Robust Simulation of Stochastic Systems

Zhaolin Hu
School of Economics and Management, Tongji University, Shanghai 200092, China, huzhaolin@gmail.com

L. Jeff Hong
Department of Economics and Finance; and Department of Management Sciences, College of Business, City University of Hong Kong, Kowloon, Hong Kong, China, jeffhong@cityu.edu.hk

As submitted on October 23, 2015

Abstract

Simulation is an important tool for studying stochastic systems. A very first step of this approach is to specify a distribution for the random input. This is called input modeling. Input models are important because they affect simulation results and thus real decisions. However, specifying a distribution precisely is typically difficult and even impossible in practice. The issue is called input uncertainty in the simulation literature, which has been considered and studied extensively in recent years. In this paper we study the input uncertainty issue when using simulation to estimate important performance measures such as expectation, probability, and value-at-risk. We propose a robust simulation (RS) approach, which assumes the real distribution is contained in an ambiguity set constructed using statistical divergences, and estimates the worst-case value of the performance measures when the distribution varies in the ambiguity set. We show that the RS approach is computationally tractable and the corresponding results reveal important information of the stochastic systems and help decision makers make better decisions.

1 Introduction

Simulation is an important tool for studying complex stochastic systems. In a typical simulation study, a simulation model is built to approximate the logic of the real system and to map the often stochastic inputs to the outputs. For instance, in a financial simulation, the inputs may be the realizations of various risk factors and the output may be the loss given the risk factors. Put it in mathematics, we use ξ to denote the random inputs, where ξ is a k-dimensional random vector supported on $\Xi \subset \Re^k$, and use $H(\xi)$ to denote the output, where $H(\cdot)$ is a single-valued function. To run simulation experiments on the model, one needs to specify the distribution of ξ so that its random realizations may be simulated. This is known as input modeling and it is often a critical step in a simulation study.

As ξ is a random vector, the output $H(\xi)$ is also random, which creates difficulty for decision making. To resolve this difficulty, various performance measures have been proposed to map the distribution of $H(\xi)$ into a deterministic value, so that decisions may be made more easily. In this paper, we consider three important performance measures: expectation, probability, and value-at-risk (VaR). Suppose that ξ follows a distribution P_*. We denote the expectation as $\mathrm{E}_{P_*}[H(\xi)]$, where $\mathrm{E}_{P_*}[\cdot]$ indicates that the expectation is taken with respect to (w.r.t.) the distribution P_*. Expectation is often the first choice in simulation study, which measures the average value of the output. Probability is another widely used measure. It measures the chance of some random event, desired or undesired. Consider some random event $A(\xi)$, where the randomness is introduced by ξ. We denote the probability of $A(\xi)$ as $\Pr_{P_*}\{A(\xi)\}$ where $\Pr_{P_*}\{\cdot\}$ means the probability is taken w.r.t. P_*. For instance, suppose the loss of a financial activity is $H(\xi)$. Then $A(\xi) := \{H(\xi) \le v\}$ is the event that the loss does not exceed the given threshold v and $\Pr_{P_*}\{A(\xi)\}$ is the probability of this event. Probability is often advocated by decision makers who are risk averse. Value-at-risk (VaR), as a risk measure, was proposed in 1990s, and has been considered as a standard risk management tool in the financial industry. Mathematically, VaR is the quantile of a loss distribution. For a given confidence level $\beta \in (0,1)$, the $(1-\beta)$-VaR of a random loss $H(\xi)$ is defined as $\mathrm{VaR}_{1-\beta,P_*}(H(\xi)) := \inf\{v \in \Re : \Pr_{P_*}\{H(\xi) \le v\} \ge 1-\beta\}$, where, similarly, the subscript P_* denotes that the VaR is calculated when ξ follows P_*. For a thorough introduction and treatment of VaR, readers are referred to Jorion (2006) and Hong et al. (2014a).

It is clear that the values of all three performance measures critically depend on the true distribution P_*. However, in practical situations, P_* is typically difficult to be specified precisely. When there are data, one can specify the distribution via some statistical methods, which often contain estimation errors. When there are no data, one can specify a distribution subjectively based on some practical knowledge of the inputs, which is rarely perfect and leads to uncertainty as well. However, the uncertainty in the input distributions affects directly the quality of the estimated performance measures, and thus may lead to incorrect or inappropriate decisions. This issue is known as "input uncertainty" in the simulation literature.

Input uncertainty has for long been considered as a fundamental issue in simulation studies. According to Barton (2012), there already existed systematic discussions on this issue in the 1992 Winter Simulation Conference. Since then, a number of approaches have been developed and applied to handle the issue, including the resampling method, Bayesian approach, and approximation approach, among others (see Barton 2012). The literature on resampling approach includes Barton and Schruben (1993, 2001), Freimer and Schruben (2002) and many others, the literature on Bayesian approach includes Chick (1997, 2001), Zouaoui and Wilson (2004), and Biller and Corlu (2011)), and the literature on approximation approach includes Cheng and Holland (1997) and Ng and Chick (2006). In recent years, metamodeling techniques, such as stochastic kriging, have been incorporated in various approaches to make them easier to use (see, for instance, Barton et al. (2014) and Xie et al. (2014)). For a broader and deeper review of the study on input uncertainty, we refer the readers to Henderson (2003), Barton (2012), Barton et al. (2014), Xie et al. (2014)

and references therein.

In this paper, we follow the convention of the economics literature (see, e.g., Ellsberg (1961) and Epstein (1999)) and use the term "ambiguity" to denote the phenomenon that a distribution cannot be fully specified. In contrast to the existing literature, we propose a robust simulation (RS) approach to handling input uncertainty. The RS approach models the input uncertainty by constraining the distribution in an ambiguity set and then estimates the worst-case value of the performance measure when the distribution varies in the ambiguity set. The size of the ambiguity set reflects the decision maker's knowledge of the input distribution. It can be some confidence region of the real distribution constructed from data or some subjective set reflecting the decision maker's understanding of the uncertainty. The use of a robust approach when facing ambiguity is consistent to the concept of "ambiguity aversion" documented in the economics literature. For instance, Ellsberg (1961) pointed out, when facing ambiguity, it is reasonable for a conservative person to consider the worst-case scenario, and Epstein (1999) provided empirical evidence to support the argument. The use of worst-case analysis when facing ambiguity is a common approach adopted by some other fields as well. In financial risk management, for instance, Artzner et al. (1999) proposed the notion of coherent risk measure that is defined as the worst-case mean performance in an ambiguity set. In optimization, the distributionally robust optimization approach is used to handle the uncertainty in optimization models, see, e.g., Delage and Ye (2010) and Ben-Tal et al. (2013).

The concept of RS used in this paper, i.e., applying a worst-case analysis on a simulation model within an ambiguity set, was also used by Hu et al. (2012) who considered the input uncertainty in environmental policy simulation. They suggested estimating the worst-case values of the performance measures of different environmental policies and selecting the policy with the best worst-case performance. Fan et al. (2013) extended the concept to robust simulation ranking and selection, which selects the alternative with the best worst-case mean performance from a group of alternatives. However, to model the input uncertainty, they only considered a finite number of scenarios of the input distributions. Other concepts of robustness have also been used in simulation studies. For instance, Sanchez (2000) and Pierreval and Durieux-Paris (2007) considered robust design in simulation studies, and Dellino et al. (2012) considered robustness in simulation optimization. All these authors used the term "robustness" according to Taguchi's view on input uncertainty.

An important issue in RS is how to specify an ambiguity set. In a practical simulation study, the true input distribution P_* is typically unknown, and we often specify a nominal distribution P_0, by either statistical fitting methods or subjective choice. Then, a natural approach to quantifying ambiguity is to consider some level of perturbation or deviation from the nominal distribution. This motivates us to model the distribution ambiguity based on the likelihood ratio (LR) function of the true and nominal distributions. We consider two classes of constraints imposed on the LR function: band constraints and ϕ-divergence constraints. The band constraints define a uniform band of the LR function, and the ϕ-divergence constraints require the ϕ-divergence between the true and nominal distributions (which is the expectation of a convex function of the LR) be no more than a positive constant. Notice that ϕ-divergence is an important class of distance measures between two distributions. It includes the widely used Kullback-Leibler (KL) divergence, χ^2-distance, Hellinger distance, Variation distance, Burg entropy, and many others. The concept of ϕ-divergence was introduced systematically by Pardo (2006), and was used by Ben-Tal et al. (2013) to model distribution ambiguity in the context of distributionally robust optimization. Our work was inspired by the sequence of papers, including Ben-Tal and Teboulle (2007), Ben-Tal et al. (2010) and Ben-Tal et al. (2013).

The major advantage for using LR constraints in modeling ambiguity set is its mathematical tractability. Given this type of ambiguity sets, using the duality theory, the RS problems may be reformulated into *convex* optimization problems for all three types of performance measures, i.e., expectation, probability and VaR. Another natural way of modeling ambiguity set is to specify a distribution family and to constrain the parameters. Hu et al. (2012) used this approach. However, the resulted optimization problems are in general non-convex and the variance of the performance estimators may be enlarged by the change-of-measure technique that is used to handle the LR function. When using our approach, however, the resulted optimization problems are convex and solvable using typical stochastic programming techniques, and the variance induced by the LR can be handled more efficiently. Actually, as will be seen, our approach absorbs the LR term and estimation of LR is no longer required.

