Noname manuscript No.
(will be inserted by the editor)

A Primal-Dual Algorithm for Risk Minimization

Drew P. Kouri · Thomas M. Surowiec

Received: date / Accepted: date

Abstract In this paper, we develop an algorithm to efficiently solve risk-averse optimization problems posed in reflexive Banach space. Such problems often arise in many practical applications as, e.g., optimization problems constrained by partial differential equations with uncertain inputs. Unfortunately, for many popular risk models including the coherent risk measures, the resulting risk-averse objective function is nonsmooth. This lack of differentiability complicates the numerical approximation of the objective function as well as the numerical solution of the optimization problem. To address these challenges, we propose a primal-dual algorithm for solving large-scale nonsmooth risk-averse optimization problems. This algorithm is motivated by the classical method of multipliers and by epigraphical regularization of risk measures. As a result, the algorithm solves a sequence of smooth optimization problems using derivative-based methods. We prove convergence of the algorithm even when the subproblems are solved inexactly and conclude with numerical examples demonstrating the efficiency of our method.

Drew P. Kouri
Optimization and Uncertainty Quantification, MS-1320, Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185-1320
E-mail: dpkouri@sandia.gov
DPK's research was sponsored by DARPA EQUiPS grant SNL 014150709, AFOSR grant F4FGA09135G001 and Sandia National Laboratories LDRD "Risk-Adaptive Experimental Design for High-Consequence Systems".
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 Energys 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.

Thomas M. Surowiec
FB12 Mathematik und Informatik, Philipps-Universität Marburg, Hans-Meerwein-Straße 6, Marburg, Germany D-35032
E-mail: surowiec@mathematik.uni-marburg.de
TMS's research was sponsored by the DFG grant no. SU 963/1-1 "Generalized Nash Equilibrium Problems with Partial Differential Operators: Theory, Algorithms, and Risk Aversion".

Keywords Risk-Averse Optimization, Coherent Risk Measures, Stochastic Optimization, Method of Multipliers

Mathematics Subject Classification (2010) 49M29, 49M37, 65K10, 90C15, 93E20

1 Introduction.

A common approach to modeling risk preference in decision theory, operations research, and stochastic programming is to employ risk measures. This often results in optimization problems of the form

$$\min_{z\in\mathcal{Z}_{\rm ad}} \{f(z) + \mathcal{R}(F(z))\}, \tag{1}$$

where $z \in \mathcal{Z}_{\rm ad}$ is a feasible decision, $f(z)$ is a deterministic cost, $F(z)$ is an uncertain cost, and $\mathcal{R}$ is a functional used to model the decision maker's risk profile, i.e., a measure of risk. However, many choices of $\mathcal{R}$ that are intuitively appealing to practitioners, e.g., mean-plus-semideviation, average value-at-risk, and (buffered) probabilities, are inherently nonsmooth. As a result, it is common to use nonsmooth optimization methods including, e.g., subgradient methods [19,48], stochastic approximation [33,35,36], stochastic mirror descent [15,29], stochastic quasigradient methods [14], or bundle methods [8,31] to solve (1). On the other hand, one could employ a smoothing approach, in which they replace $\mathcal{R}$ with a smooth approximation $\mathcal{R}_\varepsilon$ and solve (1) using existing optimization solvers for smooth problems [24,26].

When F is nonconvex, empirical evidence suggests that existing nonsmooth optimization methods result in large iteration counts [24,26]. This issue is especially important when the evaluation of $F(z)$ is computationally expensive. For example, this is the case for optimization problems constrained by nonlinear partial differential equations (PDE) with uncertain inputs; cf. the discussion and numerical examples in [21,22,24,26,25]. In contrast, the smoothing approaches developed in [24,26] perform well for fixed smoothing parameters ε. However, the problems become increasingly hard to solve as ε approaches zero and an analytical parameter-update strategy analogous to [18] remains elusive for the general case.

To remedy these issues, we propose a primal-dual algorithm inspired by the classical method of multipliers [17,37]. See [32] for a recent analysis of primal-dual methods in deterministic finite-dimensional optimization and [45] for a primal-dual method to handle constraints in risk neutral optimization. Unlike these previous works, our algorithm is applicable to both finite and infinite dimensional stochastic programs such as problems constrained by PDEs with uncertain inputs. Our approach for proving convergence utilizes the theory of epi-regularized risk measures introduced in [26] and permits inexact subproblems solves (i.e., inexact minimizers or stationary points).

The subsequent sections are structured as follows. In Section 2, we provide the problem formulation and the necessary data assumptions. We then introduce the primal-dual algorithm in Section 3. In Section 4, we prove convergence results for ϵ-minimizers and ϵ-stationary points. Under additional assumptions, we also prove convergence of the sequence of multipliers and boundedness of the iterates. In Section 5, we demonstrate the utility of the algorithm first by applying it to several

widely used risk measures and afterwards, to general coherent risk measures. In Section 6, we suggest a practical stopping criterion along with a parameter update strategy. We then demonstrate the performance of the algorithm in Section 7, where we solve multiple risk-averse PDE-constrained optimization problems using five different risk measures. Finally, in Section 8, we discuss the results and suggest future directions.

2 Problem Formulation and Assumptions.

Depending on the choice of $\mathcal{R}$, one can often reformulate (1) as an optimization problem over a larger space (see, e.g., the optimized certainty equivalents [5], the polyhedral [13] and extended polyhedral risk measures [16], and the risk quadrangle [42]). As such, we will focus on the efficient solution of the general minimization problem

$$\min_{x \in \mathcal{X}_{\text{ad}}} \{g(x) + \Phi(G(x))\} \tag{P}$$

where g is a deterministic objective function, G is an uncertain objective function, and Φ is a functional that maps random variables into the real numbers.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space where Ω is the set of outcomes, $\mathcal{F} \subseteq 2^\Omega$ is a σ-algebra of events and $\mathbb{P} : \mathcal{F} \to [0,1]$ is a probability measure. We denote the expectation of a random variable $Y : \Omega \to \mathbb{R}$ by $\mathbb{E}[Y] := \int_\Omega Y(\omega)\, \mathrm{d}\mathbb{P}(\omega)$, we denote "almost surely with respect to $\mathbb{P}$" simply by a.s., and we denote the space of random variables with finite second moment (defined up to a $\mathbb{P}$-null set) by

$$\mathcal{Y} := L^2(\Omega, \mathcal{F}, \mathbb{P}).$$

We make the following assumptions on the problem data in (P).

Assumption 1 *The optimization space $\mathcal{X}$ is a reflexive Banach space and the feasible set $\mathcal{X}_{\text{ad}} \subseteq \mathcal{X}$ is nonempty, convex and closed.*

Assumption 2 *The deterministic objective function $g : \mathcal{X} \to \mathbb{R}$ is weakly lower semicontinuous and the uncertain objective function $G : \mathcal{X} \to \mathcal{Y}$ is completely continuous, i.e.,*

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

Here, "$\rightharpoonup$" denotes convergence with respect to the weak topology and "$\to$" denotes convergence with respect to the strong topology.

Assumption 3 *The functional $\Phi : \mathcal{Y} \to \mathbb{R}$ is convex, positively homogeneous and monotonic with respect to the pointwise ordering on $\mathcal{Y}$, i.e., for X, $X' \in \mathcal{Y}$*

$$X \leq X' \text{ a.s.} \quad \Longrightarrow \quad \Phi(X) \leq \Phi(X').$$

As a consequence of Assumption 3 and the Fenchel-Moreau Theorem [47, Th. 6.5], there exists a nonempty, convex, closed and bounded set $\mathfrak{A} \subseteq \{\theta \in \mathcal{Y} \,|\, \theta \geq 0 \ \text{ a.s.}\}$ such that

$$\Phi(X) = \sup_{\theta \in \mathfrak{A}} \mathbb{E}[\theta X] \quad \forall\, X \in \mathcal{Y}. \tag{2}$$

In fact, Φ is continuous and subdifferentiable [47, Prop. 6.6], and $\mathfrak{A} = \partial\Phi(0)$. Given this relationship, we can equivalently formulate the minimization problem (P) as the minimax problem

$$\min_{x\in\mathcal{X}_{\rm ad}} \sup_{\theta\in\mathfrak{A}} \ell(x,\theta) \quad \text{where} \quad \ell(x,\theta) := g(x) + \mathbb{E}[\theta G(x)]. \tag{3}$$

When $\ell(\cdot,\theta)$ is convex, one can solve (3) using stochastic approximation methods as in [33]. However, the goal of this paper is to develop an algorithm that can exploit additional regularity of g and G in (3).

The following assumption ensures existence of minimizers of (P).

Assumption 4 *There exists a constant $\gamma \in \mathbb{R}$ such that the lower γ-level set of* (P)*,*

$$\mathrm{lev}_\gamma\,(\mathrm{P}) := \{x \in \mathcal{X} \mid g(x) + \Phi(G(x)) \leq \gamma\} \cap \mathcal{X}_{\rm ad}, \tag{4}$$

is nonempty and bounded.

As a consequence of Assumptions 1–4, the minimization problem (P) has a solution (cf. the Direct Method of the Calculus of Variations [3]).

Remark 1 The γ-level set in (4) is bounded if, e.g., $\mathcal{X}_{\rm ad}$ is bounded or if $\{g(x) + \Phi(G(x))\}$ is coercive. Moreover, $\{g(x) + \Phi(G(x))\}$ is coercive if, e.g., g is coercive and there exists a random variable $G_0 \in \mathcal{Y}$ such that $G(x) \geq G_0$ a.s. for all $x \in \mathcal{X}_{\rm ad}$. In this case, the monotonicity of Φ ensures that

$$g(x) + \Phi(G(x)) \geq g(x) + \Phi(G_0) \quad \forall\, x \in \mathcal{X}_{\rm ad},$$

the right-hand side of which is coercive.

3 Primal-Dual Algorithm for Risk Minimization.

In this section, we develop a primal-dual algorithm motivated by the classical method of multipliers [17,37]. Using the minimax formulation (3) as a guide, we define the generalized augmented Lagrangian functional as

$$L(x,\lambda,r) := \max_{\theta\in\mathfrak{A}} \left\{ \ell(x,\theta) - \frac{1}{2r}\mathbb{E}[(\lambda-\theta)^2] \right\} \tag{5}$$

(see [39] for details on the dual formulation of the classical augmented Lagrangian). As a result of the Fenchel conjugate of the infimal convolution [3, Th. 9.4.1] and the Fenchel-Moreau Theorem [3, Th. 9.3.2], $L(x,\lambda,r)$ can be rewritten as

$$\begin{aligned} L(x,\lambda,r) &= g(x) + \max_{\theta\in\mathfrak{A}} \left\{ \mathbb{E}[\theta G(x)] - \frac{1}{2r}\mathbb{E}[(\lambda-\theta)^2] \right\} \\ &= g(x) + \min_{Y\in\mathcal{Y}} \left\{ \Phi(G(x)-Y) + \mathbb{E}[\lambda Y] + \frac{r}{2}\mathbb{E}[Y^2] \right\}. \end{aligned} \tag{6}$$

That is, the augmented Lagrangian is merely the objective function in (P) where Φ is replaced by the epi-regularization of Φ [26] with the regularizer $\Psi_{r,\lambda}(Y) = \mathbb{E}[\lambda Y] + \frac{r}{2}\mathbb{E}[Y^2]$, i.e., $\Phi(X) \approx \widehat{\Phi}(X,\lambda,r)$ where

$$\widehat{\Phi}(X,\lambda,r) := \min_{Y\in\mathcal{Y}} \left\{ \Phi(X-Y) + \mathbb{E}[\lambda Y] + \frac{r}{2}\mathbb{E}[Y^2] \right\}.$$

Notice that the first equality in (6) is the dual form of the epi-regularized risk measure $\widehat{\Phi}(\cdot, \lambda, r)$ (see [26, Eq. 5]).

