Optimization of PDEs with Uncertain Inputs

Drew P. Kouri and Alexander Shapiro

Abstract Uncertainty pervades nearly all science and engineering applications including the optimal control and design of systems governed by partial differential equations (PDEs). In many applications, it is critical to determine optimal solutions that are resilient to the inherent uncertainty in unknown boundary conditions, inaccurate coefficients, and unverifiable modeling assumptions. In this tutorial, we develop a general theory for PDE-constrained optimization problems in which inputs or coefficients of the PDE are uncertain. We discuss numerous approaches for incorporating risk preference and conservativeness into the optimization problem formulation, motivated by concrete engineering applications. We conclude with a discussion of nonintrusive solution methods and numerical examples.

1 Introduction

Optimization problems constrained by partial differential equations (PDEs) arise in a number of science and engineering applications as optimal control and design problems. More often than not, the governing physical equations (PDEs) are fraught with uncertainty including uncertain coefficient, unknown boundary and initial conditions, and unverifiable modeling assumptions. When uncertainty exists, it is critical to determine optimal solutions that account for and in some sense are resilient to this uncertainty.

D. P. Kouri (✉)
Center for Computing Research, Sandia National Laboratories, Albuquerque, NM 87185-9999, USA
e-mail: dpkouri@sandia.gov

A. Shapiro
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA
e-mail: ashapiro@isye.gatech.edu

H. Antil et al. (eds.), *Frontiers in PDE-Constrained Optimization*, The IMA Volumes in Mathematics and its Applications 163,
https://doi.org/10.1007/978-1-4939-8636-1_2

Such problems arise for example in the topological design of elastic structures [5, 67, 77, 78]. Recently, topology optimization has gained increased interest due to the emergence of additive manufacturing technologies [56, 109]. There are many uncertainties associated with additively manufactured components such as random grain structures [2, 21], unknown internal forces due to, e.g., residual stresses [52], and potentially variable operating conditions such as external loads. The target then is to design a structure that is, for example, maximally stiff and in some sense reliable given the uncertain material properties and loads. Another common application is the secondary oil recovery phase in petroleum engineering. In this example, an oil company may choose to inject water or other solvents into a reservoir to increase pressure and produce more oil. Of course the subsurface rock properties are unknown but may be estimated from core samples, flow and pressure history [40, 73, 118], or seismic imaging [65, 101, 111]. The optimization problem is to determine the well locations and injection rates that maximize the net present value of the reservoir [3, 10, 102, 119]. However, the optimal rates should be resilient to the inherent uncertainties of the subsurface.

The purpose of this chapter is to review concepts from stochastic programming [25, 55, 75, 90, 108] that play fundamental roles in formulating PDE-constrained optimization problems in a rigorous and physically meaningful (application relevant) manner. In particular, we discuss the basic extension from deterministic PDE-constrained optimization to optimization of PDEs with uncertain inputs by introducing conditions on the deterministic objective function and PDE solution that ensure a well-defined stochastic problem. When the PDE has uncertain inputs, the associated state (PDE solution) becomes a random field. Substituting the random field solution into the objective function results in a random objective function. In order to solve this problem, we must replace the random objective function with a scalar quantity. There are a number of approaches for doing this. In particular, we discuss risk measures [4, 99, 115], probabilistic functions [76, 81, 93, 114], and distributionally robust optimization [15, 107, 121].

In addition to problem formulation, we discuss the challenges associated with the numerical solution of such problems. Many stochastic formulations result in nonsmooth objective functions which motivate new research on rapidly converging nonsmooth optimization algorithms that can exploit structures inherent to PDE-constrained optimization. We present three classical approaches for approximating and solving stochastic optimization problems: stochastic approximation [80, 89, 91], sample average and quadrature approximation [61, 62, 87, 106], and the progressive hedging algorithm [96].

The remainder of this chapter is structured as follows. We first discuss tensor products of Banach spaces. Such spaces play a central role in the functional analytic framework for PDE-constrained optimization under uncertainty. Next, we provide a general problem formulation and, under certain assumptions, show the existence of minimizers as well as first-order necessary optimality conditions. We demonstrate these results on the standard linear-elliptic quadratic control problem. In the following section, we discuss specific problem formulations including risk measures, probabilistic functions, and distributionally robust optimization. We

then introduce three basic numerical methods: stochastic approximation, sample average and quadrature approximation, and the progressive hedging algorithm. We briefly discuss convergence of these methods and conclude with a numerical demonstration.

2 Tensor Product Spaces

Let $(\Omega, \mathscr{F})$ be a measurable space where Ω is the set of possible outcomes and $\mathscr{F}$ is a σ-algebra of events. We denote the expected value of a random variable $X : \Omega \to \mathbb{R}$ with respect to a probability measure $P : \mathscr{F} \to [0, 1]$ defined on the measurable space $(\Omega, \mathscr{F})$ by

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

We denote the usual Lebesgue space of $r \in [1, \infty)$ integrable real-valued functions (defined up to a set of P-measure zero) by

$$L^r(\Omega, \mathscr{F}, P) := \left\{\theta : \Omega \to \mathbb{R} : \theta \text{ is } \mathscr{F}\text{-measurable}, \mathbb{E}_P[|\theta|^r] < \infty\right\}.$$

If $r = \infty$, then

$$L^\infty(\Omega, \mathscr{F}, P) := \left\{\theta : \Omega \to \mathbb{R} : \theta \text{ is } \mathscr{F}\text{-measurable}, \operatorname*{ess\,sup}_{\omega\in\Omega} |\theta(\omega)| < \infty\right\}.$$

The Lebesgue spaces defined on $(\Omega, \mathscr{F}, P)$ are Banach spaces and serve as natural spaces for real-valued random variables (i.e., $\mathscr{F}$-measurable functions). In the context of PDE-constrained optimization with uncertain inputs, the PDE solutions will be Sobolev space-valued random elements, which motivate the use of tensor-product vector spaces. Given any real Banach space V, the tensor-product vector space associated with $L^r(\Omega, \mathscr{F}, P)$ and V is

$$L^r(\Omega, \mathscr{F}, P) \otimes V := \operatorname{span}\left\{\theta v \,:\, \theta \in L^r(\Omega, \mathscr{F}, P),\ v \in V\right\},$$

i.e., the linear span of all products of elements of $L^r(\Omega, \mathscr{F}, P)$ and V. In general, there are many norms associated with $L^r(\Omega, \mathscr{F}, P) \otimes V$, including the natural projective and injective norms (cf. [35] and [100]). In this work, we restrict our attention to the so-called Bochner norms

$$\begin{cases} \|u\|_{L^r(\Omega,\mathscr{F},P)\otimes V} = \mathbb{E}_P[\|u\|_V^r]^{\frac{1}{r}} & \text{if } 1 \le r < \infty, \\ \|u\|_{L^\infty(\Omega,\mathscr{F},P)\otimes V} = \operatorname*{ess\,sup}_{\omega\in\Omega} \|u(\omega)\|_V & \text{if } r = \infty. \end{cases}$$

The space $L^r(\Omega, \mathscr{F}, P) \otimes V$ endowed with the corresponding Bochner norm is not complete and hence is not a Banach space. However, the completion of

$L^r(\Omega, \mathscr{F}, P) \otimes V$ with respect to its Bochner norm is isomorphic to the Bochner space

$$L^r(\Omega, \mathscr{F}, P; V) := \left\{u : \Omega \to V : u \text{ is strongly } \mathscr{F}\text{-measurable}, \mathbb{E}_P[\|u\|_V^r] < \infty\right\}$$

if $r \in [1, \infty)$ and

$$L^\infty(\Omega, \mathscr{F}, P; V) := \left\{u : \Omega \to V : u \text{ is strongly } \mathscr{F}\text{-measurable}, \operatorname*{ess\,sup}_{\omega \in \Omega} \|u(\omega)\|_V < \infty\right\}$$

if $r = \infty$ (again functions in $L^r(\Omega, \mathscr{F}, P; V)$ are defined up to a set of measure zero) [35, Sect. 7.1]. Here, a function $u : \Omega \to V$ is strongly $\mathscr{F}$-measurable if there exists a sequence of V-valued simple (piecewise constant, countably-valued) functions defined on sets in $\mathscr{F}$ that converges to u P-almost everywhere (P-a.e.) [53, Def. 3.5.4].

It is worth pointing out that the tensor-product vector space $L^r(\Omega, \mathscr{F}, P) \otimes V$ consists of functions

$$u = \sum_{i=1}^{N} \theta_i v_i, \quad \theta_i \in L^r(\Omega, \mathscr{F}, P), \quad v_i \in V, \quad i = 1, \ldots, N$$

for some $N \in \mathbb{N}$, and thus provides a natural approximation space for functions in $L^r(\Omega, \mathscr{F}, P; V)$. This fact is exploited by many uncertainty quantification methods. In particular, polynomial chaos [58, 122], stochastic Galerkin [8, 9], tensor decomposition, [47, 59] and other projection-based methods for approximating PDEs with uncertain inputs decompose the PDE solution into sums of random and spatial components. These two components are then approximated separately using, e.g., polynomial approximation in $L^r(\Omega, \mathscr{F}, P)$ and finite elements in V.

3 Problem Formulation

In this section, we provide the general formulation of our optimization problem. Let U and Z be real reflexive Banach spaces, and let Y be a real Banach space. Here U denotes the deterministic state space, Z denotes the space of optimization variables (i.e., controls, designs, etc.), and Y denotes the PDE residual space. The optimization variables $z \in Z$ are always deterministic and represent a control or design that must be implemented prior to observing the randomness in the system. Stochastic controls do however arise in time-dependent decision processes and multistage stochastic programs in which case the concept of time consistency plays a central role. Time consistency is based on the famous quotation of Bellman: *"An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the*

state resulting from the first decision" [13]. In this review, we restrict our attention to optimization problems constrained by steady (i.e., time-independent, stationary) PDEs. For a more detailed discussion of dynamic stochastic programs (without PDEs) and time consistency, we direct the interested reader to [108, Ch. 6.8].

Before describing the optimization problem, we assume that the uncertainty in the PDE constraint is represented by a finite random vector $\xi : \Omega \to \Xi$ where $\Xi := \xi(\Omega) \subseteq \mathbb{R}^m$ with $m \in \mathbb{N}$ (i.e., ξ is a $\mathscr{F}$-measurable vector-valued function). In the literature, this is called the *finite-dimensional noise assumption* [7, 83] and facilitates numerical approximations such as polynomial chaos and stochastic collocation [7, 9, 58, 83]. Such a finite-dimensional representation is often achieved using a truncated Karhunen–Loève expansion [57, 69]. More importantly, this assumption permits a change of variables in which the PDE and objective function depend only on the "deterministic" parameters $\xi \in \Xi$. This change of variables transforms our original uncertainty model defined on the probability space $(\Omega, \mathscr{F}, P)$ to a model defined on the probability space $(\Xi, \mathscr{B}, \mathbb{P})$ where $\mathscr{B} \subseteq 2^{\Xi}$ is the σ-algebra generated by the sets $\xi^{-1}(A)$ for $A \in \mathscr{F}$ and $\mathbb{P} := P \circ \xi^{-1}$ is the probability law of ξ. In this new setting, we define the Bochner and Lebesgue spaces analogously to the definitions in Section 2. Throughout, we will abuse notation and let ξ denote the random variable $\xi(\omega)$ as well as its realizations. Recently, researchers in uncertainty quantification have developed and analyzed methods for handling infinite-dimensional uncertainties, e.g., $\xi(\omega)$ is a sequence of real numbers for each $\omega \in \Omega$. For example, see [31]. Since all practical computational methods for solving PDEs with uncertain inputs and their corresponding optimization problems require a finite (i.e., computer) representation of the uncertainty, we restrict our attention to the finite-dimensional noise setting. Finally, it is worth noting that no result in this section requires the finite-dimensional noise assumption. However, we work under this assumption to simplify the presentation in the forthcoming sections.

Now, let $Z_{\text{ad}} \subseteq Z$ be a closed convex subset of optimization variables, let $e : U \times Z_{\text{ad}} \times \Xi \to Y$ denote, e.g., a PDE in weak form, and consider the equality constraint

$$e(u, z, \xi(\omega)) = 0. \tag{1}$$

The goal of this article is to understand and solve general stochastic optimization problems with the form

$$\min_{z \in Z_{\text{ad}}} \left\{ \mathfrak{J}(z) := \mathscr{R}(J(S(z;\xi),\xi)) + \wp(z) \right\} \tag{2}$$

where $\mathscr{R}$ is a functional that maps random variables on $(\Xi, \mathscr{B})$ into the real numbers, $J : U \times \Xi \to \mathbb{R}$ is the *uncertain objective function*, $\wp : Z \to \mathbb{R}$ is a control penalty, and $S(z;\cdot) : \Xi \to U$ satisfies $e(S(z;\xi), z, \xi) = 0$ for $\mathbb{P}$-almost every $\xi \in \Xi$ (or equivalently $e(S(z;\xi(\omega)), z, \xi(\omega)) = 0$ for P-almost every $\omega \in \Omega$). Throughout, we denote the *reduced uncertain objective function* by

$$\mathscr{J}(z) := J(S(z;\xi),\xi). \tag{3}$$

Note that $\mathscr{J}(z)$ is also a function of ξ and hence is viewed as a random variable mapping Z_{ad} into a space of real-valued random variables on $(\Xi, \mathscr{B})$.

To ensure the PDE constraint $e(u,z,\xi) = 0$ is well posed, we require that it is uniquely solvable and the solution is in $L^q(\Xi,\mathscr{B},\mathbb{P};U)$ for some $q \in [1,\infty]$. We make this statement rigorous in the following assumption.

Assumption 1 (Properties of the Solution Map) *For each $z \in Z_{\text{ad}}$, there exists a unique mapping $S(z;\cdot) : \Xi \to U$ that solves $e(S(z;\xi),z,\xi) = 0$ for $\mathbb{P}$-almost all $\xi \in \Xi$ and satisfies the following properties:*

1. **Measurability:** *$S(z;\cdot) : \Xi \to U$ is strongly $\mathscr{B}$-measurable for all $z \in Z_{\text{ad}}$.*
2. **Growth Condition:** *There exists $q \in [1,\infty]$, a nonnegative random variable $C \in L^q(\Xi,\mathscr{B},\mathbb{P})$, and a nonnegative increasing function $\varrho : [0,\infty) \to [0,\infty)$ such that*

$$\|S(z;\xi)\|_U \leq C(\xi)\varrho(\|z\|_Z)$$

   for $\mathbb{P}$-almost all $\xi \in \Xi$ and for all $z \in Z_{\text{ad}}$.
3. **Continuity:** *S satisfies the continuity property*

$$z_n \rightharpoonup z \quad \text{in } Z_{\text{ad}} \quad \Longrightarrow \quad S(z_n;\cdot) \rightharpoonup S(z;\cdot) \quad \text{in } U, \mathbb{P}\text{-a.e.}$$

Assumptions 1.1–2 ensure that $S : Z_{\text{ad}} \to L^q(\Xi,\mathscr{B},\mathbb{P};U)$. Additionally, Assumption 1 combined with the Lebesgue Dominated Convergence Theorem ensure S is weakly continuous from Z into $L^q(\Xi,\mathscr{B},\mathbb{P};U)$ [63, Sect. 2.2]. We similarly assume there exists $p \in [1,\infty]$ such that the reduced uncertain objective function satisfies

$$\mathscr{J}(z) \in L^p(\Xi,\mathscr{B},\mathbb{P}) \quad \forall\, z \in Z_{\text{ad}}.$$

To simplify notation, we denote the realization of $\mathscr{J}(z)$ at ξ, i.e., $[\mathscr{J}(z)](\xi)$, by $\mathscr{J}(z,\xi)$. For example, the authors in [63] postulate the following assumptions on the uncertain objective function.