Even though the RS approach needs only to find the worst-case performance in the ambiguity set, it can be used to find the best-case performance as well. Then, the interval formed by the best and worst values reveals how the uncertainty of the input distribution leads to the uncertainty of the output performance measure. It contains important information even for decision makers who are not ambiguity averse. For instance, a decision maker may look at the difference between the worst (or best) performance and the nominal performance to understand the potential loss (or gain) caused by the input uncertainty and make decisions based on it (or even decide to collect more input data to reduce the input uncertainty). Moreover, if the ambiguity set is a $1-\alpha$ confidence region of the input distribution, then the interval is also a $1-\alpha$ confidence interval of the performance measure without considering the simulation error. Note that the simulation error is typically controllable by setting the simulation effort. In this paper, we do not consider the simulation error, and our analysis is focused on the input error, i.e., error caused by input uncertainty.

The RS approach proposed in our paper has a major limitation. Our convex reformulations of the RS problems can only be applied to static simulation models where only one observation of the input distribution is needed to run the simulation. For instance, the risk of a portfolio depends on the realization of the risk factors at the end of the time horizon, or the atmosphere carbon dioxide concentration depends on the realization of the environment parameters in the DICE environmental simulation model (Hu et al. 2012). However, our reformulation cannot be applied to dynamic simulation models where multiple independent observations from the same uncertain input distribution are needed to run the simulation. For instance, in queueing models, to simulate a sample path, a sequence of independent and identically distributed (i.i.d.) random observations needs to be generated from the same uncertain service-time distribution. Our approach has difficulty in handling this situation because the LR function is more complicated. However, the general concept of RS is still applicable to this situation. Recently, Lam (2012) developed an asymptotic expansion of the worst-case mean performance based on the upper bound of the KL divergence in the ambiguity set, and his approach applies to dynamic simulation models. It might be possible to connect his result to the RS formulation in the case of dynamic simulation models. We propose it for future research.

The rest of this paper is organized as follows. In Section 2, we discuss the concept of RS. In Sections 3 to 5, we consider RS of expectation, probability and VaR respectively. In Section 6 we discuss how to specify the ambiguity set in real situations. Several examples are studied in Section 7. The paper is concluded in Section 8, with the proofs in Appendix.

2 Robust Simulation Formulation

To introduce the formulation of RS, we consider the expectation performance measure. Suppose that the true input distribution P_* is known. Then, the performance measure that we are interested in is $\mu_{P_*} = \mathrm{E}_{P_*}[H(\xi)]$, which may be estimated using a typical sample-mean estimator $\hat{\mu}_{P_*}$, i.e.,

$$\hat{\mu}_{P_*} = \frac{1}{N}\sum_{i=1}^{N} H(\xi_i), \tag{1}$$

where $\{\xi_1, \cdots, \xi_N\}$ is an i.i.d. sample of ξ generated from the distribution P_*. It is worthwhile noting that $H(\cdot)$ typically has no closed-form expression and its structural information (e.g., convexitiy or even monotonicity) is rarely known to the simulation modelers. By the strong law of large numbers (Durrett 2005), we have $\hat{\mu}_{P_*} \to \mu$ with probability 1 as $N \to \infty$.

In a simulation study, the true input distribution P_* is typically unknown. We select (or estimate) a nominal distribution P_0 to approximate P_* and use $\mu_{P_0} = \mathrm{E}_{P_0}[H(\xi)]$ to approximate μ_{P_*}. Then, an i.i.d. sample of $\{\xi_1, \cdots, \xi_N\}$ is generated from the distribution P_0 and μ_{P_0} is estimated by

$$\hat{\mu}_{P_0} = \frac{1}{N}\sum_{i=1}^{N} H(\xi_i). \tag{2}$$

It is important to note that, even though the right-hand sides of Equations (1) and (2) are the same, the observations $\{\xi_1, \cdots, \xi_N\}$ are generated from different distributions. Then, we often use $\hat{\mu}_{P_0}$ to approximate μ_{P_*}. In this case, however, we introduce an approximation error that cannot be removed by increasing the sample size N, i.e., as $N \to \infty$, $\hat{\mu}_{P_0}$ converges to μ_{P_0} but not μ_{P_*}. This is known as input uncertainty in the simulation literature.

In this paper, we suppose that the simulation modelers (or the decision makers) may provide a set $\mathbb{P}$, called an ambiguity set, so that $P_* \in \mathbb{P}$ (or at least $P_* \in \mathbb{P}$ with a high level of confidence). Let $\mu_P = \mathrm{E}_P[H(\xi)]$ be the expected performance under the input distribution P. Then, the RS approach suggests to estimate the worst-case μ_P when $P \in \mathbb{P}$. Without loss of generality, we suppose that a small μ_{P_*} is desirable, e.g., it may denote the average cost or average loss. Then, we may formulate the RS problem as the following optimization problem:

$$\underset{P \in \mathbb{P}}{\text{maximize}} \;\; \mathrm{E}_P[H(\xi)]. \tag{3}$$

If a large μ_{P_*} is desirable, then the RS approach needs to find $\inf_{P\in\mathbb{P}} \mathrm{E}_P[H(\xi)]$, which equals to $-\sup_{P\in\mathbb{P}} \mathrm{E}_P[-H(\xi)]$. Therefore, we may reformulate the problem in the form of (3) by adding a negative sign to $H(\xi)$. It is also worthwhile noting that, if $\mathbb{P}$ is a $1-\alpha$ confidence set of P_*, the interval $[\inf_{P\in\mathbb{P}} \mu_P, \ \sup_{P\in\mathbb{P}} \mu_P]$ is also a $1-\alpha$ confidence interval of μ_{P_*}.

2.1 A Change-of-Measure Reformulation

To solve the optimization problem (3) is not easy. The major difficulty is that there is no closed-form expression of $H(\cdot)$ and almost no structural information of it either. This makes RS problems drastically different from robust optimization problems, where the objective function has not only a closed-form expression but also various structural properties, e.g., linear, convex quadratic etc. Another difficulty is that, to evaluate $\mathrm{E}_P[H(\xi)]$, one has to specify an input distribution so that $H(\xi)$ may be observed. Therefore, it is not easy to separate the optimization process and the simulation process. One way to resolve these difficulties is to use a simulation optimization algorithm (e.g., a stochastic approximation algorithm or a random search algorithm). However, these algorithms require simulating $\mathrm{E}_P[H(\xi)]$ at different choices of P and they typically have a slow rate of convergence.

To resolve these difficulties in a different way, note that we may write

$$\mu_P = \mathrm{E}_P[H(\xi)] = \int_\Xi H(z)dP(z) = \int_\Xi H(z)\frac{dP(z)}{dP_0(z)}dP_0(z) = \mathrm{E}_{P_0}[H(\xi)L_P(\xi)], \tag{4}$$

where $L_P(z) = dP(z)/dP_0(z)$ is known as the likelihood ratio (LR) in the simulation literature, as long as the distribution P is absolutely continuous w.r.t. P_0, i.e., for every measurable set A, $P_0(A) = 0$ implies $P(A) = 0$. Notice that, in the right-hand side of the last equality of Equation (4), the expectation is taken with respect to P_0. Therefore, we may use an i.i.d. sample

$\{\xi_1, \cdots, \xi_N\}$ generated from the input distribution P_0 to simulate a sample of $H(\xi)$, denoted by $\{H(\xi_1), \cdots, H(\xi_N)\}$, and to calculate a sample of $L_P(\xi)$, denoted by $\{L_P(\xi_1), \cdots, L_P(\xi_N)\}$, for any $P \in \mathbb{P}$. Then, μ_P may be estimated by

$$\hat{\mu}_P = \frac{1}{N} \sum_{i=1}^{N} H(\xi_i) L_P(\xi_i). \tag{5}$$

This is a critical step in the RS formulation because the observations $\{H(\xi_1), \cdots, H(\xi_N)\}$ are now constants so they no longer change w.r.t. the input distribution P. Therefore, the unavailability of the closed-form expression of $H(\cdot)$ is no longer important. Moreover, in practice, a robust study is typically used as a supplement to, instead of a replacement of, the nominal study. Therefore, a sample of $\{H(\xi_1), \cdots, H(\xi_N)\}$ is typically available under the nominal input distribution P_0. Therefore, Equation (5) allows the estimation of μ_P for any $P \in \mathbb{P}$ without running additional simulation experiments.

2.2 Modeling Ambiguity Set

There are generally two approaches to modeling the ambiguity set $\mathbb{P}$ in Problem (3), a parametric one and a nonparametric one. The parametric approach first specifies the distribution family of the input distribution and then specifies a set for the parameters of the distribution. Such an approach is very natural and quite appealing. It reduces the infinite dimensional optimization problem (searching for the extreme distribution among an ambiguity set) to a finite dimensional optimization problem (searching for the extreme parameters among the specified set). Applying the LR method for the parametric case, we can convert the RS problem to a stochastic optimization (SO) problem with real-valued decision variables. This approach was adopted by Hu et al. (2012) in modeling the uncertainty in the mean vector and covariance matrix of a multivariate normal distribution. However, two issues arise when using this approach. The first is that the objective function of Problem (3) may be non-concave, and the second is that the LR will affect the variance of the estimator and thus the efficiency of simulation.