As a consequence of Propositions 1 and 2 in [26], we have the following bounds

$$\Phi(X) - \frac{K^2}{r} \leq \widehat{\Phi}(X, \lambda, r) \leq \Phi(X) \quad \forall\, \lambda \in \mathfrak{A},\ r > 0 \tag{7}$$

where $K > 0$ is the smallest scalar such that $\|\theta\|_{\mathcal{Y}} \leq K$ for all $\theta \in \mathfrak{A}$ (i.e., K is the Lipschitz modulus of Φ at 0). Note that the lower bound in (7) follows from the bound

$$\inf_{\theta \in \partial\Phi(X)} \frac{1}{2}\mathbb{E}[(\theta - \lambda)^2] \leq \inf_{\theta \in \partial\Phi(X)} \frac{1}{2}(\|\theta\|_{\mathcal{Y}}^2 + \|\lambda\|_{\mathcal{Y}}^2) \leq K^2 \tag{8}$$

where the first inequality holds because θ, $\lambda \geq 0$ a.s. (see Proposition 2 in [26]). One can often tighten this bound for specific choices of Φ as we will discuss later. Additionally, $\widehat{\Phi}(\cdot, \lambda, r)$ is convex, monotonic and continuously Fréchet differentiable (see Section 3 and Corollary 2 in [26]). From (6), it is easy to see that $L(x, \cdot, r)$ is concave since $L(x, \lambda, r)$ is the infimum over $Y \in \mathcal{Y}$ of a function that is linearly parametrized by λ. Furthermore, the maximum in (5) is uniquely attained at

$$\Lambda(x, \lambda, r) := \mathbf{P}_{\mathfrak{A}}(rG(x) + \lambda) \tag{9}$$

where $\mathbf{P}_{\mathfrak{A}} : \mathcal{Y} \to \mathcal{Y}$ is the projection onto the convex set $\mathfrak{A}$, i.e.,

$$\|\lambda - \mathbf{P}_{\mathfrak{A}}(\lambda)\|_{\mathcal{Y}} = \min_{\theta \in \mathfrak{A}} \|\lambda - \theta\|_{\mathcal{Y}}.$$

In fact, $\Lambda(x, \lambda, r)$ is the Fréchet derivative of $\widehat{\Phi}(\cdot, \lambda, r)$ at $G(x)$ [26, Th. 2]. Substituting $\Lambda(x, \lambda, r)$ in the right-hand side of (5), adding and subtracting $rG(x)$ in the quadratic term, and then expanding the quadratic term gives

$$\begin{aligned} L(x, \lambda, r) &= \ell(x, \Lambda(x, \lambda, r)) - \frac{1}{2r}\mathbb{E}[(\lambda - \Lambda(x, \lambda, r))^2] \\ &= g(x) + \mathbb{E}[\lambda G(x)] + \frac{r}{2}\mathbb{E}[G(x)^2] - \frac{1}{2r}\mathbb{E}[\{(\mathrm{Id} - \mathbf{P}_{\mathfrak{A}})(rG(x) + \lambda)\}^2] \end{aligned}$$

where Id denotes the identity operator on $\mathcal{Y}$. This equivalent form relates $L(x, \lambda, r)$ to the traditional augmented Lagrangian. For example, in equality constrained problems, $\mathfrak{A} = \mathcal{Y}$ (i.e., the multipliers are unconstrained) and the final term in $L(x, \lambda, r)$ is zero.

To conclude, we note that $L(x, \cdot, r)$ is continuously Fréchet differentiable with gradient

$$\nabla_\lambda L(x, \lambda, r) = (\Lambda(x, \lambda, r) - \lambda)/r. \tag{10}$$

To see this let $t > 0$ and fix $\delta\lambda \in \mathcal{Y}$. It follows from the definition of L that

$$\begin{aligned} \frac{t}{r}\mathbb{E}[(\Lambda(x, \lambda, r) - \lambda)\delta\lambda] - \frac{t^2}{r}\|\delta\lambda\|_{\mathcal{Y}}^2 &\leq L(x, \lambda + t\delta\lambda, r) - L(x, \lambda, r) \\ &\leq \frac{t}{r}\mathbb{E}[(\Lambda(x, \lambda + t\delta\lambda, r) - \lambda)\delta\lambda] - \frac{t^2}{r}\|\delta\lambda\|_{\mathcal{Y}}^2. \end{aligned}$$

Combining the above bounds with the continuity of $\mathbf{P}_{\mathfrak{A}}$ [4, Prop. 4.8] ensures that $L(x, \cdot, r)$ is Gâteaux differentiable with continuous derivative (10). Therefore, as a

consequence of the Mean Value Theorem, $L(x,\cdot,r)$ is continuously Fréchet differentiable (cf. [9, Pg. 35-36]). Based on these properties of L, we have the following generalization of the classical method of multipliers listed as Algorithm 1. As written, Algorithm 1 is not implementable since we have not specified what "not converged" and "approximately minimizes" mean or how to update the penalty parameter r_k. In addition, we have not discussed how to approximate the expectation in the definition of L. We address these issues in Section 6 and provide a concrete specification of Algorithm 1 based on the forthcoming analysis.

Algorithm 1 Primal-Dual Risk Minimization

Initialize: Given $x_0 \in \mathcal{X}_{\rm ad}$, $r_0 > 0$ and $\lambda_0 \in \mathfrak{A}$.
While("Not Converged")

1. Compute $x_{k+1} \in \mathcal{X}_{\rm ad}$ that approximately minimizes $L(\cdot,\lambda_k,r_k)$.
2. Update the dual variable $\lambda_{k+1} = \Lambda(x_{k+1},\lambda_k,r_k)$ using (9).
3. Update the penalty parameter $r_{k+1} > 0$.

End While

We emphasize that Algorithm 1 solves a sequence of smooth approximations to the optimization problem (P) and that each smooth approximation is generated by a specific epi-regularized risk measure [26], which depends on the dual information λ_k and the penalty parameter r_k. In contrast, the methods described in [24,26], choose a smooth approximation of the risk measure with fixed smoothing parameter $\varepsilon > 0$ and solve the resulting approximation of (P). As shown in Theorem 6 of [26], under certain assumptions, this approach yields an approximation to the optimal solution with error that is of order $\sqrt{\varepsilon}$. To ensure that the computed solution is sufficiently accurate, one must choose ε sufficiently small, which often leads to ill-conditioning as the smooth approximate risk measure converges to the original nonsmooth risk measure. As a result, the approximate optimization problem becomes increasingly difficult to solve, warranting continuation on ε. To this end, Algorithm 1 can be viewed as a systematic and rigorous continuation procedure that does not simply modify the smoothing parameter (i.e., $\varepsilon = 1/r_k$), but also adapts the dual variables λ_k.

4 Convergence Analysis.

In this section, we prove multiple convergence results for Algorithm 1. For the first result, we require the following terminology. An ϵ-minimizer, $\epsilon \geq 0$, of (P) is any $x \in \mathcal{X}_{\rm ad}$ such that

$$g(x) + \Phi(G(x)) - \epsilon \leq \min_{y\in\mathcal{X}_{\rm ad}} \{g(y) + \Phi(G(y))\}.$$

Similarly, an ϵ-minimizer, $\epsilon \geq 0$, of $L(\cdot,\lambda,r)$ over $\mathcal{X}_{\rm ad}$ is any $x \in \mathcal{X}_{\rm ad}$ such that

$$L(x,\lambda,r) - \epsilon \leq \inf_{y\in\mathcal{X}_{\rm ad}} L(y,\lambda,r).$$

Theorem 1 *Consider the optimization problem* (P) *and let Assumptions 1–4 hold. Let* $\{(x_k, \lambda_k, r_k)\} \subset \mathcal{X}_{\text{ad}} \times \mathcal{Y} \times (0, \infty)$ *denote the sequence of iterates produced by Algorithm 1 with* $r_k \to r^\star > 0$ *(possibly infinity) and* x_k *satisfies*

$$L(x_k, \lambda_{k-1}, r_{k-1}) - \epsilon_k \leq \inf_{x \in \mathcal{X}_{\text{ad}}} L(x, \lambda_{k-1}, r_{k-1}) \tag{11}$$

for some sequence $\{\epsilon_k\} \subset [0, \infty)$ *with* $\epsilon_k \to \epsilon^\star$ *(possibly zero). Then, any weak accumulation point* $x^\star$ *of* $\{x_k\}$ *satisfies*

$$g(x^\star) + \Phi(G(x^\star)) - \left(\tfrac{K^2}{r^\star} + \epsilon^\star\right) \leq \min_{x \in \mathcal{X}_{\text{ad}}} \{g(x) + \Phi(G(x))\}$$

where $K > 0$ *is such that* $\|\theta\|_{\mathcal{Y}} \leq K$ *for all* $\theta \in \mathfrak{A}$*. That is,* $x^\star$ *is a* $(\frac{K^2}{r^\star} + \epsilon^\star)$*-optimal solution.*

Proof Suppose $x^\star$ is a weak accumulation point of $\{x_k\}$, then there exists a subsequence $\{x_{k_j}\}$ such that $x_{k_j} \rightharpoonup x^\star$ in $\mathcal{X}_{\text{ad}}$. Combining (7), the continuity of Φ, the weak lower semicontinuity of g, and the complete continuity of G yields

$$\begin{aligned} g(x) + \Phi(G(x)) &\geq \liminf_{k_j \to \infty} L(x, \lambda_{k_j - 1}, r_{k_j - 1}) \\ &\geq \liminf_{k_j \to \infty} \{L(x_{k_j}, \lambda_{k_j - 1}, r_{k_j - 1}) - \epsilon_{k_j - 1}\} \\ &\geq \liminf_{k_j \to \infty} \{g(x_{k_j}) + \Phi(G(x_{k_j})) - \tfrac{K^2}{r_{k_j - 1}} - \epsilon_{k_j - 1}\} \\ &\geq g(x^\star) + \Phi(G(x^\star)) - \left(\tfrac{K^2}{r^\star} + \epsilon^\star\right) \end{aligned}$$

for all $x \in \mathcal{X}_{\text{ad}}$ as desired. □

To apply Theorem 1, we must determine if an accumulation point of $\{x_k\}$ exists. In the next result, we provide sufficient conditions on (P) that ensure that the sequence $\{x_k\}$ is bounded.

