Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

DISTRIBUTIONALLY ROBUST OPTIMIZATION FOR INPUT MODEL UNCERTAINTY IN SIMULATION-BASED DECISION MAKING

Soumyadip Ghosh
Mark S. Squillante

Mathematical Sciences, IBM Research
Thomas J. Watson Research Center
Yorktown Heights, NY 20198, USA

ABSTRACT

We consider a new approach to solve distributionally robust optimization formulations that address nonparametric input model uncertainty in simulation-based decision making problems. Our approach for the minimax formulations applies stochastic gradient descent to the outer minimization problem and efficiently estimates the gradient of the inner maximization problem through multi-level Monte Carlo randomization. Leveraging theoretical results that shed light on why standard gradient estimators fail, we establish the optimal parameterization of the gradient estimators of our approach that balances a fundamental tradeoff between computation time and statistical variance. We apply our approach to nonconvex portfolio choice modeling under cumulative prospect theory, where numerical experiments demonstrate the significant benefits of this approach over previous related work.

1 INTRODUCTION

Consider the following general formulations of distributionally robust optimization (DRO) in the context of decision making. Let $\mathbb{X}$ denote a sample space, P a probability distribution on $\mathbb{X}$, and $\Theta \subseteq \mathbb{R}^d$ a parameter space. Define $L_P(\theta) := \mathbb{E}_P[l(\theta,\xi)]$ to be the expectation, with respect to (w.r.t.) P, of an objective function in terms of a risk of loss function $l : \Theta \times \mathbb{X} \to \mathbb{R}$ to be minimized over parameters $\theta \in \Theta$ given (data) samples $\xi \in \mathbb{X}$. Define the worst-case expected risk of loss $R(\theta) := \mathbb{E}_{P^*(\theta)}[l(\theta,\xi)] = \sup_{P\in\mathscr{P}}\{L_P(\theta)\}$, which maximizes the risk of loss L_P over a well-defined set of measures $\mathscr{P}$. This set typically takes the form $\mathscr{P} = \{P \,|\, D(P,P_b) \le \rho, \int dP(\xi) = 1, P(\xi) \ge 0\}$, where $D(\cdot,\cdot)$ is a distance function on a set or space of probability distributions on $\mathbb{X}$ and where the constraints limit the feasible candidates to be within a distance ρ of a base P_b. We seek parameters $\theta^*_{rob} \in \Theta$ that, for a given $\mathscr{P}$, solve the DRO problem formulated as

$$\theta^*_{rob} = \arg\min_{\theta\in\Theta} \Big\{ R(\theta) \Big\} = \arg\min_{\theta\in\Theta} \Big\{ \sup_{P\in\mathscr{P}} \{L_P(\theta)\} \Big\}. \tag{1}$$

Simulation-based optimization represents an important approach to solving decision-making problems in which nonparametric input model uncertainty is a predominant concern. This uncertainty arises when only a finite set of observations $\mathscr{N} = \{\xi_n, n = 1,\ldots,N\}$ are available to characterize the inputs of the simulation model that estimates the risk of loss function l. A confidence interval (CI) constructed for the expected risk of loss using as an input model the equal-weight *empirical* distribution $U_N = \{1/N\}$ can provide poor coverage of the true value. A rich literature exists on constructing CIs using boot-strapping methods to incorporate the impact of input model uncertainty (Barton et al. 2014). Lam (2019) shows how ϕ-divergence balls centered at $P_b = U_N$ with appropriately chosen radius ρ can construct robust risk of loss $R(\theta)$ as in (1) to obtain an asymptotically valid CI for $L_{P_0}(\theta)$ (for a fixed θ), where P_0 is the true unknown distribution generating the input samples $\mathscr{N}$. In terms of simulation optimization, the equal-weight empirical distribution U_N over $\mathscr{N}$ is the nonparametric maximum likelihood estimator (Owen 2001) of the (unknown) distribution underlying the datasets, which motivates the standard practice of minimizing the *empirical risk of loss* $L_{U_N}(\cdot)$ over Θ. The DRO philosophy seeks to extend the input model uncertainty analysis to simulation optimization by instead picking θ^*_{rob} as the best parameter. In practice, estimating with DRO formulations (1) amounts to dynamically re-weighing the data using the solution $P^*(\theta)$ to the inner maximization at each parameter $\theta \in \Theta$. Unlike the equal emphasis placed by $L_{U_N}(\theta)$ on all observed data, the $R(\theta)$ in DRO formulations sets these weights to emphasize data that experience high risk of loss at θ. Hence, this approach explicitly treats the ambiguity in the identity of P_0, since in general $U_N \neq P_0$.

This problem is also studied in the statistical learning setting where the best model parameters θ of a statistical model is sought given only a finite *training* dataset $\mathscr{N}$ and the model is then used for inference over other test datasets, all of which are typically assumed to be identically distributed. In real-world settings, the training dataset and any dataset to which the trained model is applied are finite sets sampled from the same underlying distribution P_0. While popular model selection techniques, such as cross-validation (Stone 1974)), seek to improve the estimation error between training and testing datasets, they are often computationally prohibitive and lack rigorous guarantees. The DRO formulation (1) with the empirical distribution U_N over the finite training dataset as the base distribution P_b has been proposed as a promising alternative approach. In particular, Blanchet et al. (2016) show for Wasserstein distance metrics that, with an appropriately chosen value of constraint parameter ρ, there exists a $P \in \mathscr{P}$ which leads to the same optimal decision θ^* as P_0. with high probability; and Namkoong and Duchi (2017) establish similar results for ϕ-divergence measures.

Our study concerns efficiently finding solutions of (1) as a fundamental approach to simulation-based decision making. The key obstacle is the minimax formulation, and specifically the inner maximization over probability sets $\mathscr{P}$. In some cases, its solution is explicitly available; e.g., $\mathscr{P}$ constrained by certain instances of the Wasserstein distance, studied by Blanchet et al. (2016) and Sinha et al. (2017), admit an explicit characterization of the robust objective $\mathbb{E}_{P^*(\theta)}[l(\theta,\xi)]$. However, such reductions do not hold in general, and they require solving a convex nonlinear program (Esfahani and Kuhn 2018). Namkoong and Duchi (2017) show that the inner maximization with χ^2-divergence constraints can be efficiently solved, while Hu and Hong (2012) show the same for Kullback-Leibler (KL) divergence. We therefore focus on the general ϕ-divergence distance function $D_\phi(P,P_b) = \mathbb{E}_{P_b}[\phi(\frac{dP}{dP_b})]$, where $\phi(s)$ is a nonnegative convex function taking a value of 0 only at $s=1$. The modified χ^2 and KL divergences are given by $\phi(s) = (s-1)^2$ and $\phi(s) = s\log s - s + 1$, respectively. Define the N-sized vector $P := (p_n)$ and set the base $P_b = U_N$. We then have

$$L_P(\theta) = \sum_{n=1}^{N} p_n l(\theta,\xi_n) \quad \text{and} \quad \mathscr{P} = \Big\{P \,\Big|\, D_\phi(P,U_N) = \frac{1}{N}\sum_{n=1}^{N} \phi(Np_n) \le \rho,\ \sum_{n=1}^{N} p_n = 1, p_n \ge 0, \forall n\Big\}.$$