To understand these two issues, we consider a special case where ξ follows an exponential family distribution. By Casella and Berger (2002), its distribution function (i.e., mass function if ξ is discrete and density function if ξ is continuous) may be written as

$$p_\theta(z) = h(z)c(\theta) \exp\left\{ \sum_{i=1}^{k} w_i(\theta) t_i(z) \right\},$$

where θ is the uncertain parameter. For instance, if ξ follows an exponential distribution with rate θ, then

$$p_\theta(z) = \theta \exp\{-\theta z\},$$

and if ξ follows a k-dimensional multivariate normal (MVN) distribution with mean vector μ and covariance matrix Σ (and the precision matrix $\Omega = \Sigma^{-1}$), then

$$p_{\mu,\Omega}(z) = (2\pi)^{-k/2}|\Omega|^{1/2} \exp\left\{-\frac{1}{2}(z-\mu)^{\mathrm{T}}\Omega(z-\mu)\right\}.$$

Therefore, for an exponential family distribution, the LR function may be written as

$$L_P(z) = \frac{p_\theta(z)}{p_{\theta_0}(z)} = \frac{c(\theta)}{c(\theta_0)} \exp\left\{\sum_{i=1}^{k} t_i(z)\left[w_i(\theta) - w_i(\theta_0)\right]\right\}.$$

Now we look at the first issue, i.e., Problem (3) may be non-concave. Notice that many of the exponential family distributions are quasi-concave in θ, and so do the LR functions. For instance, if ξ follows an exponential distribution, it is easy to see that

$$\log p_\theta(z) = -\theta z + \log\theta,$$

which is concave in θ. Then, $p_\theta(z)$ and $L_P(z)$ is log-concave, thus quasi-concave, in θ. If ξ follows a MVN distribution, then

$$\log p_{\mu,\Omega}(z) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}(z-\mu)^{\mathrm{T}}\Omega(z-\mu) + \frac{1}{2}\log|\Omega|.$$

Then, $p_{\mu,\Omega}(z)$ as well as $L_P(z)$ are both log-concave, thus quasi-concave, in μ and in Ω. However, it is known that the sum of quasi-concave functions are in general not quasi-concave (Boyd and Vandenberghe 2004). Therefore, $\mu_P = \mathrm{E}_{P_0}[H(\xi)L_P(\xi)]$ is in general not a quasi-concave function of θ even if $H(\xi) \geq 0$ for all ξ. Therefore, Problem (3) is in general a non-convex optimization problem and this creates difficulty in solving the problem. For instance, the optimization algorithm of Hu et al. (2012) only guarantees to find a local optimal solution instead of a global optimal one. However, the concept of robustness in Problem (3) is no longer clear enough if the local optimal solution is not global optimal.

Now we look at the second issue, i.e., the variance of $\hat{\mu}_P$ will be affected by the LR. The change-of-measure approach used in this paper has also been used in simulation studies to conduct importance sampling (IS) to reduce the variance of the estimators. Even though a carefully designed IS scheme may be very helpful in variance reduction, it is well known that a careless use of the approach may increase the variance or even lead to an infinite variance (Law and Kelton 2000). When using the change-of-measure approach in RS with a parametric modeling of the ambiguity, we encounter this problem as well. To illustrate the issue, we examine $\mathrm{E}_{P_0}\left[L_p^2(\xi)\right]$. Under the exponential family of distributions,

$$\mathrm{E}_{P_0}\left[L_p^2(\xi)\right] = \mathrm{E}_{P_0}\left[\left(\frac{dP(\xi)}{dP_0(\xi)}\right)^2\right] = \frac{c^2(\theta)}{c(\theta_0)}\int_\Xi \exp\left\{\sum_{i=1}^{k} t_i(z)\left[2w_i(\theta) - w_i(\theta_0)\right]\right\} dz.$$

Then it is clear that, when Ξ is unbounded above, $t_i(z)$ is a non-negative polynomial function of z and $2w_i(\theta) - w_i(\theta_0) > 0$, $\mathrm{E}_{P_0}\left[L_p^2(\xi)\right] = \infty$. For instance, when ξ follows an exponential distribution, $\mathrm{E}_{P_0}\left[L_p^2(\xi)\right] = \infty$ if $\theta < \frac{1}{2}\theta_0$, and when ξ follows a single-variate normal distribution with a fixed mean μ but an uncertain variance σ^2, then $\mathrm{E}_{P_0}\left[L_p^2(\xi)\right] = \infty$ if $\sigma^2 > 2\sigma_0^2$. These results suggest that when implementing the LR method, one needs to select a "good" distribution as the sampling distribution to avoid the variance blow up phenomenon. However, even a good distribution is chosen, it can not be avoided that the LR will introduce extra variance during simulation.

Both of these issues are not easily solvable. Therefore, in this paper, we suggest to use a nonparametric approach that builds an ambiguity set directly on the LR function in the respective functional space. We show in the rest of this paper that the resulted optimization problem is convex no matter what is the sign of $H(\xi)$ and we can avoid estimating the LR in simulation. Notice that, by the definition of the LR function, it is a good candidate for measuring the perturbation/deviation of the true distribution to the nominal one. In this paper, we use two different classes of constraints on the LR to model the ambiguity. The first is called band constraints. Specifically, we consider a convex function $\varphi : \Re \to \Re$, and construct the constraint

$$\varphi(L) \leq \rho, \tag{6}$$

where ρ is a positive constant. To guarantee that the nominal distribution satisfies (6), we impose the regularity condition for φ that $\varphi(1) \leq \rho$. Because φ is convex and finite valued, the constraint (6) defines a closed interval for L. Furthermore, a finite number of constraints taking the form of (6) still define a closed interval. Therefore, using the band constraints we are arriving at a set of p such that the LR falls in an interval, i.e., $a \leq L \leq b$ for some $0 \leq a < 1 < b \leq \infty$ (We omit the degenerate case where $a = 1 = b$). Although L is itself a function of ξ, (6) requires the constraint be satisfied for all ξ.

The second class is called ϕ-divergence constraints. Specifically, consider a convex function ϕ on $\Re$ and construct the constraint $\mathrm{E}_{P_0}[\phi(L)] \leq \eta$. Imposing some minor regularity conditions on ϕ, we are arriving at the famous ϕ-divergence, which has been used frequently in statistics to measure the distance of a distribution to another one. Therefore, imposing constraints on the LR function using ϕ-divergence admits a clear statistical and practical meaning. Following the definition of Pardo (2006) and Ben-Tal et al. (2013), a ϕ-divergence function is a convex function for $t > 0$, satisfying $\phi(1) = 0$, $0\phi(a/0) := a \lim_{t\to\infty} \phi(t)/t$ for $a > 0$, and $0\phi(0/0) := 0$. For P and P_0 introduced above, the ϕ-divergence from P to P_0 is defined as

$$D_\phi(P\|P_0) = \int_\Xi p_0(z)\phi\left(\frac{p(z)}{p_0(z)}\right) \mathrm{d}z = \mathrm{E}_{P_0}\left[\phi\left(\frac{p(\xi)}{p_0(\xi)}\right)\right] = \mathrm{E}_{P_0}\left[\phi\left(L\right)\right]. \tag{7}$$

Similarly, we may understand the integral in (7) as the summation if P_0 is a discrete distribution, and as a mixture of integral and summation if P_0 is a mixed distribution. It can be shown that

$D(P\|P_0) \geq 0$ and the equality holds if and only if $p(z) = p_0(z)$ almost surely (a.s.) under P_0. We now construct a neighborhood $D_\phi(P||P_0) \leq \eta$, which from (7) yields a ϕ-divergence constraint $\mathrm{E}_{P_0}[\phi(L)] \leq \eta$. As can be seen, instead of requiring L satisfy a constraint for all ξ in band constraints, in ϕ-divergence constraints, one only requires L satisfy a constraint on average.

Combining the two classes of constraints, we construct the following ambiguity set of P in terms of the LR function L:

$$\mathbb{L} = \{L \in \mathbb{L}(a,b) : \mathrm{E}_{P_0}[L] = 1, \mathrm{E}_{P_0}[\phi_i(L)] \leq \eta_i, i = 1, \cdots, m\}, \tag{8}$$

where $\mathbb{L}(a,b) := \{L : a \leq L \leq b \text{ a.s.}\}$, and the constants a, b and $\eta_i, i = 1, 2, \cdots, m$ are indexes of ambiguity that control the size of $\mathbb{L}$. In what follows, we discuss how to conduct RS for different performance measures with the ambiguity set $\mathbb{L}$.

3 Expectation Performance Measure

We start from the expectation performance measure. Suppose the random output of the system that we are interested in is $H(\xi)$. For simplicity of the notation, we suppress the dependence of H on ξ. Based on the discussion in Section 2, we may formulate the RS problem (3) as

$$\underset{L\in\mathbb{L}}{\text{maximize}}\ \mathrm{E}_{P_0}[HL], \tag{9}$$

where $\mathbb{L}$ is defined in Equation (8). We can then rewrite Problem (9) as

$$\begin{aligned} &\underset{L\in\mathbb{L}(a,b)}{\text{maximize}} && \mathrm{E}_{P_0}[HL] \\ &\text{subject to} && \mathrm{E}_{P_0}[\phi_i(L)] \leq \eta_i, i = 1, \cdots, m,\ \mathrm{E}_{P_0}[L] = 1. \end{aligned} \tag{10}$$

Problem (10) is a functional optimization problem with L being the decision variable. It is not difficult to verify that (10) is a convex optimization problem. One standard approach to handling such constrained functional optimization problem is to use the Lagrangian duality, see, for instance, Ben-Tal et al. (2010). We construct the Lagrangian functional associated with Problem (10):

$$\begin{aligned} \ell_0(\lambda, \alpha, L) &:= \mathrm{E}_{P_0}[HL] - \sum_{i=1}^m \alpha_i (\mathrm{E}_{P_0}[\phi_i(L)] - \eta_i) + \lambda(\mathrm{E}_{P_0}[L] - 1) \\ &= \mathrm{E}_{P_0}\left[(H + \lambda) L - \sum_{i=1}^m \alpha_i \phi_i(L)\right] + \sum_{i=1}^m \alpha_i \eta_i - \lambda. \end{aligned}$$

Then Problem (10) is equivalent to

$$\underset{L\in\mathbb{L}(a,b)}{\text{maximize}}\ \underset{\lambda\in\Re,\alpha\geq 0}{\text{minimize}}\ \ell_0(\lambda, \alpha, L). \tag{11}$$

Interchanging the maximum and minimum in Problem (11), we obtain the Lagrangian dual of Problem (11):

$$\underset{\lambda\in\Re,\alpha\geq 0}{\text{minimize}}\ \underset{L\in\mathbb{L}(a,b)}{\text{maximize}}\ \ell_0(\lambda,\alpha,L). \tag{12}$$

The major concern about the primal and dual problems above are whether they have the same optimal value. Fortunately, the duality gap turns out to be zero. We summarize the result in the following theorem, whose proof can be found in Appendix.