Proposition 1 *Let the assumptions of Theorem 1 hold and suppose there exists* $\gamma^\star > 0$ *satisfying*

$$\gamma^\star > \left(\min_{x \in \mathcal{X}_{\text{ad}}} \{g(x) + \Phi(G(x))\}\right) + \frac{K^2}{r^\star} + \epsilon^\star$$

such that the $\gamma^\star$*-lower level set of* (P), $\text{lev}_{\gamma^\star}$ (P) *in* (4)*, is bounded. Then, the sequence of iterates* $\{x_k\}$ *produced by Algorithm 1 with stopping condition* (11) *is bounded.*

Proof By (7) and the ϵ_k-optimality of x_k, we have that

$$\begin{aligned} g(x) + \Phi(G(x)) &\geq L(x, \lambda_k, r_k) \geq L(x_{k+1}, \lambda_k, r_k) - \epsilon_k \\ &\geq g(x_{k+1}) + \Phi(G(x_{k+1})) - \epsilon_k - \frac{K^2}{r_k} \end{aligned}$$

for any $x \in \mathcal{X}_{\text{ad}}$. Minimizing the upper bound over $x \in \mathcal{X}_{\text{ad}}$ then ensures that $x_k \in \text{lev}_{\gamma^\star}$ (P) for sufficiently large k. Since $\text{lev}_{\gamma^\star}$ (P) is bounded, so is $\{x_k\}$. □

In addition to the convergence of the primal variables $\{x_k\}$, it is often important to characterize the convergence of the dual variables $\{\lambda_k\}$. Before doing this, we prove the following important properties of the dual objective function. For this result, we define the set $\mathfrak{S} \subseteq \mathcal{Y}$ by

$$\mathfrak{S} := \left\{ \theta \in \mathcal{Y} \;\middle|\; \operatorname*{arg\,min}_{x \in \mathcal{X}_{\rm ad}} \ell(x, \theta) \neq \emptyset \right\}$$

and we employ the notation $\delta_{\mathfrak{A}} : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ to denote the indicator function of the nonempty, closed and convex set $\mathfrak{A}$ (i.e., $\delta_{\mathfrak{A}}(\lambda) = 0$ if $\lambda \in \mathfrak{A}$ and $\delta_{\mathfrak{A}}(\lambda) = +\infty$ if $\lambda \notin \mathfrak{A}$).

Lemma 1 *Let Assumptions 1–4 hold and consider the dual problem associated with* (P)*, i.e.,*

$$\max_{\theta \in \mathfrak{A}} v(\theta) \quad \textit{where} \quad v(\theta) := \inf_{x \in \mathcal{X}_{\rm ad}} \ell(x, \theta). \tag{D}$$

The dual objective function v in (D) *is concave and upper semicontinuous. In addition, $-v$ is subdifferentiable for all $\theta \in \mathfrak{S}$. Finally, if there exists $\lambda_0 \in \mathfrak{A}$ such that $v(\lambda_0) > -\infty$, then the subdifferential $\partial(-v + \delta_{\mathfrak{A}})$ is maximal monotone.*

Proof We first note that v is concave since it is the minimum value of $\ell(\cdot, \lambda)$ over the convex set $\mathcal{X}_{\rm ad}$ and $\ell(x, \cdot)$ is affine. To prove that v is upper semicontinuous, let $\lambda_k \rightharpoonup \lambda$ in $\mathcal{Y}$, then by definition of v we have that

$$v(\lambda_k) \leq \ell(x, \lambda_k) = \ell(x, \lambda) + \mathbb{E}[(\lambda_k - \lambda) G(x)] \quad \forall\, x \in \mathcal{X}_{\rm ad}.$$

Fixing $x \in \mathcal{X}_{\rm ad}$ and passing to the upper limit in the above inequality yields

$$\limsup_{k \to \infty} v(\lambda_k) \leq \ell(x, \lambda) \quad \forall\, x \in \mathcal{X}_{\rm ad}.$$

Taking the infimum on the right-hand side over $x \in \mathcal{X}_{\rm ad}$ proves the desired result. To prove subdifferentiability of $-v$, let $\theta \in \mathfrak{S}$, $\eta \in \mathcal{Y}$, and $x_\theta \in \mathcal{X}_{\rm ad}$ be a minimizer of $\ell(\cdot, \theta)$ in $\mathcal{X}_{\rm ad}$. By the definition of v, we have that

$$v(\eta) - v(\theta) \leq g(x_\theta) + \mathbb{E}[\eta G(x_\theta)] - g(x_\theta) - \mathbb{E}[\theta G(x_\theta)] = \mathbb{E}[(\eta - \theta) G(x_\theta)].$$

The previous inequality clearly holds for any minimizer x_θ and all $\eta \in \mathcal{Y}$. Hence,

$$-G(x_\theta) \in \partial(-v)(\theta) \quad \forall\, x_\theta \in \operatorname*{arg\,min}_{x \in \mathcal{X}_{\rm ad}} \ell(x, \theta).$$

To conclude, since $-v + \delta_{\mathfrak{A}}$ is proper (i.e., $-v(\lambda_0) + \delta_{\mathfrak{A}}(\lambda_0) = -v(\lambda_0) < +\infty$), convex and lower semicontinuous, its subdifferential is maximal monotone by Theorem A in [38]. □

Remark 2 Under the properties discussed in Remark 1, $v(\lambda) > -\infty$ for all $\lambda \in \mathcal{Y}$ with $\lambda \geq 0$ a.s. (in particular, $\lambda \in \mathfrak{A}$). That is, $\ell(\cdot, \lambda)$ is weakly lower semicontinuous and either $\mathcal{X}_{\rm ad}$ is bounded or $\ell(\cdot, \lambda)$ is coercive since $\ell(x, \lambda) \geq g(x) + \mathbb{E}[\lambda G_0]$ for all $x \in \mathcal{X}$ and $\lambda \geq 0$ a.s. Therefore, a minimizer of $\ell(\cdot, \lambda)$ exists in $\mathcal{X}_{\rm ad}$ for any $\lambda \in \mathcal{Y}$ with $\lambda \geq 0$ a.s. Notably, $\mathfrak{A} \subseteq \mathfrak{S}$.

The subsequent result was motivated by Proposition 6 in [39] and demonstrates that Algorithm 1 produces a weakly convergent sequence of dual variables.

Theorem 2 *Let the assumptions of Lemma 1 hold and suppose Algorithm 1 is executed with* $\{r_k\} \subset [r, \infty)$ *for some* $r > 0$ *and* $x_{k+1} \in \mathcal{X}_{\mathrm{ad}}$ *satisfying*

$$L(x_{k+1}, \lambda_k, r_k) - \frac{\epsilon_k^2}{2r_k} \leq \inf_{x \in \mathcal{X}_{\mathrm{ad}}} L(x, \lambda_k, r_k) \quad \text{with} \quad \sum_{k=0}^{\infty} \epsilon_k < \infty,\ \epsilon_k \geq 0. \tag{12}$$

Furthermore, assume that for each k

$$\begin{aligned} &\inf_{x \in \mathcal{X}_{\mathrm{ad}}} \sup_{\theta \in \mathfrak{A}} \left\{ \ell(x, \theta) - \frac{1}{2r_k} \mathbb{E}[(\theta - \eta)^2] \right\} \\ &\qquad = \sup_{\theta \in \mathfrak{A}} \inf_{x \in \mathcal{X}_{\mathrm{ad}}} \left\{ \ell(x, \theta) - \frac{1}{2r_k} \mathbb{E}[(\theta - \eta)^2] \right\} \quad \forall\, \eta \in \mathcal{Y}. \end{aligned} \tag{13}$$

Then the sequence $\{\lambda_k\}$ *converges weakly to* $\lambda^\star \in \mathfrak{A}$*, which is a solution to the dual problem* (D).

Proof As previously shown, the functional $L(x_{k+1}, \cdot, r_k)$ is continuously Fréchet differentiable at λ_k with gradient given by (10). Since $L(x_{k+1}, \cdot, r_k)$ is concave, the gradient in (10) satisfies the following inequality

$$L(x_{k+1}, \lambda_k, r_k) + \frac{1}{r_k} \mathbb{E}[(\lambda_{k+1} - \lambda_k)(\eta - \lambda_k)] \geq L(x_{k+1}, \eta, r_k)$$

for all $\eta \in \mathcal{Y}$. Moreover, by the definition of L, we have that

$$L(x_{k+1}, \eta, r_k) \geq \inf_{x \in \mathcal{X}_{\mathrm{ad}}} L(x, \eta, r_k) = \inf_{x \in \mathcal{X}_{\mathrm{ad}}} \sup_{\theta \in \mathfrak{A}} \left\{ \ell(x, \theta) - \frac{1}{2r_k} \mathbb{E}[(\theta - \eta)^2] \right\}$$

for all $\eta \in \mathcal{Y}$. Therefore, we obtain the lower bound

$$L(x_{k+1}, \eta, r_k) \geq \sup_{\theta \in \mathfrak{A}} \left\{ v(\theta) - \frac{1}{2r_k} \mathbb{E}[(\theta - \eta)^2] \right\}$$

for all $\eta \in \mathcal{Y}$. By Lemma 1, $\partial(-v + \delta_{\mathfrak{A}})$ is a maximal monotone operator and we can apply the proximal point algorithm [40] to maximize v over $\mathfrak{A}$. We denote the associated proximal point operator with penalty parameter r_k by

$$P_k(\eta) := \arg\max_{\theta \in \mathfrak{A}} \left\{ v(\theta) - \frac{1}{2r_k} \mathbb{E}[(\theta - \eta)^2] \right\}.$$

Notice that these definitions and the above estimates yield

$$L(x_{k+1}, \eta, r_k) \geq v(P_k(\lambda_k)) - \frac{1}{2r_k} \mathbb{E}[(P_k(\lambda_k) - \eta)^2] \quad \forall\, \eta \in \mathcal{Y}$$

and by (13), we have that

$$\inf_{x \in \mathcal{X}_{\mathrm{ad}}} L(x, \lambda_k, r_k) = v(P_k(\lambda_k)) - \frac{1}{2r_k} \mathbb{E}[(P_k(\lambda_k) - \lambda_k)^2].$$

Combining all of the above estimates then produces

$$\begin{aligned} &L(x_{k+1}, \lambda_k, r_k) - \inf_{x \in \mathcal{X}_{\mathrm{ad}}} L(x, \lambda_k, r_k) \\ &\quad \geq \frac{1}{2r_k} \mathbb{E}[(P_k(\lambda_k) - \lambda_k)^2 - (P_k(\lambda_k) - \eta)^2 - 2(\lambda_{k+1} - \lambda_k)(\eta - \lambda_k)] \end{aligned}$$

for all $\eta \in \mathcal{Y}$. The right-hand side above can be directly maximized over $\eta \in \mathcal{Y}$, which yields the tight lower bound

$$L(x_{k+1}, \lambda_k, r_k) - \inf_{x \in \mathcal{X}_{\rm ad}} L(x, \lambda_k, r_k) \geq \frac{1}{2r_k} \mathbb{E}[(P_k(\lambda_k) - \lambda_{k+1})^2].$$

The stopping condition (12) then ensures that

$$\mathbb{E}[(P_k(\lambda_k) - \lambda_{k+1})^2]^{\frac{1}{2}} \leq \epsilon_k \quad \text{with} \quad \sum_{k=0}^{\infty} \epsilon_k < \infty.$$

Therefore, Theorem 1 in [40] and the boundedness of $\mathfrak{A}$ ensure that λ_k converges weakly to a maximizer of the dual problem as desired. □

Remark 3 Equation (13) holds if, for each k and fixed $\eta \in \mathcal{Y}$, there exists a saddle point $(\bar{x}_k, \bar{\lambda}_k) \in \mathcal{X}_{\rm ad} \times \mathfrak{A}$, i.e., $(\bar{x}_k, \bar{\lambda}_k)$ satisfies

$$\begin{aligned} \ell(\bar{x}_k, \theta) - \frac{1}{2r_k}\mathbb{E}[(\theta - \eta)^2] &\leq \ell(\bar{x}_k, \bar{\lambda}_k) - \frac{1}{2r_k}\mathbb{E}[(\bar{\lambda}_k - \eta)^2] \qquad \forall\, \theta \in \mathfrak{A} \\ &\leq \ell(x, \bar{\lambda}_k) - \frac{1}{2r_k}\mathbb{E}[(\bar{\lambda}_k - \eta)^2] \qquad \forall\, x \in \mathcal{X}_{\rm ad}. \end{aligned}$$

On the other hand, if $\ell(\cdot, \theta)$ is quasiconvex for each $\theta \in \mathfrak{A}$, i.e., for all $\alpha \in \mathbb{R}$, the lower level sets $\{x \in \mathcal{X} \,|\, \ell(x, \theta) \leq \alpha\}$ are convex, then Sion's theorem [49] ensures that (13) holds. To justify this, we note that $\ell(x, \cdot) - \frac{1}{2r_k}\mathbb{E}[(\cdot - \eta)^2]$ is concave and weakly upper semicontinuous over the weakly compact convex set $\mathfrak{A}$ and $\ell(\cdot, \theta) - \frac{1}{2r_k}\mathbb{E}[(\theta - \eta)^2]$ is quasiconvex and weakly lower semicontinuous over the closed convex set $\mathcal{X}_{\rm ad}$.