The robustness of simulation models to input model uncertainty is well established as a critically important problem in simulation-based optimization. Glasserman and Xu (2013) use relative entropy (equivalent to KL divergence) to constrain model distance in studying a portfolio optimization problem with a convex objective function. They employ a parametric model for the distributions in $\mathscr{P}$ and propose a simulation approach tailored to the formulation to characterize worst-case model errors in portfolio allocation; they then apply their approach to various problems including robust portfolio risk measurement. Hu and Hong (2012) study DRO formulations based on distance functions defined by KL divergence arising from input model uncertainty in financial decision-making problems. They show that when the empirical loss $\mathbb{E}l(\theta,\xi)$ is convex, the minimax formulation can be reduced to a single layer minimization optimization problem with a moderately augmented decision space. This is in general agreement with previous work by Ben-Tal et al. (2013) for general D_ϕ-constrained DRO problems, who take a similar Lagrangian dual algorithm approach to reduce the inner concave problem to a convex minimization problem. Hu and Hong (2012) cover popular risk measures such as conditional value-at-risk (CVaR), but importantly their analysis also extends to special cases of nonconvex risk measures, such as value-at-risk (VaR), that can be formulated as chance-constraints. Many simulation optimization problems of interest in theory and practice, however, involve more general (nonconvex) objective functions, including those based on observed human behaviors exhibiting greater risk adversity to losses than to gains (Kahneman and Tversky 1979). Of particular interest herein are optimization problems based on cumulative prospect theory in which the objective function is concave for gains, convex for losses, and steeper for losses than for gains (Tversky and Kahneman 1992). He and Zhou (2011) formulate and derive an analytical treatment of single-period portfolio choice optimization within the context of cumulative prospect theory and a financial market consisting of one risky asset and one riskless asset. Our approach to solving (1) in this paper extends to such general nonconvex problems, and we demonstrate the benefit of our approach using numerical experiments over a robust version of this nonconvex portfolio optimization problem. Our DRO formulation includes a more general class of optimization problems and a more general class of distances functions than Hu and Hong (2012) and Glasserman and Xu (2013), as well as a more general class of minimax optimization problems and a higher dimensional simulation-based decision making approach than He and Zhou (2011).

Our DRO framework utilizes a stochastic approximation approach to iterating the decision parameter θ in Θ and subsumes the inner problem into that of estimating the gradient of the robuse loss $R(\theta)$. Since $R(\theta)$ is an extreme value function (that is, it is the optimal value of a maximization problem), unbiased estimation of its value and derivatives at any θ requires an expensive computation over an optimization problem of large dimension N. This is indeed the approach of Namkoong and Duchi (2017). We devise in this paper a new small-sample stochastic estimator of the gradient of $R(\theta)$ that is unbiased. Our construction of this estimator is based on a novel approach called Multi-level Monte Carlo (MLMC) that has been recently introduced in the research literature by Giles (2008) to eliminate the bias exhibited in typical numerical methods for stochastic problems. This technique has been adapted to various unconstrained stochastic optimization contexts by Blanchet and Glynn (2015) and Blanchet, et al. (2017). The strategy is to randomize the choice of the subsample size M_t over which the estimate $\nabla_\theta \hat{R}(\theta)$ is constructed in each iteration t in a manner that allows for $M_t = N$ only with a small probability. While MLMC randomization scheme of Giles potentially incurs a high computational cost and the randomized M_t imply a possibly larger variance for the gradient estimator, our careful analysis in Theorems 4 and 5 addresses these two competing concerns. Moreover, such results provide the range of randomization parameter values that balances these competing objectives, leading to an efficient unbiased estimator of $\nabla_\theta R(\theta)$. We then establish in Theorem 6 convergence of a stochastic gradient descent (SGD) algorithm with this gradient estimator for nonconvex smooth objectives $l(\theta,\xi)$ by exploiting standard tools.

A superficially similar approach was independently and simultaneously developed by Levy et al. (2020). Our approach, however, fundamentally differs from their procedure in that we are the first to solve the DRO formulation by assembling the mini-batch through sampling subsets M_t *without replacement* from the training dataset, whereas in strong contrast sampling *with replacement* is employed in Levy et al. (2020). While our general theory in the without-replacement sampling approach is harder to establish because the individual samples in the mini-batch are not distributionally independent, we are able to provide a crucial bound on the variance experienced by our gradient estimator that further leads to establishing stronger theoretical results on the convergence of our algorithm. On the other hand, Levy et al. (2020) show that the variance of their $R(\theta)$ estimator may not vanish with increasing batch size, leading to poorer guarantees on convergence. Hence the estimation is significantly improved for the same computational effort by only considering unique samples ξ and the corresponding losses $l(\theta,\xi)$ in approximating $R(\theta)$. Indeed, a set of size M sampled with replacement has on average only about $N(1-e^{-M/N})$ unique support points, and thus estimation using with-replacement sampled $\mathscr{M}$ may lead to wasted computational effort.

In Ghosh et al. (2021), we present an alternative scheme to tackle the efficient estimation of an optimal solution to (1). There, in each iteration, the gradient $\nabla_\theta R(\theta)$ is estimated with a bias, and the algorithm parameters are controlled in a way that ensures the mean squared error of the gradient estimates reduces to zero as iterates progress, which ensures convergence. The method presented here, on the other hand, ensures that the gradient $\nabla_\theta R(\theta)$ is efficiently estimated *without bias* in each iteration. This yields a standard SGD form of (1) w.r.t. the variables θ, and in turn the convergence of the method is analyzed by standard SGD theory. In particular, when losses $l(\theta,\xi)$ are strongly convex, this MLMC method will produce ε-optimal solutions with the fastest work-complexity of $O(\varepsilon^{-1})$ for any desired $\varepsilon > 0$, as established by SGD theory. In Ghosh et al. (2021), a similar rate of convergence is established, but with a caveat that this is achievable only for $\varepsilon \geq \underline{\varepsilon} = O(1/N)$, a small price paid for the bias in estimation of $\nabla_\theta R(\theta)$ in every iteration.

We conduct a broad collection of numerical experiments over the nonconvex portfolio allocation problems grounded in prospect theory that were mentioned earlier. In particular, we compare the performance of our algorithm against those proposed by Namkoong and Duchi (2017), who consider full dataset computations in each iteration to estimate $\nabla_\theta R(\theta)$. The experiments show that our algorithm maintains a performance level similar to this method while being far more economical in the number of samples accessed for the gradient computation steps. In addition, our results shed important light on the fact that the choice of the best ball size parameter ρ can interact non-trivially with the problem formulation and dataset characteristics. While useful progress has been made (Lam 2019; Blanchet et al. 2016) in predicting the appropriate choice of ρ that maximizes coverage of the true unknown input model in certain subclasses of decision problems, a general approach for simulation optimization problems remains an open question of high interest.