Theorem 1. *The optimal values of Problems (11) and (12) are equal. The optimal value of Problem (12) is attained at some* $\lambda^*\in\Re$ *and* $\alpha^*\geq 0$.

Theorem 1 guarantees that, to solve (11) it suffices to solve (12). Let $v(\lambda,\alpha)$ denote the optimal value of the inner maximization problem of (12), i.e.,

$$v(\lambda,\alpha)=\sup_{L\in\mathbb{L}(a,b)}\ \ell_0(\lambda,\alpha,L).$$

The following proposition states that we can put the supremum into the expectation in the expression of $v(\lambda,\alpha)$. The proof of Proposition 1 is included in Appendix.

Proposition 1. *For any* $\lambda\in\Re,\alpha\geq 0$,

$$v(\lambda,\alpha)=\mathrm{E}_{P_0}\left[\sup_{L\in\mathbb{L}(a,b)}\left\{(H+\lambda)\,L-\sum_{i=1}^{m}\alpha_i\phi_i\,(L)\right\}\right]+\sum_{i=1}^{m}\alpha_i\eta_i-\lambda. \tag{13}$$

To simplify $v(\lambda,\alpha)$, we define an auxiliary function

$$\Psi(s,\alpha)=\sup_{t\in\mathbb{L}(a,b)}\left\{st-\sum_{i=1}^{m}\alpha_i\phi_i\,(t)\right\}. \tag{14}$$

It is not difficult to see that $\Psi(s,\alpha)$ is a well defined deterministic function. Moreover, we have the following proposition, whose proof is included in Appendix.

Proposition 2. $\Psi(s,\alpha)$ *is convex in* (s,α) *and non-decreasing in* s*, and it satisfies* $\Psi(s,\alpha)\geq s$.

Proposition 2 summarizes important properties of $\Psi(s,\alpha)$. We will frequently refer to this proposition in the following analysis. With the theory built above, it is easy to prove the following theorem, which summarizes the main result on RS of expectations.

Theorem 2. *The optimal value of Problem (9) is equal to that of the following problem*

$$\underset{\lambda\in\Re,\alpha\geq 0}{\mathit{minimize}}\ \mathrm{E}_{P_0}\left[\Psi(H+\lambda,\alpha)\right]+\sum_{i=1}^{m}\alpha_i\eta_i-\lambda. \tag{15}$$

Remark 1. *The strong duality between (11) and (12) allows us to transform the RS problem to its dual. However, it is possible that the optimal values of both problems are infinite. If the random output H is bounded or $b < \infty$, the optimal values are finite and in such a case we can use any ϕ-divergence. Nevertheless, when H is unbounded (e.g., it follows a normal distribution) and $b = \infty$, the optimal values may be infinite. In this case one needs to select an appropriate ϕ-divergence such that the RS problem has a finite optimal value. This phenomenon, from another angle, shows that different ϕ-divergences may constrain the distribution in different manners (e.g., in different aspects of constraining the tail of the distribution). One of the merits of introducing the uniform bound $[a, b]$ is that it makes the RS problem always meaningful, no matter what ϕ-divergence we use. A more detailed discussion on this issue can be found in the last paragraph of Section 3.1.1.*

Compared to (9), (15) becomes much more specific, as the expectation is now taken w.r.t. an explicit distribution, i.e., the nominal distribution P_0. Moreover, Proposition 2 guarantees that the problem is a convex optimization problem. Therefore, (15) is much easier to solve than the original functional optimization problem. Next, we discuss how to solve Problem (15).

3.1 Solution Methods

Various techniques have been developed for SO problems. Among them the sample average approximation (SAA) method and the stochastic approximation (SA) method are widely used, see, e.g., Shapiro et al. (2014). The idea of SAA is to approximate the SO problem by a deterministic sample problem and then implement deterministic optimization techniques to solve the sample problem. The SA method mimics the steepest decent (ascent) method and iteratively updates the solution based on the sample taken at each iteration. Both methods have their advantages/disadvantages and applicability.

The difficulty of Problem (15) depends on the function $\Psi(\cdot, \cdot)$. If the expression of $\Psi(\cdot, \cdot)$ can be derived analytically, then SAA may be directly applied. When a closed form of $\Psi(\cdot, \cdot)$ is unavailable, it becomes difficult to apply SAA, and in such circumstance SA is often a better choice. In this subsection, we first show that, for the ambiguity set defined by only one divergence, we can usually obtain the closed-form expression for $\Psi(\cdot, \cdot)$. For this class, we use SAA to solve it. For the general case, we suggest using SA algorithms.

3.1.1 Sample Average Approximation

Suppose $\{\xi_1, \xi_2, \cdots, \xi_N\}$ is an i.i.d. sample generated from P_0. SAA suggests using the following sample problem

$$\underset{\lambda \in \Re, \alpha \geq 0}{\text{minimize}} \ \frac{1}{N} \sum_{j=1}^{N} \Psi(H(\xi_j) + \lambda, \alpha) + \alpha\eta - \lambda \tag{16}$$

Table 1: Some ϕ-Divergence Functions and Their Conjugates

Divergence	$\phi(t), t \geq 0$	$\phi^*(s)$
Kullback-Leibler	$t\log t - t + 1$	$e^s - 1$
Burg entropy	$-\log t + t - 1$	$-\log(1-s), s < 1$
J-divergence	$(t-1)\log t$	No closed form
χ^2-distance	$(t-1)^2$	$\begin{cases} -1 & s < -2 \\ s + s^2/4 & s \geq -2 \end{cases}$
Neyman χ^2-distance	$\frac{(t-1)^2}{t}$	$2 - 2\sqrt{1-s}, s \leq 1$
Hellinger distance	$(\sqrt{t}-1)^2$	$\frac{s}{1-s}, s < 1$
χ-distance of order $\theta > 1$	$\lvert t-1\rvert^\theta$	$\begin{cases} -1 & s < -\theta \\ s + (\theta-1)\left(\frac{\lvert s\rvert}{\theta}\right)^{\theta/(\theta-1)} & s \geq -\theta \end{cases}$
Variation distance	$\lvert t-1\rvert$	$\begin{cases} -1 & s < -1 \\ s & -1 \leq s \leq 1 \end{cases}$
Cressie-Read	$\frac{t^\theta - \theta t + \theta - 1}{\theta(\theta-1)}, \theta \neq 0, 1$	$\begin{cases} -1/\theta & (\theta-1)s + 1 < 0 \\ \frac{1}{\theta}\left(1 - s(1-\theta)\right)^{\theta/(\theta-1)} - \frac{1}{\theta} & (\theta-1)s + 1 \geq 0 \end{cases}$

to approximate (15). With some regularity conditions on $\Psi(\cdot,\cdot)$, the optimal solutions and optimal value of (16) converge to those of (15) as $N \to \infty$. We refer readers to Shapiro et al. (2014) for details.

Note that the implementation of SAA relies on efficient solution methods for the sample problem (16). Consider the following special case of $\mathbb{L}$:

$$\mathbb{L}_\phi = \left\{L \in \mathbb{L}(0,+\infty) : \mathrm{E}_{P_0}[L] = 1, \mathrm{E}_{P_0}[\phi(L)] \leq \eta\right\}. \tag{17}$$

Then $\mathbb{L}_\phi$ contains the distributions whose distance (measured by ϕ) to the nominal distribution P_0 is within a constant η. Note that $\mathbb{L}_\phi$ is often the most natural choice of $\mathbb{L}$ in practice. Let $\phi^*(s) = \sup_{t\geq 0}\{st - \phi(t)\}$. It is the conjugate of the ϕ-divergence. Table 1, extracted from Ben-Tal et al. (2013), summarizes the conjugates of various ϕ-divergence measures.

For the ambiguity set $\mathbb{L}_\phi$, by Equation (14),

$$\Psi(s,\alpha) = \sup_{t\geq 0}\{st - \alpha\phi(t)\} = \alpha\phi^*\left(\frac{s}{\alpha}\right).$$

Then, Problem (15) takes the following form

$$\underset{\lambda\in\Re,\alpha\geq 0}{\text{minimize}} \ \mathrm{E}_{P_0}\left[\alpha\phi^*\left(\frac{H(\xi)+\lambda}{\alpha}\right)\right] + \alpha\eta - \lambda, \tag{18}$$

with the sample problem being

$$\underset{\lambda\in\Re,\alpha\geq 0}{\text{minimize}} \ \frac{1}{N}\sum_{j=1}^{N}\alpha\phi^*\left(\frac{H(\xi_j)+\lambda}{\alpha}\right) + \alpha\eta - \lambda. \tag{19}$$

From Table 1, we see that, for most of the divergences, the conjugate function ϕ^* has a closed-form expression. In these cases, we can design efficient procedures to solve the deterministic convex sample problem (19). To illustrate this, we consider two examples, KL divergence and χ^2-distance, both of which are widely-used statistical distance measures.