For our final results, we assume that G and g are continuously Fréchet differentiable on some open set containing $\mathcal{X}_{\rm ad}$. In this case, for any $x \in \mathcal{X}_{\rm ad}$, $\lambda \in \mathfrak{A}$ and $r > 0$, we have that the objective function $L(\cdot, \lambda, r)$ is continuously Fréchet differentiable with derivative

$$L_x'(x, \lambda, r) = g'(x) + \mathbb{E}[\Lambda(x, \lambda, r) G'(x)]$$

(see, e.g., Theorem 2 and Corollary 2 in [26]). To define stationary and approximate stationary points, we recall the definitions of the normal and ϵ-normal, $\epsilon \geq 0$, cones,

$$\begin{aligned} N(x, \mathcal{X}_{\rm ad}) &:= \left\{ \theta \in \mathcal{X}^* \,|\, \langle \theta, y - x \rangle_{\mathcal{X}^*, \mathcal{X}} \leq 0 \quad \forall\, y \in \mathcal{X}_{\rm ad} \right\} \\ N_\epsilon(x, \mathcal{X}_{\rm ad}) &:= \left\{ \theta \in \mathcal{X}^* \,|\, \langle \theta, y - x \rangle_{\mathcal{X}^*, \mathcal{X}} \leq \epsilon \|y - x\|_{\mathcal{X}} \quad \forall\, y \in \mathcal{X}_{\rm ad} \right\}, \end{aligned}$$

respectively. We say that a point $x \in \mathcal{X}_{\rm ad}$ is a stationary point of the subproblem

$$\min_{x \in \mathcal{X}_{\rm ad}} L(x, \lambda, r) \tag{14}$$

if $-L_x'(x, \lambda, r) \in N(x, \mathcal{X}_{\rm ad})$, i.e.,

$$\langle g'(x) + \mathbb{E}[\Lambda(x, \lambda, r) G'(x)], y - x \rangle_{\mathcal{X}^*, \mathcal{X}} \geq 0 \quad \forall\, y \in \mathcal{X}_{\rm ad}.$$

Similarly, we say that $x \in \mathcal{X}_{\rm ad}$ is a stationary point of (P) if there exists $\theta \in \partial\Phi(G(x))$ such that

$$\langle g'(x) + \mathbb{E}[\theta G'(x)], y - x \rangle_{\mathcal{X}^*, \mathcal{X}} \geq 0 \quad \forall\, y \in \mathcal{X}_{\rm ad}.$$

Furthermore, we say that $x \in \mathcal{X}_{\text{ad}}$ is an ϵ-stationary point of (14), $\epsilon \geq 0$, if $-L'_x(x, \lambda, r) \in N_\epsilon(x, \mathcal{X}_{\text{ad}})$, i.e.,

$$\langle g'(x) + \mathbb{E}[\Lambda(x, \lambda, r)G'(x)], y - x\rangle_{\mathcal{X}^*,\mathcal{X}} \geq -\epsilon\|y - x\|_{\mathcal{X}} \quad \forall\, y \in \mathcal{X}_{\text{ad}}$$

and $x \in \mathcal{X}_{\text{ad}}$ is an ϵ-stationary point of (P) if there exists $\theta \in \partial\Phi(G(x))$ such that

$$\langle g'(x) + \mathbb{E}[\theta G'(x)], y - x\rangle_{\mathcal{X}^*,\mathcal{X}} \geq -\epsilon\|y - x\|_{\mathcal{X}} \quad \forall\, y \in \mathcal{X}_{\text{ad}}.$$

This brings us to our next result.

Theorem 3 *Let Assumption 1–4 hold and suppose G is Fréchet differentiable with completely continuous derivative G'. Suppose further that $g = g_1 + g_2$ where g_1 is convex, Fréchet differentiable with weak-to-weak* sequentially continuous derivative g_1', i.e.,*

$$x_k \rightharpoonup x \ \text{ in } \mathcal{X} \quad \Longrightarrow \quad g_1'(x_k) \rightharpoonup^* g_1'(x) \ \text{ in } \mathcal{X}^*,$$

and g_2 is Fréchet differentiable with completely continuous derivative g_2'. Moreover, suppose $\{(x_k, \lambda_k, r_k)\}$ is a sequence of iterates generated by Algorithm 1 such that x_k are ϵ_k-stationary points of $L(\cdot, \lambda_{k-1}, r_{k-1})$ over $\mathcal{X}_{\text{ad}}$ with $\{\epsilon_k\} \subset [0, \infty)$, $\epsilon_k \to 0$, and $\{r_k\} \subset (0, \infty)$ and $r_k \to \infty$. Then any weak accumulation point of $\{x_k\}$ is a stationary point of (P).

Proof Let $x^\star \in \mathcal{X}_{\text{ad}}$ be a weak accumulation point and let $\{x_{k_j}\}$ denote a subsequence of $\{x_k\}$ that converges weakly to $x^\star$. The complete continuity of G ensures that $G(x_{k_j}) \to G(x^\star)$ in $\mathcal{Y}$. Moreover, since $\mathfrak{A}$ is bounded, we have that there exists a weakly converging subsequence $\{\lambda_{k_{j_l}}\}$ with weak limit $\lambda^\star \in \mathcal{Y}$. Propositions 1 and 2 in [26] then ensure that

$$\begin{aligned}\Phi(Y) \geq \widehat{\Phi}(Y, \lambda_{k_{j_l}-1}, r_{k_{j_l}-1}) &\geq \widehat{\Phi}(G(x_{k_{j_l}}), \lambda_{k_{j_l}-1}, r_{k_{j_l}-1}) + \mathbb{E}[\lambda_{k_{j_l}}(Y - G(x_{k_{j_l}}))] \\ &\geq \Phi(G(x_{k_{j_l}})) - \frac{K^2}{r_{k_{j_l}-1}} + \mathbb{E}[\lambda_{k_{j_l}}(Y - G(x_{k_{j_l}}))]\end{aligned}$$

for all $Y \in \mathcal{Y}$. Passing to the limit, then ensures that $\lambda^\star \in \partial\Phi(G(x^\star))$. Combining these facts with the stated assumptions on g and G ensure that

$$\begin{aligned}0 &= \lim_{k_{j_l}\to\infty} -\epsilon_{k_{j_l}}\|x - x_{k_{j_l}}\|_{\mathcal{X}} \\ &\leq \liminf_{k_{j_l}\to\infty} \langle g'(x_{k_{j_l}}) + \mathbb{E}[\lambda_{k_{j_l}} G'(x_{k_j})], x - x_{k_{j_l}}\rangle_{\mathcal{X}^*,\mathcal{X}} \\ &= \langle g'(x^\star) + \mathbb{E}[\lambda^\star G'(x^\star)], x - x^\star\rangle_{\mathcal{X}^*,\mathcal{X}} \\ &\quad + \liminf_{k_j\to\infty} \langle (g_1'(x_{k_{j_l}}) - g_1'(x^\star)), x - x_{k_{j_l}}\rangle_{\mathcal{X}^*,\mathcal{X}} \\ &\leq \langle g'(x^\star) + \mathbb{E}[\lambda^\star G'(x^\star)], x - x^\star\rangle_{\mathcal{X}^*,\mathcal{X}} \quad \forall\, x \in \mathcal{X}_{\text{ad}}.\end{aligned}$$

The first equality follows since $\{x_{k_{j_l}}\}$ converges weakly and hence $\|x_{k_{j_l}}\|_{\mathcal{X}}$ is bounded. The first inequality follows since $\{x_{k_{j_l}}\}$ are $\epsilon_{k_{j_l}}$-stationary points. The

second equality follows from the complete continuity of g_2' and G'. The final inequality follows from the convexity of g_1 and the weak-to-weak* continuity of g_1'. That is, since g_1 is convex, g_1' is maximal monotone. Therefore,

$$\begin{aligned}\langle g_1'(x_{k_{j_l}}) - g_1'(x^\star), x - x_{k_{j_l}}\rangle_{\mathcal{X}^*,\mathcal{X}} &= \langle g_1'(x_{k_{j_l}}) - g_1'(x^\star), x - x^\star\rangle_{\mathcal{X}^*,\mathcal{X}} \\ &\quad - \langle g_1'(x_{k_{j_l}}) - g_1'(x^\star), x_{k_{j_l}} - x^\star\rangle_{\mathcal{X}^*,\mathcal{X}} \\ &\leq \langle g_1'(x_{k_{j_l}}) - g_1'(x^\star), x - x^\star\rangle_{\mathcal{X}^*,\mathcal{X}}\end{aligned}$$

for all $x \in \mathcal{X}_{\rm ad}$. The right-hand side of the upper bound converges to zero due to the weak-to-weak* continuity of g_1'. Hence, $x^\star$ is a stationary point of (P). □

Our final result of this section provides assumptions that ensure that the sequence of ϵ-stationary points in Theorem 3 is bounded.

Proposition 2 *Let the assumptions of Proposition 1 hold and suppose Algorithm 1 produces iterates $\{x_k\}$ that are ϵ_k-stationary points of $L(\cdot,\lambda_{k-1},r_{k-1})$ over $\mathcal{X}_{\rm ad}$ with $\{\epsilon_k\} \subseteq [0,\infty)$, $\epsilon_k \to \epsilon^\star \geq 0$, $\{r_k\} \subseteq (0,\infty)$, and $r_k \to r^\star > 0$. Moreover, assume for each $k = 1,2,\ldots$ that there exists a minimizer $x_k^\star$ of $L(\cdot,\lambda_{k-1},r_{k-1})$ over $\mathcal{X}_{\rm ad}$ and a constant $c_k \geq c_0 > 0$ satisfying*

$$\langle L_x'(x_k^\star,\lambda_{k-1},r_{k-1}) - L_x'(x_k,\lambda_{k-1},r_{k-1}), x_k^\star - x_k\rangle_{\mathcal{X}} \geq c_k\|x_k^\star - x_k\|^2. \tag{15}$$

Then, the sequence of iterates $\{x_k\}$ is bounded.