Assumption 2 (Properties of the Objective Function) *There exists $1 \leq p < \infty$ such that the function $J : U \times \Xi \to \mathbb{R}$ satisfies:*

1. ***Carathéodory:*** *J is a Carathéodory function, i.e., $J(\cdot,\xi)$ is continuous for $\mathbb{P}$-almost every $\xi \in \Xi$ and $J(u,\cdot)$ is $\mathscr{B}$-measurable for all $u \in U$.*
2. ***Growth Condition:*** *If $q < \infty$, then there exists $a \in L^p(\Xi,\mathscr{B},\mathbb{P})$ with $a \geq 0$ $\mathbb{P}$-a.e. and $c > 0$ such that*

$$|J(u,\xi)| \leq a(\xi) + c\|u\|_U^{q/p} \quad \forall\, u \in U \text{ and } \mathbb{P}\text{-almost all } \xi \in \Xi$$

If $q = \infty$, then for all $c > 0$ there exists $\gamma_c \in L^p(\Xi, \mathscr{B}, \mathbb{P})$ such that

$$|J(u, \xi)| \leq \gamma_c(\xi) \quad \mathbb{P}\text{-a.e. } \xi \quad \forall\, u \in U,\ \|u\|_U \leq c.$$

3. ***Convexity:*** *$J(\cdot, \xi)$ is convex for $\mathbb{P}$-almost every $\xi \in \Xi$.*

Assumptions 2.1–2 combined with Krasnosel'skii's Theorem [116, Thm. 19.1] ensure that the uncertain objective function $u \mapsto J(u, \cdot)$ is continuous from $L^q(\Xi, \mathscr{B}, \mathbb{P}; U)$ into $L^p(\Xi, \mathscr{B}, \mathbb{P})$.

3.1 Existence of Minimizers and Optimality Conditions

In this section, we present one set of assumptions on $\mathscr{R}$ that ensure the existence of minimizers of (2). In addition, when a minimizer of (2) exists, we characterize the first-order necessary optimality conditions that it satisfies.

Theorem 1 *Let Assumptions 1 and 2 hold, and define $\mathscr{X} := L^p(\Xi, \mathscr{B}, \mathbb{P})$ where $p \in [1, \infty)$ is defined in Assumption 2. Moreover, suppose that $\wp : Z \to \mathbb{R}$ is weakly lower semicontinuous and $\mathscr{R} : \mathscr{X} \to \mathbb{R}$ is convex, and satisfies the monotonicity property: for any $X,\ X' \in \mathscr{X}$,*

$$X \leq X' \quad \mathbb{P}\text{-a.e.} \qquad \Longrightarrow \qquad \mathscr{R}(X) \leq \mathscr{R}(X'). \tag{4}$$

Finally, assume that the level set $\{z \in Z_{\mathrm{ad}} : \mathfrak{J}(z) \leq \gamma\}$ is nonempty and bounded for some $\gamma \in \mathbb{R}$. Then problem (2) has an optimal solution, i.e., there exists $z_\star \in Z_{\mathrm{ad}}$ such that $\mathfrak{J}(z_\star) \leq \mathfrak{J}(z)$ for all $z \in Z_{\mathrm{ad}}$.

Proof Since $\mathscr{R}$ is finite, convex, and satisfies (4), it is continuous and subdifferentiable [108, Prop. 6.6]. The Fenchel–Young inequality then ensures that

$$\mathscr{R}(\mathscr{J}(z)) \geq \mathbb{E}[\theta\, \mathscr{J}(z)] - \mathscr{R}^*(\theta) \quad \forall\, z \in Z_{\mathrm{ad}},\ \theta \in \operatorname{dom} \mathscr{R}^* \tag{5}$$

where

$$\mathscr{R}^*(\theta) = \sup_{X \in \mathscr{X}} \{\mathbb{E}[\theta X] - \mathscr{R}(X)\}$$

is the Legendre–Fenchel transformation of $\mathscr{R}$ and

$$\operatorname{dom} \mathscr{R}^* := \{\theta \in \mathscr{X}^* : \mathscr{R}^*(\theta) < \infty\}$$

is the effective domain of $\mathscr{R}^*$. Equality in (5) holds if and only if $\theta \in \partial \mathscr{R}(\mathscr{J}(z))$ [6, Prop. 9.5.1]. Now, owing to (4), $\theta \in \operatorname{dom} \mathscr{R}^*$ satisfies $\theta \geq 0$ $\mathbb{P}$-a.e. [108, Thm. 9.3.5]. Therefore, Assumption 2 and Krasnosel'skii's Theorem ensure that $u \mapsto J(u, \cdot)$ is continuous and hence $u \mapsto \mathbb{E}[\theta J(u, \cdot)]$ is convex and continuous.

Therefore, $u \mapsto \mathbb{E}[\theta J(u,\cdot)]$ is weakly lower semincontinuous [26, Thm. 2.23], which when combined with the weak continuity of $z \mapsto S(z;\cdot)$ ensures that $z \mapsto \mathbb{E}[\theta \mathscr{J}(z)]$ is weakly lower semicontinuous. Thus, for any sequence $\{z_n\} \subset Z_{\text{ad}}$ that weakly converges to $z \in Z_{\text{ad}}$, we have that

$$\begin{aligned}\liminf_{n\to\infty} \mathscr{R}(\mathscr{J}(z_n)) &\geq \liminf_{n\to\infty} \mathbb{E}[\theta \mathscr{J}(z_n)] - \mathscr{R}^*(\theta) \\ &\geq \mathbb{E}[\theta \mathscr{J}(z)] - \mathscr{R}^*(\theta) = \mathscr{R}(\mathscr{J}(z)) \quad \forall\, \theta \in \partial\mathscr{R}(\mathscr{J}(z)),\end{aligned}$$

which implies that $z \mapsto \mathscr{R}(\mathscr{J}(z))$ is weakly lower semicontinuous. Since $\wp$ is also weakly lower semicontinuous, $\mathfrak{J}$ is as well. Moreover, the minimization is performed over a bounded weakly closed level set in the reflexive Banach space Z, which implies the level set is weakly compact. Under these conditions, the direct method of the calculus of variations [6, Thm. 3.2.1] applies and ensures the existence of a minimizer. □

Since minimizers exist, it is natural to ask what the first-order necessary optimality conditions are. The following theorem characterizes the optimality conditions when J, $\wp$, and S are differentiable. For this result, we denote the space of bounded linear operators from a Banach space A to a Banach space B by $\mathscr{L}(A,B)$. Moreover, by $T_{Z_{\text{ad}}}(z)$ and $N_{Z_{\text{ad}}}(z)$, we denote the tangent and normal cones, respectively, to the (convex) set Z_{ad} at $z \in Z_{\text{ad}}$. We say that a function $f : Z \to \mathbb{R}$ is continuously differentiable if it possesses a derivative $f'(\cdot)$ in the sense of Gâteaux and $f'(\cdot)$ is continuous. It follows then by the mean value theorem that f is differentiable in the sense of Fréchet, e.g., [26, pp. 35–36]. It is said that f is (Gâteaux) directionally differentiable at $z \in Z$ if the directional derivative $f'(z,h) := \lim_{t\downarrow 0}[f(z+th) - f(z)]/t$ exists for all $h \in Z$. Note that if f is convex and continuous, then it is locally Lipschitz [30, Prop. 2.2.7] and $f'(z,\cdot)$ is a Hadamard directional derivative [105, Prop. 3.5].

Theorem 2 *Let the assumptions of Theorem 1 hold. In addition, suppose there exists an open set $V \subseteq Z$ with $Z_{\text{ad}} \subseteq V$ such that $z \mapsto S(z;\cdot) : V \to L^q(\Xi,\mathscr{B},\mathbb{P};U)$ is continuously differentiable with derivative*

$$S'(z;\cdot) \in \mathscr{L}(Z, L^q(\Xi,\mathscr{B},\mathbb{P};U)),$$

$u \mapsto J(u,\cdot) : L^q(\Xi,\mathscr{B},\mathbb{P};U) \to L^p(\Xi,\mathscr{B},\mathbb{P})$ is continuously differentiable with derivative

$$J'(u,\cdot) \in \mathscr{L}(L^q(\Xi,\mathscr{B},\mathbb{P};U), L^p(\Xi,\mathscr{B},\mathbb{P})),$$

and $\wp : Z \to \mathbb{R}$ is continuously differentiable with derivative $\wp'(z) \in Z^*$. Then if $z_\star \in Z_{\text{ad}}$ is a minimizer of $\mathfrak{J}$ over Z_{ad}, the following first-order optimality conditions hold: $\exists\, \theta \in \partial\mathscr{R}(\mathscr{J}(z_\star))$ such that

$$\left\langle \mathbb{E}\left[\theta\, S'(z_\star;\cdot)^* J'(S(z_\star;\cdot),\cdot)\right] + \wp'(z_\star),\, h\right\rangle_{Z^*,Z} \geq 0, \quad \forall\, h \in T_{Z_{\text{ad}}}(z_\star). \tag{6}$$

Proof Let us note that if $z_\star$ is an optimal solution of problem (2), then necessarily the directional derivatives $\mathfrak{J}'(z_\star, h) \geq 0$ for all $h \in T_{Z_{\text{ad}}}(z_\star)$. Since $\wp$ is differentiable, it follows that $\wp'(z_\star, h) = \langle \wp'(z_\star), h \rangle_{Z^*,Z}$. Also under the stated assumptions, $\mathscr{J}$ is continuously differentiable with derivative

$$\mathscr{J}'(z) = J'(S(z;\cdot),\cdot)S'(z;\cdot) \in \mathscr{L}(Z, L^p(\Xi, \mathscr{B}, \mathbb{P})).$$

Now since $\mathscr{R}$ is continuous, it is subdifferentiable and its (Hadamard) directional derivatives are given by

$$\mathscr{R}'(\mathscr{J}(z_\star), H) = \sup_{\theta \in \partial \mathscr{R}(\mathscr{J}(z_\star))} \mathbb{E}[\theta H] \quad \forall\, H \in \mathscr{X},$$

cf. [108, Thm. 6.10]. By the chain rule for directional derivatives, it follows that

$$\mathfrak{J}'(z_\star, h) = \sup_{\theta \in \partial \mathscr{R}(\mathscr{J}(z_\star))} \left\langle \mathbb{E}[\theta S'(z_\star;\cdot)^* J'(S(z_\star;\cdot),\cdot)] + \wp'(z_\star),\ h \right\rangle_{Z^*,Z}. \tag{7}$$

The function $\phi(\cdot) := \mathfrak{J}'(z_\star, \cdot)$ is convex and positively homogeneous. Moreover, the condition that $\phi(h) \geq 0$ for all $h \in T_{Z_{\text{ad}}}(z_\star)$ means that $h = 0$ is a minimizer of $\phi(h)$ subject to $h \in T_{Z_{\text{ad}}}(z_\star)$. This in turn means that $0 \in \partial\phi(0) + N_{Z_{\text{ad}}}(z_\star)$, which by (7) is equivalent to condition (6). □

Under appropriate differentiability assumptions on the PDE constraint function e, one can show that $\Lambda_\star = S'(z_\star;\cdot)^* J'(S(z_\star;\cdot),\cdot)$ is related to the solution to the adjoint equation. Informally, if the assumptions of the Implicit Function Theorem hold, then $\Lambda_\star = e_z(S(z_\star;\xi), z_\star, \xi)^* \lambda_\star$ where $\lambda_\star$ solves the adjoint equation

$$e_u(S(z_\star;\xi), z_\star, \xi)^* \lambda_\star(\xi) = -J_u(S(z_\star;\xi),\xi)$$

for $\mathbb{P}$-almost all $\xi \in \Xi$. See [61–64] for PDE-constrained optimization examples for which this holds.

3.2 Linear Elliptic Optimal Control

For this example, we assume Ξ is an m-fold Cartesian product of compact intervals and $\mathbb{P}$ is absolutely continuous with respect to the m-dimensional Lebesgue measure. Let $D \subset \mathbb{R}^d$ with $d \in \mathbb{N}$ be an open bounded Lipschitz domain, and define $U = H_0^1(D)$, $Y = U^* = H^{-1}(D)$, and $Z = L^2(D)$. Given the continuous matrix-valued function $A : \Xi \to \mathbb{R}^{d\times d}$ with $A(\xi) = A(\xi)^\top$ for all $\xi \in \Xi$, we define the parametrized linear elliptic PDE as the variational problem: find $u : \Xi \to U$ that solves

$$\langle e(u, z, \xi), v\rangle_{U^*,U} := \int_D (A(\xi)\nabla u(\xi, x)) \cdot \nabla v(x)\,\mathrm{d}x - \int_D z(x)v(x)\,\mathrm{d}x = 0 \tag{8}$$

for all $v \in U$ and fixed $z \in Z$. If there exist constants $0 < \underline{c} \leq \overline{c} < \infty$ such that

$$\underline{c} \leq \frac{x^\top A(\xi)x}{x^\top x} \leq \overline{c} \quad \forall\, x \in \mathbb{R}^d \setminus \{0\},\ \xi \in \Xi \tag{9}$$

then the Lax–Milgram Lemma [28] ensures the existence of a unique solution $S(z;\xi)$ to (8) for each $z \in Z$ and all $\xi \in \Xi$. Additionally, (9) and Poincaré's inequality guarantee the existence of a positive constant $C = C(D, \underline{c})$ such that

$$\|S(z;\cdot)\|_U \leq C\|z\|_Z \quad \forall \xi \in \Xi.$$

This and the linearity of the PDE then imply that $S(\cdot;\xi)$ is a bounded linear operator for all $\xi \in \Xi$ and since Z is compactly embedded into Y [1], $S(\cdot;\xi)$ is completely continuous for all $\xi \in \Xi$. Recall that an operator W mapping a Banach space X into another Banach space Y is completely continuous if

$$x_k \rightharpoonup x \ \text{ in } X \quad \Longrightarrow \quad W(x_k) \to W(x) \ \text{ in } Y.$$

In particular, all compact operators are completely continuous [33, Prop. 3.3]. Finally, $S(z;\cdot)$ is continuous and hence strongly $\mathscr{B}$-measurable since $A(\cdot)$ is continuous. Therefore, Assumption 1 is satisfied and since C is independent of $\xi \in \Xi$, we have that $S(z;\cdot) \in L^\infty(\Xi, \mathscr{B}, \mathbb{P}; U)$ for all $z \in Z$.

Now, let $\beta > 0$ and $u_d \in L^2(D)$ be a desired profile. We consider the PDE-constrained optimization problem

$$\min_{z \in Z} \mathscr{R}\left(\frac{1}{2}\|S(z;\xi) - u_d\|^2_{L^2(D)}\right) + \frac{\beta}{2}\|z\|^2_{L^2(D)} \tag{10}$$

where $S(z;\xi)$ solves (8) for fixed $\xi \in \Xi$ and $z \in Z$. The uncertain objective function and control penalty are

$$J(u,\xi) = \frac{1}{2}\|u - u_d\|^2_{L^2(D)} \quad \text{and} \quad \wp(z) = \frac{\beta}{2}\|z\|^2_{L^2(D)}.$$

J clearly satisfies Assumption 2 and therefore is continuous from $L^q(\Xi, \mathscr{B}, \mathbb{P}; U)$ into $L^p(\Xi, \mathscr{B}, \mathbb{P})$ for any $q \geq 2$ and $p \leq q/2$. Hence, Theorem 1 holds for any $\mathscr{R} : L^p(\Xi, \mathscr{B}, \mathbb{P}) \to \mathbb{R}$ that is convex and satisfies the monotonicity property (4).