Example 3.1. *For KL divergence, Problem (18) becomes*

$$\underset{\lambda\in\Re,\alpha\geq 0}{\text{minimize}}\ \mathrm{E}_{P_0}\left[\alpha\exp\left\{\frac{H(\xi)+\lambda}{\alpha}\right\}\right]+\alpha\left(\eta-1\right)-\lambda,$$

and the corresponding sample problem becomes

$$\underset{\lambda\in\Re,\alpha\geq 0}{\text{minimize}}\ \frac{1}{N}\sum_{j=1}^{N}\alpha\exp\left\{\frac{H(\xi_j)+\lambda}{\alpha}\right\}+\alpha(\eta-1)-\lambda.$$

By introducing auxiliary decision variables $z_j, j=1,\cdots,N$, we can convert the problem to

$$\begin{aligned}&\text{minimize} && \frac{1}{N}\sum_{j=1}^{N}z_j+\alpha(\eta-1)-\lambda\\ &\text{subject to} && \alpha\exp\left\{\frac{H(\xi_j)+\lambda}{\alpha}\right\}\leq z_j,\ z_j\in\Re,\ j=1,\cdots,N,\ \lambda\in\Re,\ \alpha\geq 0.\end{aligned}$$

This is a perspective-of-exponential program and can be solved efficiently by CVX, a software package for convex optimization (Grant and Boyd 2013).

Example 3.2. *For χ^2-distance, Problem (18) takes the following form*

$$\underset{\lambda\in\Re,\alpha\geq 0}{\text{minimize}}\ \frac{1}{4\alpha}\mathrm{E}_{P_0}\left[\left([H+\lambda+2\alpha]^+\right)^2\right]+\alpha(\eta-1)-\lambda,$$

where $[z]^+=\max\{z,0\}$. The corresponding sample problem takes the following form

$$\underset{\lambda\in\Re,\alpha\geq 0}{\text{minimize}}\ \frac{1}{N}\sum_{j=1}^{N}\frac{1}{4\alpha}\left([H(\xi_j)+\lambda+2\alpha]^+\right)^2+\alpha(\eta-1)-\lambda.$$

By introducing auxiliary decision variables $z_j, j=1,\cdots,N$, we can reformulate the problem as

$$\begin{aligned}&\text{minimize} && \frac{1}{N}\sum_{j=1}^{N}\frac{z_j^2}{4\alpha}+\alpha(\eta-1)-\lambda\\ &\text{subject to} && H(\xi_j)+\lambda+2\alpha\leq z_j,\ z_j\geq 0,\ j=1,\cdots,N,\ \lambda\in\Re,\ \alpha\geq 0.\end{aligned}$$

This is a sum-of-quadratic-over-linear program, which can also be solved efficiently by CVX (Grant and Boyd 2013).

Comparing the KL divergence with the χ^2-distance, we find that the KL divergence requires the existence of the moment generating function of the random variable H, whereas the χ^2-distance only requires the existence of the second moment of H. Thus in terms of modeling, χ^2-distance is less restrictive than KL divergence. Now we discuss further the issue raised in Remark 1 using the

ϕ-divergence examples in Table 1. For a function f defined on $\Re$, let $\text{dom}f = \{s : f(s) < +\infty\}$, which is called the effective domain of f. From Table 1 we can find that for some conjugate functions, the effective domain is $\Re$, but for others, the effective domain is only a proper subset of $\Re$. Consider Problem (18). Let H_u be the essential supremum of $H(\xi)$ under measure P_0, i.e.,

$$H_u = \inf \left\{ t \in \Re : \Pr_{P_0} \left\{ H(\xi) > t \right\} = 0 \right\}.$$

If H is unbounded from above, then $H_u = +\infty$. In this case, suppose $\text{dom}\phi^*$ is only a proper subset of $\Re$. Then for any α and λ, the objective in (18) is $+\infty$. This shows that using such ϕ-divergence always leads to a $+\infty$ worst case. Therefore, for the unbounded case, we need to use those ϕ-divergences for which $\text{dom}\phi^* = \Re$. As can be seen, for the most widely used KL divergence and χ^2-distance, $\text{dom}\phi^* = \Re$. Furthermore, for the class of χ-distance of order $\theta > 1$ and the class of Cressie-Read divergence, the conjugates also have effective domain $\Re$. This suggests the decision makers have abundant choices. If decision makers want to use those ϕ-divergences for which $\text{dom}\phi^* \neq \Re$, they may have to impose certain finite bounds a and b on L. However, this usually results in situations where a closed-form expression of $\Psi(\cdot,\cdot)$ is difficult to derive.

3.1.2 Stochastic Approximation

One of the merits of SA, compared to SAA, is that it does not require a closed-form expression of $\Psi(\cdot,\cdot)$. Therefore, when it is difficult to derive the expression of $\Psi(\cdot,\cdot)$, we often resort to SA. There have been numerous SA procedures in the literature. In this paper, we suggest using the robust stochastic approximation (RSA) procedure proposed by Nemirovski et al. (2009) to solve (15). To describe the procedure, we introduce some notation for (15). Let $x = (\lambda, \alpha)$ denote the decision vector. Suppose Θ is a compact set that includes the optimal solution, and $G(x, \xi)$ is the stochastic subgradient of the objective function. Let $\Pi_\Theta(x)$ denote the projection of x onto Θ. Suppose the number of allowed iterations is N. The RSA procedure is as follows (in the form of Ghadimi and Lan (2015)).

Robust Stochastic Approximation (RSA)

Step 0. Let $x_0 \in \Theta$ be given.

Step k. For $k = 0, 1, \cdots, N-1$, generate ξ_k, and set $x_{k+1} = \Pi_\Theta\left(x_k - \gamma_k G(x_k, \xi_k)\right)$ for some $\gamma_k \in (0, +\infty)$.

Output: $\bar{x}_N = \frac{\sum_{k=1}^N \gamma_k x_k}{\sum_{k=1}^N \gamma_k}$.

To implement RSA, we need to provide the step size γ_k and the stochastic subgradient $G(x_k, \xi_k)$. Nemirovski et al. (2009) suggest several choices for γ_k given N. Suppose $st - \sum_{i=1}^m \alpha_i \phi_i(t)$ is strictly convex in t, which is often guaranteed by the strict convexity of ϕ_i. Then, there is a unique optimal solution t^* for (14). It follows from Danskin Theorem (Shapiro et al. 2014) that $\Psi(s,\alpha)$ is differentiable and

$$\nabla\Psi(s,\alpha) = \nabla\left\{st - \sum_{i=1}^{m}\alpha_i\phi_i\left(t\right)\right\}\Bigg|_{t=t^*}.$$

The stochastic subgradient $G(x,\xi)$ then becomes a stochastic gradient and can be computed accordingly. For further properties (e.g., convergence) and implementations of RSA, we refer the readers to Nemirovski et al. (2009).

4 Probability Performance Measure

We next consider the probability performance measure. Suppose $A(\xi)$ is the set of events of concern. Depending on it is the set of events that are desirable or undesirable, RS considers one of the following quantities,

$$P_l := \inf_{P\in\mathbb{P}}\ \mathrm{Pr}_P\left\{A(\xi)\right\} \quad \text{and} \quad P_u := \sup_{P\in\mathbb{P}}\ \mathrm{Pr}_P\left\{A(\xi)\right\}.$$

Let $A^c(\xi)$ denote the compliment of the set $A(\xi)$. Then,

$$\inf_{P\in\mathbb{P}}\ \mathrm{Pr}_P\left\{A(\xi)\right\} = \inf_{P\in\mathbb{P}}\ 1-\mathrm{Pr}_P\left\{A^c(\xi)\right\} = 1-\sup_{P\in\mathbb{P}}\ \mathrm{Pr}_P\left\{A^c(\xi)\right\}. \tag{20}$$

The relation suggests it suffices to consider either the supremum or the infimum. In what follows we focus on P_u.

Let $\mathbb{1}_{\{A(\xi)\}}$ denote the indicator function, which equals 1 if $A(\xi)$ happens and 0 otherwise. Then $\mathrm{Pr}_P\left\{A(\xi)\right\}$ can be rewritten as $\mathrm{E}_P\left[\mathbb{1}_{\{A(\xi)\}}\right]$. For simplicity of notation, we abbreviate $\mathbb{1}_{\{A(\xi)\}}$ by $\mathbb{1}$. Furthermore, we consider the ambiguity set $\mathbb{L}$ defined in Equation (8). Then, P_u corresponds to the following optimization problem

$$\underset{L\in\mathbb{L}}{\text{maximize}}\ \ \mathrm{E}_{P_0}\left[\mathbb{1}L\right], \tag{21}$$

which may be placed within the framework of RS of expectations studied in Section 3.

To simplify the notation, we let $\kappa = \mathrm{Pr}_{P_0}\left\{A(\xi)\right\}$. It is the probability of $A(\xi)$ under the nominal distribution P_0. Note that when $H=\mathbb{1}$,

$$\mathrm{E}_{P_0}\left[\Psi(H+\lambda,\alpha)\right] = \Psi(1+\lambda,\alpha)\kappa + \Psi(\lambda,\alpha)\left(1-\kappa\right).$$

Then, by Theorem 2, we obtain the following result on Problem (21).

Theorem 3. *The optimal value of Problem (21) (i.e., P_u) is equal to that of the following problem*

$$\underset{\lambda\in\Re,\alpha\geq 0}{\textit{minimize}}\ \ \Psi(1+\lambda,\alpha)\kappa + \Psi(\lambda,\alpha)\left(1-\kappa\right) + \sum_{i=1}^{m}\alpha_i\eta_i - \lambda.$$

Theorem 3 builds that the maximal probability is equal to the optimal value of an optimization problem with real decision variables. It clearly shows that estimating the maximum of the probability can be accomplished by estimating the probability under the nominal distribution and by solving a simple optimization problem. Now we discuss in detail how to solve the problem. To unify the analysis, we define for each $y \in [0,1]$,

$$Z(\lambda, \alpha, y) := y\Psi(1+\lambda, \alpha) + (1-y)\Psi(\lambda, \alpha) + \sum_{i=1}^{m} \alpha_i \eta_i - \lambda. \tag{22}$$

Construct the following problem

$$\underset{\lambda \in \Re, \alpha \geq 0}{\text{minimize}} \; Z(\lambda, \alpha, y), \tag{23}$$

and denote its optimal value by $v^*(y)$. Then clearly, $P_u = v^*(\kappa)$. Suppose we have computed the value of κ. We only need to solve (23) for $y = \kappa$. It follows from Proposition 2 that for any given $y \in [0,1]$, $Z(\lambda, \alpha, y)$ is convex in (λ, α), and thus (23) is a convex optimization problem. Because the function $\Psi(s, \alpha)$ is defined by (14) which is in the form of a supremum, we can obtain the dual of (14) and build the corresponding strong duality. This yields the following corollary.