Proof Since $x_k \in \mathcal{X}_{\rm ad}$ is an ϵ_k-stationary point and $x_k^\star \in \mathcal{X}_{\rm ad}$ is optimal, we have that

$$\langle L_x'(x_k,\lambda_{k-1},r_{k-1}), x_k^\star - x_k\rangle_{\mathcal{X}} \geq -\epsilon_k\|x_k^\star - x_k\|_{\mathcal{X}}$$

and

$$\langle L_x'(x_k^\star,\lambda_{k-1},r_{k-1}), x_k - x_k^\star\rangle_{\mathcal{X}} \geq 0.$$

Combining these two inequalities and applying (15) gives

$$\begin{aligned}\epsilon_k\|x_k^\star - x_k\|_{\mathcal{X}} &\geq \langle L_x'(x_k^\star,\lambda_{k-1},r_{k-1}) - L_x'(x_k,\lambda_{k-1},r_{k-1}), x_k^\star - x_k\rangle_{\mathcal{X}} \\ &\geq c_k\|x_k^\star - x_k\|_{\mathcal{X}}^2.\end{aligned}$$

By the arguments in the proof of Proposition 1, $\{x_k^\star\}$ is bounded. Therefore, for fixed $\epsilon > 0$,

$$c_0\|x_k^\star - x_k\|_{\mathcal{X}} \leq \epsilon^\star + \epsilon$$

for sufficiently large k. Hence, the sequence of iterates $\{x_k\}$ is bounded. □

Remark 4 Note that the bound (15) holds if G is convex with respect to the pointwise ordering on $\mathcal{Y}$ and g is strongly convex. In this case, $\widehat{\Phi}(G(x),\lambda,r)$ is convex in x since $\widehat{\Phi}(\cdot,\lambda,r)$ is convex and monotonic (see Section 3 in [26]). Therefore, $L(\cdot,\lambda,r)$ is strongly convex.

5 Applications of Algorithm 1.

In this section, we discuss Algorithm 1 for concrete choices of Φ. We begin by describing the algorithm for $\Phi(X) = \mathbb{E}[(X)_+]$ where $(x)_+ := \max\{0, x\}$. This choice of Φ is sufficient to represent at least four common risk measures: the mean-plus-semideviation of order 1

$$\mathcal{R}(X) = \mathbb{E}[X] + c\mathbb{E}[(X - \mathbb{E}[X])_+], \;\; c > 0,$$

the mean-plus-semideviation-from-target of order 1

$$\mathcal{R}(X) = \mathbb{E}[X] + c\mathbb{E}[(X - t)_+], \;\; c > 0, \;\; t \in \mathbb{R},$$

a convex combination of the expected value and the average value-at-risk

$$\mathcal{R}(X) = (1 - t)\mathbb{E}[X] + t \inf_{a \in \mathbb{R}} \left\{ a + \tfrac{1}{1-\beta}\mathbb{E}[(X - a)_+] \right\}, \;\; \beta \in (0, 1), \;\; t \in (0, 1],$$

and the buffered probability of exceedance

$$\mathcal{R}(X) = \inf_{a \geq 0} \mathbb{E}[(a(X - \tau) + 1)_+], \;\; \tau \in \mathbb{R}.$$

Moreover, in this case, Algorithm 1 is closely related to the traditional method of multipliers. To conclude this section, we discuss the application of Algorithm 1 to general coherent measures of risk.

5.1 The Positive Part Function.

To begin, let $\Phi(X) = \mathbb{E}[(X)_+]$. In this setting, Φ satisfies all previously stated assumptions and is given by (2) with

$$\mathfrak{A} = \{\theta \in \mathcal{Y} \mid \theta \in [0, 1] \;\text{ a.s.}\}.$$

Upon substituting this specific Φ into (5), we obtain

$$\widehat{\Phi}(X, \lambda, r) = \inf_{Y \in \mathcal{Y}} \mathbb{E}[(X - Y)_+ + \lambda Y + \tfrac{r}{2}(Y)^2].$$

Since $\mathcal{Y}$ is decomposable in the sense of Definition 14.59 in [44], Theorem 14.60 in [44] then ensures that we can interchange the expectation and the minimization operations to obtain

$$\widehat{\Phi}(X, \lambda, r) = \mathbb{E}\left[\inf_{y \in \mathbb{R}} \{(X - y)_+ + \lambda y + \tfrac{r}{2} y^2\} \right]. \tag{16}$$

The scalar minimization problem inside the expectation in (16) has a unique solution given pointwise by

$$[y^\star(X, \lambda, r)](\omega) = \begin{cases} -\lambda(\omega)/r & \text{if } rX(\omega) + \lambda(\omega) < 0 \\ X(\omega) & \text{if } 0 \leq rX(\omega) + \lambda(\omega) \leq 1 \\ (1 - \lambda(\omega))/r & \text{if } 1 < rX(\omega) + \lambda(\omega) \end{cases}.$$

Substituting $y^\star(X,\lambda,r)$ into the expression for $\widehat{\Phi}$ yields $\widehat{\Phi}(X,\lambda,r) = \mathbb{E}[\phi(X,\lambda,r)]$ where the scalar function $\phi : \mathbb{R}\times\mathbb{R}\times(0,\infty)\to\mathbb{R}$ is given by

$$\begin{aligned}\phi(x,t,r) &:= \frac{1}{2r}\{(rx+t)_+^2 - (rx+(t-1))_+^2 - t^2\}\\ &= \begin{cases} -\frac{1}{2r}t^2 & \text{if } rx+t<0\\ \frac{r}{2}x^2+tx & \text{if } 0\le rx+t\le 1\\ \frac{1}{r}\{(rx+t)-\frac{1}{2}(t^2+1)\} & \text{if } 1<rx+t. \end{cases}\end{aligned}$$

As mentioned earlier, it is often possible to tighten the lower bound in (7) for specific Φ. To this end, recall that the subgradients of Φ are given by

$$\theta\in\partial\Phi(Y) \quad\Longrightarrow\quad \theta(\omega)\in\begin{cases}\{0\} & \text{if } Y(\omega)<0\\ [0,1] & \text{if } Y(\omega)=0\\ \{1\} & \text{if } Y(\omega)>0\end{cases} \quad \text{for almost all } \omega\in\Omega.$$

As a result of (8) and the specific form of $\partial\Phi$, the constant K^2 that appears in (7) can be replaced by $\frac{1}{2}$.

To conclude, we note that Algorithm 1 is equivalent to the method of multipliers applied to the following smooth reformulation of (P),

$$\min_{x\in\mathcal{X}_{\rm ad},\, s,\,\eta\in\mathcal{Y}} \{g(x)+\mathbb{E}[\eta]\} \quad \text{subject to} \quad G(x)-\eta+s=0,\ \ \eta\ge 0,\ \ s\ge 0 \text{ a.s.}$$

Applying the method of multipliers to the equality constraint and explicitly handling the bound constraints, we obtain bound-constrained subproblems of the form

$$\min_{x\in\mathcal{X}_{\rm ad}}\left\{g(x)+\min_{s\ge 0,\,\eta\ge 0}\left\{\mathbb{E}[\lambda(G(x)-\eta+s)]+\frac{r}{2}\|G(x)-\eta+s\|_{\mathcal{Y}}^2\right\}\right\}$$

for fixed $\lambda\in\mathcal{Y}$ and $r>0$. The inner bound-constrained minimization problem is strongly convex and has the unique solution defined pointwise by

$$\eta^\star(x,\lambda,r) = \frac{1}{r}\left(rG(x)+\lambda-1\right)_+ \quad\text{and}\quad s^\star(x,\lambda,r) = \frac{1}{r}\left(-(rG(x)+\lambda)\right)_+.$$

Substituting these expressions into the augmented Lagrangian then produce the smooth objective function $L(x,\lambda,r) = \{g(x)+\mathbb{E}[\phi(G(x),\lambda,r)]\}$.

5.2 Coherent Risk Measures.

Let $\mathcal{R}:\mathcal{Y}\to\mathbb{R}$ be a coherent risk measure [2]. That is, $\mathcal{R}$ satisfies the following four axioms: For X, $X'\in\mathcal{Y}$,

(C1) *Convexity:* If $0\le t\le 1$, then $\mathcal{R}(tX+(1-t)X')\le t\mathcal{R}(X)+(1-t)\mathcal{R}(X')$;
(C2) *Monotonicity:* If $X\le X'$ a.s. then $\mathcal{R}(X)\le\mathcal{R}(X')$;
(C3) *Translation Equivariance:* If $t\in\mathbb{R}$, then $\mathcal{R}(X+t)=\mathcal{R}(X)+t$;
(C4) *Positive Homogeneity:* If $t\ge 0$, then $\mathcal{R}(tX)=t\mathcal{R}(X)$.

Since $\mathcal{R}$ satisfies the assumptions from the previous section (i.e., convexity, monotonicity and positive homogeneity), we can directly apply Algorithm 1 to solve

$$\min_{z \in \mathcal{Z}_{\text{ad}}} \{f(z) + \mathcal{R}(F(z))\}. \tag{17}$$

where $\mathcal{Z}_{\text{ad}}$, f and F satisfy similar assumptions as $\mathcal{X}_{\text{ad}}$, g and G, respectively (see Section 2). By the Fenchel-Moreau Theorem [47, Th. 6.5], we again have that

$$\mathcal{R}(X) = \sup_{\theta \in \mathfrak{D}} \mathbb{E}[\theta X]$$

where $\mathfrak{D} \subseteq \{\theta \in \mathcal{Y} \,|\, \mathbb{E}[\theta] = 1, \;\; \theta \geq 0 \text{ a.s.}\}$ is a closed, convex and bounded set. Since the augmented Lagrangian (5) requires the projection onto $\mathfrak{D}$, it is often simpler to decompose $\mathfrak{D}$ as $\mathfrak{D} = \mathfrak{A} \cap \{\theta \in \mathcal{Y} \,|\, \mathbb{E}[\theta] = 1\}$ where $\mathfrak{A}$ ignores the constraint $\mathbb{E}[\theta] = 1$. In this case, we can employ Lagrangian duality to obtain

$$\begin{aligned}\mathcal{R}(X) &= \sup_{\theta \in \mathfrak{A}} \inf_{t \in \mathbb{R}} \{\mathbb{E}[\theta X] + t(1 - \mathbb{E}[\theta])\} \\ &= \inf_{t \in \mathbb{R}} \left\{ t + \sup_{\theta \in \mathfrak{A}} \mathbb{E}[\theta(X - t)] \right\} = \inf_{t \in \mathbb{R}} \{t + \Phi(X - t)\}\end{aligned}$$

where Φ is equal to the support function of $\mathfrak{A}$ (see (2)). Note that the interchange of infimum and supremum in the second equality is justified if, e.g., there exists $\delta > 0$ such that

$$(-\delta, \delta) \subseteq \{\mathbb{E}[\theta] - 1 \,|\, \theta \in \mathfrak{A}\} \subseteq \mathbb{R}.$$

See [24] for similar arguments applied to the average value-at-risk. We can then formulate (17) as (P) with $\mathcal{X}_{\text{ad}} = \mathbb{R} \times \mathcal{Z}_{\text{ad}}$, $g(x) = f(z) + t$, and $G(x) = F(z) - t$ where $x = (t, z)$. In this case, we can directly apply Algorithm 1 to the modified problem. The potential benefit of this approach is simplifying the projection required by Algorithm 1.