2 ALGORITHM AND ANALYSIS

We now present our subgradient descent approach for efficiently solving the general DRO minimax optimization problem with our algorithm in Figure 1, comprising SGD-like iterations for the outer minimization problem in (1):

$$\theta_{t+1} \;=\; \theta_t - \gamma_t \nabla_\theta \hat{R}_t(\theta_t) \;=\; \theta_t - \gamma_t G_t(\theta_t), \tag{2}$$

where γ_t is the step size, $\hat{R}_t(\cdot)$ is a stochastic approximation of the robust loss $R(\cdot)$ from the inner maximization over D_ϕ-constrained $\mathscr{P}$, and $G_t(\theta_t) := \nabla_\theta \hat{R}_t(\theta_t)$. This view of (1) allows us to depart from the convex-concave formulations of Ben-Tal et al. (2013) and Hu and Hong (2012), and to consider nonconvex risk of loss functions l, as long as the subgradient $G_t(\cdot)$ approximates the gradient $\nabla_\theta R(\cdot)$ sufficiently well. In this section, we consider the inner maximization and outer minimization of our algorithm in Figure 1 followed by a theoretical analysis of this algorithm.

2.1 Inner Maximization

Recall that $R(\theta)$ is the optimal value of the inner maximization problem. Define the set $\Theta_\varnothing := \{\theta : l(\theta,\xi_{n_1}) = l(\theta,\xi_{n_2}),\, \forall n_1,n_2\}$ and, for a small $\varsigma > 0$, define the ς-neighborhood of $\Theta_\varnothing$ as $\Theta_{\varnothing,\varsigma} := \cup_{\theta_o\in\Theta_\varnothing}\{\theta : \|\theta-\theta_o\|_2 < \varsigma\}$. Exploiting Danskin's Theorem, Proposition 1 shows the existence of $\nabla_\theta R(\theta)$ assuming that formulation (1) precludes $\Theta_{\varnothing,\varsigma}$ in order to avoid a degenerate inner maximization objective function that does not depend on the decision variables p_n, in which case the entire feasible set is optimal.

Proposition 1 (Ghosh et al. 2021, Proposition 1) Let the feasible region Θ be compact and assume $\Theta \subseteq \Theta^c_{\varnothing,\varsigma}$, for a small $\varsigma > 0$. Further suppose ϕ in the D_ϕ-constraint has strictly convex level sets, and let $\rho < \bar{\rho}(N,\phi) = \left(1-\frac{1}{N}\right)\phi\left(\frac{N}{N-1}\right) + \frac{1}{N}\phi(0)$. Then: (i) the optimal solution P^* of $R(\theta) = \sup_{P\in\mathscr{P}}\{L_P(\theta)\}$ is unique and the gradient is given by $\nabla_\theta R(\theta) := \sum_{n\in\mathscr{N}} p^*_n(\theta)\nabla_\theta l(\theta,\xi_n)$; and (ii) for all ρ, the $\nabla_\theta R(\theta)$ is a sub-gradient of $R(\theta)$.

We then construct the estimate $\hat{R}_M(\theta)$ in (2) from the inner maximization problem restricted only to a subset $\mathscr{M}$ of size $|\mathscr{M}| = M$ of the full dataset $\mathscr{N}$ of size N. Defining $[N] := \{1,\ldots,N\}$, $P := (p_m)$ of dimension M and objective coefficients $z_m := l(\theta,\xi_m)$, consider

$$\hat{R}_M(\theta) \;=\; \max_{P=(p_m)} \sum_{m\in\mathscr{M}} p_m z_m \qquad \text{s.t.} \sum_{m\in\mathscr{M}} \phi(Mp_m) \leq M\rho_M,\; \sum_{m\in\mathscr{M}} p_m = 1,\, p_m \geq 0, \tag{3}$$

where the uncertainty radius $\rho_M = \rho + \eta_M$ now changes with the subsample size M, motivated by Theorem 2 discussed below, and where $\eta_M = c(1/M - 1/N)^{(1-\delta)/2}$ for small positive constants c,δ. Suppose $P^*_M(\theta)$

```
1:  procedure GILES SSD($\gamma_t, r, \theta_0$)
2:
3:      for  $t = 1, 2, \ldots$  do
4:          Sample $\tau_t$ from truncated geometric $q_k$
5:          Sample set $\mathscr{M}_{\tau_t} \subseteq \mathscr{N}$ uniformly without replacement
6:          Solve problems (3) to obtain estimates $\nabla_\theta \hat{R}_\tau(\theta_t)$, $\nabla_\theta \hat{R}_{\tau,l}(\theta_t)$ and $\nabla_\theta \hat{R}_{\tau,r}(\theta_t)$
7:          Sample singleton $\{\xi_{1,t}\}$ and set $\nabla_\theta \hat{R}_1(\theta) \leftarrow \nabla_\theta l(\theta_t, \xi_{1,t})$
8:          Set $\Delta_{\tau_t}(\theta_t)$, $G_t(\theta_t)$ from (4)
9:          Set $\theta_{t+1} \leftarrow \theta_t - \gamma_t G_t(\theta_t)$
10:         If stopping criterion satisfied, then
11:             Set $T \leftarrow t$ and break
12:         Increment $t \leftarrow t + 1$
13:     return $\theta_T$
14: end procedure
```

(a) Outer Minimization, where input parameters include step size γ, sampling parameter r, initial iterate θ_0.

```
1:  procedure INNERMAX($\mathscr{Z}, \mathscr{M}, \rho$)
2:      $M \leftarrow |\mathscr{M}|$, base $P_b = \{\frac{1}{M}, \forall m \in \mathscr{M}\}$
3:      $\bar{z} \leftarrow \max_m \{z_m \mid z_m \in \mathscr{Z}\}$
4:      $\mathscr{M}' \leftarrow \{m \in \mathscr{M} : z_m = \bar{z}\}$ and $M' \leftarrow |\mathscr{M}'|$
5:      $P' \leftarrow \{\frac{1}{M'} \mathbb{I}\{m \in \mathscr{M}'\}, \forall m \in \mathscr{M}\}$
6:      If $D_\phi(P^*, P_b) \le \rho$ then
7:          $P^* \leftarrow P'$ and return $P^*$
8:      for $\alpha \in [0, \bar{\alpha}]$  do
9:          for  $\lambda \in [\underline{\lambda}, \bar{\lambda}]$  do
10:             $\mathscr{M}' \leftarrow \{m \mid \lambda \le z_m - \alpha\phi'(0)\}$
11:             $P' \leftarrow \left\{\frac{1}{M}(\phi')^{-1}(\frac{z_m - \lambda}{\alpha}), m \in \mathscr{M}'\right\}$
12:             If $\sum_m p'_m = 0$, then
13:                 $P^*(\alpha) \leftarrow P'$, and break
14:         If $D_\phi(P^*(\alpha), P_b) = \rho$, then
15:             $P^* \leftarrow P^*(\alpha)$ and break
16:     return $P^*$
17: end procedure
```

(b) Inner Maximization, where input parameters include loss values $\mathscr{Z}$, subsampled support $\mathscr{M}$, D_ϕ constraint ρ.