In addition, since $e(\cdot,\cdot,\xi)$ is continuous and linear in u and z for all $\xi \in \Xi$, it is continuously Fréchet differentiable in u and z for all $\xi \in \Xi$, and again by the Lax–Milgram Lemma the state Jacobian is boundedly invertible for all $\xi \in \Xi$. Furthermore, the control Jacobian is a bounded linear operator for all $u \in U$, $z \in Z$ and $\xi \in \Xi$. In fact, $e_z(u,z,\xi)$ is independent of u, z, and ξ. Therefore, $S(\cdot;\xi)$ is continuously Fréchet differentiable for all $\xi \in \Xi$ and the derivative satisfies: For any $h \in Z$, $d = S'(z;\cdot)h : \Xi \to U$ solves the sensitivity equation

$$\int_D (A(\xi)\nabla d(\xi,x)) \cdot \nabla v(x)\,\mathrm{d}x - \int_D h(x)v(x)\,\mathrm{d}x = 0 \quad \forall\, v \in U$$

Since the sensitivity equation is identical to (8), we have that $d = S'(z;\cdot)h = S(h;\cdot) \in L^\infty(\Xi, \mathscr{B}, \mathbb{P}; U)$ for all $h \in Z$. Returning to the objective function, J and $\wp$ are clearly continuously Fréchet differentiable and thus Theorem 2 holds for any $\mathscr{R}$ satisfying the stated assumptions. Moreover, the adjoint equation corresponding to (10), at fixed $z \in Z$, is: find $\lambda : \Xi \to U$ such that

$$\int_D (A(\xi)\nabla\lambda(\xi,x)) \cdot \nabla v(x)\,\mathrm{d}x = -\int_D (S(z;\xi)(x) - u_d(x))v(x)\,\mathrm{d}x \quad \forall\, v \in U.$$

Note again that the above analysis ensures $\lambda \in L^\infty(\Xi, \mathscr{B}, \mathbb{P}; U)$.

4 Choosing the Functional $\mathscr{R}$

Under the assumptions of Section 3 (or similar assumptions), the stochastic PDE-constrained optimization problem

$$\min_{z \in Z_{\mathrm{ad}}} \mathscr{R}\left(\mathscr{J}(z)\right) + \wp(z) \tag{11}$$

where $\mathscr{R} : L^p(\Xi, \mathscr{B}, \mathbb{P}) \to \mathbb{R}$ is well-defined, but ambiguous since $\mathscr{R}$ is not explicitly specified. In traditional stochastic programming, $\mathscr{R}$ is taken to be the expected value, i.e., $\mathscr{R} = \mathbb{E}_{\mathbb{P}}$. This results in a *risk neutral* formulation of (11) for which the optimal solutions minimize $\mathscr{J}(z)$ on average. The risk neutral formulation is often not conservative enough for high-consequence applications because the average behavior of a system does not provide a sufficient proxy for variability or low probability and tail events. This motivates the use of *risk measures*. Another popular class of cost surrogates are the *probabilistic functions*. This class seeks to minimize the probability of undesirable events occurring. The use of the expectation, risk measures, and probabilistic functions is justified when the probability law $\mathbb{P}$ is known but can lead to nonsensical, even dangerous, results if $\mathbb{P}$ is unknown and estimated from noisy or incomplete data. In the subsequent sections, we will review both cases of known and unknown probability law. When the probability law is known, we simplify notation and denote $\mathbb{E} = \mathbb{E}_{\mathbb{P}}$.

It is worth mentioning that (11) is only one of many meaningful problem formulations for PDE-constrained optimization. In many applications, constraints in addition to the objective function are uncertain. In this case, we must handle the uncertainty in the constraints in a rigorous and physically relevant way. Popular approaches in stochastic programming include: chance (probabilistic) constraints (see, e.g., [81]) and stochastic dominance constraints (see, e.g., [38]). Chance constraints seek to ensure that the probability of an uncertain quantity of interest exceeding a prescribed threshold is below some nominal value (e.g., the probability that a bridge collapses is smaller than 10^{-3} percent). Stochastic dominance constraints, on the other hand, aim to ensure that our uncertain quantity of interest is in

some sense preferred over a predefined uncertain benchmark value. Since a rigorous treatment of these concepts in PDE-constrained optimization is still an open area of research, we restrict our attention to problems of the type (11). We do, however, introduce and discuss the notions of stochastic orders and stochastic dominance in the coming subsection.

4.1 Risk-Averse Optimization

When the probability law of the random vector ξ is known, we can use any of the multitudes of risk measures to complete the problem definition in (11). A particularly important class of risk measures is the class of *coherent* risk measures [4]. To simplify notation, we denote $\mathscr{X} := L^p(\Xi, \mathscr{B}, \mathbb{P})$. A function $\mathscr{R} : \mathscr{X} \to \mathbb{R}$ is a coherent risk measure if it satisfies:

(R1) **Subadditivity:** For all $X,\ X' \in \mathscr{X}$, $\mathscr{R}(X + X') \leq \mathscr{R}(X) + \mathscr{R}(X')$;
(R2) **Monotonicity:** If $X,\ X' \in \mathscr{X}$ satisfy $X \leq X'$ $\mathbb{P}$-a.e., then $\mathscr{R}(X) \leq \mathscr{R}(X')$;
(R3) **Translation Equivariance:** For all $X \in \mathscr{X}$ and $t \in \mathbb{R}$, $\mathscr{R}(X+t) = \mathscr{R}(X)+t$;
(R4) **Positive Homogeneity:** For all $X \in \mathscr{X}$ and $t \geq 0$, $\mathscr{R}(tX) = t\mathscr{R}(X)$.

Note that axiom (R1) and (R4) imply convexity of $\mathscr{R}$ and convexity plus (R4) imply subadditivity of $\mathscr{R}$. Therefore, axiom (R1) is typically replaced by

(R1′) **Convexity:** For all $X,\ X' \in \mathscr{X}$ and $t \in [0, 1]$

$$\mathscr{R}(tX + (1-t)X') \leq t\mathscr{R}(X) + (1-t)\mathscr{R}(X').$$

In the context of physical applications, $\mathscr{R}(X)$ should inherit the units of X. In which case, (R4) ensures that a change of the units of X results in a consistent change of the units of $\mathscr{R}(X)$. Additionally, (R3) ensures that deterministic quantities, such as the control penalty $\wp$ in (11), do not contribute to the overall risk. In fact, (R3) combined with (R4) ensure that deterministic quantities are riskless, i.e., $\mathscr{R}(t) = t$ for all $t \in \mathbb{R}$.

The axioms for coherent risk measures result in many desirable properties of $\mathscr{R}$. Any functional $\mathscr{R} : \mathscr{X} \to \mathbb{R}$ satisfying axioms (R2) and (R1′) is continuous in the norm topology of the space $\mathscr{X} = L^p(\Xi, \mathscr{B}, \mathbb{P})$ (see Proposition 6.6 in [108]). Therefore, the Fenchel–Moreau theorem [6, Thm. 9.3.5] ensures that $\mathscr{R}$ is equal to its biconjugate function,

$$\mathscr{R}(X) = \sup_{\theta \in \mathscr{X}^*} \{\mathbb{E}[\theta X] - \mathscr{R}^*(\theta)\}, \tag{12}$$

where $\mathscr{R}^* : \mathscr{X}^* \to \mathbb{R} \cup \{+\infty\}$ is the Legendre–Fenchel transformation of $\mathscr{R}$, i.e.,

$$\mathscr{R}^*(\theta) = \sup_{X \in \mathscr{X}} \{\mathbb{E}[\theta X] - \mathscr{R}(X)\}.$$

Clearly, the set $\mathscr{X}^*$ in the representation (12) can be replaced by

$$\mathrm{dom}(\mathscr{R}^*) = \{\theta \in \mathscr{X}^* : \mathscr{R}^*(\theta) < +\infty\}.$$

In this setting, one can further show that (R2) and (R3) hold if and only if for all $\theta \in \mathrm{dom}(\mathscr{R}^*)$ we have that $\theta \geq 0$ $\mathbb{P}$-a.e. and $\mathbb{E}[\theta] = 1$. That is, $\mathrm{dom}(\mathscr{R}^*)$ is a subset of the probability density functions in $\mathscr{X}^*$. Finally, (R4) holds if and only if $\mathscr{R}^*(\theta) = 0$ for all $\theta \in \mathrm{dom}(\mathscr{R}^*)$. See [108, Th. 6.5] for a proof of these results. In fact, Theorem 6.7 in [108] ensures that a risk measure $\mathscr{R}$ is coherent if and only if it has the equivalent form

$$\mathscr{R}(X) = \sup_{\theta \in \mathfrak{A}} \mathbb{E}[\theta X] \tag{13}$$

where $\mathfrak{A} \subset \mathscr{X}^*$ is a convex, bounded, and weakly* closed subset of probability density functions, i.e., $\mathfrak{A} = \mathrm{dom}(\mathscr{R}^*)$.

In addition to the axioms for coherent risk measures, a fundamentally important property of $\mathscr{R}$ is law invariance. We say that two random variables are *distributionally equivalent*, denoted $X \overset{D}{\sim} X'$, if their cumulative distribution functions (cdf) $\Psi_X(t) = \mathbb{P}(X \leq t)$ and $\Psi_{X'}(t) = \mathbb{P}(X' \leq t)$ are equal for all $t \in \mathbb{R}$. A functional $\mathscr{R} : \mathscr{X} \to \mathbb{R}$ is then said to be *law invariant* if

$$X \overset{D}{\sim} X' \quad \Longrightarrow \quad \mathscr{R}(X) = \mathscr{R}(X') \tag{14}$$

for any two random variables $X, X' \in \mathscr{X}$. In words, property (14) ensures that $\mathscr{R}$ is only a function of the cdf $\Psi_X(t) = \mathbb{P}(X \leq t)$ for any random variable X. For example, this excludes the scenario in which $\mathscr{R}(X) \neq \mathscr{R}(X')$ where X and X' are distributionally equivalent discrete random variables whose atoms are ordered differently.

Another important notion in stochastic optimization is that of *stochastic dominance*. A random variable X dominates another random variable X' with respect to the *first stochastic order* if

$$\Psi_X(t) \leq \Psi_{X'}(t) \quad \forall\, t \in \mathbb{R}. \tag{15}$$

We denote the relation (15) by $X \succeq_{(1)} X'$. Similarly, X dominates X' with respect to the *second stochastic order* if

$$\int_{-\infty}^{t} \Psi_X(\eta)\, \mathrm{d}\eta \leq \int_{-\infty}^{t} \Psi_{X'}(\eta)\, \mathrm{d}\eta \quad \forall\, t \in \mathbb{R}. \tag{16}$$

Owing to Fubini's theorem [45, Thm. 2.37], it is straightforward to show that

$$\int_{-\infty}^{t} \Psi_X(\eta)\, \mathrm{d}\eta = \mathbb{E}\left[\int_{-\infty}^{t} \mathbb{1}_{X \leq \eta}\, \mathrm{d}\eta\right] = \mathbb{E}[(t - X)_+]$$

where, for any $E \in \mathscr{B}$, $\mathbb{1}_E(\xi) = 1$ if $\xi \in E$ and $\mathbb{1}_E(\xi) = 0$ otherwise, and $(x)_+ = \max\{0, x\}$. Therefore, (16) is equivalent to the condition

$$\mathbb{E}[(t - X)_+] \leq \mathbb{E}[(t - X')_+] \quad \forall\, t \in \mathbb{R}.$$

We denote the relation (16) by $X \succeq_{(2)} X'$. If $(\Xi, \mathscr{B}, \mathbb{P})$ is nonatomic and $\mathscr{R}$ is law invariant, then the following two results hold: (i) the implication $X \succeq_{(1)} X' \implies \mathscr{R}(X) \geq \mathscr{R}(X')$ holds if and only if $\mathscr{R}$ satisfies the monotonicity condition (R2) [108, Th. 6.50]; (ii) if $\mathscr{R}$ satisfies conditions (R1′), (R2), and (R3), then $-X' \succeq_{(2)} -X$ implies $\mathscr{R}(X) \geq \mathscr{R}(X')$ [108, Th. 6.51]. These two properties demonstrate that law invariant coherent risk measures $\mathscr{R}$ *prefer dominated random variables* and thus are critical in reducing uncertainty (i.e., variability) in the optimized system. On the other hand, as previously noted, one could define risk aversion via stochastic dominance constraints instead of risk measures. For example, suppose $\bar{z}$ is known to produce an acceptable objective value $\mathscr{J}(\bar{z})$. One could then incorporate a constraint of the form

$$\mathscr{J}(\bar{z}) \succeq_{(1)} \mathscr{J}(z) \quad \text{or} \quad -\mathscr{J}(z) \succeq_{(2)} -\mathscr{J}(\bar{z}).$$

For more information of stochastic dominance constraints, see [38].

Example 1 (Mean-Plus-Deviation) A common risk measure in engineering applications, motivated by Markowitz's pioneering work in portfolio optimization [74], is the mean-plus-deviation risk measure

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

for $p \in [1, \infty)$. Clearly, $\mathscr{R}$ is naturally defined and real valued on $\mathscr{X} = L^p(\Xi, \mathscr{B}, \mathbb{P})$ and is law invariant, convex, positively homogeneous, and translation equivariant. Unfortunately, $\mathscr{R}$ is not monotonic and can lead to the paradoxical scenario where one position is always smaller than another, but the larger position has smaller risk. In the context of finance, the risk measure $\mathscr{R}$ can lead to the selection of portfolios that have smaller risk and smaller returns. See [108, Ex. 6.62] for a simple example of this undesirable situation. The lack of monotonicity results from $\mathscr{R}$ equally penalizing the deviation below and above the expected value. In terms of minimization, one prefers large deviation below the expected value since this could lead to better than expected performance. A related law-invariant risk measure that is coherent is the mean-plus-upper-semideviation risk measure

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

Note that this risk measure only penalizes deviation in excess of the expected value. Since this $\mathscr{R}$ is coherent, it can be represented as in (13) with *risk envelope*

$$\operatorname{dom}(\mathscr{R}^*) = \left\{\theta \in \mathscr{X}^* : \theta = 1 + \theta' - \mathbb{E}[\theta'],\ \|\theta'\|_{\mathscr{X}^*} \leq c,\ \theta' \geq 0\ \mathbb{P}\text{-a.e.}\right\}.$$

See [108, Ex. 6.23] for more details.

Example 2 (Conditional Value-at-Risk) The conditional value-at-risk[1] (CVaR) is a coherent risk measure that has recently received much attention [64, 94, 115]. CVaR at confidence level $\alpha \in (0, 1)$ is defined as

$$\mathscr{R}(X) = \mathrm{CVaR}_\alpha(X) := \inf_{t\in\mathbb{R}} \left\{ t + \frac{1}{1-\alpha}\mathbb{E}[(X-t)_+] \right\}, \tag{17}$$

which naturally acts on random variables in $\mathscr{X} = L^1(\Xi, \mathscr{B}, \mathbb{P})$. If the random variable X is continuously distributed, then $\mathrm{CVaR}_\alpha(X)$ is the expectation of X conditioned on the event that X is larger than its α-quantile, i.e.,

$$\mathrm{CVaR}_\alpha(X) = \mathbb{E}[X | X > \Psi_X^{-1}(\alpha)].$$

In the financial literature, the quantile $\Psi_X^{-1}(\alpha)$ is called the Value-at-Risk. Moreover, when $\alpha = 0$ we have that $\mathrm{CVaR}_0(X) = \mathbb{E}[X]$ and

$$\lim_{\alpha\uparrow 1} \mathrm{CVaR}_\alpha(X) = \operatorname{ess\,sup} X.$$

Since CVaR is coherent, it can be represented as in (13) with risk envelope

$$\mathrm{dom}(\mathscr{R}^*) = \left\{ \theta \in L^\infty(\Xi, \mathscr{B}, \mathbb{P}) \,:\, \mathbb{E}[\theta] = 1,\ 0 \le \theta \le (1-\alpha)^{-1}\ \mathbb{P}\text{-a.e.} \right\}.$$

See [108, Ex. 6.19] for more details.