To demonstrate this approach, we consider the coherent risk measure

$$\mathcal{R}(X) = \inf_{t \in \mathbb{R}} \left\{ t + \sigma \| (X - t)_+ \|_{\mathcal{Y}} \right\}$$

for $\sigma > 1$, which is a special case of the higher moment coherent risk measures [28] and more generally the transformed norm risk measures [10]. The associated set $\mathfrak{D}$ for this choice of $\mathcal{R}$ is

$$\mathfrak{D} = \{\theta \in \mathcal{Y} \,|\, \mathbb{E}[\theta] = 1, \;\; \theta \geq 0 \text{ a.s.}, \;\; \|\theta\|_{\mathcal{Y}} \leq \sigma\}$$

[10, Sect. 8.2]. We can decompose $\mathfrak{D}$ as above with

$$\mathfrak{A} = \{\theta \in \mathcal{Y} \,|\, \theta \geq 0 \text{ a.s.}, \;\; \|\theta\|_{\mathcal{Y}} \leq \sigma\} \quad \text{and} \quad \Phi(X) = \sigma \| (X)_+ \|_{\mathcal{Y}}.$$

Moreover, the projection onto $\mathfrak{A}$ is given by

$$\mathbf{P}_{\mathfrak{A}}(X) = \begin{cases} (X)_+ & \text{if } \| (X)_+ \|_{\mathcal{Y}} \leq \sigma \\ \frac{\sigma (X)_+}{\| (X)_+ \|_{\mathcal{Y}}} & \text{otherwise} \end{cases}.$$

Substituting this projection into (5) gives the following the expression for $\widehat{\Phi}$,

$$\widehat{\Phi}(X, \lambda, r) = \begin{cases} \frac{r}{2} \| (X + \lambda/r)_+ \|_{\mathcal{Y}}^2 & \text{if } \| (X + \lambda/r)_+ \|_{\mathcal{Y}} \leq \frac{\sigma}{r} \\ \sigma \| (X + \lambda/r)_+ \|_{\mathcal{Y}} - \frac{\sigma^2}{2r} & \text{otherwise} \end{cases}.$$

6 Practical Implementation Details.

In this section, we describe termination conditions for the outer iteration in Algorithm 1 ("while" loop) as well as for the subproblem solves (step 1). We conclude with a discussion on updating the penalty parameter (step 3). Our motivation for each of these steps is the following characterization of the first-order necessary optimality conditions for (P).

Proposition 3 *Suppose there exists* $(\bar{x},\bar{\lambda})\in\mathcal{X}\times\mathcal{Y}$ *such that*

$$\bar{x}\in\operatorname*{arg\,min}_{x\in\mathcal{X}_{\rm ad}}\ell(x,\bar{\lambda})\quad\textit{and}\quad\bar{\lambda}\in\operatorname*{arg\,max}_{\theta\in\mathfrak{A}}\ell(\bar{x},\theta).$$

Then, $(\bar{x},\bar{\lambda})\in\mathcal{X}_{\rm ad}\times\mathfrak{A}$ *satisfies*

$$-L_x'(\bar{x},\bar{\lambda},r)\in N(\bar{x},\mathcal{X}_{\rm ad})\quad\textit{and}\quad\bar{\lambda}=\mathit{\Lambda}(\bar{x},\bar{\lambda},r)$$

for arbitrary fixed $r>0$.

Proof Since $\ell(\bar{x},\cdot)$ is concave and differentiable, we have that $-\nabla_\lambda\ell(\bar{x},\cdot)$ is maximally monotone and $\bar{\lambda}$ solves the following fixed point problem

$$\nabla_\lambda\ell(\bar{x},\bar{\lambda})\in N(\bar{\lambda},\mathfrak{A})\quad\Longleftrightarrow\quad\bar{\lambda}\in\bar{\lambda}+r(N(\bar{\lambda},\mathfrak{A})-\nabla_\lambda\ell(\bar{x},\bar{\lambda}))$$

for arbitrary fixed $r>0$. Therefore, $\bar{\lambda}$ is a maximizer in (5) and $\bar{\lambda}=\mathit{\Lambda}(\bar{x},\bar{\lambda},r)$. This ensures that

$$L(x,\bar{\lambda},r)\geq\ell(x,\bar{\lambda})\geq\ell(\bar{x},\bar{\lambda})=L(\bar{x},\bar{\lambda},r)\quad\forall\,x\in\mathcal{X}_{\rm ad}.$$

That is, $\bar{x}$ is a minimizer of $L(\cdot,\bar{\lambda},r)$ over $\mathcal{X}_{\rm ad}$. Since $L(\cdot,\bar{\lambda},r)$ is continuously Fréchet differentiable, $\bar{x}$ satisfies the first-order necessary optimality conditions $-L_x'(\bar{x},\bar{\lambda},r)\in N(\bar{x},\mathcal{X}_{\rm ad})$ as desired. □

In the forthcoming discussion, we avoid complications with the normal cone representation in Proposition 3 and assume that $\mathcal{X}$ is a Hilbert space. In this case, the condition $-L_x'(\bar{x},\bar{\lambda},r)\in N(\bar{x},\mathcal{X}_{\rm ad})$ is equivalent to the condition

$$\bar{x}=\mathbf{P}_{\mathcal{X}_{\rm ad}}(\bar{x}-\gamma\nabla_x L(\bar{x},\bar{\lambda},r)),\ \ \gamma>0,$$

where $\nabla_x L(x,\lambda,r)\in\mathcal{X}$ denotes the gradient of $L(\cdot,\lambda,r)$ (see, e.g., Example 23.4 and Theorem 26.2 in [4]). Here, $\mathbf{P}_{\mathcal{X}_{\rm ad}}:\mathcal{X}\to\mathcal{X}$ denotes the projection onto $\mathcal{X}_{\rm ad}$. This and the results in Proposition 3 suggest that it is reasonable to terminate Algorithm 1 based on the following criteria

$$\|x_{k+1}-\mathbf{P}_{\mathcal{X}_{\rm ad}}(x_{k+1}-\nabla_x L(x_{k+1},\lambda_k,r_k))\|_{\mathcal{X}}\leq\tau_x \tag{18a}$$

$$\|\lambda_k-\lambda_{k+1}\|_{\mathcal{Y}}\leq\tau_\lambda \tag{18b}$$

where $\tau_x>0$, $\tau_\lambda>0$ are user-defined input parameters. In a similar fashion, we terminate the inexact optimization in step 1 of Algorithm 1 (i.e., ϵ_k from Theorem 3) if

$$\|x_{k+1}-\mathbf{P}_{\mathcal{X}_{\rm ad}}(x_{k+1}-\nabla_x L(x_{k+1},\lambda_k,r_k))\|_{\mathcal{X}}\leq\tau_{x,k} \tag{19}$$

and we increase the penalty parameter r_k from step 3 of Algorithm 1 if

$$\|\lambda_k - \lambda_{k+1}\|_{\mathcal{Y}} > \tau_{\lambda,k}. \tag{20}$$

Here, $\tau_{x,k} > 0$ and $\tau_{\lambda,k} > 0$ are user-defined sequences. A simple choice to update r_k would be to set $r_{k+1} = \rho_r r_k$ for some $\rho_r > 1$ if (20) is satisfied. Similarly, we can decrease $\tau_{x,k}$ and $\tau_{\lambda,k}$ as $\tau_{x,k+1} = \rho_x \tau_{x,k}$ for some $0 < \rho_x < 1$ and $\tau_{\lambda,k+1} = \rho_\lambda \tau_{\lambda,k}$ for some $0 < \rho_\lambda < 1$. To motivate this update strategy further, note that if $\mathbb{P}$ is a discrete probability measure supported at a finite number of atoms, then $\mathcal{Y}$ is finite dimensional. In this case, the strong and weak topologies on $\mathcal{Y}$ coincide and Theorem 2 ensures that the sequence of multipliers produced by Algorithm 1 converge with respect to the norm topology to a maximizer of the dual problem. Therefore, (20) will be satisfied for k sufficiently large. Additionally, notice that (20) is related to the classical equality-constrained method of multipliers in which case $(\lambda_{k+1} - \lambda_k)/r_k$ is the constraint violation (see, e.g., [6,11] for additional details). The above considerations give rise to Algorithm 2

Algorithm 2 Practical Primal-Dual Risk Minimization

Initialize: Given $x_0 \in \mathcal{X}_{\text{ad}}$, $r_0 \in (0,\infty)$, $\lambda_0 \in \mathfrak{A}$, $\rho_x \in (0,1)$, $\rho_\lambda \in (0,1)$, $\rho_r \in (1,\infty)$, $0 < \tau_x < \tau_{x,0}$, and $0 < \tau_\lambda < \tau_{\lambda,0}$.
For $k = 0, 1, 2, \ldots$

1. Compute $x_{k+1} \in \mathcal{X}_{\text{ad}}$ satisfying (19);
2. Set $\lambda_{k+1} = \Lambda(x_{k+1}, \lambda_k, r_k)$;
3. **If** (18) is satisfied
   Return x_{k+1};
   End If
4. **If** (20) is satisfied
   Set $r_{k+1} = \rho_r r_k$;
   End If
5. Set $\tau_{x,k+1} = \rho_x \tau_{x,k}$ and $\tau_{\lambda,k+1} = \rho_\lambda \tau_{\lambda,k}$.

End For

To apply Algorithm 2, we must approximate the expectations defining the augmented Lagrangian L. A straightforward approach to numerically solving these subproblems is to use sample average approximation (SAA) with a fixed set of samples and then approximately solve the resulting optimization problem using a nonlinear optimization solver. See, e.g., [47] for large-sample statistical analysis of the resulting SAA solution. If the augmented Lagrangian L is the expectation of a scalar function of the random cost (cf. Section 5.1), then we could solve the subproblem using stochastic approximation [33,35,36], stochastic mirror descent [15,29], stochastic quasigradient methods [14], or progressive hedging [43]. Moreover, if the integrand is sufficiently regular with respect to the random inputs, we could employ deterministic quadrature, such as the adaptive sparse-grid techniques in [20–22]. See [23] for an overview of these methods applied to stochastic PDE-constrained optimization.