Figure 1: Giles sampled subgradient descent algorithm.

is an optimal solution to (3). Then a valid subgradient for $\hat{R}_M(\theta)$ is obtained as an expression analogous to that in Proposition 1(i) under appropriate substitutions w.r.t. $P^*_M(\theta)$ and $\mathscr{M}$. This procedure is used in Step 6 of our algorithm in Figure 1a(a) to obtain each of the gradient estimators of the subsampled sets $\mathscr{M}_\tau$, $\mathscr{M}_{\tau,l}$ and $\mathscr{M}_{\tau,r}$.

Next, we consider the exact solution to the inner maximization problem presented in our algorithm in Figure 1b(b). This procedure obtains the optimal primal and dual variables that solves (3) for various ϕ-divergences by solving the equivalent Lagrangian formulations. This procedure is thoroughly analyzed in Ghosh et al. (2021), who take the same basic approach for such problems pursued by others (see, e.g., Ben-Tal et al. (2013), Ghosh and Lam (2018), Namkoong and Duchi (2017), Hu and Hong (2012)) and adapt it to the inner maximization formulation over D_ϕ-constrained $\mathscr{P}$. Ghosh et al. (2021), Proposition 2 establish a worst-case computational complexity bound of $O(M \log M + (\log(\frac{1}{\varepsilon}))^2)$ for finding an ε-optimal solution to (3) when applying our algorithm in Figure 1b(b) to any ϕ-divergence. Since the machine-precision ε is set to a fixed arbitrarily small value independent of any other parameter of the formulation or algorithm (e.g., M, N, ρ), we follow Ghosh et al. (2021) and assume that our algorithm in Figure 1b(b) returns the exact unique solution $(P^*, \alpha^*, \lambda^*)$ to (3) with computational complexity bounded by $O(M \log M)$.

Now, we consider the bias and variance induced by the approach of subsampling the full support using the subgradient approximation $\nabla_\theta \hat{R}_M(\theta)$ to the true gradient $\nabla_\theta R(\theta)$. Further let $\mathbb{E}_M$ and $\mathbb{P}_M$ denote expectation and probability w.r.t. the random set $\mathscr{M}$, respectively.

Theorem 2 (Ghosh et al. 2021, Corollary 1, Theorem 4) Suppose the ϕ-divergence satisfies uniformly for all s and $\zeta < \zeta_0$ the continuity condition $|\phi(s(1+\zeta)) - \phi(s)| \le \kappa_1 \zeta \phi(s) + \kappa_2 \zeta$, for constants $\zeta_0, \kappa_1, \kappa_2 > 0$. Further suppose the assumptions of Proposition 1 hold. Define $\eta_M = c(\frac{1}{M} - \frac{1}{N})^{(1-\delta)/2}$ for small constants $c, \delta > 0$, and set the D_ϕ-target in (3) to be $\rho_M = \rho + \eta_M$. Then, for the estimate $\nabla_\theta \hat{R}_M(\theta)$ calculated over the *sampled-without-replacement* set $\mathscr{M}$ of size M, we have

$$||\mathbb{E}_M[\nabla_\theta \hat{R}_M(\theta)] - \nabla_\theta R(\theta)||_2^2 = O(\eta_M^2) \quad \text{and} \quad \mathbb{E}_M[||\nabla_\theta \hat{R}_M(\theta) - \nabla_\theta R(\theta)||_2^2] = O(\eta_M^{2/(1-\delta)}).$$

Theorem 2 shows that a squared bias of order $\eta_M^2 \approxeq M^{-1}$ is incurred when a fixed subsample of size M is used in creating the gradient estimate. The variance incurred is also of a similar order, and hence so is the mean squared error. Levy et al. (2020), on the other hand, are not able to establish the variance bound for their MLMC-based procedure for DRO, which leads to significantly slower convergence of their procedure.

These results are broadly in line with what is expected for sample average approximated solutions of stochastic optimization problems. It is important to note, however, that the result in Theorem 2 is distinct because of the set $\mathscr{M}$ being sampled uniformly-without-replacement from $\mathscr{N}$, rather than the more easily analyzed with-replacement sampling procedure that is broadly followed. The proof in Ghosh et al. (2021) requires analogous probability tools for without-replacement sampling to establish the result.

2.2 Outer Minimization

The bound on the bias presented in Theorem 2 vanishes only as $M \uparrow N$ as the iterate $t \to \infty$. This motivated the approach of Ghosh et al. (2021) to progressively reduce the bias by increasing the subsample set size M_t as t grows, consequently also increasing the computation time. The maximum size N is then hit after a large number of iterations T, at which point their algorithm switches to the deterministic optimization of Namkoong and Duchi (2017). Ghosh et al. (2021) establish the convergence of their algorithm and show how growth parameters of the per-iterate samples M_t can be carefully chosen to minimize the overall computation time needed to converge to the optimal solution θ^*_{rob}.

We propose here a fundamentally different approach to solving the outer minimization that eliminates bias without the concomitant increase in computation by adding an MLMC randomization step; see Giles (2008) and Blanchet and Glynn (2015) for a basic introduction to MLMC randomization. For ease of exposition, we henceforth assume the training set is such that $N = 2^K + 1$ for an integral value of K, noting that the general case is easily handled by appropriately adjusting our algorithm in Figure 1a(a) above. Let τ be a discrete random variable taking values in $[K] := \{1, \dots, K\}$. The random variable τ is sampled geometrically using the probability mass function

$$q_k := P(\tau = k) = r^{k-1} \frac{1-r}{1-r^{K+1}}, \qquad k \in [K].$$

Let $M_\tau = 2^\tau$ be a subset size associated with τ, and $\mathscr{M}_\tau$ the corresponding subset sampled uniformly without replacement from the full dataset $\mathscr{N}$. Partition $\mathscr{M}_\tau$ into two equal-sized subsets $\mathscr{M}_{\tau,l}$ and $\mathscr{M}_{\tau,r}$, each of size $2^{\tau-1}$. To simplify notation, we denote the robust loss calculated over $\mathscr{M}_\tau$, $\mathscr{M}_{\tau,l}$ and $\mathscr{M}_{\tau,r}$ by $\hat{R}_\tau$, $\hat{R}_{\tau,l}$ and $\hat{R}_{\tau,r}$, respectively. Define

$$\Delta_\tau(\theta) := \nabla_\theta \hat{R}_\tau(\theta) - \frac{\left(\nabla_\theta \hat{R}_{\tau,l}(\theta) + \nabla_\theta \hat{R}_{\tau,r}(\theta)\right)}{2} \qquad \text{and} \qquad G(\theta) := \nabla_\theta \hat{R}_1(\theta) + \frac{\Delta_\tau(\theta)}{q_\tau}, \tag{4}$$

where $\nabla_\theta \hat{R}_1(\theta)$ is the gradient computed from a singleton subset, as per the definition (3) with $M = 1$. (Note that the inner maximization in this one-sample set $\{\xi_1\}$ is degenerate and hence $\nabla_\theta \hat{R}_1(\theta) = \nabla_\theta l(\theta, \xi_1)$.) Then the estimator $G(\theta)$ provides an unbiased estimate for $\nabla_\theta R(\theta)$, in the style of Giles (2008), under very general conditions on $l(\cdot, \xi)$. In fact, this allows our algorithm in Figure 1 above to be used under general conditions when the risk of loss functions $l(\cdot, \xi_n)$ are nonconvex.