Corollary 1. *Suppose that the intersection of the relative interiors of the effective domains of ϕ_i is nonempty. Then*

$$\Psi(s, \alpha) = \inf_{\mu_1 \geq 0, \mu_2 \leq 0, \sum_{i=1}^m s_i - \mu_1 - \mu_2 = s} \left\{ \sum_{i=1}^{m} \alpha_i \phi_i^* \left(\frac{s_i}{\alpha_i} \right) - a\mu_1 - b\mu_2 \right\}.$$

Corollary 1 essentially generalizes Corollary 4 of Ben-Tal et al. (2013). For completeness, we include the proof of the corollary in Appendix. From Corollary 1, we immediately have that, with the assumption in Corollary 1 satisfied, (23) is equivalent to the following problem:

$$\begin{aligned} \text{minimize} \quad & y \left[\sum_{i=1}^{m} \alpha_i \phi_i^* \left(\frac{s_i}{\alpha_i} \right) - a\mu_1 - b\mu_2 \right] + (1-y) \left[\sum_{i=1}^{m} \alpha_i \phi_i^* \left(\frac{t_i}{\alpha_i} \right) - a\nu_1 - b\nu_2 \right] + \sum_{i=1}^{m} \alpha_i \eta_i - \lambda \\ \text{subject to} \quad & \sum_{i=1}^{m} s_i - \mu_1 - \mu_2 = 1 + \lambda, \sum_{i=1}^{m} t_i - \nu_1 - \nu_2 = \lambda, \\ & \lambda \in \Re, \alpha \geq 0, \mu_1 \geq 0, \mu_2 \leq 0, \nu_1 \geq 0, \nu_2 \leq 0, s_i \in \Re, t_i \in \Re, i = 1, 2, \cdots, m. \end{aligned}$$

For each given y, the problem above is a convex optimization problem. With the conjugates ϕ_i^* given explicitly, it can be solved via conventional convex optimization software packages.

Note that for the ambiguity set $\mathbb{L}_\phi$ defined in Equation (17), Problem (23) can be further simplified as

$$\underset{\lambda \in \Re, \alpha \geq 0}{\text{minimize}} \quad y\alpha\phi^* \left(\frac{1+\lambda}{\alpha} \right) + (1-y)\alpha\phi^* \left(\frac{\lambda}{\alpha} \right) + \alpha\eta - \lambda. \tag{24}$$

Problem (24) has a similar structure as Problem (13) of Ben-Tal et al. (2013). Similarly, it can be reformulated into simpler optimization problems such as conic quadratic program (CQP) and linear program (LP) for various ϕ-divergences, including KL divergence, χ^2-distance, Neyman χ^2-distance, Hellinger distance, χ-distance of order $\theta > 1$, and Cressie-Read divergence. For illustration, we consider the examples of KL divergence and χ^2-distance.

Example 4.1. *For KL divergence, it is easily shown that (24) can be reformulated as*

$$\begin{aligned}
\textit{minimize} \quad & \frac{1}{2} y z_1 + \frac{1}{2}(1-y) z_2 + \alpha\eta - \lambda \\
\textit{subject to} \quad & \alpha \exp\left\{\frac{1+\lambda}{\alpha}\right\} \leq z_1, \ \alpha \exp\left\{\frac{\lambda}{\alpha}\right\} \leq z_2, \ \lambda \in \Re, \ \alpha \geq 0.
\end{aligned}$$

This is a simple perspective-of-exponential program and thus can be solved easily.

Example 4.2. *For χ^2-distance, based on the sum-of-quadratic-over-linear program reformulation in Section 3.1.1, we can further convert (24) equivalently to the following*

$$\begin{aligned}
\textit{minimize} \quad & \frac{1}{4} y_i \mu_3 + \frac{1}{4}(1-y_i)\mu_4 + \alpha(\eta - 1) - \lambda \\
\textit{subject to} \quad & \sqrt{\mu_1^2 + \frac{1}{4}(\alpha - \mu_3)^2} \leq \frac{1}{2}(\alpha + \mu_3), \\
& \sqrt{\mu_2^2 + \frac{1}{4}(\alpha - \mu_4)^2} \leq \frac{1}{2}(\alpha + \mu_4), \\
& \mu_1 \geq 1 + \lambda + 2\alpha, \mu_2 \geq \lambda + 2\alpha, \alpha \geq 0, \mu_1 \geq 0, \mu_2 \geq 0, \lambda, \mu_3, \mu_4 \in \Re.
\end{aligned}$$

This is a simple conic program and thus can be solved readily.

As mentioned, Equation (20) shows that we can use a similar procedure to compute P_l. The procedure requires the input value $\Pr_{P_0}\{A^c(\xi)\}$. Note that $\Pr_{P_0}\{A^c(\xi)\} = 1 - \kappa$. We have $\sup_{P\in\mathbb{P}} \Pr_P\{A^c(\xi)\} = v^*(1-\kappa)$. Therefore, we can also compute P_l by only simulating κ (which is done for P_u). We summarize the result in the following theorem.

Theorem 4. *Suppose that the ambiguity set is $\mathbb{L}$. Then $P_l = 1 - v^*(1-\kappa)$.*

5 Value-at-Risk Performance Measure

We now turn our discussion to VaR. The distribution mis-specification issue for VaR has received much attention in real applications. In insurance and actuarial literature, people have suggested considering the worst-case VaR and using it to guide the reservation of capital to protect against both risks and uncertainties, see, for instance, Wang et al. (2013) and references therein.

Suppose $H(\xi)$ is the random loss. To ease the analysis, we assume $H(\xi)$ is a continuous random variable. Because VaR is a risk measure that is used to quantify the undesired risk, decision makers usually only concern the worst case, i.e., the maximal VaR. But for completeness, we still consider the two quantities, the minimal VaR and the maximal VaR

$$V_l := \inf_{P\in\mathbb{P}} \mathrm{VaR}_{1-\beta,P}(H(\xi)) \quad \text{and} \quad V_u := \sup_{P\in\mathbb{P}} \mathrm{VaR}_{1-\beta,P}(H(\xi)).$$

Consider the maximum and the corresponding optimization problem

$$\underset{P\in\mathbb{P}}{\text{maximize}}\ \ \text{VaR}_{1-\beta,P}(H(\xi)). \tag{25}$$

From the definition of VaR, it is not difficult to verify Problem (25) can be rewritten as

$$\begin{aligned} &\underset{t\in\Re}{\text{minimize}} \quad t \\ &\text{subject to} \quad \Pr_P\left\{H(\xi)-t\le 0\right\}\ge 1-\beta, \quad \forall\ P\in\mathbb{P}. \end{aligned} \tag{26}$$

The constraint in (26) is equivalent to

$$\sup_{P\in\mathbb{P}}\ \mathrm{E}_P\left[\mathbb{1}_{\{H(\xi)-t>0\}}\right]\le\beta.$$

Suppose that we consider the ambiguity set $\mathbb{L}$ defined in Equation (8). Then, the constraint is equivalent to

$$\sup_{L\in\mathbb{L}}\ \mathrm{E}_{P_0}\left[\mathbb{1}_{\{H(\xi)-t>0\}}L\right]\le\beta. \tag{27}$$

Let $\kappa(t)=\Pr_{P_0}\left\{H(\xi)-t>0\right\}$. From Theorem 3 we have that (27) is equivalent to

$$\inf_{\lambda\in\Re,\alpha\ge 0}\ Z(\lambda,\alpha,\kappa(t))\le\beta \tag{28}$$

where $Z(\cdot,\cdot,\cdot)$ is defined in Equation (22). Note that $\inf_{\lambda\in\Re,\alpha\ge 0}Z(\lambda,\alpha,\kappa(t))$ is nondecreasing in $\kappa(t)$. Thus (28) is equivalent to $\kappa(t)\le\beta_l$ for some β_l. We summarize the results in the following theorem, with the complete proof in Appendix.

Theorem 5. *Suppose that the ambiguity set is $\mathbb{L}$. Then*

$$V_u = VaR_{1-\beta_l,P_0}(H(\xi)) \tag{29}$$

with

$$\beta_l=\sup\left\{y:\inf_{\lambda\in\Re,\alpha\ge 0}Z(\lambda,\alpha,y)\le\beta\right\}. \tag{30}$$

Similarly, we have the following theorem for V_l and the proof is included in Appendix.

Theorem 6. *Suppose that the ambiguity set is $\mathbb{L}$. Then*

$$V_l = VaR_{\beta_u,P_0}(H(\xi)), \tag{31}$$

where

$$\beta_u=\inf\left\{y:\inf_{\lambda\in\Re,\alpha\ge 0}Z(\lambda,\alpha,y)\ge 1-\beta\right\}. \tag{32}$$

Theorems 5 and 6 show that the maximal and minimal VaRs are also VaRs with different confidence levels. It is not difficult to verify that $\beta_l\le\beta$ and $\beta_u\le 1-\beta$. This shows that, to compensate the distributional robustness in the specification of the VaRs, a certain amount of allowed error probability needs to be given up. From Theorems 5 and 6, we see that risk and ambiguity are interrelated in an interesting way.