Example 3 (Higher-Moment Coherent Risk) CVaR was extended in [66] to the higher-moment coherent risk measure (HMCR),

$$\mathscr{R}(X) = \inf_{t\in\mathbb{R}} \left\{ t + \frac{1}{1-\alpha}\mathbb{E}[(X-t)_+^p]^{\frac{1}{p}} \right\},$$

with $p \in (1, \infty)$. HMCR is a law-invariant coherent risk measure and is finite for random variables in $\mathscr{X} = L^p(\Xi, \mathscr{B}, \mathbb{P})$ (see [37] for a thorough analysis of HMCR). Since HMCR is coherent, it can be represented as in (13) with risk envelope

$$\mathrm{dom}(\mathscr{R}^*) = \left\{ \theta \in \mathscr{X}^* \,:\, \mathbb{E}[\theta] = 1,\ \theta \ge 0\ \mathbb{P}\text{-a.e.},\ \|\theta\|_{\mathscr{X}^*} \le \frac{1}{1-\alpha} \right\}.$$

This risk envelope was determined in [29, Sect. 5.3.1] for the more general class of *transformed norm risk measures*. Note that HMCR and CVaR coincide if $p = 1$ and thus so do their risk envelopes.

[1] Also called Average Value-at-Risk, Expected Shortfall, Expected Tail Loss and Superquantile.

Example 4 (Entropic Risk) The entropic risk measure is defined as

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

and is finite for random variables in $\mathscr{X} = L^\infty(\Xi, \mathscr{B}, \mathbb{P})$. The entropic risk is convex, monotonic, and translation equivariant but is not positively homogeneous and therefore is not coherent. The name entropic risk comes from the Legendre–Fenchel transformation of $\mathscr{R}$. Since the topological dual space of $\mathscr{X} = L^\infty(\Xi, \mathscr{B}, \mathbb{P})$ is difficult to handle in practice, it is natural to view $\mathscr{X}$ and $L^1(\Xi, \mathscr{B}, \mathbb{P})$ as paired, locally convex topological vector spaces where $\mathscr{X}$ is equipped with the weak* topology and $L^1(\Xi, \mathscr{B}, \mathbb{P})$ is equipped with the norm topology (see, e.g., [108, Sect. 6.3] for a discussion of essentially bounded random variables). In this setting, one can show that the Legendre–Fenchel transformation of $\mathscr{R}$ is

$$\mathscr{R}^*(\theta) = \sup_{X \in \mathscr{X}} \{\mathbb{E}[\theta X] - \mathscr{R}(X)\} = \sigma^{-1}\mathbb{E}[\theta \log(\theta)]$$

when $\theta \in L^1(\Xi, \mathscr{B}, \mathbb{P})$ satisfying $\theta \geq 0$ $\mathbb{P}$-a.e. and $\mathbb{E}[\theta] = 1$. This is the negative of Shannon's entropy, i.e., the Kullback–Leibler divergence (up to the scaling by σ^{-1}). See [108, Ex. 6.20] for more details.

4.2 Probabilistic Optimization

As with risk measures, we assume in this section that $\mathbb{P}$ is known. In many applications, it is extremely important that an optimal control or design reduces the probability that the event

$$\left\{\xi \in \Xi \,:\, [\mathscr{J}(z)](\xi) > \tau\right\} \tag{18}$$

for some prescribed threshold $\tau \in \mathbb{R}$ occurs. For example, the event (18) could signify the failure of a structure. This naturally leads to the probabilistic objective function

$$\mathscr{R}(\mathscr{J}(z)) = \mathbb{P}(\mathscr{J}(z) > \tau) = \mathbb{E}[\mathbb{1}_{\mathscr{J}(z)>\tau}]. \tag{19}$$

Recall the definition of $\mathbb{1}_E$ from Section 4.1. Much work has been devoted to probabilistic optimization including the derivation of derivative formulas for this choice of $\mathscr{R}$ [76, 98, 113, 114, 117]. The functional $\mathscr{R}$ is only differentiable under certain assumptions which may be difficult to verify in the context of PDE-constrained optimization. For example, the authors in [117] require that $\xi \mapsto [\mathscr{J}(z)](\xi)$ is convex with respect to ξ and that the random vector ξ is Gaussian. Moreover, many differentiation formulas are stated in finite dimensions and it is unclear whether or not these formulas hold in infinite dimensions. Additional complications arise

when estimating probabilistic functions. See [93] for a detailed discussion of the challenges associated with estimation and optimization of probabilistic functions. Finally, $\mathscr{R}$ only quantifies the "number" of scenarios for which $\mathscr{J}(z) > \tau$ but ignores the magnitudes of these scenarios. This could lead to a situation where the optimal controls or designs result in a small probability of (18) occurring, but all scenarios in (18) have large magnitude. For example, (18) could represent any failure (no matter how minor) of the system whereas large-magnitude scenarios signal catastrophic failure.

For these reasons, the authors of [93] developed the concept of buffered probabilities. Roughly speaking, the buffered probability is one minus the inverse of $\alpha \mapsto \mathrm{CVaR}_\alpha(X)$. Let $X \in \mathscr{X} = L^1(\Xi, \mathscr{B}, \mathbb{P})$ be a nondegenerate (i.e., nonconstant) random variable, then $\alpha \mapsto \mathrm{CVaR}_\alpha(X)$ is continuous and nondecreasing for $\alpha \in [0, 1)$ and strictly increasing for $\alpha \in [0, 1 - \pi_\infty)$ where

$$\pi_\infty = \pi_\infty(X) = \mathbb{P}(\{\xi \in \Xi : X(\xi) = \operatorname{ess\,sup} X\})$$

[94]. Therefore, an inverse to $\alpha \mapsto \mathrm{CVaR}_\alpha(X) : [0, 1) \to [\mathbb{E}[X], \operatorname{ess\,sup} X)$ exists. Now, suppose X is degenerate, i.e., there exists $t \in \mathbb{R}$ such that $X = t$ $\mathbb{P}$-a.e., then $\mathrm{CVaR}_\alpha(X) = t$ for any $\alpha \in [0, 1)$ by axioms (R3) and (R4) in Section 4.1 and thus the inverse is not defined. Using these properties of CVaR, we define the buffered probability that a nondegenerate random variable X exceeds the threshold τ as $\bar{p}_\tau(X)$ where $\alpha = 1 - \bar{p}_\tau(X)$ solves

$$\tau = \mathrm{CVaR}_\alpha(X).$$

It is not hard to show that $\bar{p}_\tau(X) \geq \mathbb{P}(X > \tau)$. Moreover, if X is continuously distributed then the buffered probability is $\bar{p}_\tau(X) = \mathbb{P}(X > \tau_X)$ where τ_X solves

$$\mathbb{E}[X | X > \tau_X] = \tau.$$

In this case, τ_X is the $\alpha = 1 - \bar{p}_\tau(X)$ quantile of X. One can think of τ_X as defining a "buffer" or "safety" zone around the event (18) defined via the average of scenarios in the upper tail. Figure 1 contains a comparison of the buffered probability and the usual probability for a normally distributed random variable X. The blue line corresponds to the cdf Ψ_X while the red line corresponds to the inverse of $\alpha \mapsto \mathrm{CVaR}_\alpha(X)$, denoted $\overline{\Psi}_X(\tau)$.

It was shown in [71] that for $\tau < \operatorname{ess\,sup} X$ the buffered probability has the convenient optimization formulation

$$\bar{p}_\tau(X) = \inf_{t \geq 0} \mathbb{E}[(t(X - \tau) + 1)_+]. \tag{20}$$

This permits the optimization of $z \mapsto \bar{p}_\tau(\mathscr{J}(z))$ over Z_{ad} to be reformulated as the optimization of $(z, t) \mapsto \mathbb{E}[(t(\mathscr{J}(z) - \tau) + 1)_+]$ over the augmented space $Z_{\mathrm{ad}} \times [0, \infty)$. The objective function in the later expression is the composition of a convex

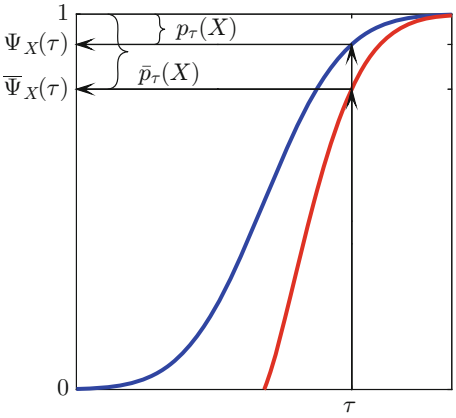

Fig. 1 A comparison of the probability that X exceeds τ, $p_\tau(X)$, and the buffered probability that X exceeds τ, $\bar{p}_\tau(X)$. The blue line is Ψ_X whereas the red line is the inverse of $\alpha \mapsto \text{CVaR}_\alpha(X)$, denoted $\overline{\Psi}_X$

function with our random variable objective function. In addition, the authors of [71] show that $X \mapsto \bar{p}_\tau(X)$ is a lower semicontinuous, quasi-convex, and monotonic function (i.e., satisfies (R2) in Section 4.1). Finally, if $X \mapsto \bar{p}_\tau(X)$ is considered as a function on $L^2(\Xi, \mathscr{B}, \mathbb{P})$, one can show that it is the minimal upper bound for $\mathbb{P}(X > \tau)$ among all quasi-convex, lower semicontinuous law-invariant functions acting on elements of $L^2(\Xi, \mathscr{B}, \mathbb{P})$ [71, Prop. 3.12]. This optimality result is related to the results in [81] in which the authors seek an optimal convex approximation for chanced constrained optimization problems.

4.3 Distributionally Robust Optimization

Often the true probability law $\mathbb{P}$ of the random inputs ξ is not known but estimated from noisy and incomplete data. In this case, making a decision based solely on an estimate of $\mathbb{P}$ can be catastrophic if the estimate does not accurately characterize the statistical behavior of the true underlying distribution. In such scenarios, we must be "averse" to the risk associated with our lack of knowledge of the true underlying probability distribution. This motivates the *distributionally robust* approach to stochastic programming of optimizing the "worst expectation"

$$\min_{z \in Z_{\text{ad}}} \left\{ \mathfrak{J}(z) := \sup_{P \in \mathfrak{M}} \mathbb{E}_P\big[\mathscr{J}(z)\big] + \wp(z) \right\}, \tag{21}$$

where $\mathfrak{M}$ is a specified set of admissible probability measures defined on the measurable space $(\Xi, \mathscr{B})$ and

$$\mathscr{R}(X) := \sup_{P \in \mathfrak{M}} \mathbb{E}_P[X] \tag{22}$$

is the associated risk functional. The set $\mathfrak{M}$ is often called the *ambiguity set*. For more information on robust optimization see, e.g., [14, 23, 41, 107] and the references therein.

In the setting of distributionally robust optimization, we often have partial information regarding the probability law $\mathbb{P}$. Using this information, we can employ Bayesian analysis to determine a single posterior distribution for ξ (see, e.g., [19]), which we can then use to formulate and solve a risk-averse (Section 4.1) or probabilistic (Section 4.2) optimization problem. Although Bayes' rule provides an analytic expression for the posterior distribution, the posterior distribution often does not have a practical (i.e., implementable on a computer) representation. Moreover, Bayesian analysis relies on subjective beliefs encoded in the chosen prior distribution for ξ. Therefore, if the prior distribution is chosen incorrectly, any decision made using the posterior distribution may result in unexpected, undesirable outcomes. There are a number of ways to circumvent this potential pitfall such as, e.g., uninformative priors or robust Bayesian analysis. Robust Bayesian analysis generates a family of posterior distributions using predefined families of noise and prior distributions [18, 20]. In the context of the distributionally robust optimization problem (21), we can incorporate this family of posterior distributions within the ambiguity set $\mathfrak{M}$.

In addition to the previously described robust Bayesian approach, there are two somewhat different methods for constructing the ambiguity set $\mathfrak{M}$. In one approach, we assume that there is a specified reference probability measure $\mathbb{P}_0$ and that the set $\mathfrak{M}$ consists of probability measures in some sense close to $\mathbb{P}_0$. If we assume further that $\mathfrak{M}$ is a set of probability measures that are absolutely continuous with respect to the reference probability measure $\mathbb{P}_0$, then as a consequence of the Radon–Nikodym theorem [45], for every $Q \in \mathfrak{M}$ there exists a $\mathscr{B}$-measurable function $\theta : \Xi \to \mathbb{R}$ such that $\mathrm{d}Q = \theta \mathrm{d}\mathbb{P}_0$. That is, with the set $\mathfrak{M}$ is associated the set of densities $\mathfrak{A} = \{\theta = \mathrm{d}Q/\mathrm{d}\mathbb{P}_0 : Q \in \mathfrak{M}\}$. Assuming that $\mathfrak{A} \subset \mathscr{X}^*$ where $\mathscr{X} = L^p(\Xi, \mathscr{B}, \mathbb{P}_0)$ with $1 \leq p < \infty$, the corresponding functional

$$\mathscr{R}(X) = \sup_{\theta \in \mathfrak{A}} \mathbb{E}[\theta X] \tag{23}$$

becomes a coherent risk measure defined on $\mathscr{X}$. By the duality relation (13), there is a one-to-one correspondence between coherent risk measures and distributionally robust functionals of the form (23).

Another common approach is to define $\mathfrak{M}$ through moment matching. This approach was pioneered by Scarf [103]. For moment matching, we assume that K moments of ξ are specified (e.g., estimated from data), and the ambiguity set is defined as

$$\mathfrak{M} := \left\{ Q : \mathscr{B} \to [0,1] \,:\, Q(\Xi) = 1,\ \mathbb{E}_Q[\psi_k(\xi)] \leq m_k,\ k = 1, \ldots, K \right\}, \tag{24}$$

where ψ_k are real-valued $\mathscr{B}$-measurable functions and $m_k \in \mathbb{R}$. For example, setting $\psi_k(\xi) := e_k^\top \xi$ where e_k denotes the m-vector of zeros with one in the kth position

(i.e., the kth component of ξ) for $k = 1, \ldots, m$ would produce the mean value in each direction of Ξ. The moment matching problem is naturally posed in the uniform closure of the space of continuous random variables with compact support, $\mathscr{X} = C_0(\Xi)$, whose topological dual space, by the Riesz Representation Theorem (see, e.g., [45, Th. 7.17] or [6, Th. 2.4.6]), is isometrically isomorphic to the Banach space of signed regular Borel measures endowed with the total variation norm (i.e., $\Xi \subseteq \mathbb{R}^m$ is a locally compact Hausdorff space). Note that if Ξ is compact, then $C_0(\Xi) = C(\Xi)$.