2.3 Theoretical Analysis

We next establish various mathematical properties for our DRO approach, including properties of $G(\theta)$ and convergence. This analysis provides theoretical justification for our algorithm and its parameter settings. To start, we can easily establish the unbiasedness of the estimator $G(\theta)$ given in (4). Let $\mathbb{E}_\tau$ denote expectation under the joint distribution of the randomizing variable τ and the subsampled set $\mathscr{M}_\tau$.

Proposition 3 We have that $\mathbb{E}_\tau[G(\theta)] = \nabla_\theta R(\theta)$.

Proof. Recall that $\hat{R}_M(\theta)$ is the robust loss estimate constructed by subsampling an M-sized subset of the training set. We then obtain

$$\mathbb{E}_\tau[G(\theta)] = \sum_{k=1}^{K} q_k \cdot \mathbb{E}\left(\frac{\Delta_k(\theta)}{q_k} + \nabla_\theta \hat{R}_1(\theta)\right) = \mathbb{E}\nabla_\theta \hat{R}_1(\theta) + \sum_{k=1}^{K}\left(\mathbb{E}\nabla_\theta \hat{R}_{2^\tau}(\theta) - \mathbb{E}\nabla_\theta \hat{R}_{2^{\tau-1}}(\theta)\right) = \nabla_\theta R(\theta),$$

where the telescoping sums in the last equality cancel out to leave only the leading term for $k = K$, with $\hat{R}_{2^K}(\theta) = \hat{R}_N(\theta) = R(\theta)$ by assumption. □

Observe that the computation time in sampling the zero-bias estimator now depends on the randomizer τ because we solve inner maximization problems of size $M_\tau = 2^\tau$. Let T denote the total computation time involved in constructing one replication of the Giles estimator $G(\theta)$. We then establish the following result.

Theorem 4 When $r < 1/2$, the expected computation time is bounded by a quantity independent of the dataset size $N > 1$: $\mathbb{E}_\tau[T] \leq \frac{4C'r(1-r)\log 2}{(1-2r)^2}$, where constant C' is finite.

Proof. The computation time in calculating Δ_k for a fixed k lies mainly with estimating the solutions to the four inner maximization problems over subsets of size up to $M_k = 2^k$ within ε-accuracy. Recall from Ghosh et al. (2021), Proposition 2 that the computation time in obtaining approximations to a problem of size M is $O(M \log M)$. Hence, we have

$$\begin{aligned}
\mathbb{E}_\tau T &\leq C' \sum_{k=1}^{K} q_k \cdot 2^k k \log 2 = \frac{C'(1-r)\log 2}{1-r^K} \sum_{k=1}^{K} k(2r)^k \\
&\leq 2C'(1-r)\log 2 \sum_{k=1}^{\infty} k(2r)^k = 2C'r(1-r)\log 2 \sum_{k=1}^{\infty} \frac{d}{dr}((2r)^k) = 2C'r(1-r)\log 2 \, \frac{d}{dr}\left(\sum_{k=1}^{\infty}(2r)^k\right) \\
&= 2C'r(1-r)\log 2 \, \frac{d}{dr}\left(\frac{1}{1-2r}\right) = \frac{4C'r(1-r)\log 2}{(1-2r)^2}.
\end{aligned}$$

Here, the first inequality uses the above $O(M \log M)$ computational complexity result. The second inequality holds because $K > 1$ and $r < 1/2$ by assumption. The interchange of the derivative and the infinite sum is justified by the convergence of the sum (because $2r < 1$). □

It is also important to ensure that the injection of extraneous randomness via τ does not result in a very large variance of the estimator $G(\theta)$. We now establish that the variance of our Giles estimator is bounded given the variance result in Theorem 2.

Theorem 5 Suppose the assumptions of Theorem 2 hold, and further suppose that, for each ξ, the loss function $l(\cdot,\xi)$: (i) is L-Lipschitz smooth, i.e., has gradients $\nabla_\theta l(\cdot,\xi)$ that are L-Lipschitz continuous; and (ii) has a finite minima $l^*(\xi)$ over $\theta \in \Theta$. Then, with $r \in (1/4, 1/2)$, the variance of $G(\theta)$ is bounded from above by a quantity independent of the dataset size $N > 1$:

$$\mathbb{E}_\tau\left[\|G(\theta) - \nabla_\theta R(\theta)\|_2^2\right] \leq C(r) < \infty.$$

Proof Sketch of Theorem 5: Space considerations limit us to the brief proof sketch below; please refer to Ghosh and Squillante (2020) for the complete proof. We will first show that the result in the statement can be derived if $\mathbb{E}\left[\|\Delta_k\|_2^2\right] \leq C\,(2^{-k} - 2^{-K})^2$ for a fixed k and a constant $C < \infty$. Noting that

$\|a+b\|^2 \leq 2(\|a\|^2+\|b\|^2)$, we have

$$\begin{aligned}
\mathbb{E}_\tau\|G(\theta)\|_2^2 &= \sum_{k=1}^{K} q_k \ \mathbb{E}\left\|\frac{\Delta_k}{q_k}+\nabla_\theta \hat{R}_1(\theta)\right\|_2^2 \leq 2\mathbb{E}\|\nabla_\theta \hat{R}_1(\theta)\|_2^2+2\sum_{k=1}^{K}\frac{1}{q_k}\mathbb{E}\,\|\Delta_k\|_2^2 \\
&\leq 2\mathbb{E}\|\nabla_\theta \hat{R}_1(\theta)\|_2^2+\frac{2(1-r)}{1-r^K}\sum_{k=1}^{K} r^{-k}\left(\frac{1}{2^k}-\frac{1}{2^K}\right)^2 \leq 2\mathbb{E}\|\nabla_\theta \hat{R}_1(\theta)\|_2^2+\frac{2(1-r)}{1-r^K}\sum_{k=1}^{K} r^{-k}2^{-2k} \\
&\leq 2\mathbb{E}\|\nabla_\theta \hat{R}_1(\theta)\|_2^2+\frac{4(1-r)}{1-1/(4r)} = C(r).
\end{aligned}$$

The final inequality follows from the assumption that $K > 1$ and $r \in (1/4, 1/2)$, whereas the first term is bounded via the variance bound from Theorem 2. The order of the variance $\mathbb{E}\left[\|\Delta_k\|_2^2\right]$ of the individual difference terms in the Giles estimator is established by relating the variance of the gradients $\nabla_\theta \hat{R}_M(\theta)$ to those of the corresponding robust loss estimates $\hat{R}_M(\theta)$ and using the gradient-smoothness assumption in the statement of the proof; please refer to Ghosh and Squillante (2020) for further details. □

Now, we turn to analyze the convergence of our algorithm in Figure 1, for which a common requirement is that the objective $R(\theta)$ be Lipschitz smooth. The gradient $\nabla_\theta R(\theta)$ of the optimal value of an optimization problem is in general not Lipschitz if the objective function is Lipschitz. Consider, for example, a linear objective $l(\theta, \xi_n) = \theta^t \xi_i$, which implies $R(\theta) = \max_p \sum_i p_i \theta^t \xi_i$. When maximized over a *polyhedral* constraint set (e.g., the probability simplex constraints of (1)) the 0-Lipschitzness of the l are not preserved for R, because in this case the optimal solutions P^* are picked from the discrete set of vertices of the polyhedron and thus $\nabla_\theta R(\theta)$ is piecewise discontinuous.