7 Numerical Results.

In this section, we demonstrate the performance of Algorithm 2 on multiple convex and nonconvex PDE-constrained optimization problems. We solve each problem for a variety of risk measures including the mean-plus-semideviation of order 1 (`MPSD`), the mean-plus-semideviation-from-target of order 1 (`MPSDFT`), a convex combination of the expectation and the average value-at-risk (`AVAR`), the higher moment coherent risk measure of order 2 (`HMCR`), and the buffered probability of exceedance (`BPOE`). The parameters for each example's risk measures are listed in Table 1. Each of these risk measures permit the use of the analysis in Section 5.1. Our implementation of Algorithm 2 uses sample-based approximation of the augmented Lagrangian subproblems and is available in the Rapid Optimization Library (ROL) [27]. ROL is a C++ library for solving constrained optimization problems with special interfaces for simulation-constrained and stochastic optimization. To approximately solve the subproblems in step 1 of Algorithm 2, we use a matrix-free trust-region Newton method (cf. Algorithms 6.1.1 and 7.5.1 in [12] for unconstrained problems and [30] for bound constrained problems). Since the function $L(\cdot,\lambda,r)$ is in general not twice continuously differentiable, we use generalized Hessians based on the Newton derivative of $\nabla_x L(x,\lambda,r)$ [50] to formulate the quadratic trust-region subproblem model. The computational costs at each iteration of this method are: (i) an evaluation of $L(x,\lambda,r)$; (ii) an evaluation of the gradient $\nabla_x L(x,\lambda,r)$; and (iii) the iterative solution of the trust-region subproblem, which requires the application of the generalized Hessian to a vector at each iteration. At the cost of additional augmented Lagrangian subproblem iterations, we can reduce the computational burden of the trust-region subproblem solver by replacing the generalized Hessian with a secant approximation [34]. Since the possibilities for subproblem solvers are essentially limitless, we restrict our numerical experiments to the aforementioned trust-region Newton method.

`name`	`example 7.1`	`example 7.2`	`example 7.3`
`MPSD`	$c = 0.95$	$c = 0.95$	$c = 0.95$
`MPSDFT`	$c = 0.95$, $t = 0.2$	$c = 0.95$, $t = 5$	$c = 0.95$, $t = 0.01$
`AVAR`	$\beta = 0.9$, $t = 0.75$	$\beta = 0.9$, $t = 0.75$	$\beta = 0.9$, $t = 0.75$
`HMCR`	$\sigma = 10$	$\sigma = 10$	$\sigma = 10$
`BPOE`	$\tau = 0.7$	$\tau = 6$	$\tau = 0.01$

Table 1: Parameters for the five choices of $\mathcal{R}$ in the subsequent examples.

For each example below, we provide a table summarizing the performance of Algorithm 2. In these tables, `iter` is the number of iterations required by Algorithm 2 to meet the prescribed termination criteria (18), `nfval` is the total number of evaluations of $L(x,\lambda,r)$, `ngrad` is the total number of evaluations of the gradient of $L(\cdot,\lambda,r)$, and `subiter` is the total number of subproblem iterations from step 1 in Algorithm 2. For each example, we set $\tau_x = 10^{-8}$, $\tau_\lambda = 10^{-6}$, $\rho_x = 0.1$, $\rho_\lambda = 0.1$ and $\rho_r = 10$. We compare the performance of Algorithm 2 with ROL's implementation of the nonconvex trust-region bundle method from [46]. We terminate the bundle method when the aggregate subgradient and linearization error are below 10^{-8}. For convex problems, this ensures that the solution is ϵ-optimal with $\epsilon = 10^{-8}$. The performance of this algorithm is summarized in the final

two columns of each table: `iter` is the number of bundle iterations and `neval` is the number of objective function (gradient) evaluations. Note that the objective function and gradient are evaluated the same number of times for the bundle method, whereas these numbers may differ for the trust-region method depending on how many steps are rejected. ROL's implementation of the bundle method only solves unconstrained problems. Therefore, we do not compare Algorithm 2 with the bundle method when auxiliary constraints are present (i.e., when using the buffered probability or in our second example which has bound constraints on the optimization variables).

Finally, we provide a comparison of Algorithm 2 with the epi-regularization [26] of the risk measure `AVAR` from Table 1. Our epi-regularization approach mimics the quadratic penalty algorithm described in [7, Prop. 4.2.2]. In particular, at iteration k of this method, we compute an iterate x_{k+1} that satisfies (19) with $\lambda_k = 0$, $r_k = 10^k$, and $\tau_{x,k} = 10^{-(k+2)}$. Roughly speaking, this approach performs continuation on the sequence of epi-regularized risk measures. We compute x_{k+1} by applying the trust-region Newton methods described above. We refer to this algorithm as `Epi-Reg` and note that Theorem 1 applies to `Epi-Reg` since the proof only requires boundedness of λ_k. Consequently, Theorem 1 suggests (at least for convex problems) that weak accumulation points are ϵ-minimizers. For the comparison of Algorithm 2 with `Epi-Reg`, we tabulate `iter`, `nfval`, `ngrad` and `subiter` as before. We also include `nhess`, which is the number of applications of the generalized Hessian—the dominant cost when solving the trust-region subproblem.

7.1 Optimization of a 1D Linear Elliptic PDE.

We consider the optimal control of a linear elliptic PDE with discontinuous conductivity first analyzed in [21,22] for risk-neutral optimization and subsequently studied in [24] for risk-averse optimization using the average value-at-risk. Let $\alpha = 10$, $D = (-1, 1)$, $\mathcal{Z} = L^2(D)$ and consider the optimization problem

$$\min_{z\in\mathcal{Z}} \left\{ \mathcal{R}\left(\frac{1}{2}\int_D (S(z)-1)^2 \,\mathrm{d}x \right) + \frac{\alpha}{2}\int_D z^2 \,\mathrm{d}x \right\} \tag{21}$$

where $u = S(z) : \Omega \to U = H_0^1(D)$ solves the weak form of

$$-\partial_x\left(\epsilon(\omega)\partial_x u(\omega)\right) = f(\omega) + z \qquad \text{in } D \text{ a.s.} \tag{22a}$$

$$[u(\omega)](-1) = 0, \quad [u(\omega)](1) = 0 \qquad \text{a.s.} \tag{22b}$$

Here, $H^1(D)$ denotes the usual Sobolev space of $L^2(D)$ functions whose weak derivatives are also $L^2(D)$ functions and $H_0^1(D)$ denotes the subspace of $H^1(D)$ consisting of functions with boundary trace equal to zero [1]. The uncertain diffusivity coefficients ϵ are discontinuous where the location of the discontinuity is uncertain and the random force is given by a Gaussian function with uncertain center. See [21] for more details on the existence and regularity properties of the solution map $S(z)$.

We discretize the state and controls using continuous piecewise linear finite elements built on different meshes of 256 intervals. These meshes are described in [21]. We approximate with respect to the uncertainty using sample average approximation with 10,000 Monte Carlo samples. In Table 2, we summarize the performance

of Algorithm 2 on (21) for the five choices of $\mathcal{R}$. The parameters for each choice of $\mathcal{R}$ are listed in Table 1. In particular, notice that Algorithm 2 required at most 38 total subproblem solves and at most 49 evaluations of $L(x, \lambda, r)$. When comparing the performance of Algorithm 2 to the bundle method, we see between a 3.8 and 18.7–fold reduction in the number of objective function and gradient evaluations (i.e., solves of (22) and the associated adjoint equation). In Table 3, we compare Algorithm 2 with `Epi-Reg`. For this problem, Algorithm 2 provides a modest benefit over `Epi-Reg` in all metrics. We emphasize that Algorithm 2 requires 10 fewer function evaluations, resulting in 100,000 fewer PDE solves. The same holds for the reduction in gradient and generalized Hessian applications. In total, Algorithm 2 requires 340,000 fewer PDE solves than `Epi-Reg`.

	Algorithm 2				Bundle	
name	iter	nfval	ngrad	subiter	iter	neval
MPSD	7	14	14	7	31	208
MPSDFT	7	11	11	4	24	206
AVAR	7	23	23	16	39	88
HMCR	6	16	15	10	40	104
BPOE	11	49	36	38	---	---

Table 2: Summary of the performance of Algorithm 2 and the bundle method applied to (21) for five choices of $\mathcal{R}$.

AVAR	iter	nfval	ngrad	nhess	subiter
Algorithm 2	7	23	23	90	16
Epi-Reg	8	33	29	99	25

Table 3: Comparison between Algorithm 2 and epi-regularization with continuation for (21) using `AVAR` from Table 1.

7.2 Optimization of a 2D Linear Elliptic PDE.

Let $D = (0,1)^2$ denote the physical domain. We consider the contaminant mitigation problem from [25]:

$$\min_{z \in \mathcal{Z}_{\mathrm{ad}}} \left\{ \mathcal{R}\left(\frac{\kappa_s}{2} \int_D S(z)^2 \, \mathrm{d}x \right) + \kappa_c \|z\|_1 \right\} \tag{23}$$

where $\kappa_s = 10^5$, $\kappa_c = 1$ and $S(z) = u : \Omega \to H^1(D)$ solves the weak form of

$$-\nabla \cdot (\epsilon(\omega)\nabla u) + \mathbb{V}(\omega) \cdot \nabla u = f(\omega) - Bz \qquad \text{in } D, \text{ a.s.} \tag{24a}$$

$$u = 0 \qquad \text{on } \Gamma_d = \{0\} \times (0,1), \text{ a.s.} \tag{24b}$$

$$\epsilon(\omega)\nabla u \cdot n = 0 \qquad \text{on } \partial D \setminus \Gamma_d, \text{ a.s.} \tag{24c}$$

The control space is $\mathcal{Z} = \mathbb{R}^9$ with feasible set $\mathcal{Z}_{\mathrm{ad}} = \{z \in \mathcal{Z} \,|\, 0 \le z \le 1\}$. See [25] for the explicit forms of the PDE coefficients ϵ, $\mathbb{V}$, f and B as well as existence and regularity properties of the solution $S(z)$.

We discretized this problem in space using Q1 finite elements on a uniform mesh of 4096 quadrilaterals. Furthermore, we approximate with respect to the uncertainty using sample average approximation with 10,000 Monte Carlo samples. In Table 4, we summarize the performance of Algorithm 2 on (23) for the five choices of $\mathcal{R}$. The parameters for each risk measure are listed in Table 1. In particular, notice that Algorithm 2 required at most 63 total subproblem solves and at most 72 evaluations of $L(x, \lambda, r)$. In Table 5, we compare Algorithm 2 with `Epi-Reg`. For this problem, Algorithm 2 provides considerable benefit over `Epi-Reg` in all metrics. In particular, Algorithm 2 requires 3,760,000 fewer PDE solves than `Epi-Reg`. This disparity arises from the ill-conditioning of the epi-regularized risk measure as r_k increases, resulting in a substantial increase in the number of trust-region subproblem solver iterations (i.e., generalized Hessian applications).

	Algorithm 2			
name	iter	nfval	ngrad	subiter
MPSD	5	10	10	5
MPSDFT	6	13	13	7
AVAR	9	35	30	26
HMCR	7	25	24	18
BPOE	9	72	41	63

Table 4: Summary of the performance of Algorithm 2 applied to (23) for five choices of $\mathcal{R}$.

AVAR	iter	nfval	ngrad	nhess	subiter
Algorithm 2	9	35	30	138	26
Epi-Reg	10	80	45	296	70

Table 5: Comparison between Algorithm 2 and epi-regularization with continuation for (23) using `AVAR` from Table 1.