When the ambiguity set $\mathfrak{M}$ is defined by the moment constraints (24), evaluation of the respective functional $\mathscr{R}(X)$, defined as the optimal value of the maximization problem given by the right-hand side of (22), is known as the *problem of moments*. It is possible to show that it suffices to perform the maximization in (22) with respect to probability measures $P \in \mathfrak{M}$ with support having at most $K + 1$ points [97] (see also Proposition 6.66 and Theorem 7.37 in [108]). That is, $\mathscr{R}(\mathscr{J}(z))$ is equal to the optimal value of the following program:

$$\begin{aligned} \max_{\xi_1,\ldots,\xi_{K+1}\in\Xi,\ \alpha\in\mathbb{R}^{K+1}_+} \quad & \sum_{i=1}^{K+1} \alpha_i \mathscr{J}(z,\xi_i) \\ \text{s.t.} \quad & \sum_{i=1}^{K+1} \alpha_i \psi_k(\xi_i) \le m_k,\ k = 1,\ldots,K,\ \sum_{i=1}^{K+1} \alpha_i = 1 \end{aligned} \tag{25}$$

where $\mathbb{R}_+ := [0, +\infty)$. Furthermore, the (Lagrangian) dual of the optimization problem (25) can be written as the following semi-infinite program:

$$\begin{aligned} \min_{\mu\in\mathbb{R}\times\mathbb{R}^K_+} \quad & \mu_0 + \sum_{k=1}^{K} m_k\mu_k \\ \text{s.t.} \quad & \mu_0 + \sum_{k=1}^{K} \mu_k\psi_k(\xi) \ge \mathscr{J}(z,\xi),\ \xi \in \Xi. \end{aligned} \tag{26}$$

Under mild regularity conditions, there is no duality gap between problems (25) and (26), and hence $\mathscr{R}(\mathscr{J}(z))$ is equal to the optimal value of the dual problem (26). One such regularity condition is that the set Ξ is nonempty and compact, and the functions ψ_k, $k = 1, \ldots, K$, and $\mathscr{J}(z, \cdot)$ are continuous on Ξ. Consequently, the respective minimax problem (21) can be written as the following semi-infinite optimization problem:

$$\begin{aligned} \min_{z\in Z_{\mathrm{ad}},\ \mu\in\mathbb{R}\times\mathbb{R}^K_+} \quad & \mu_0 + \sum_{k=1}^{K} m_k\mu_k + \wp(z) \\ \text{s.t.} \quad & \mu_0 + \sum_{k=1}^{K} \mu_k\psi_k(\xi) \ge \mathscr{J}(z,\xi),\ \xi \in \Xi. \end{aligned} \tag{27}$$

In general, solving semi-infinite programs of the form (27) is not easy. In some rather specific cases, (27) can be formulated as a semi-definite programming

problem and solved efficiently [24, 36]. Also a number of specialized algorithms were suggested to solve the moment-matching problem in, e.g., [43, 44, 46].

From the point of view of risk measures $\mathscr{R} : \mathscr{X} \to \mathbb{R}$, with $\mathscr{X} = L^p(\Xi, \mathscr{B}, \mathbb{P}_0)$, the concept of law invariance is a natural one. It ensures that $\mathscr{R}(X)$ can be considered as a function of the cdf $\Psi_X(t) = \mathbb{P}_0(X \leq t)$ associated with X. In the distributionally robust setting, it makes sense to talk about law invariance when the ambiguity set consists of probability measures absolutely continuous with respect to a specified reference probability measure $\mathbb{P}_0$ and the corresponding functional $\mathscr{R}$ is defined in the form (23). It is natural to say that the respective ambiguity set $\mathfrak{A}$, of density functions, is *law invariant* (with respect to the reference probability measure $\mathbb{P}_0$) if $\theta \in \mathfrak{A}$ and $\theta' \overset{D}{\sim} \theta$ implies that $\theta' \in \mathfrak{A}$.

Theorem 3 ([107]) *Consider a set $\mathfrak{A} \subset \mathscr{X}^*$ of density functions and the respective functional $\mathscr{R}$ defined in (23). If the set $\mathfrak{A}$ is law invariant, then the functional $\mathscr{R}$ is law invariant. Conversely, if the functional $\mathscr{R}$ is law invariant and the set $\mathfrak{A}$ is convex and weakly* closed, then $\mathfrak{A}$ is law invariant.*

We can define a large class of law invariant ambiguity sets $\mathfrak{A}$ using the concept of ϕ-divergence [34, 79]. Consider a convex lower semicontinuous function $\phi : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ such that $\phi(1) = 0$ and $\phi(x) = +\infty$ for $x < 0$, and define $\mathfrak{A}$ as the set of density functions $\theta \in \mathscr{X}^*$ satisfying the constraint $\mathbb{E}_{\mathbb{P}_0}[\phi(\theta)] \leq \epsilon$ for some $\epsilon > 0$. For example, let $\phi(x) = x \ln x - x + 1$ for $x \geq 0$, and $\phi(x) = +\infty$ for $x < 0$. Then for a probability measure Q absolutely continuous with respect to $\mathbb{P}_0$ and density function $\theta = \mathrm{d}Q/\mathrm{d}\mathbb{P}_0$, we have that $\mathbb{E}_{\mathbb{P}_0}[\theta] = 1$ and hence

$$\mathbb{E}_{\mathbb{P}_0}[\phi(\theta)] = \mathbb{E}_{\mathbb{P}_0}[\theta \ln \theta] = \mathbb{E}_{\mathbb{P}_0}\left[\frac{\mathrm{d}Q}{\mathrm{d}\mathbb{P}_0} \ln \theta\right] = \mathbb{E}_Q[\ln \theta]$$

is the Kullback–Leibler divergence of Q from $\mathbb{P}_0$. As another example for $\alpha \in [0, 1)$, let $\phi(x) = 0$ for $x \in [0, (1-\alpha)^{-1}]$, and $\phi(x) = +\infty$ otherwise. Then for any $\epsilon \geq 0$, the corresponding set $\mathfrak{A}$ consists of density functions θ such that $\theta \leq (1-\alpha)^{-1}$. In that case, the corresponding functional $\mathscr{R}$ becomes the CVaR_α. For many other examples of ϕ-divergence functionals, we refer to [16, 70].

Employing Lagrange multipliers, it is possible to show that the functional $\mathscr{R}$ associated with the ϕ-divergence ambiguity set can be written as

$$\mathscr{R}(X) = \inf_{\mu \geq 0,\, \nu} \{\mu\epsilon + \nu + \mathbb{E}_{\mathbb{P}_0}[(\mu\phi)^*(X - \nu)]\}, \tag{28}$$

e.g., [16, 107]. Here $(\mu\phi)^*(y) = \sup_{x \in \mathbb{R}}\{yx - (\mu\phi)(x)\}$ is the Legendre–Fenchel transformation of $(\mu\phi)$. For the specific case of the Kullback–Leibler divergence, this can be simplified to

$$\mathscr{R}(X) = \inf_{\mu \geq 0} \left\{\mu\epsilon + \mu \ln \mathbb{E}_{\mathbb{P}_0}[\exp(\mu^{-1}X)]\right\}.$$

For the ϕ-divergence ambiguity set, the respective distributionally robust problem (21) can be written as the following stochastic programming problem:

$$\min_{z \in Z_{\text{ad}},\, \mu \geq 0,\, \nu} \mu\epsilon + \nu + \mathbb{E}_{\mathbb{P}_0}[(\mu\phi)^*(\mathscr{J}(z) - \nu)] + \wp(z). \tag{29}$$

We note that the function $(\mu\phi)^*$ is convex and hence problem (29) is convex provided that $\mathscr{J}(\cdot, \xi)$, $\wp$ and the set Z_{ad} are convex. Such problems can be solved by, e.g., Monte Carlo randomization algorithms. We will discuss this further in Section 5.

To conclude this discussion, we point out that the authors of [121] introduce a specific class of ambiguity sets that permit a reformulation of the inner maximization problem to a conic programming problem. The assumptions required for this reformulation are likely not satisfied for general nonlinear, nonconvex PDE-constrained optimization problems, motivating the need for new approximation techniques and optimization algorithms for solving (21).

5 Methods for Expectation-Based Optimization

In general, we cannot apply rapidly converging derivative-based optimization algorithms to solve (2) because the functional $\mathscr{R}$ and hence the composite function $\mathscr{R} \circ \mathscr{J}$ are often not continuously differentiable even if the underlying uncertain reduced objective function is. This issue is critical in determining the practicality of solving (2) since traditional nonsmooth optimization algorithms typically require a number of assumptions that are not satisfied in PDE-constrained optimization (e.g., convexity) and typically exhibit linear or sublinear convergence rates.

With these issues in mind, we restrict our attention to the expectation-based functionals $\mathscr{R}$ of the form

$$\mathscr{R}(X) = \inf_{t \in T} \mathbb{E}[v(X, t)]$$

where $v : \mathbb{R} \times \mathbb{R}^K \to \mathbb{R}$ and $T \subseteq \mathbb{R}^K$, $K \in \mathbb{N}$, is a closed convex set. This is a sufficiently rich class of functionals $\mathscr{R}$ that includes risk neutral $\mathscr{R} = \mathbb{E}$, the conditional value-at-risk (17), the probabilistic objective (19), the buffered probability (20), and the ϕ-divergence distributionally robust objective (28). In general, this class of functionals $\mathscr{R}$ includes the optimized certainty equivalent risk measures [17] and the expectation quadrangle risk measures [95]. To simplify notation, we denote $x = (z, t)$ for $z \in Z$ and $t \in \mathbb{R}^K$, $X = Z \times \mathbb{R}^K$ and $X_{\text{ad}} = Z_{\text{ad}} \times T$. The corresponding PDE-constrained optimization problem is

$$\min_{x=(z,t) \in X_{\text{ad}}} \mathbb{E}[v(\mathscr{J}(z), t)] + \wp(z). \tag{30}$$

For such problems, the composite objective function $h(x) := \mathbb{E}[v(\mathscr{J}(z), t)]$ inherits the differentiability properties of $v(\mathscr{J}(z), t)$ (e.g., [108, Sect. 7.2.4]). In many cases, the function v introduces nonsmoothness into the problem. For example, if $\mathscr{R} = \text{CVaR}_\alpha$, then $v(X, t) = \{t + (1 - \alpha)^{-1}(X - t)_+\}$ with $T = \mathbb{R}$ and if $\mathscr{R}$ is the buffered probability, then $v(X, t) = (t(X - \tau) + 1)_+$ with $T = [0, \infty)$. After fully discretizing (30), one could solve the resulting nonsmooth nonlinear optimization problem using, e.g., bundle methods [72]. We point out that there recently have been attempts to solve risk-averse optimization problems by smoothing CVaR (see [88] for finite-dimensional problems and [64] for PDE-constrained problems). One complication of smoothing approaches is that the gradient of the smoothed risk measure may become unstable as the smoothing is refined (i.e., as the smooth approximation approaches the original nonsmooth quantity), potentially leading to poor convergence of derivative-based optimization algorithms.

The growing interest in uncertainty quantification has led to the development of a multitude of methods for approximating the solution of PDEs with uncertain inputs. These methods can generally be partitioned into two classes: (i) intrusive methods and (ii) nonintrusive methods. Nonintrusive methods treat the deterministic PDE solver as a "black box," whereas intrusive methods require a reformulation of the deterministic PDE solver. Intrusive methods often approximate the solution of a PDE with uncertain inputs by projecting the solution or the PDE residual onto a finite-dimensional subspace such as a space of polynomials. Projection methods include, e.g., stochastic Galerkin and polynomial chaos methods [8, 9, 58, 122] (although there are nonintrusive forms of polynomial chaos [68]). On the other hand, nonintrusive approaches propagate a finite set of samples of ξ through the PDE. One then approximates the PDE solution field using interpolation or approximates integrated quantities such as moments using numerical integration. Some common choices for generating samples of ξ are (quasi) Monte Carlo [39], stochastic collocation on, e.g., sparse grids, [48, 49, 83–86, 110] and stochastic reduced order models [50, 51, 120]. In addition to these well-established methods, there has been much recent work devoted to low-rank tensor decomposition for parametrized PDE solutions [47, 59, 104]. In general, the approximation quality for polynomial-based uncertainty quantification methods is highly dependent on the choice of the approximation space, the dimension of Ξ, and the regularity of the PDE solution with respect to the random inputs.

The incorporation of uncertainty quantification methods within PDE-constrained optimization is an important and open area of research. Any feasible optimization method should be *mesh independent* in the sense that the convergence behavior does not depend on the size of the resulting discretizated problem (with respect to both the spatial domain and Ξ). Additionally, methods should exploit any structures inherent to the problem such as, e.g., adjoints, differentiability, and the optimality conditions in Theorem 2. Recently, numerous authors have applied intrusive and nonintrusive methods to approximate risk neutral optimization problems constrained by PDEs with uncertain inputs. Such problems were efficiently solved in [61, 62] using a trust-region algorithm to guide adaptive sparse grids for approximating the

objective function and its gradient. Similarly, [60] introduces a multilevel sparse grid approach that works well for some linear-quadratic and nonlinear control problems. Furthermore, the authors in [27] solve the risk neutral problem using sparse grids and reduced order models, whereas the authors of [112] solve this problem by combining nonintrusive polynomial chaos with sequential quadratic programming (SQP). Finally, the authors of [47] develop a semismooth Newton solver based on low-rank tensor decomposition to solve the risk neutral problem. Unfortunately, when v in (30) is not differentiable (e.g., minimizing CVaR or the buffered probability), the aforementioned trust-region, SQP, and semismooth Newton algorithms do not apply.

Given the myriad of possible approximations and algorithms for solving (30), we restrict our attention to three nonintrusive sampling approaches: the stochastic approximation algorithm, sample average and quadrature approximation, and the progressive hedging algorithm. We do not intend for this to be a complete list of possible solution techniques, but rather a review of classical methods in stochastic programming that may be applicable in PDE-constrained optimization. For each method, we provide an overview and highlight the challenges associated with the method in the context of PDE-constrained problems.

In the subsequent subsections, we assume X is a Hilbert space with inner product $\langle x, y\rangle_X$ and norm $\|x\|_X = \sqrt{\langle x, x\rangle_X}$. Moreover, we denote the uncertain composite objective function by $H(x,\xi) = v(\mathscr{J}(z,\xi), t)$ and the (deterministic) composite objective function by $h(x) = \mathbb{E}[H(x,\cdot)]$. We further denote the gradient or any subgradient (when $H(\cdot,\xi)$ is convex) of $H(\cdot,\xi)$ by $G(\cdot,\xi)$. To simplify the presentation, we ignore the control penalty term $\wp(z)$. However, all algorithms and results apply if $\wp(z)$ is included.

5.1 Stochastic Approximation

The *stochastic approximation* (SA) method was originally developed by Robbins and Monro in [91]. The method is based on the projected (sub)gradient method. The projection operator $\Pi : X \to X_{\text{ad}}$, onto the set $X_{\text{ad}} \subset X$, is defined as

$$\Pi(y) := \operatorname*{arg\,min}_{x \in X_{\text{ad}}} \|y - x\|_X.$$

Since X is a Hilbert space and X_{ad} is closed and convex, $\Pi(y)$ is uniquely defined for all $y \in X$ [12, Th. 3.14], and $y \mapsto \Pi(y)$ is nonexpansive [12, Prop. 4.8]. At the kth step of SA with the current iteration point x_k, the algorithm computes the next iteration point as

$$x_{k+1} = \Pi\left(x_k - \gamma_k G(x_k, \xi^k)\right). \tag{31}$$

Here $\gamma_k > 0$ are chosen step sizes and ξ^k is a realization of the random vector ξ typically generated by Monte Carlo sampling techniques. The random samples ξ^k, $k = 1, 2, \ldots$, are independent and generated according to the specified distribution of the random vector ξ. Therefore, each iteration point x_k is a random vector depending on the history of random samples $(\xi^1, \ldots, \xi^k)$. Note that each iteration requires a single state and adjoint solve corresponding to the random sample ξ^k. Although per-iteration cost of SA is low, the convergence (which is probabilistic) is heavily dependent on the convexity of $H(\cdot, \xi)$ and the choice of stepsize γ_k.