Our assumptions from Proposition 1 yield an inner maximization with a non-zero linear objective over a *strictly convex* feasible set. The desired smoothness can then be obtained with some additional conditions on the loss functions $l(\theta, \xi_i)$. Ghosh et al. (2021), in Proposition 5, note that the Lipschitzness of $\nabla_\theta R(\theta)$ follows from the Hessian $\nabla_\theta^2 l(\theta, \xi)$ being bounded in norm, which is often satisfied by the optimization problems of interest.

The combination of the Lipschitzness of the robust loss function $R(\theta)$, the finiteness of the variance of the Giles gradient estimator $G(\theta)$, and the finite expected computation time to obtain the estimator enables us to now apply the standard SGD convergence machinery in establishing the convergence of (2) to first-order optimal solutions. We exploit the following result adapted to our problem and its setting.

Theorem 6 (Bottou et al. 2016, Theorem 4.9) Suppose the robust loss objective $R(\theta)$ satisfies: (i) A lower bound R_{inf} exists for the robust loss function $R(\theta) \geq R_{\text{inf}},\ \forall \theta \in \Theta$; (ii) The gradient $\nabla_\theta R(\theta)$ is L-Lipschitz. Further suppose the estimator $G(\theta)$ of the gradient $\nabla_\theta R(\theta)$ is unbiased and has variance bounded above by a constant $C < \infty$. Choosing the step size sequence γ_t to satisfy $\sum_t \gamma_t \to \infty$ and $\sum_t \gamma_t^2 < \infty$, we then have

$$\liminf_{t\to\infty} \mathbb{E}\left[\|\nabla_\theta R(\theta_t)\|_2^2\right] = 0.$$

3 NUMERICAL EXPERIMENTS

In this section we present numerical experiments evaluating our new DRO approach and its above theoretical properties applied in the context of optimal nonlinear portfolio allocation. Motivated by input uncertainty, Section 3.1 presents a DRO version of this general nonlinear problem. Section 3.2 describes empirical results that compare our approach against previous related work in the research literature.

3.1 Nonlinear Utility Portfolio Optimization

Since the seminal work of Markowitz (1952), optimal portfolio allocation has been a foundational problem formulation for decision making under uncertainty. In this formulation, a decision maker chooses the fractional allocation $\theta \in \mathbb{R}^d$ of a portfolio over d financial instruments, where $\xi \in \mathbb{R}^d$ represents the (single-period) outcome of the relative increase in the value of the instruments. The allocation decision θ requires careful balancing of the risk of large losses against the expected profitability of the portfolio $\theta^\top \xi$. Supposing $\xi \sim P_0$, the original formulation of Markowitz chose the θ^* that maximized $\mathbb{E}_{P_0}[\theta^\top \xi]$ subject to an upper bound on the variance $\mathrm{Var}_{P_0}[\theta^\top \xi]$. Under specific assumptions on the distributional form of P_0, this formulation has an equivalent convex representation. For general P_0, a relaxation that takes a soft-penalty form of the variance constraint into the objective can be further well-approximated using convex optimization techniques. Over time, this version of the portfolio allocation problem has been generalized to consider other measures of risk, including VaR (Pflug 2000) and coherent risk measures (Artzner et al. 1999) such as CVaR.

The uncertainty in specifying the input distribution can have a significant impact on the actual portfolio performance. The instruments ξ in a typical real-world scenario are significantly correlated, and usually there is a lack of adequate data to obtain a well-fit estimate of the high-dimensional joint distribution P_0. Consequently, the portfolio allocation problem has been increasingly studied from the DRO perspective, where the worst-case loss $R(\theta)$ suffered over an appropriately defined uncertainty ball is minimized. (Indeed, Artzner et al. (1999) show that many coherent risk measures can be characterized as the robust loss $R(\theta)$ arising from a carefully defined loss $l(\theta, \xi)$ and ball $\mathscr{P}$.) As described in the introduction, this includes the work of Glasserman and Xu (2013) and Hu and Hong (2012) based on portfolio risk minimization over KL divergence balls. Notably, in addition to many convex risk measures, Hu and Hong (2012) also model nonconvex VaR objectives.

Our approach in this paper explicitly allows for the loss function $l(\theta, \xi)$ to be nonconvex in θ, and in Theorem 6 we establish convergence of our algorithm to local minima under minimal regularity conditions. To demonstrate the power of our DRO approach, we consider a large class of nonconvex portfolio allocation objectives that arise from the field of prospect theory in economics. Kahneman and Tversky (1979) started this body of work, awarded the 2002 Nobel prize in economics, by proposing the study of S-shaped utility functions of financial returns using econometric experiments. They posited that the nonlinear form $l(\theta, \xi) = -\underline{C}(r_0 - \theta^\top \xi)^\alpha \mathbb{I}\{r_0 \geq \theta^\top \xi\} + (\theta^\top \xi - r_0)^\beta \mathbb{I}\{r_0 < \theta^\top \xi\}$ conforms better to actual human preferences, where r_0 is a reference return value (such as the risk-free return rate). The parameter $\underline{C}$ is typically greater than 1, encoding a larger disutility for losses compared to gains. In addition, $\alpha, \beta < 1$ which implies that users are more sensitive to losses (and gains) around the reference (certain) rate of return than to larger lower-probability losses (and gains) in the tails of the return distribution. Tversky and Kahneman (1992) explain that $\alpha = \beta = 0.88$ and $\underline{C} = 2.25$ adequately match experimental observations. Our experiments will therefore be performed using these parameter values for the S-shaped utility function given above for $l(\theta, \xi)$. He and Zhou (2011) analyze the S-utility portfolio allocation problem that minimizes the expectation $\mathbb{E}[l(\theta, \xi)]$ for the two-instrument case (e.g., risk-free vs. an index fund) and show that even this small problem requires non-trivial conditions to be satisfied for the well-posedness of this problem.

3.2 Algorithmic Performance

A large collection of numerical experiments were conducted to empirically evaluate our new unbiased Giles estimator sampled subgradient descent algorithm (*Giles*) in comparison with the full-support gradient algorithm (*FG*) of Namkoong and Duchi (2017). Each numerical experiment solves the DRO formulation (1) of minimizing the worst expected S-shaped portfolio return utility for the nonlinear $l(\theta, \xi)$ defined above over a χ^2-divergence ball, namely D_ϕ with $\phi(s) = (s-1)^2$. The portfolio allocations θ are further constrained to be within the simplex $\{\sum_{i=1}^d \theta_i = 1,\ \theta_i \geq 0\}$. The reference return r_0 is assumed to be a risk-free (i.e., zero-variance) return of 0.01 percent per annum. This instrument is taken as an option in the portfolio of instruments ξ. We consider portfolios of size $d = 20$, 50 and 75, comprising stocks picked equally from each of the constituent sectors of the S&P 500 index. The base input model P_0 is constructed from two datasets: 1258 daily returns recorded for the calendar years 2008 to 2012; and 1008 daily returns recorded for the calendar years 2018 to 2021. In every experiment, the dataset was split to randomly select 90% of