5.1 Computation of New Confidence Level

To simulate the $(1-\beta_l)$-VaR and β_u-VaR, we still need to obtain the new confidence levels β_l and β_u, which are defined by (30) and (32) respectively. Below we discuss how to search β_l and β_u numerically. Because the two quantities are defined by the same optimization structure, we only consider β_l.

Note that $\beta_l \in [0,1]$. Therefore, we only need to seek β_l from $[0,1]$ and β_l is equal to the optimal value of the following optimization problem:

$$\underset{0\le y\le 1, \lambda\in\Re, \alpha\ge 0}{\text{maximize}} \quad y \quad \text{subject to} \quad Z(\lambda,\alpha,y) \le \beta, \tag{33}$$

where $Z(\cdot,\cdot,\cdot)$ is defined by (22). From Section 4, we know that, for any given $y \in [0,1]$, $Z(\lambda,\alpha,y)$ is convex in (λ,α). Furthermore, from Proposition 2, it is not difficult to show that $Z(\lambda,\alpha,y)$ is nondecreasing in y. The nice structures motivate us to design the following Bisection Search procedure to solve Problem (33).

Bisection Search

Step 0. Set $i=0$. Set $y_l := 0$ and $y_u := 1$.

Step i. Set $y_i = \frac{y_l+y_u}{2}$ and solve the following problem to obtain its optimal value v:

$$\underset{\lambda\in\Re, \alpha\ge 0}{\text{minimize}} \quad Z(\lambda,\alpha,y_i).$$

If $v \le \beta$, update $y_l =: y_i$, otherwise, update $y_u =: y_i$. Set $i = i+1$.

For the Bisection Search procedure, we have the following convergence result whose proof is straightforward and thus is omitted.

Proposition 3. *When the Bisection Search procedure is used to solve Problem (33), $y_i \to \beta_l$ as $i \to \infty$ and for any ε, $|y_i - \beta_l| \le \varepsilon$ whenever $i > -\log_2 \varepsilon$.*

Proposition 3 shows that the sequence $\{y_i\}$ generated by the Bisection Search procedure converges to the optimal value of Problem (33), i.e., β_l, and the rate of convergence is in an exponential order. To implement the Bisection Search procedure, we need to solve a sequence of convex optimization problems in Step i. Because the problem in Step i is an instance of Problem (23), it may be solved readily.

6 Construction of Ambiguity Set

We have discussed the computation methods for RS. The remaining question is how to specify the ambiguity set. In this paper we only discuss an objective approach, that is, we specify the size of ambiguity set based on the data information to obtain a confidence region. Suppose that an i.i.d. sample of ξ, denoted by $\{\xi_1, \ldots, \xi_n\}$, from its true distribution is available. We focus on $\mathbb{L}_\phi$ which is most natural for decision makers. Even though ϕ-divergence is used widely, setting $\mathbb{L}_\phi$ based on a sample of observations is not trivial. It is a difficult statistical problem.

We first consider the case where ξ is a discrete random variable (or vector) that has a known finite set of support $\{x_1, x_2, \ldots, x_s\}$. Ben-Tal et al. (2013) studied this case and built an approximate confidence region for the probability mass function p_* based on asymptotic results of ϕ-divergence. Basically, given the sample $\{\xi_1, \ldots, \xi_n\}$, one can estimate $p_*(x_i)$ by just counting how many observations in the sample that equals x_i and dividing it by n for any $i = 1, \ldots, s$. Denote the estimated probability mass function as $\hat{p}$. Then, Ben-Tal et al. (2013) showed that

$$\frac{2n}{\phi''(1)} D_\phi(p_*, \hat{p}) \Rightarrow \chi^2_{s-1}, \quad \text{as} \;\; n \to \infty, \tag{34}$$

where χ^2_{s-1} is a χ^2-distribution with $s-1$ degrees of freedom. Then, an approximate $1-\alpha$ confidence region for p_* is

$$\left\{ p \in \Re^s : p \geq 0, \sum_{i=1}^{s} p_i = 1, D_\phi(p, \hat{p}) \leq \eta \right\},$$

where $\eta = \frac{\phi''(1)}{2n} \chi^2_{s-1,1-\alpha}$. The result shows that the ϕ-divergence from the true distribution to the estimated distribution follows an asymptotic chi-square distribution. The rate of convergence is of order $1/n$, which is consistent with the conventional rate of $1/\sqrt{n}$, because ϕ-divergence is a square of the conventional distance. We can then use $\hat{p}$ as the nominal distribution P_0 with the same ϕ-divergence and same η in the ambiguity set $\mathbb{L}_\phi$. Then, $\mathbb{L}_\phi$ is an approximate $1 - \alpha$ confidence region of the true distribution, expressed in the LR function. This case, where the distribution of ξ is supported on s scenarios, is called the base case by Ben-Tal et al. (2013). In their paper, they also considered more general cases and discussed how to improve the approximate confidence region using correction parameters.

When ξ is a continuous random variable (or vector), constructing an ambiguity set $\mathbb{L}_\phi$ that is similar to the discrete case turns out to be quite difficult. To solve the problem, we propose to use a parametric distribution as a bridge. Suppose that the unknown true density function of ξ is p_{θ^*}, which is known to be a member of a parametric family, denoted as $\{p_\theta\}_{\theta \in \Theta}$. Suppose we have a set of i.i.d. sample $\{\xi_1, \xi_2, \cdots, \xi_n\}$ from ξ. Then, based on the sample, we can obtain the maximum likelihood estimator (MLE) of θ^*, denoted as $\hat{\theta}$. We make some assumptions in Appendix A.7. Based on Theorem 9.1 of Pardo (2006), we have the following proposition. The proof of the proposition follows from the proof of Theorem 9.1 of Pardo (2006) with some modifications. We put it in Appendix.

Proposition 4. *Suppose that Assumptions (i1)-(i5) and (ii1)-(ii2) in Appendix A.7 are satisfied.*

Then

$$\frac{2n}{\phi''(1)} D_\phi(p_{\theta^*}, p_{\hat\theta}) \Rightarrow \chi^2_d, \quad \textit{as} \quad n \to \infty,$$

where χ^2_d *denotes the chi-square distribution with degree of freedom* d *and* d *is the degree of freedom of* θ *(i.e., the number of free elements in the vector* θ*).*

Note that the result in (34) for the discrete case may be viewed as a special case of Proposition 4, because the discrete distribution may be viewed as a parametric distribution with $s-1$ free parameters (i.e., $p_*(x_1), \ldots, p_*(x_{s-1})$).

Based on Proposition 4, given n observations of ξ, an approximate $1-\alpha$ confidence region for p_{θ^*} is

$$\mathrm{CR}_{p_\theta} := \left\{p_\theta : \theta \in \Theta, D_\phi(p_\theta, p_{\hat\theta}) \le \eta\right\},$$

where $\eta = \frac{\phi''(1)}{2n}\chi^2_{d,1-\alpha}$. However, CR_{p_θ} is not in the form of $\mathbb{L}_\phi$ as the distributions in CR_{p_θ} are restricted to the parametric family. Replacing p_θ with a general density $p \in \mathbb{D}$ in CR_{p_θ}, we obtain the following ambiguity set

$$\mathrm{CR}_p := \left\{p : p \in \mathbb{D}, D_\phi(p, p_{\hat\theta}) \le \eta\right\},$$

which may be converted to $\mathbb{L}_\phi$ by setting $p_{\hat\theta}$ as the nominal distribution. Because CR_{p_θ} is a subset of CR_p, CR_p is also an approximate $1-\alpha$ confidence region for p_{θ^*}. We use CR_p in the RS analysis. It is easy to see that, when using CR_p to replace CR_{p_θ}, we convexify the ambiguity set and make the RS problem tractable. The cost is that using the larger ambiguity set CR_p in our RS analysis may yield a conservative lower bound and upper bound for the real performance measure. Fortunately, the set CR_p still shrinks in the order of $1/n$ and the size of CR_p may be controlled by the sample size n.

When there is no data for specifying the ambiguity set, we can absorb some expert opinion to the ambiguity set. For instance, suppose that we believe the random vector ξ follows a parametric density p_{θ^*} and the nominal estimate (or the best guess) of θ^* is θ_0. If it is believed that $\theta^* \in \Theta$, then we may let $\eta = \sup_{\theta\in\Theta} D_\phi(p_\theta, p_{\theta_0})$ and construct the following ambiguity set

$$\mathrm{CR}_p := \{p : p \in \mathbb{D}, D_\phi(p, p_{\theta_0}) \le \eta\},$$

which includes all p_θ with $\theta \in \Theta$. This implies $p_{\theta^*} \in \mathrm{CR}_p$. We can then covert CR_p to the ambiguity set $\mathbb{L}_\phi$. In such approach, the problem becomes whether η can be computed efficiently. Because the ϕ-divergence may be derived analytically for many parametric distributions, computing η then becomes a deterministic optimization problem that may be easy to solve.

7 Numerical Illustrations

In this section, we consider two test examples from health care management and financial risk management. We conduct numerical experiments to illustrate the implementation of our RS approach.

7.1 Response of Emergency Medical Service

Consider a simplified model on emergency medical service (EMS) operation. Suppose the emergency call may occur at any point $\xi = (\xi_1, \xi_2)$ of a geographic region (the whole plane, measured by km) with a joint normal distribution $P := N(\mu, \Sigma)$. Suppose there are five EMS stations located at $O(0,0), A(12,0), B(0,12), C(-12,0)$, and $D(0,-12)$. Once a call arrives, an ambulance in the nearest station will be equipped to serve it. Let $(\zeta_1(\xi), \zeta_2(\xi))$ denote the location of the response station for ξ. For simplicity, assume the speed of any ambulance is a constant $v = 40$km/h. Then, the response time for ξ is

$$H(\xi) = v^{-1}\sqrt{(\xi_1 - \zeta_1(\xi))^2 + (\xi_2 - \zeta_2(\xi))^2}.$$

We are primarily interested in two measures, the expected response time $\mathrm{E}_P[H(\xi)]$ and the percentage of late calls. The percentage of late calls for EMS is often concerned in health care management practice. A call is taken to be late if the response time exceeds a threshold, see, e.g., Maxwell et al. (2014). In this example, we consider the nine minutes threshold. We are interested in the late percentage

$$\mathrm{Pr}_P\{H(\xi) > 9/60\}.$$

While it is not impossible to derive analytical expressions for the expectation and the probability, we estimate them via simulation.