In the classical SA method, the step size is chosen to be $\gamma_k := \kappa/k$, where $\kappa > 0$ is a fixed constant. To analyze this method, we make the following assumptions:

(i) There exists a constant $M > 0$ such that

$$\mathbb{E}\left[\|G(x, \cdot)\|_X^2\right] \leq M^2, \quad x \in X_{\text{ad}}. \tag{32}$$

(ii) The function $h(x) = \mathbb{E}[H(x, \cdot)]$ is Fréchet differentiable and strongly convex, i.e., there exists $c > 0$ such that

$$h(x') \geq h(x) + \langle \nabla h(x), x' - x \rangle_X + \frac{1}{2} c \|x' - x\|_X^2 \quad \forall\, x,\ x' \in X.$$

Given these assumptions, problem (30) has a unique optimal solution $x_\star$. This result follows from the Direct Method of the Calculus of Variations (i.e., the strong convexity plus the continuity of h ensure the weak lower semicontinuity and coercivity of h). It is possible to show (cf. [80] for finite dimensional X) that for $\kappa > 1/(2c)$,

$$\mathbb{E}\left[\|x_k - x_\star\|_X^2\right] = O(k^{-1}). \tag{33}$$

That is, after k iterations, the expected error of the current solution in terms of the distance to the optimal solution $x_\star$ is of order $O(k^{-1/2})$. Moreover, if $\nabla h(x)$ is Lipschitz continuous and $x_\star \in X_{\text{ad}}$ satisfies $\nabla h(x_\star) = 0$, then (as a consequence of the Mean Value Theorem) we have

$$\mathbb{E}\left[h(x_k) - h(x_\star)\right] = O(k^{-1}). \tag{34}$$

For general convergence results of SA in Hilbert space, see [11].

Under the above assumptions (i) and (ii), the classical SA method produces iterates converging to the optimal solution. However, the method is very sensitive to choice of the step sizes and the convergence can be very slow. A simple example in [80] demonstrates that minimization of a deterministic quadratic function of one variable by the classical SA method can be extremely slow for a wrong choice of the constant κ. Moreover without strong convexity, the step sizes $\gamma_k = \kappa/k$ can result in disastrously slow convergence for any choice of the constant κ.

Another problem with (sub)gradient type algorithms is the possibility of different scales for the components of the vector x. Suppose that the space $X = \mathbb{R}^n$ is equipped with the standard Euclidean inner product $\langle x, y\rangle_X = x^\top y$ and consider the minimization of the (deterministic) quadratic function $h(x) = \frac{1}{2}x^\top Qx$ with Q being an $n \times n$ symmetric positive definite matrix. If the matrix Q is ill conditioned, then for any choice of the step sizes γ_k the SA algorithm will typically produce a zigzag trajectory, resulting in very slow convergence to the optimal solution.

Further, step sizes of order $O(k^{-1})$ could be too small to attain a reasonable rate of convergence, while taking larger step sizes, say of order $O(k^{-1/2})$, may result in no convergence of the algorithm. In order to resolve this problem, it was suggested in [82] (for finite-dimensional problems) to take larger step sizes and to use appropriate averages of the iterates x_k rather than these points themselves. It was shown in [89] that under the assumptions (i) and (ii), this strategy of taking larger step sizes and averaging automatically achieves the asymptotically optimal convergence rate. We follow [80] in analysis of this approach referred to as the robust SA method. Although the results in [80] are for finite dimensional X, it may be possible to extend them to the more general Hilbert space setting. We assume below that the function $h(x)$ is convex continuous, but not necessary strongly convex or differentiable, and that $\mathbb{E}[G(x, \cdot)]$ is a subgradient of h at x, i.e., $\mathbb{E}[G(x, \cdot)] \in \partial h(x)$. We also assume that condition (32) holds and the set X_{ad} is bounded.

For $1 < i < k$, together with the iterates x_k, consider the averages $\hat{x}_{ik} := \sum_{j=i}^k \nu_j x_j$ with weights $\nu_\ell := (\sum_{j=i}^k \gamma_j)^{-1}\gamma_\ell$. Note that $\nu_\ell > 0$ and $\sum_{j=i}^k \nu_j = 1$. We have then the following estimate: [80, p. 1580]

$$\mathbb{E}\big[h(\hat{x}_{ik}) - h(x_\star)\big] \le \frac{4D^2 + M^2 \sum_{j=i}^k \gamma_j^2}{2\sum_{j=i}^k \gamma_j} \quad \text{for } 1 < i < k, \tag{35}$$

where $D := \max_{x \in X_{\mathrm{ad}}} \|x - x_1\|_X$ (since it is assumed that the set X_{ad} is bounded, the constant D is finite). In particular, consider the strategy of fixing in advance the number of iterations N and the constant step sizes $\gamma_k = \gamma$, $k = 1, \ldots, N$. Then it follows from (35) that

$$\mathbb{E}\big[h(\hat{x}_{1N}) - h(x_\star)\big] \le \frac{4D^2 + M^2 N\gamma}{2N\gamma}. \tag{36}$$

Minimization of the right-hand side of (36) over $\gamma > 0$ suggests the optimal constant step size is

$$\gamma := \frac{2D}{M\sqrt{N}}, \tag{37}$$

providing the corresponding error estimate

$$\mathbb{E}\big[h(\hat{x}_{1N}) - h(x_\star)\big] \le \frac{2DM}{\sqrt{N}}. \tag{38}$$

Another possible strategy is to take step sizes of order $O(k^{-1/2})$, specifically

$$\gamma_k := \frac{\theta D}{M\sqrt{k}} \tag{39}$$

for some $\theta > 0$. Choosing i as a fixed fraction of N, i.e., setting $i = \lceil rN \rceil$ for some $r \in (0, 1)$, leads to the estimate

$$\mathbb{E}\big[h(\hat{x}_{iN}) - h(x_\star)\big] \leq C(r) \max\{\theta, \theta^{-1}\} \frac{DM}{\sqrt{N}}, \tag{40}$$

where $C(r)$ is a constant depending only on r.

The estimates (38) and (40) suggest the average error of the objective function to be of order $O(N^{-1/2})$. This could be compared with the estimate (34) of order $O(N^{-1})$. However, the error bounds (38) and (40) do not require differentiability or strong convexity of h. Additionally, scaling the step size in the robust SA algorithm by $\theta > 0$ has only a moderate effect on the bound (40), i.e., $\max\{\theta, \theta^{-1}\}$. Therefore, the robust SA method is considerably less sensitive to the choice of step sizes than the classical SA method. Nevertheless, the choice is still crucial for convergence of the algorithm and, unfortunately, the stepsize formulas (37) and (39) involve constants M, D, and the scaling factor θ that are often impossible to determine for PDE-constrained optimization problems.

5.2 Sample Average and Quadrature Approximation

Both the *sample average approximation* (SAA) and the deterministic quadrature approach result in approximations of the expectation in (30). As such, these methods are not algorithms for solving (30). The idea of the SAA method is to use equally probable random samples $\xi^1, \ldots, \xi^N$ to approximate the "true" optimization problem (30), whereas the quadrature approach aims to approximate the expectation in (30) using deterministic quadrature defined by N abscissae $\{\xi^1, \ldots, \xi^N\}$ and their corresponding weights $\{w^1, \ldots, w^N\}$. Both the SAA and quadrature approximations to (30) have the form

$$\min_{x \in X_{\text{ad}}} \left\{ \hat{h}_N(x) := \sum_{j=1}^{N} p^j H(x, \xi^j) \right\} \tag{41}$$

where $p^j = N^{-1}$ for SAA and $p^j = w^j$ for the quadrature approach. In the context of PDE-constrained optimization, (41) is a deterministic optimization problem with N PDE constraints. Therefore, any solution method for (41) should be mesh independent to avoid convergence issues associated with the dimension of the fully discretized problem.

There are advantages and disadvantages of the SA versus SAA or the quadrature approach. In finite dimensions, estimates of the sample size N needed to attain a specified accuracy of computed solutions are similar for both the SAA and the SA methods (cf., [108, Ch. 5]). SA is a simple algorithm requiring evaluation of a *single* (sub)gradient $G(x_j, \xi^j)$ at each iteration step, while SAA and the quadrature approach are not algorithms – the constructed problem (41) still has to be solved by a numerical procedure. Depending on the choice of algorithm for solving (41), each involved iteration can be considerably more expensive than in the SA method. For example, evaluation of the gradient (or a subgradient) of $\hat{h}_N$ at a given point x requires the calculation of *all* $G(x, \xi^j)$, $j = 1, \ldots, N$. On the other hand, SAA and the quadrature approach, combined with a good numerical optimization algorithm, may overcome the difficulties of the choice of step sizes that plagues the SA method. Also SAA and the quadrature approach are more receptive to parallelization, e.g., the (sub)gradients $G(x, \xi^j)$, $j = 1, \ldots, N$ can be computed in parallel as opposed to the sequential nature of the SA method. However, additional difficulty may arise for the quadrature approximation if the weights w^j are not all positive as with, e.g., sparse grids [48, 49, 85, 86, 110]. The presence of negative weights may adversely influence a numerical optimization solver by changing the sign associated with the objective sample $H(x, \xi^j)$.

Given the similarities between SAA and the quadrature approach, we can characterize the error committed through the approximation of (30) using the same techniques. For the subsequent analysis, we assume $x \mapsto H(x, \xi)$ is continuously Fréchet differentiable for each $\xi \in \Xi$, ensuring that h and $\hat{h}_N$ are continuously Fréchet differentiable. If h is strongly convex, then we can characterize the errors between the true optimal solution $x_\star \in X_{\text{ad}}$ and the approximate solution $x_N \in X_{\text{ad}}$. Namely, strong convexity implies there exists $c > 0$ such that

$$c\|x_\star - x_N\|_X^2 \leq \langle \nabla h(x_\star) - \nabla h(x_N), x_\star - x_N \rangle_X.$$

Similar to Theorem 2, the optimality conditions for h and $\hat{h}_N$ over X_{ad} are

$$\langle \nabla h(x_\star), x - x_\star \rangle_X \geq 0 \quad \forall\, x \in X_{\text{ad}} \quad \text{and} \quad \langle \nabla \hat{h}_N(x_N), x - x_N \rangle_X \geq 0 \quad \forall\, x \in X_{\text{ad}},$$

respectively. Since $x_\star,\ x_N \in X_{\text{ad}}$, we have that

$$\langle \nabla h(x_\star), x_\star - x_N \rangle_X \leq 0 \leq \langle \nabla \hat{h}_N(x_N), x_\star - x_N \rangle_X.$$

This relation and the Cauchy–Schwarz inequality ensure that

$$c\|x_\star - x_N\|_X \leq \|\nabla \hat{h}_N(x_N) - \nabla h(x_N)\|_X = \left\| \sum_{j=1}^{N} p^j G(x_N, \xi^j) - \mathbb{E}[G(x_N, \cdot)] \right\|_X. \tag{42}$$

Therefore, the right-hand side of (42) is simply the error associated with approximately integrating the gradient of $H(x_N, \cdot)$ and thus the error will be dictated by the

approximation quality of the points $(\xi^1, \ldots, \xi^N)$ and weights $(p^1, \ldots, p^N)$. In the context of quadrature approximation, this error depends heavily on the regularity of, e.g., the adjoint state with respect to ξ, the dimension of Ξ, and the polynomial order of the quadrature rule (see, for example, [83, 84, 86]). Thus, the convergence rate of the optimal solutions for the quadrature approximation may be algebraic, even exponential, if the gradients G are sufficiently regular with respect to ξ. On the other hand, the convergence rate for SAA is probabilistic since $(\xi^1, \ldots, \xi^N)$ are random realizations of ξ and will likely recover the Monte Carlo rate of convergence $O(N^{-1/2})$ [39].

5.3 Progressive Hedging

The progressive hedging algorithm [96], originally introduced for dynamic stochastic programs, employs a sample-based decomposition of (30). As in Section 5.2, we consider the approximate optimization problem (41) where $(\xi^1, \ldots, \xi^N)$ are fixed scenarios of the uncertain inputs ξ with associated probabilities $(p^1, \ldots, p^N)$ (i.e., $p^j \geq 0$ for all j and $p^1 + \ldots + p^N = 1$). As discussed in Section 5.2, we can exploit parallelism in (41) by evaluating $\hat{h}_N$ and its derivatives in parallel. By assigning a separate optimization variable x^j for each ξ^j (i.e., we allow x^j to *anticipate* the scenario ξ^j), the progressive hedging algorithm further exploits parallel computations at each iteration by concurrently solving a deterministic PDE-constrained optimization problem for each scenario ξ^j.

To describe the progressive hedging algorithm, we first reformulate (41) as

$$\min_{x_j, x \in X_{\text{ad}}} \sum_{j=1}^{N} p^j H(x^j, \xi^j) \quad \text{subject to} \quad x^j = x, \ \ j = 1, \ldots, N. \tag{43}$$

Here, the objective function is the sum of decoupled, scenario-specific objective functions, whereas the constraint ensures that we recover a solution to (41). We call the deterministic variable x an *implementable* solution. We then relax the equality constraint for each j using the augmented Lagrangian penalty function

$$\ell_r^j(x^j, x, \mu^j) = H(x^j, \xi^j) + \langle \mu^j, x^j \rangle_X + \frac{r}{2} \|x^j - x\|_X^2, \quad r > 0,$$

where the multipliers $\{\mu^1, \ldots, \mu^N\}$ are called an *information price system* in [96] and are required to satisfy

$$\sum_{j=1}^{N} p^j \mu^j = 0.$$

Taking the expectation of ℓ_r^j then yields the full Augmented Lagrangian for (43). In light of this, we can describe the progressive hedging algorithm as follows. Given the kth iteration points $x_k^j \in X_{\text{ad}}$ and $\mu_k^j \in X$ for $j = 1, \ldots, N$, and the current implementable solution $x_k = \sum_{j=1}^N p^j x_k^j$:

1. Compute the scenario-dependent solutions x_{k+1}^j, $j = 1, \ldots, N$ by minimizing $\ell_r^j(\cdot, x_k, \lambda_k^j)$ concurrently, i.e.,

$$x_{k+1}^j \in \underset{x^j \in X_{\text{ad}}}{\arg\min}\ \ell_r^j(x^j, x_k, \lambda_k^j),\ \ j = 1, \ldots, N; \tag{44}$$

2. Aggregate x_{k+1}^j to compute the current implementable solution x_{k+1}, i.e.,

$$x_{k+1} = \sum_{j=1}^N p^j x_{k+1}^j;$$

3. Update the multiplier estimates for fixed $x = x_{k+1}$ and $x^j = x_{k+1}^j$, $j = 1, \ldots, N$, as

$$\mu_{k+1}^j = \mu_k^j + r(x_{k+1}^j - x_{k+1}),\ \ j = 1, \ldots, N. \tag{45}$$

Clearly, all steps of this algorithm are parallelizable with the exception of the second (i.e., aggregation) step.

The convergence theory for the progressive hedging algorithm, as set fourth in [96], is restricted to finite dimensions. When $H(\cdot, \xi)$ is convex, the progressive hedging algorithm converges under specified stopping rules for approximately solving (44) (see Equation 5.35 and Theorem 5.4 in [96]). In fact, the convergence theory in the convex case is based on the convergence theory for the proximal point algorithm [92] applied to a certain saddle function. As the authors in [42] point out, the progressive hedging algorithm can be seen as a special case of Douglas–Rachford splitting and thus inherits the Hilbert space convergence theory. On the other hand, Theorem 6.1 in [96] demonstrates that if $H(\cdot, \xi)$ is not convex and X is finite dimensional, then if the sequences of iterates x_k^j and multipliers μ_k^j converge, where x_k^j are only required to be δ-locally optimal for fixed $\delta > 0$, then these sequences converge to a stationary point of the original problem (30). Given the relations between the progressive hedging and Augmented Lagrangian algorithms, it may be possible to extend the convergence analysis for Augmented Lagrangian for infinite-dimensional nonconvex problems (see, e.g., [54, Chapt. 3]).

To conclude, one potential inefficiency of the progressive hedging algorithm is the typically slow convergence rate. For example, if X is finite dimensional, $H(\cdot, \xi)$ is convex quadratic, and X_{ad} is convex polyhedral, then Theorem 5.2 in [96] ensures that the progressive hedging algorithm will converge at a linear rate. One can potentially overcome this by increasing the penalty parameter r at each

iteration (see, e.g., Theorem 2 in [92] where superlinear convergence for convex problems is shown using the proximal point algorithm). In any case, the convergence of the progressive hedging algorithm is strongly dependent on the penalty parameter r which is difficult to select a priori, especially for nonconvex problems. Another possibility to enhance the convergence rate is to replace (45) with a "second-order" multiplier update (see, e.g., [22, Ch. 2.3.2] and [54, Chapt. 6.2] for second-order multiplier updates in the context of the Augmented Lagrangian algorithm).

