf (1958), A min-max solution of an inventory problem, in *Studies in The Mathematical Theory of Inventory and Production*, Stanford University Press, pp. 201–209.
V. Schulz and C. Schillings (2013), Optimal aerodynamic design under uncertainty, in *Management and Minimisation of Uncertainties and Errors in Numerical Aerodynamics: Results of the German collaborative project MUNA* (B. Eisfeld *et al.*, eds), Springer, pp. 297–338.
A. Shapiro (2017), Distributionally robust stochastic programming, *SIAM J. Optim.* **27**, 2258–2275.
A. Shapiro, D. Dentcheva and A. Ruszczyński (2014), *Lectures on Stochastic Programming: Modeling and Theory*, Vol. 9 of MPS/SIAM Series on Optimization, second edition, SIAM.
R. C. Smith (2014), *Uncertainty Quantification: Theory, Implementation, and Applications*, Vol. 12 of Computational Science & Engineering, SIAM.
S. A. Smolyak (1963), Quadrature and interpolation formulas for tensor products of certain classes of functions, *Dokl. Akad. Nauk SSSR* **148**, 1042–1045.
J. E. Spingarn (1983), Partial inverse of a monotone operator, *Appl. Math. Optim.* **10**, 247–265.
J. E. Spingarn (1985), Applications of the method of partial inverses to convex programming: Decomposition, *Math. Program.* **32**, 199–223.

A. L. Teckentrup, R. Scheichl, M. B. Giles and E. Ullmann (2013), Further analysis of multilevel Monte Carlo methods for elliptic PDEs with random coefficients, *Numer. Math.* **125**, 569–600.

H. Tiesler, R. M. Kirby, D. Xiu and T. Preusser (2012), Stochastic collocation for optimal control problems with stochastic PDE constraints, *SIAM J. Control Optim.* **50**, 2659–2682.

P. L. Toint (1988), Global convergence of a class of trust-region methods for nonconvex minimization in Hilbert space, *IMA J. Numer. Anal.* **8**, 231–252.

F. Tröltzsch (2010), *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, Vol. 112 of Graduate Studies in Mathematics, American Mathematical Society.

S. Uryasev (1994), Derivatives of probability functions and integrals over sets given by inequalities, *J. Comput. Appl. Math.* **56**, 197–223.

S. Uryasev (1995), Derivatives of probability functions and some applications, *Ann. Oper. Res.* **56**, 287–311.

S. Uryasev, ed. (2000), *Probabilistic Constrained Optimization: Methodology and Applications*, Vol. 49 of Nonconvex Optimization and its Applications, Kluwer Academic.

S. Uryasev and R. T. Rockafellar (2001), Conditional value-at-risk: Optimization approach, in *Stochastic Optimization: Algorithms and Applications* (S. Uryasev and P. M. Pardalos, eds), Vol. 54 of Applied Optimization, Kluwer, pp. 411–435.

A. Van Barel and S. Vandewalle (2019), Robust optimization of PDEs with random coefficients using a multilevel Monte Carlo method, *SIAM/ASA J. Uncertain. Quantif.* **7**, 174–202.

A. Van Barel and S. Vandewalle (2021), MG/OPT and multilevel Monte Carlo for robust optimization of PDEs, *SIAM J. Optim.* **31**, 1850–1876.

J. von Neumann and O. Morgenstern (2007), *Theory of Games and Economic Behavior*, sixtieth-anniversary edition, Princeton University Press.

A. J. Wathen (2015), Preconditioning, *Acta Numer.* **24**, 329–376.

T. Wen and M. J. Zahr (2023), A globally convergent method to accelerate large-scale optimization using on-the-fly model hyperreduction: Application to shape optimization, *J. Comput. Phys.* **484**, art. 112082.

D. Xiu (2010), *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton University Press.

M. J. Zahr and C. Farhat (2015), Progressive construction of a parametric reduced-order model for PDE-constrained optimization, *Int. J. Numer. Methods Engrg* **102**, 1111–1135.

M. J. Zahr, K. T. Carlberg and D. P. Kouri (2019), An efficient, globally convergent method for optimization under uncertainty using adaptive model reduction and sparse grids, *SIAM/ASA J. Uncertain. Quantif.* **7**, 877–912.

M. Zandvliet, G. Van Essen and D. Brouwer (2008), Adjoint-based well-placement optimization under production constraints, *SPE J.* **13**, 392–399.

M. Zhou, B. S. Lazarov and O. Sigmund (2014), Topology optimization for optical projection lithography with manufacturing uncertainties, *Appl. Optics* **53**, 2720–2729.

J. C. Ziems (2013), Adaptive multilevel inexact SQP-methods for PDE-constrained optimization with control constraints, *SIAM J. Optim.* **23**, 1257–1283.

J. C. Ziems and S. Ulbrich (2011), Adaptive multilevel inexact SQP methods for PDE-constrained optimization, *SIAM J. Optim.* **21**, 1–40.

Z. Zou, D. P. Kouri and W. Aquino (2017), An adaptive sampling approach for solving PDEs with uncertain inputs and evaluating risk, in *19th AIAA Non-Deterministic Approaches Conference*, AIAA SciTech Forum, art. AIAA 2017-1325.

Z. Zou, D. P. Kouri and W. Aquino (2018), A locally adapted reduced basis method for solving risk-averse PDE-constrained optimization problems, in *2018 AIAA Non-Deterministic Approaches Conference*, AIAA SciTech Forum, art. AIAA 2018-2174.

Z. Zou, D. P. Kouri and W. Aquino (2019), An adaptive local reduced basis method for solving PDEs with uncertain inputs and evaluating risk, *Comput. Methods Appl. Mech. Engrg* **345**, 302–322.

Z. Zou, D. P. Kouri and W. Aquino (2022), A locally adapted reduced-basis method for solving risk-averse PDE-constrained optimization problems, *SIAM/ASA J. Uncertain. Quantif.* **10**, 1629–1651.