7.3 Optimization of Burger's Equation

To conclude, we consider the optimization of the steady viscous Burger's equation with uncertain coefficients. Burger's equation is nonlinear which results in a nonconvex optimization problem. This problem was first studied in [21,22] for risk-neutral optimization and later studied in [24] for risk-averse optimization using the average value-at-risk. Let $\alpha = 10^{-3}$, $D = (0,1)$, $\mathcal{Z} = L^2(0,1)$. We consider the optimization problem

$$\min_{z \in \mathcal{Z}} \left\{ \mathcal{R}\left(\frac{1}{2} \int_D (S(z) - 1)^2 \, \mathrm{d}x \right) + \frac{\alpha}{2} \int_D z^2 \, \mathrm{d}x \right\} \tag{25}$$

where $u = S(z) : \Omega \to U = H^1(D)$ solves the weak form of

$$-\nu(\omega)\partial_{xx} u(\omega) + u(\omega)\partial_x u(\omega) = f(\omega) + z \qquad \text{in } D, \text{ a.s.} \tag{26a}$$

$$[u(\omega)](0) = d_0(\omega), \quad [u(\omega)](1) = d_1(\omega) \qquad \text{a.s.} \tag{26b}$$

For the explicit form of the coefficients ν, f, d_0 and d_1 as well as a thorough analysis of the solution to (26), including existence and regularity of $S(z)$, see [21].

Following [21], we discretize the state and controls using continuous piecewise linear finite elements on different mesh of 256 intervals. We solve the discretized nonlinear PDE using Newton's method globalized with a backtracking line search. We approximate with respect to the uncertainty using sample average approximation with 10,000 Monte Carlo samples. In Table 6, we summarize the performance of Algorithm 2 on (21) for the five choices of $\mathcal{R}$. The parameters for each choice of $\mathcal{R}$ are listed in Table 1. In particular, notice that Algorithm 2 required at most 71 total subproblem solves and at most 79 evaluations of $L(x, \lambda, r)$. When comparing the performance of Algorithm 2 to the bundle method, we see between a 2.4 and 7.2–fold reduction in the number of objective function and gradient evaluations (i.e., solves of (26) and the associated adjoint equation). In Table 7, we compare Algorithm 2 with `Epi-Reg`. For this problem, Algorithm 2 provides considerable benefit over `Epi-Reg` in all metrics. In fact, Algorithm 2 requires 260,000 fewer nonlinear PDE solves and 1,270,000 fewer linearized PDE solves than `Epi-Reg`.

	Algorithm 2				Bundle	
name	iter	nfval	ngrad	subiter	iter	neval
MPSD	12	31	26	19	51	176
MPSDFT	9	17	17	8	53	123
AVAR	8	46	44	38	69	197
HMCR	8	79	73	71	84	182
BPOE	9	52	42	43	---	---

Table 6: Summary of the performance of Algorithm 2 and the bundle method applied to (25) for five choices of $\mathcal{R}$.

AVAR	iter	nfval	ngrad	nhess	subiter
Algorithm 2	8	46	44	128	38
Epi-Reg	8	72	63	182	64

Table 7: Comparison between Algorithm 2 and epi-regularization with continuation for (25) using `AVAR` from Table 1.

8 Conclusions and Outlook.

Motivated by the classical method of multipliers, we have developed a new algorithm for nonsmooth risk minimization. At each iteration of our algorithm, one must approximately solve smooth problems using, e.g., rapidly-converging derivative-based optimization methods. We have demonstrated that our algorithm converges when the subproblem solver returns either approximate minimizers or approximate stationary points and that the sequence of dual variables converges to a solution of the associated dual problem. We have demonstrated our method on multiple examples from PDE-constrained optimization. On each example, our

method requires a modest number of iterations and hence a modest number of PDE solves when compared with traditional nonsmooth methods.

Our current approach is based on the assumption that Φ is finite, monotonic, and positive homogeneous. We believe these assumptions can be relaxed. The complication with allowing Φ to be extended real valued is that the associated set $\mathfrak{A}$ may no longer be bounded. For example, if Φ is the indicator function of the set $\{\theta \in \mathcal{Y} \,|\, \theta \leq 0 \text{ a.s.}\}$, then Φ is monotonic and positive homogeneous. Moreover, $\mathfrak{A} = \{\theta \in \mathcal{Y} \,|\, \theta \geq 0 \text{ a.s.}\}$, which is not bounded. On the other hand, if Φ is not monotonic, then $\theta \in \mathfrak{A}$ is not guaranteed to satisfy $\theta \geq 0$ a.s. Finally, if Φ is not positive homogeneous, then dual representation (2) is replaced by

$$\Phi(X) = \sup_{\theta \in \mathfrak{A}} \{\mathbb{E}[\theta X] - \Phi^*(\theta)\}$$

where Φ^* is the Fenchel conjugate of Φ. The addition of Φ^* complicates the derivation of an explicit maximizer in (5). By extending our algorithm for more general Φ, it may be possible to apply our approach to solve extended nonlinear programming problems [41].

References

1. R. A. Adams, *Sobolev Spaces*, Academic Press, Orlando, San Diego, New-York,..., 1975.
2. P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent measures of risk*, Math. Finance, 9 (1999), pp. 203–228.
3. H. Attouch, G. Buttazzo, and G. Michaille, *Variational analysis in Sobolev and BV spaces*, vol. 6 of MPS/SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.
4. H. H. Bauschke and P. L. Combettes, *Convex Analysis and Montone Operator Theory in Hilbert Space*, CMS Books in Mathematics, Springer New York, 2011.
5. A. Ben-Tal and M. Teboulle, *An old-new concept of convex risk measures: The optimized certainty equivalent*, Mathematical Finance, 17 (2007), pp. 449–476.
6. D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York,. London, Paris, San Diego, San Francisco, 1982.
7. ———, *Nonlinear programming*, Athena Scientific Optimization and Computation Series, Athena Scientific, Belmont, MA, second ed., 1999.
8. J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical optimization: Theoretical and practical aspects*, Universitext, Springer-Verlag, Berlin, second ed., 2006.
9. J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer Verlag, Berlin, Heidelberg, New York, 2000.
10. P. Cheridito and T. Li, *Dual characterization of properties of risk measures on Orlicz hearts*, Mathematics and Financial Economics, 2 (2008), p. 29.
11. A. R. Conn, N. I. M. Gould, and P. L. Toint, *LANCELOT: A FORTRAN package for large scale nonlinear optimization with simple bounds*, Springer Series in Computational Mathematics, Vol. 17, Springer Verlag, Berlin, Heidelberg, New York, 1992.
12. ———, *Trust–Region Methods*, SIAM, Philadelphia, 2000.
13. A. Eichhorn and W. Römisch, *Polyhedral risk measures in stochastic programming*, SIAM Journal on Optimization, 16 (2005), pp. 69–95.
14. Y. Ermoliev, *Stochastic quasigradient methods and their application to system optimization*, Stochastics, 9 (1983), pp. 1–36.
15. V. Guigues, *Multistep stochastic mirror descent for risk-averse convex stochastic programs based on extended polyhedral risk measures*, Mathematical Programming, 163 (2017), pp. 169–212.
16. V. Guigues and W. Römisch, *Sampling-based decomposition methods for multistage stochastic programs based on extended polyhedral risk measures*, SIAM Journal on Optimization, 22 (2012), pp. 286–312.

17. M. R. Hestenes, *Multiplier and gradient methods*, Journal of Optimization Theory and Applications, 4 (1969), pp. 303–320.
18. M. Hintermüller and K. Kunisch, *Path-following methods for a class of constrained minimization problems in function space*, SIAM Journal on Optimization, 17 (2006), pp. 159–187.
19. K. C. Kiwiel, *Methods of descent for nondifferentiable optimization.* Lecture Notes in Mathematics. 1133. Berlin etc.: Springer-Verlag, 1985.
20. D. P. Kouri, *A multilevel stochastic collocation algorithm for optimization of PDEs with uncertain coefficients*, SIAM/ASA Journal on Uncertainty Quantification, 2 (2014), pp. 55–81.
21. 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 (2013), pp. A1847–A1879.
22. ———, *Inexact objective function evaluations in a trust-region algorithm for PDE-constrained optimization under uncertainty*, SIAM Journal on Scientific Computing, 36 (2014), pp. A3011–A3029.
23. D. P. Kouri and A. Shapiro, *Optimization of PDEs with Uncertain Inputs*, Springer New York, New York, NY, 2018, pp. 41–81.
24. D. P. Kouri and T. M. Surowiec, *Risk-averse PDE-constrained optimization using the conditional value-at-risk*, SIAM Journal on Optimization, 26 (2016), pp. 365–396.
25. ———, *Existence and optimality conditions for risk-averse PDE-constrained optimization*, SIAM/ASA Journal on Uncertainty Quantification, 6 (2018), pp. 787–815.
26. ———, *Epi-regularization of risk measures*, Mathematics of Operations Research, (2019). 10.1287/moor.2019.1013.
27. D. P. Kouri, G. von Winckel, and D. Ridzal, *ROL: Rapid Optimization Library.* `https://trilinos.org/packages/rol`, 2017.
28. P. A. Krokhmal, *Higher moment coherent risk measures*, Quantitative Finance, 7 (2007), pp. 373–387.
29. G. Lan, A. Nemirovski, and A. Shapiro, *Validation analysis of mirror descent stochastic approximation method*, Mathematical Programming, 134 (2012), pp. 425–458.
30. C.-J. Lin and J. J. Moré, *Newton's method for large bound-constrained optimization problems*, SIAM Journal on Optimization, 9 (1999), pp. 1100–1127.
31. M. M. Mäkelä and P. Neittaanmäki, *Nonsmooth optimization : Analysis and algorithms with applications to optimal control*, World Scientific Publishing Co., Inc., River Edge, NJ, 1992.
32. A. Nedić and A. Ozdaglar, *Approximate primal solutions and rate analysis for dual subgradient methods*, SIAM Journal on Optimization, 19 (2009), pp. 1757–1780.
33. A. Nemirovski, A. B. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optimization, 19 (2009), pp. 1574–1609.
34. J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Verlag, Berlin, Heidelberg, New York, second ed., 2006.
35. B. T. Polyak, *A new method of stochastic approximation type*, Avtomat. i Telemekh., (1990), pp. 98–107.
36. B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.
37. M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, Optimization, (1969), pp. 283–298.
38. R. T. Rockafellar, *On the maximal monotonicity of subdifferential mappings.*, Pacific J. Math., 33 (1970), pp. 209–216.
39. ———, *Augmented lagrangians and applications of the proximal point algorithm in convex programming*, Mathematics of Operations Research, 1 (1976), pp. 97–116.
40. ———, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.
41. ———, *Extended nonlinear programming*, in Nonlinear optimization and related topics, Springer, 2000, pp. 381–399.
42. 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 (2013), pp. 33 – 53.
43. R. T. Rockafellar and R. J.-B. Wets, *Scenarios and policy aggregation in optimization under uncertainty*, Mathematics of Operations Research, 16 (1991), pp. 119–147.
44. ———, *Variational Analysis*, Springer Verlag, Berlin, Heidelberg, New York, 1998.

45. L. L. Sakalauskas, *Nonlinear stochastic programming by Monte-Carlo estimators*, European Journal of Operational Research, 137 (2002), pp. 558 – 573.
46. H. Schramm and J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*, SIAM Journal on Optimization, 2 (1992), pp. 121–152.
47. 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.
48. N. Z. Shor, *Minimization Methods for Non-differentiable Functions*, Springer-Verlag, New York, 1985.
49. M. Sion, *On general minimax theorems*, Pacific J. Math., 8 (1958), pp. 171–176.
50. M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, SIAM, Philadelphia, 2011.