6 Numerical Example

To demonstrate the various stochastic programming formulations discussed in Section 4, we consider the problem of optimally mitigating a contamination by injecting chemicals at specified locations that dissolve the contaminant. We model the contaminant transport using the steady advection diffusion equation. Clearly, uncertainties arise in nearly all coefficients such as the velocity field (e.g., wind) and the contaminant source locations and magnitudes. This example was first considered in [64]. Let $D = (0, 1)^2$ denote the physical domain and $U = H^1(D)$ be the space of contaminant concentrations. The target optimization problem is

$$\min_{z \in Z_{\text{ad}}} \mathscr{R}\left(\frac{\kappa_s}{2}\int_D S(z;\cdot)^2 \,\mathrm{d}x\right) + \wp(z) \tag{46}$$

where $\kappa_s > 0$ and $S(z;\cdot) = u : \Xi \to U$ solves the weak form of the advection-diffusion equation

$$-\nabla\cdot(\epsilon(\xi)\nabla u) + \mathbb{V}(\xi)\cdot\nabla u = f(\xi) - Bz \qquad \text{in } D \tag{47a}$$

$$u = 0 \qquad \text{on } \Gamma_d \tag{47b}$$

$$-\epsilon(\xi)\nabla u \cdot n = 0 \qquad \text{on } \Gamma_n \tag{47c}$$

where the Neumann boundary is $\Gamma_n := \{1\}\times(0,1)$ and the Dirichlet boundary is $\Gamma_d := \partial D \setminus \Gamma_n$. The control space (the space of mitigating chemical concentrations) is $Z = \mathbb{R}^9$ with admissible control set $Z_{\text{ad}} := \left\{z \in \mathbb{R}^9 : 0 \le z \le 1\right\}$ and control cost

$$\wp(z) := \kappa_c\|z\|_1 = \kappa_c\sum_{k=1}^{9}|z_k|, \quad \kappa_c > 0.$$

The controls are applied using the operator $B \in \mathscr{L}(Z, L^\infty(D))$ given by

$$(Bz)(x) = \sum_{k=1}^{9} z_k \exp\left(-\frac{(x-p_k)^\top(x-p_k)}{2\sigma^2}\right)$$

Table 1 Predetermined contaminant mitigating control locations

Source	1	2	3	4	5	6	7	8	9
x_1	0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75
x_2	0.25	0.25	0.25	0.50	0.50	0.50	0.75	0.75	0.75

where p_k are predetermined control locations and $\sigma = 0.05$. That is, we model the control mechanism as Gaussians sources with magnitude dictated by z. The control locations are tabulated in Table 1.

The PDE coefficients ϵ, $\mathbb{V}$, and f are random fields. The diffusivity is given by

$$\epsilon(x, \xi) = 0.5 + c \exp(\delta(x, \xi))$$

where the specific form of δ is described in [83, Sect. 4, Eqs. 4.2–4.4]. Associated with δ are 10 random variables, $(\xi_1, \ldots, \xi_{10})$, uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$. The constant $c > 0$ is chosen to be the reciprocal of the maximum of $\exp(\delta)$. Clearly, ϵ satisfies: $\exists\, 0 < \epsilon_0 \leq \epsilon \leq \epsilon_1 < \infty$ for all $x \in D$ and $\xi_i \in [-\sqrt{3}, \sqrt{3}]$, $i = 1, \ldots, 10$. Moreover, the velocity field $\mathbb{V}$ is

$$\mathbb{V}(x, \xi) = \begin{bmatrix} \xi_{12} - \xi_{11} x_1 \\ \xi_{11} x_2 \end{bmatrix}$$

where ξ_{11} is uniformly distributed on $[0, 5]$, and ξ_{12} is uniformly distributed on $[5, 10]$. The two extreme cases of $\mathbb{V}$ are depicted in Figure 2. $\mathbb{V}$ is divergence free and satisfies $\mathbb{V} \cdot n \geq 0$, where n is the outward unit normal vector on the Neumann boundary. Finally, f is the sum of five Gaussian sources whose locations, widths, and magnitudes are random, i.e., f is described by 25 uniform random variables $(\xi_{13}, \ldots, \xi_{37})$. This results in a total of 37 random variables associated with the PDE (47). As shown in [64], this example satisfies the assumptions of Theorems 1 and 2 and thus a minimizing control exists and it satisfies the first-order necessary conditions in Theorem 2.

We approximate the contaminant mitigation problem using SAA with $N = 800$ Monte Carlo samples. For $\mathscr{R}$, we chose risk neutral (RN), entropic risk (ER) with $\sigma = 1$, CVaR with $\alpha = 0.95$, a convex combination of expectation and CVaR

$$\mathscr{R}(X) = \beta \mathbb{E}[X] + (1 - \beta) \mathrm{CVaR}_\alpha(X)$$

with $\alpha = 0.95$ and $\beta = 0.5$ (MCVaR), buffered probability with threshold $\tau = 6$ (BP), and KL-divergence distributionally robust optimization with threshold $\epsilon = 0.1$ (KL). Additionally, we solved the *mean value problem* (MV) in which we replaced ξ with $\mathbb{E}[\xi]$ and solved the corresponding deterministic control problem. For RN, ER, KL, and MV, we solved the resulting nonlinear program using a trust-region Newton method [32]; while for CVaR, MCVaR, and BP, we combined the aforementioned trust-region method with an adaptation of the smoothing approach described in [64]. Figure 3 depicts the optimal control sources and Table 2 includes the optimal control magnitudes. We excluded the MV control from Figure 3 due

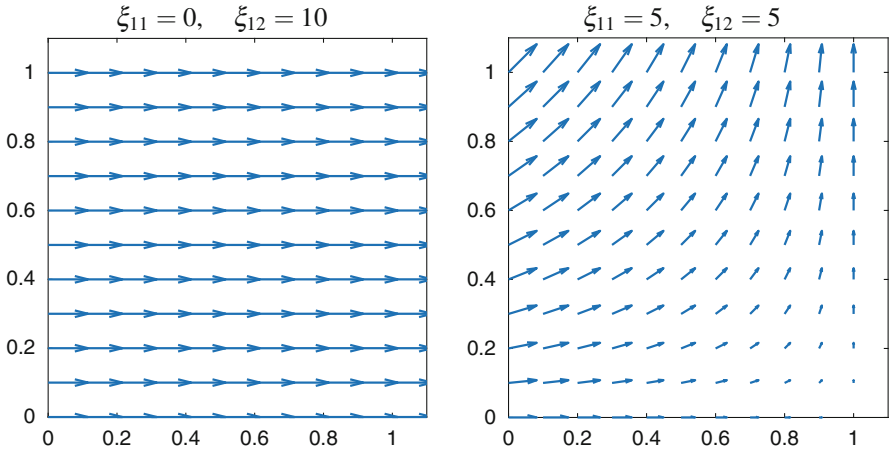

Fig. 2 Left: The vector field $\mathbb{V}$ with $\xi_{11} = 0$ and $\xi_{12} = 10$. Right: The vector field $\mathbb{V}$ with $\xi_{11} = 5$ and $\xi_{12} = 5$

Table 2 Optimal contaminant mitigating controls using different functionals $\mathscr{R}$. MV refers to the deterministic problem in which the random inputs are replaced with their expected values. RN refers to risk neutral and ER refers to the entropic risk with $\sigma = 1$. For CVaR, we set $\alpha = 0.95$ and for the "mixture of CVaRs" (MCVaR), we set $\alpha = 0.95$ and $\beta = 0.5$. For the "buffered probability of exceedance" (bPOE), we set the threshold $\tau = 6$ and for the KL-divergence distributionally robust problem, we set the threshold $\epsilon = 0.1$.

$\mathscr{R}$	1	2	3	4	5	6	7	8	9	Cost
MV	–	0.23	–	–	1.00	–	–	–	–	1.23
RN	–	0.27	–	–	1.00	–	–	–	–	1.27
CVaR	0.42	1.00	0.15	0.81	1.00	1.00	–	–	–	4.37
MCVaR	–	1.00	–	0.33	1.00	0.59	–	–	–	2.92
ER	0.33	1.00	1.00	0.55	1.00	1.00	–	–	–	4.88
BP	0.02	1.00	–	0.56	1.00	0.91	–	–	–	3.49
KL	–	1.00	–	–	1.00	–	–	–	–	2.00

to its similarity with the RN control. For the given parameter specifications, ER produced the most conservative control, whereas RN and MV produce the least conservative. However, conservativeness results in a more expensive control. This fact is depicted in Figure 4. Figure 4 includes the cdfs of the uncertain objective function $\mathscr{J}(z)$ (left) and the full objective function $\mathscr{J}(z) + \wp(z)$ (right) evaluated at the different optimal controls. The left image clearly demonstrates that more conservative approaches reduce variability and produce uncertain objective values that dominate (in the sense of the first stochastic order) those of the RN and MV approaches. On the other hand, the right image emphasizes the increased cost of being conservative. As seen in the right image, the RN and MV controls outperform the other controls in terms of total cost for more than 60% of scenarios.

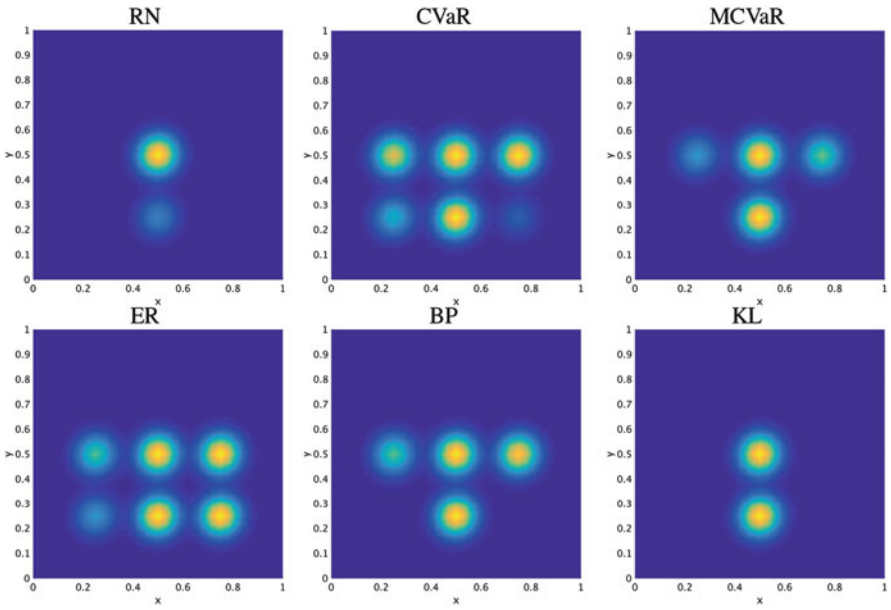

Fig. 3 The optimal controls computed using risk neutral (RN), CVaR, a mixture of expectation and CVaR (MCVaR), entropic risk (ER), buffered probability (BP) and KL-divergence distributionally robust optimization (KL)

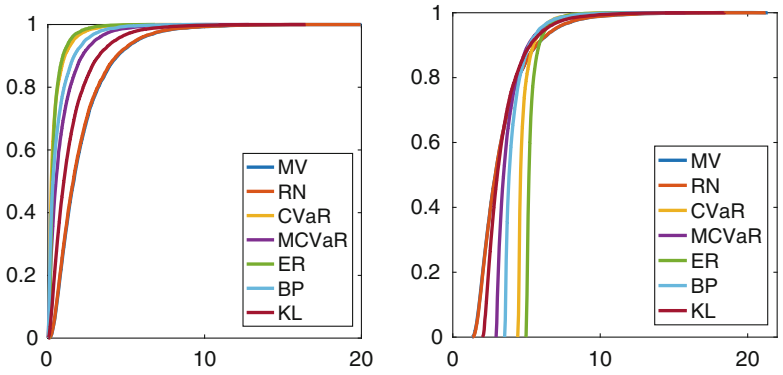

Fig. 4 Left: Cumulative distribution functions of the random variable objective function, $\mathscr{J}(z)$, evaluated at the different optimal controls. Right: Cumulative distribution functions of $\mathscr{J}(z) + \wp(z)$ evaluated at the different optimal controls

7 Conclusions

In this chapter, we reviewed a set of stochastic programming tools for formulating and solving optimization problems constrained by PDEs with uncertain coefficients. For the problem formulation, we discussed risk measures, probabilistic

optimization, and distributionally robust optimization. Each of these approaches can be justified within the context of the physical application. When the underlying probability law of the random coefficients is known, risk-averse and probabilistic optimization provide a natural foundation for incorporating conservativeness in the optimization problem formulation. However, such approaches are unjustified and may lead to arbitrarily poor solutions if the underlying probability law is unknown. In this scenario, one often has noisy, incomplete data describing the distribution of uncertain coefficients which can be used to define an ambiguity set of "feasible" distributions. This leads naturally to distributionally robust optimization in which we minimize the worst-case expectation over the ambiguity set.

For solution approaches, we discussed stochastic approximation (SA), sample average approximation (SAA), deterministic quadrature approximation, and the progressive hedging algorithm. Each approach has particular downsides. The SA approach is a simple optimization algorithm but requires convexity to guarantee convergence, which is probabilistic. The SAA approach approximates the expected value in the objective function using a sample average (e.g., Monte Carlo). The resulting approximate problem is then solved using nonlinear programming algorithms. SAA exhibits dimension-independent convergence, but the convergence is probabilistic with rate $1/\sqrt{N}$. Similar to SAA, the deterministic quadrature approach approximates the expected value using quadrature. The resulting problem is again solved with a nonlinear programming method. This approach requires sufficient regularity (with respect to the random inputs) to obtain rapidly decaying approximation error. Finally, the progressive hedging algorithm employs a sample-based decomposition of the optimization problem and the controls which permits the concurrent solution of deterministic PDE-constrained optimization problems at every iteration. For convex problems, convergence is guaranteed in Hilbert space; however, the convergence rate can be linear or worse.

Common among many stochastic optimization problems is the challenge of minimizing a nonsmooth objective function. In particular, the typical slow convergence rates of nonsmooth optimization algorithms may render the solution of PDE-constrained optimization under uncertainty computationally infeasible. Efficiently solving these nonsmooth problems is challenging and is an active research topic. Additional open research topics include the formulation and analysis for state-constrained problems; the incorporation of stochastic dominance and chance constraints for PDE-constrained optimization; and the formulation, analysis, and numerical solution of optimal control problems constrained by variational inequalities with uncertain inputs as well as optimal control problems constrained by dynamic stochastic PDEs.