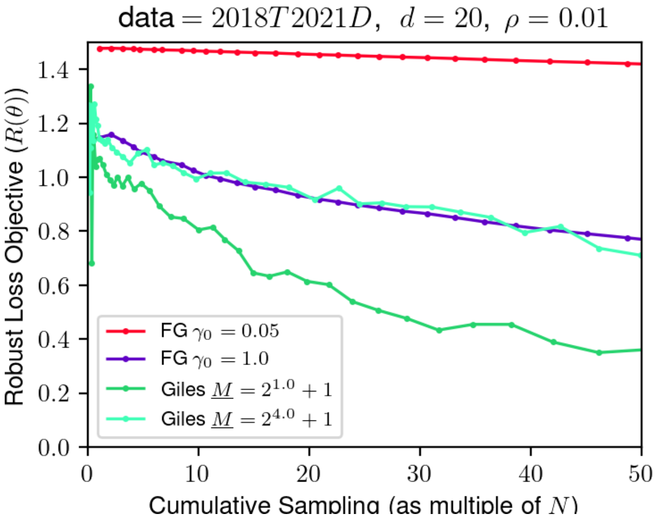

Figure 2: Comparison of Giles and FG methods on averaged robust loss $R(\theta)$ (100 replications) over cumulative samples expressed as multiples of N. Two settings of the minimum sampleset size (denoted $\underline{\mathrm{M}}$) for Giles method and two settings of maximum step-length γ_0 for FG method are presented.

Table 1: Comparison of estimated optimal robust loss $R(\theta^*)$ over various d, ρ settings, presented as a 95% confidence interval over 100 replications.

Dataset	2008-to-2012		
ρ	$d = 20$	$d = 50$	$d = 75$
0.0001	0.17±0.08	0.36±0.09	0.43±0.09
0.001	0.20±0.09	0.40±0.11	0.49±0.1
0.01	0.11±0.04	0.40±0.06	0.52±0.06
0.1	0.10±0.01	0.58±0.05	0.77±0.04
1	0.20±0.04	1.22±0.07	1.80±0.09
Dataset	2018-to-2021		
ρ	$d = 20$	$d = 50$	$d = 75$
0.0001	0.37±0.11	0.51±0.07	0.54±0.05
0.001	0.41±0.12	0.56±0.08	0.60±0.06
0.01	0.28±0.1	0.57±0.07	0.67±0.06
0.1	0.15±0.03	0.76±0.07	0.93±0.05
1	0.23±0.05	1.60±0.13	2.03±0.11

the data as the empirical distribution P_0, respectively yielding $N = 1132$ and $N = 907$ for the two datasets. Results are presented from a collection of 100 replications using this permutation procedure. The initial θ_0 for each method is sampled from $U[0,1]^d$, projected to the constraint simplex.

Blanchet et al. (2016) and Namkoong and Duchi (2017) consider at length the question of setting the parameter ρ in (1), providing a broad guideline that $\rho = O(d/N)$ for a class of statistical binary classification models. This suggests for our datasets that $\rho = O(0.01)$. We therefore experiment with $\rho \in \{0.0001, 0.001, 0.01, 0.1, 1.0\}$, covering two orders of magnitude in either direction.

Figure 2 presents our representative results for the 2018-to-2021 daily-returns dataset with $d = 20$ instruments allowed in the model, comparing the Giles method with the FG method on the robust loss objective $R(\theta)$ as iterations progress in each method. For consistency, we present our results as a function of the cumulative sample size accessed by the gradient computation steps. Two settings of the minimum sample-set size (denoted $\underline{\mathrm{M}}$) of Giles is plotted against two settings of the initial step-length (γ_0) for the FG method. As expected, FG is not competitive with Giles as a function of the computational effort expended. The Giles method does exhibit added variance, likely because of the added randomization in the gradient computation procedure; moreover, its form includes a subtraction of estimates of the gradient over nested sample sets, which might also contribute to the larger variability. On the other hand, the Giles method is able to make adequate progress even with a small minimum batch-size, indicating that the larger minimum does not necessarily contribute in higher accuracy and hence faster convergence (as a function of total computational effort).

Table 1 presents the estimated optimal robust loss $R(\theta^*)$ observed over the different ρ values for the distributional ball constraint. The three columns present results for three models, each respectively allowing $d = 20$, 50 and 75 instruments in the portfolio. Additionally, results are presented for the two datasets: 2018-to-2021 and 2008-to-2012. The two larger portfolio columns ($d = 50$ and $d = 75$) show an increasing relationship over ρ, with the large values for $\rho = 1$ possibly due to the distributional ball being too large. The behaviour of the smallest portfolio problem is interesting, in that the lowest robust loss is observed for $\rho = 0.1$. Notice further that the confidence intervals grow as ρ shrinks. Recalling that in each replication the P_0 is constituted from 100 random permutations of the input dataset, one might infer from this that the smallest ρ values yield too small of a distributional ball to adequately capture the impact of input uncertainty on this model, thus leading to higher variability in the estimated optimal robust loss $R(\theta^*)$. In addition, one would expect greater hedging opportunities with the larger portfolios.

Giles parameters. Our Giles method requires a diminishing step-size sequence, which we choose to be $\gamma_t = \gamma_0 \cdot 5000/(5000+t)$ with $\gamma_0 = 0.1$. The Giles estimator parameter r has a feasibility range of $(1/4, 1/2)$. Blanchet and Glynn (2015) consider a Giles-based estimator for unconstrained optimization problems with continuously distributed random variables, while we solve the constrained problem (3) using discrete samples from the finite training set. They show that the product $\mathbb{E}[T] \times \mathrm{Var}[G(\theta)]$, denoting the variance of a generalized central limit obeyed by the Giles estimator as a function of the computational budget, is minimized by $r^* = 2^{-3/2}$ which represents the geometric mean of the end points of the feasible region. We therefore use this choice in the experiments for our Giles method.

FG parameters. The step lengths of the FG algorithm are determined by the LBFGS-B algorithm with a maximum of γ_0 for all experiments. This parameter can have a notable impact on the speed of convergence, as our results in Figure 2 illustrate.

Other parameters. In all cases, the inner-maximization formulation is solved to within an ε-accuracy where $\varepsilon = 10^{-7}$. Parameter δ appears prominently in the definition of ρ_M and in defining the bias of the gradient estimation in Theorem 2, but the result requires only that δ be a small positive constant. We find that the numerical experiments are not sensitive to δ and therefore set $\delta = 0.01$ in defining the expanded constraint ρ_M. Each method stops if the average of the robust loss objective values of the last 20 iterates does not improve more than 1% when compared with the average of the previous 80 evaluations.

4 CONCLUSION

This paper presents a new approach to efficiently solve distributionally robust optimization formulations that address nonparametric input model uncertainty in simulation-based decision making problems. We utilize a carefully tuned multi-level Monte Carlo randomization scheme to estimate the gradient of the robust loss objective (the inner maximization in the formulation) to provide a stochastic gradient descent equivalent form of the bi-level optimization problem. This approach handles non-convex simulation objectives, and we exhibit its significant benefits over previous related work on a simulation-based analysis of portfolio optimization by demonstrating via numerical experiments how solutions to a general nonconvex portfolio choice modeling problem arising from cumulative prospect theory are efficiently obtained.