Acknowledgements This work was supported by DARPA EQUiPS grant SNL 014150709.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

1. R. A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.
2. E. Andreassen, B. S. Lazarov, and O. Sigmund. Design of manufacturable 3d extremal elastic microstructure. *Mechanics of Materials*, 69(1):1–10, 2014.
3. V. Artus, J. L. Durlofsky, J. Onwunalu, and K. Aziz. Optimization of nonconventional wells under uncertainty using statistical proxies. *Computational Geosciences*, 10(4):389–404, 2006.
4. Ph. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Math. Finance*, 9(3):203–228, 1999.
5. A. Asadpoure, M. Tootkaboni, and J. K. Guest. Robust topology optimization of structures with uncertainties in stiffness – applications to trust structures. *Computers & Structures*, 89(11–12):1131–1141, 2011.
6. H. Attouch, G. Buttazzo, and G. Michaille. *Variational analysis in Sobolev and BV spaces*, volume 6 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.
7. I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Rev.*, 52(2):317–355, 2010.
8. I. Babuška, R. Tempone, and G. E. Zouraris. Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM J. Numer. Anal.*, 42(2):800–825 (electronic), 2004.
9. I. Babuška, R. Tempone, and G. E. Zouraris. Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation. *Comput. Methods Appl. Mech. Engrg.*, 194(12–16):1251–1294, 2005.
10. W. Bangerth, H. Klie, M. F. Wheeler, P. L. Stoffa, and M. K. Sen. On optimization algorithms for the reservoir oil well placement problem. *Computational Geosciences*, 10(3):303–319, 2006.
11. K. Barty, J.-S. Roy, and C. Strugarek. Hilbert-valued perturbed subgradient algorithms. *Mathematics of Operations Research*, 32(3):551–562, 2007.
12. H. H. Bauschke and P. L. Combettes. *Convex Analysis and Montone Operator Theory in Hilbert Space*. CMS Books in Mathematics. Springer New York, 2011.
13. R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.
14. A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, 2009.
15. A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.
16. A. Ben-Tal and M. Teboulle. Penalty functions and duality in stochastic programming via phi-divergence functionals. *Mathematics of Operations Research*, 12:224–240, 1987.
17. A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.
18. J. O. Berger. The robust Bayesian viewpoint (with discussion). *Robustness of Bayesian Analysis*, pages 63–124, 1985.
19. J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, 1985.
20. J. O. Berger. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.
21. J. G. Berryman and G. W. Milton. Microgeometry of random composites and porous media. *Journal of Physics D: Applied Physics*, 21(1):87, 1988.
22. D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, New York,. London, Paris, San Diego, San Francisco, 1982.
23. D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

24. D. Bertsimas and J. Sethuraman. Moment problems and semidefinite optimization. In H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors, *Handbook of Semidefinite Programming*, pages 469–510. Kluwer Academic Publishers, 2000.
25. J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer-Verlag, New York, 1997.
26. J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Verlag, Berlin, Heidelberg, New York, 2000.
27. A. Borzì and G. von Winckel. A POD framework to determine robust controls in PDE optimization. *Comput. Vis. Sci.*, 14:91–103, 2011.
28. S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer Verlag, Berlin, Heidelberg, New York, second edition, 2002.
29. P. Cheridito and T. Li. Risk measures on Orlicz hearts. *Mathematical Finance*, 19(2):189–214, 2009.
30. F. H. Clarke. *Nonsmooth Analysis and Control Theory*. Graduate Texts in Mathematics. Springer, 1998.
31. A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best n-term Galerkin approximations for a class of elliptic sPDEs. *Foundations of Computational Mathematics*, 10(6):615–646, 2010.
32. A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust–Region Methods*. SIAM, Philadelphia, 2000.
33. J. B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York, 1985.
34. I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutato Int. Kozls*, 8, 1063.
35. A. Defant and K. Floret. *Tensor Norms and Operator Ideals*. North-Holland Mathematics Studies. Elsevier Science, 1993.
36. E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58:595–6127, 2010.
37. D. Dentcheva, S. Penev, and A. Ruszczyński. Kusuoka representation of higher order dual risk measures. *Annals of Operations Research*, 181(1):325–335, 2010.
38. D. Dentcheva and A. Ruszczyński. Optimization with stochastic dominance constraints. *SIAM Journal on Optimization*, 14(2):548–566, 2003.
39. I. T. Dimov. *Monte Carlo methods for applied scientists*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2008.
40. O. Dorn and R. Villegas. History matching of petroleum reservoirs using a level set technique. *Inverse Problems*, 24(3):035015, 2008.
41. J. Dupačová. Uncertainties in minimax stochastic programs. *Optimization*, 60(10–11):1235–1250, 2011.
42. J. Eckstein and D. P. Bertsekas. On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, Apr 1992.
43. Y. M. Ermoliev and A. A. Gaivoronski. Stochastic methods for solving minimax problems. *Cybernetics*, 19(4):550–559, 1983.
44. Y. M. Ermoliev, A. A. Gaivoronski, and C. Nedeva. Stochastic optimization problems with incomplete information on distribution functions. *SIAM Journal on Control and Optimization*, 23(5):697–716, 1985.
45. G. B. Folland. *Real analysis. Modern techniques and their applications*. Pure and Applied Mathematics (New York). John Wiley & Sons Inc., New York, second edition, 1999.
46. A. A. Gaivoronski. A numerical method for solving stochastic programming problems with moment constraints on a distribution function. *Annals of Operations Research*, 31(1):347–369, 1991.
47. S. Garreis and M. Ulbrich. Constrained optimization with low-rank tensors and applications to parametric problems with PDEs. *SIAM Journal on Scientific Computing*, 39(1):A25–A54, 2017.

48. T. Gerstner and M. Griebel. Numerical integration using sparse grids. *Numer. Algorithms*, 18(3–4):209–232, 1998.
49. T. Gerstner and M. Griebel. Dimension-adaptive tensor-product quadrature. *Computing*, 71(1):65–87, 2003.
50. M. Grigoriu. Reduced order models for random functions. application to stochastic problems. *Applied Mathematical Modelling*, 33(1):161–175, 2009.
51. M. Grigoriu. A method for solving stochastic equations by reduced order models and local approximations. *Journal of Computational Physics*, 231(19):6495–6513, 2012.
52. V. Hauk. *Structural and Residual Stress Analysis by Nondestructive Methods: Evaluation - Application - Assessment*. Elsevier Science, 1997.
53. E. Hille and R. S. Phillips. *Functional analysis and semi-groups*. American Mathematical Society Colloquium Publications, vol. 31. American Mathematical Society, Providence, R. I., 1957. rev. ed.
54. K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*. Society for Industrial and Applied Mathematics, 2008.
55. P. Kall and S. W. Wallace. *Stochastic Programming*. Wiley, Chichester etc., 1994.
56. S. Kalpakjian and S. R. Schmid. *Manufacturing Engineering and Technology*. Prentice Hall, 2010.
57. K. Karhunen. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.*, 1947(37):79, 1947.
58. G. E. Karniadakis, C.-H. Su, D. Xiu, D. Lucor, C. Schwab, and R. A. Todor. Generalized polynomial chaos solution for differential equations with random inputs. Technical Report 2005–01, Seminar for Applied Mathematics, ETH Zurich, Zurich, Switzerland, 2005.
59. B. Khoromskij and C. Schwab. Tensor-structured Galerkin approximation of parametric and stochastic elliptic PDEs. *SIAM J. Sci. Comput.*, 33(1):364–385, 2011.
60. D. P. Kouri. A multilevel stochastic collocation algorithm for optimization of PDEs with uncertain coefficients. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):55–81, 2014.
61. D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders. A trust-region algorithm with adaptive stochastic collocation for PDE optimization under uncertainty. *SIAM Journal on Scientific Computing*, 35(4):A1847–A1879, 2013.
62. D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders. Inexact objective function evaluations in a trust-region algorithm for PDE-constrained optimization under uncertainty. *SIAM Journal on Scientific Computing*, 36(6):A3011–A3029, 2014.
63. D. P. Kouri and T. M. Surowiec. Existence and optimality conditions for risk-averse PDE-constrained optimization. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):787–815, 2018.
64. D. P. Kouri and T. M. Surowiec. Risk-averse PDE-constrained optimization using the conditional value-at-risk. *SIAM Journal on Optimization*, 26(1):365–396, 2016.
65. J. R. Krebs, J. E. Anderson, D. Hinkley, R. Neelamani, S. Lee, A. Baumstein, and M. D. Lacasse. Fast full-waveform seismic inversion using encoded sources. *Geophysics*, 74(6):177–188, 2009.
66. P. A. Krokhmal. Higher moment coherent risk measures. *Quantitative Finance*, 7(4):373–387, 2007.
67. B. Lazarov, M. Schevenels, and O. Sigmund. Topology optimization considering material and geometric uncertainties using stochastic collocation methods. *Structural and Multidisciplinary Optimization*, pages 1–16, 2012. online first.
68. O. P. Le Maitre and O. M. Knio. *Spectral Methods for Uncertainty Quantification With Applications to Computational Fluid Dynamics*. Scientific Computation. Springer-Verlag, Berlin, 2010.
69. M. Loève. *Probability theory. II*. Graduate Texts in Mathematics, Vol. 46. Springer-Verlag, New York, fourth edition, 1978.
70. D. Love and G. Bayraksan. Phi-divergence constrained ambiguous stochastic programs. Technical report, Technical report, Program in Applied Mathematics, University of Arizona, 2013.

71. A. Mafusalov and S. Uryasev. Buffered probability of exceedance: mathematical properties and optimization. *SIAM Journal on Optimization,* 28(2):1077–1103, 2018.
72. M. M. Mäkelä and N. Neittaanmäki. *Nonsmooth Optimization: Analysis And Algorithms With Applications To Optimal Control*. World Scientific Publishing Company, 1992.
73. E. M. Makhlouf, W. H. Chen, M. L. Wasserman, and J. H. Seinfeld. A general history matching algorithm for three-phase, three-dimensional petroleum reserviors. *Society of Petroleum Engineers*, 1(2), 1993.
74. H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):pp. 77–91, 1952.
75. K. Marti, editor. *Stochastic Optimization. Numerical Methods and Technical Applications*. Springer, Berlin, 1992. LN in Economics and Math. Systems 379.
76. K. Marti. Differentiation formulas for probability functions: The transformation method. *Mathematical Programming*, 75:201–220, 1996.
77. K. Maute. *Topology Optimization under Uncertainty*, pages 457–471. Springer Vienna, Vienna, 2014.
78. K. Maute and D. M. Frangopol. Reliability-based design of mems mechanisms by topology optimization. *Computers & Structures*, 81(8–11):813–824, 2003.
79. T. Morimoto. Markov processes and the h-theorem. *J. Phys. Soc. Jap.*, 18(3):328–333, 1963.
80. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
81. A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.
82. A. Nemirovski and D. Yudin. On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. *Soviet Math. Dokl.*, 239:1056–1059, 1978.
83. F. Nobile, R. Tempone, and C. G. Webster. An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, 46(5):2411–2442, 2008.
84. F. Nobile, R. Tempone, and C. G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 46(5):2309–2345, 2008.
85. E. Novak and K. Ritter. High-dimensional integration of smooth functions over cubes. *Numer. Math.*, 75(1):79–97, 1996.
86. E. Novak and K. Ritter. Simple cubature formulas with high polynomial exactness. *Constr. Approx.*, 15(4):499–522, 1999.
87. B.K. Pagnoncelli, S. Ahmed, and A. Shapiro. Sample average approximation method for chance constrained programming: theory and applications. *J. Optim. Theory Appl.*, 142(2):399–416, 2009.
88. J. S. Pang and S. Leyffer. On the global minimization of the value-at-risk. *Optimization Methods and Software*, 19(5):611–631, 2004.
89. B.T. Polyak. New method of stochastic approximation type. *Automat. Remote Control*, 51:937–946, 1990.
90. A. Prékopa. Probabilistic programming. In *Stochastic programming*, volume 10 of *Handbooks Oper. Res. Management Sci.*, pages 267–351. Elsevier Sci. B. V., Amsterdam, 2003.
91. H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 9 1951.
92. R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.
93. R. T. Rockafellar and J. O. Royset. On buffered failure probability in design and optimization of structures. *Reliability Engineering & System Safety*, 95(5):499–510, 2010.
94. R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

95. R. T. Rockafellar and S. Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18(1–2):33–53, 2013.
96. R. T. Rockafellar and Roger J.-B. Wets. Scenarios and policy aggregation in optimization under uncertainty. *Math. Oper. Res.*, 16(1):119–147, 1991.
97. W. W. Rogosinski. Moments of non-negative mass. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 245(1240):1–27, 1958.
98. J. O. Royset and E. Polak. Extensions of stochastic optimization results to problems with system failure probability functions. *Journal of Optimization Theory and Applications*, 133(1):1–18, 2007.
99. A. Ruszczyński and A. Shapiro. Optimization of risk measures. In G. Calafiore and F. Dabbene, editors, *Probabilistic and Randomized Methods for Design Under Uncertainty*, pages 119–157, London, 2006. Springer Verlag.
100. R. A. Ryan. *Introduction to tensor products of Banach spaces*. Springer Monographs in Mathematics. Springer-Verlag London Ltd., London, 2002.
101. F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.
102. P. Sarma, L. J. Durlofsky, K. Aziz, and W. H. Chen. Efficient real-time reservoir management using adjoint-based optimal control and model updating. *Computational Geosciences*, 10(1):3–36, 2006.
103. H. Scarf. A min-max solution of an inventory problem. In *Studies in the Mathematical Theory of Inventory and Production*, pages 201–209. Stanford University Press, 1958.
104. C. Schwab and C. J. Gittelson. Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. *Acta Numer.*, 2011:291–467, 2011.
105. A. Shapiro. On concepts of directional differentiability. *J. Optim. Theory Appl.*, 66(3):477–487, 1990.
106. A. Shapiro. Monte Carlo sampling methods. In A. Ruszczynski and A. Shapiro, editors, *Stochastic Programming*, Handbooks in Operations Research and Management Science, Vol. 10, pages 353–425. Elsevier, 2003.
107. A. Shapiro. Distributionally robust stochastic programming. *SIAM J. Optimization*, 27(4):2258–2275, 2017.
108. A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2014.
109. O. Sigmund. Manufacturing tolerant topology optimization. *Acta Mechanica Sinica*, 25(2):227–239, 2009.
110. S. A. Smoljak. Quadrature and interpolation formulae on tensor products of certain function classes. *Soviet Math. Dokl.*, 4:240–243, 1963.
111. W. W. Symes and J. J. Carazzone. Velocity inversion by differential semblance optimization. *Geophysics*, 56(5):654–663, 1991.
112. H. Tiesler, R. M. Kirby, D. Xiu, and T. Preusser. Stochastic collocation for optimal control problems with stochastic PDE constraints. *SIAM Journal on Control and Optimization*, 50(5):2659–2682, 2012.
113. S. Uryasev. Derivatives of probability functions and integrals over sets given by inequalities. *J. Comput. Appl. Math.*, 56(1–2):197–223, 1994. Stochastic programming: stability, numerical methods and applications (Gosen, 1992).
114. S. Uryasev. Derivatives of probability functions and some applications. *Ann. Oper. Res.*, 56:287–311, 1995. Stochastic programming (Udine, 1992).
115. S. Uryasev and R. T. Rockafellar. Conditional value-at-risk: Optimization approach. In S. Uryasev and P. M. Pardalos, editors, *Stochastic optimization: algorithms and applications. Papers from the conference held at the University of Florida, Gainesville, FL, February 20–22, 2000*, volume 54 of *Appl. Optim.*, pages 411–435. Kluwer Acad. Publ., Dordrecht, 2001.

116. M. M. Vainberg. *Variational methods for the study of nonlinear operators*. Holden-Day, Inc., San Francisco, Calif.-London-Amsterdam, 1964. With a chapter on Newton's method by L. V. Kantorovich and G. P. Akilov. Translated and supplemented by Amiel Feinstein.
117. W. van Ackooij and R. Henrion. (Sub-)gradient formulae for probability functions of random inequality systems under Gaussian distribution. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):63–87, 2017.
118. B. van den Bosch and J. H. Seinfeld. History matching in two-phase petroleum reserviors: Incompressible flow. *Society of Petroleum Engineers*, 17(6), 1977.
119. G. van Essen, M. Zandvliet, P. van den Hof, O. Bosgra, and J. D. Jansen. Robust waterflooding optimization of multiple geological scenarios. *Society of Petroleum Engineers*, 14(1), 2009.
120. J. E. Warner, M. D. Grigoriu, and W. Aquino. Stochastic reduced order models for random vectors: Application to random eigenvalue problems. *Probabilistic Engineering Mechanics*, 31:1–11, 2013.
121. W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.
122. D. Xiu and G. E. Karniadakis. Modeling uncertainty in flow simulations via generalized polynomial chaos. *J. Comput. Phys.*, 187(1):137–167, 2003.