Our results shed important light on the fact that the choice of the best ball size parameter ρ can interact non-trivially with the problem formulation and dataset characteristics. A general approach for determining the best ρ that maximizes coverage of the true unknown input model for simulation optimization problems remains an open question of great interest. Note that in a subsequent analysis of our formulation, we focus on simulation outcomes $l(\theta, \xi)$ where only one copy of the input model $\xi \sim P$ is accessed. This model captures a broad class of decision making under uncertainty problems from diverse fields such as finance (portfolio modeling) and inventory (including newsvendor models, customer choice theory, and so on). A second important avenue of future work lies in the extension of these methods to analyze general simulation optimization objectives, where the simulation model may access multiple copies of the input model P. This yields a harder to solve inner problem; see, for example, Ghosh and Lam (2018).

REFERENCES

Artzner, P. F., J. M. Eber, and D. Heath. 1999. "Coherence Measures of Risk". *Mathematical Finance* 9:203–228.

Barton, R. R., B. L. Nelson, and W. Xie. 2014. "Quantifying Input Uncertainty via Simulation Confidence Intervals". *INFORMS Journal on Computing* 26(1):74–87.

Ben-Tal, A., D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. 2013. "Robust Solutions of Optimization Problems Affected by Uncertain Probabilities". *Management Science* 59(2):341–357.

Blanchet, J. and Goldfarb, D. and Iyengar, G. and Li, F. and Zhou, C. 2017. “Unbiased Simulation for Optimizing Stochastic Function Compositions”. *arXiv preprint* arXiv:1711.07564.

Blanchet, J., Y. Kang, and K. Murthy. 2016, October. “Robust Wasserstein Profile Inference and Applications to Machine Learning”. *Journal of Applied Probability* 56(3):830–857.

Blanchet, J. H., and P. W. Glynn. 2015. “Unbiased Monte Carlo for Optimization and Functions of Expectations via Multi-level Randomization”. In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, V.W.K. Chan, I.-C. Moon, T.M.K. Roeder, C. Macal and M.D. Rossetti, 3656–3667. Piscataway, New Jersey: IEEE, Inc.

Bottou, L., F. E. Curtis, and J. Nocedal. 2018. “Optimization Methods for Large-Scale Machine Learning”. *Siam Reviews*, 60(2):223-311.

Esfahani, P., and D. Kuhn. 2018. “Data-driven Distributionally Robust Optimization using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations”. *Mathematical Programming* 171:115–166.

Ghosh, S., and H. Lam. 2019. “Robust Analysis in Stochastic Simulation: Computation and Performance Guarantees”. *Operations Research*, 67(1):232–249.

Ghosh, S., and M. S. Squillante. 2020. “Unbiased Gradient Estimation for Distributionally Robust Learning”. *arXiv e-prints* 2012.12367.

Ghosh, S., M. S. Squillante, and E. Wollega. 2021. “Efficient Generalization with Distributionally Robust Learning”. In *Advances in Neural Information Processing Systems*, 34:28310–28322, Curran Associates, Inc.

Giles, M. B. 2008. “Multilevel Monte Carlo Path Simulation”. *Operations Research* 56(3):607–617.

Glasserman, P., and X. Xu. 2014. “Robust Risk Measurement and Model Risk”. *Quantitative Finance* 14(1):29–58.

He, X. D., and X. Y. Zhou. 2011. “Portfolio Choice Under Cumulative Prospect Theory: An Analytical Treatment”. *Management Science* 57:315–331.

Hu, Z., and L. J. Hong. 2012. “Kullback-Leibler Divergence Constrained Distributionally Robust Optimization”. *Optimization Online*, http://www.optimization-online.org/DB_HTML/2012/11/3677.html.

Kahneman, D., and A. Tversky. 1979. “Prospect Theory: An Analysis of Decision Under Risk”. *Econometrica* 47:263–291.

Lam, H. 2019. “Recovering Best Statistical Guarantees via the Empirical Divergence-Based Distributionally Robust Optimization”. *Operations Research* 67(4):1090–1105.

Levy, D., Y. Carmon, J. C. Duchi, and A. Sidford. 2020. “Large-Scale Methods for Distributionally Robust Optimization”. In *Advances in Neural Information Processing Systems*, 33:8847–8860, Curran Associates, Inc.

Markowitz, H. 1952. “Portfolio Selection”. *The Journal of Finance* 7(1):77–91.

Namkoong, H., and J. C. Duchi. 2017. “Variance-based Regularization with Convex Objectives”. In *Advances in Neural Information Processing Systems*, 30:2971–2980, Curran Associates, Inc.

Owen, A. B. 2001. *Empirical Likelihood*. New York: Chapman & Hall.

Pflug, G. 2000. “Some Remarks on the Value-at-Risk and Conditional Value-at-Risk”. In *Probabilistic Constrained Optimization: Methodology and Applications*, edited by S. Uryasev, 272–281. Dordrecht: Kluwer.

Sinha, A., H. Namkoong, and J. Duchi. 2017, October. “Certifying Some Distributional Robustness with Principled Adversarial Training”. In *International Conference on Learning Representations*, Apr 30 - May 3, Vancouver.

Stone, M. 1974. “Cross-Validatory Choice and Assessment of Statistical Predictions”. *Journal of Royal Statistical Society* 36:111–147.

Tversky, A., and D. Kahneman. 1992. “Advances in Prospect Theory: Cumulative Representation of Uncertainty”. *Journal of Risk and Uncertainty* 5:297–323.

AUTHOR BIOGRAPHIES

SOUMYADIP GHOSH is a member of IBM Research in the mathematics of AI division at Yorktown Heights. His research focuses on stochastic optimization and decision making under uncertainty, with applications in training large statistical learning models with improved generalization in speech recognition and natural language processing, and previously on Smarter Grid and energy analytics, cloud infrastructure, and supply chain management. He is an associate editor of Stochastic Models and has served as a committee member for INFORMS / WSC. He holds a PhD from Cornell University (Ithaca, NY, USA), an MS from Univ of Michigan (Ann Arbor, MI, USA) and a B Tech from Indian Institute of Technology (Chennai, TN, INDIA). His email address is ghoshs@us.ibm.com. His website is https://researcher.watson.ibm.com/researcher/view.php?person=us-ghoshs.

MARK S. SQUILLANTE is a Distinguished Research Staff Member and the Manager of Foundations of Probability, Dynamics and Control within Mathematical Sciences at IBM Research (Yorktown Heights, NY, USA). His research interests broadly concern mathematical foundations of the analysis, modeling and optimization of the design and control of stochastic systems, and their applications across a wide range of domains in computing, communications, science, engineering, and business. He is an elected Fellow of ACM, IEEE, INFORMS and AIAA, and currently serves as Editor-in-Chief of Stochastic Models. He received a PhD from the University of Washington (Seattle, WA, USA). His email address is mss@us.ibm.com. His website is https://researcher.watson.ibm.com/researcher/view.php?person=us-mss.