Assume the true distribution is $P = N(\mathbf{0}, 10 \times \mathbf{I})$ where $\mathbf{0}$ is the zero vector and $\mathbf{I}$ is the identity matrix. Because our random vector follows a 2 dimensional multivariate normal distribution, the degree of freedom of the chi-square distribution is $d = 5$. In the experiment, we mainly illustrate the implementation of RS, so for simplicity we assume the nominal distribution P_0 is just the true distribution P. We consider different confidence level α and sample size n and conduct RS with ambiguity set $\mathbb{L}_\phi$ where $\eta = \frac{\phi''(1)}{2n}\chi^2_{d,1-\alpha}$.

We first consider the expected response time. The nominal expected response time under P_0 is estimated to be 0.0928 hours (5.5680 minutes). We use the KL divergence and χ^2-distance to construct the ambiguity set. As shown in Section 3.1.1, for both KL divergence and χ^2-distance, the sample problem (19) can be reformulated as convex optimization problems which can be handled efficiently by CVX. In the experiment, we use Matlab to call CVX to solve the problems. We use a sample size 1000 in the SAA problem and simultaneously compute the maximal expectation (denoted by E_u) and the minimal expectation (denoted by E_l). Table 2 reports the computational results, which are averaged over five typical replications. From Table 2 we can see that input uncertainty has a significant impact on the performance. On the other hand, when the sample size n increases, the index of ambiguity η reduces, and the maximal expectation and minimal expectation become closer and closer to the nominal value, supporting that the uncertainty in output may be suppressed by reducing the input uncertainty.

Table 2: Robust Simulation for Expected Response Time (Minutes)

α	n		η	E_l	E_u		η	E_l	E_u
0.2	10	KL	0.3645	3.5183	7.7313	χ^2-	0.7289	3.5286	7.7508
0.2	50	divergence	0.0729	4.6575	6.5736	distance	0.1458	4.6384	6.5802
0.2	100		0.0365	4.8638	6.2399		0.0729	4.8958	6.2681
0.1	10		0.4618	3.3133	8.0211		0.9236	3.2495	7.9148
0.1	50		0.0924	4.5013	6.6825		0.1847	4.5157	6.7004
0.1	100		0.0462	4.8546	6.3948		0.0924	4.8348	6.4215
0.05	10		0.5535	3.0000	8.1786		1.1070	3.0631	8.1027
0.05	50		0.1107	4.4389	6.8099		0.2214	4.3545	6.7262
0.05	100		0.0554	4.7075	6.4160		0.1107	4.7449	6.4355

We next consider the late percentage. The nominal late percentage under P_0 is estimated to be 0.0912. Since the indicator function is bounded, we can use any divergences. In the experiment, we consider KL divergence, χ^2-distance, Neyman χ^2-distance and Hellinger distance. As discussed in Section 4, for both KL divergence and χ^2-distance, Problem (24) can be reformulated as optimization problems which can be solved easily. Also, it is easily shown that for Neyman χ^2-distance and Hellinger distance, (24) can be reformulated as conic quadratic programs, which can be solved readily via CVX. We conduct RS of the probability measure using the approach developed in Section 4, and report the computational results for different values of α and n in Table 3.

Table 3: Robust Simulation for Late Percentage

α	n	divergence	η	P_l	P_u	divergence	η	P_l	P_u
0.2	10	KL	0.3645	0.0000	0.4116	χ^2-distance	0.7289	0.0000	0.3370
0.2	100	divergence	0.0365	0.0248	0.1780		0.0729	0.0135	0.1689
0.2	1000		0.0037	0.0675	0.1169		0.0073	0.0666	0.1158
0.1	10		0.4618	0.0000	0.4593		0.9236	0.0000	0.3679
0.1	100		0.0462	0.0184	0.1900		0.0924	0.0037	0.1787
0.1	1000		0.0046	0.0649	0.1200		0.0092	0.0636	0.1188
0.05	10		0.5535	0.0000	0.5006		1.1070	0.0000	0.3941
0.05	100		0.0554	0.0134	0.2004		0.1107	0.0000	0.1870
0.05	1000		0.0056	0.0623	0.1231		0.0111	0.0609	0.1215
0.2	10	Neyman	0.7289	0.0093	0.5178	Hellinger	0.1822	0.0000	0.4516
0.2	100	χ^2-distance	0.0729	0.0390	0.1990	distance	0.0182	0.0292	0.1828
0.2	1000		0.0073	0.0695	0.1188		0.0018	0.0683	0.1171
0.1	10		0.9236	0.0076	0.5673		0.2309	0.0000	0.5067
0.1	100		0.0924	0.0352	0.2164		0.0231	0.0237	0.1962
0.1	1000		0.0092	0.0672	0.1227		0.0023	0.0655	0.1207
0.05	10		1.1070	0.0065	0.6054		0.2768	0.0000	0.5537
0.05	100		0.1107	0.0323	0.2315		0.0277	0.0195	0.2079
0.05	1000		0.0111	0.0652	0.1262		0.0028	0.0631	0.1239

From the table, we can see that the late percentage is quite sensitive to the input distribution, reflecting the importance of the input uncertainty issue. Similar to the expectation, when the sample size n increases, the maximal probability and minimal probability become closer and closer to the nominal value. However, we require a relatively large sample size for the input data so that the output error is small enough. This further reflects the difficulty of input modeling.

7.2 Financial Risk Model

As a second example, we consider the VaR problem of a portfolio described in Hong et al. (2014b). Suppose an investor holds k assets with random returns $\xi = (\xi_1, \cdots, \xi_k)^{\mathrm{T}}$ where ξ_j is the return of asset j. Suppose x_j is the capital invested in asset j. Then the random loss of the portfolio is $H(\xi) = -x^{\mathrm{T}}\xi$ where $x = (x_1, \cdots, x_k)^{\mathrm{T}}$. We consider the VaR of the loss, $\mathrm{VaR}_{1-\beta,P}(H(\xi))$. Again, we assume the return vector ξ follows a multivariate normal distribution $P := N(\mu, \Sigma)$. Under this assumption, the VaR admits an analytical expression:

$$\mathrm{VaR}_{1-\beta,P}(H(\xi)) = -\mu^{\mathrm{T}}x + \kappa(\beta)\sqrt{x^{\mathrm{T}}\Sigma x} \quad \text{with} \quad \kappa(\beta) = \Phi^{-1}(1-\beta), \tag{35}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cumulative distribution function. Thus, given all the parameters, the VaR can be computed directly via (35). In the experiment, we set $k = 10$, and assume $\mu = (\mu_1, \cdots, \mu_k)^{\mathrm{T}}$ evenly spread between 0.04 and 0.50 and increase with i, the standard deviation is $\mu_i + 0.05$ for all $i = 1, \cdots, k$ and the correlation between ξ_i and ξ_j is 0.35 for any $i \neq j$. We also assume the nominal distribution is the same as P.

We consider two confidence levels $\beta = 0.1, 0.05$. For the nominal distribution $N(\mu, \Sigma))$, it can be computed via (35) that $\mathrm{VaR}_{0.9,P_0}(H(\xi)) = 0.0020$ and $\mathrm{VaR}_{0.95,P_0}(H(\xi)) = 0.0725$. The values are used as benchmark in RS. We also use $\mathbb{L}_\phi$ and consider the four ϕ-divergences as for the probability performance measure in preceding example. Since $k = 10$, we have the degree of freedom $d = (k^2 + 3k)/2 = 65$. This shows the dimension of the distribution parameters is high, suggesting that a data set with a large sample size may be required to model the input uncertainty. The computational results for different combinations of α and n are summarized in Table 4. In Table 4, the β_l column computes the new confidence level and V_u is the maximal VaR, derived via (35). From Table 4 we see again the significant impact of the input uncertainty on the output performance. For the Neyman χ^2-distance and Hellinger distance, there are zeros in the β_l column, for which the VaR should be $+\infty$ (we left it as blank). The reason is that for any $y > 0$, (33) has no feasible solution. Only when $y = 0$, (33) is feasible. This phenomenon is caused by the fact that the effective domain of ϕ^* is not $\Re$. Consequently, the feasible sets in Step i of Bisection Search are different for the cases $y = 0$ and $y > 0$.

8 Conclusions

RS studied in this paper provides a new approach to analyzing input uncertainty. When the ambiguity is modeled by LR, we show that the RS problems are quite tractable for three most important performance measures: expectation, probability, and VaR. The study is just a start of RS and, even under our framework, it is far from complete. For the RS approach being a practical

Table 4: Robust Simulation for Value-at-rsik

α	n	divergence	η	β_l	V_u	divergence	η	β_l	V_u
0.1	100	KL	0.3998	7.1560e-004	0.3724	χ^2-distance	0.7997	0.0102	0.2035
0.1	1000	divergence	0.0400	0.0362	0.1020	$\beta = 0.1$	0.0800	0.0428	0.0870
0.1	10000	$\beta = 0.1$	0.0040	0.0753	0.0322		0.0080	0.0763	0.0309
0.05	100		0.4241	5.6044e-004	0.3860		0.8482	0.0097	0.2072
0.05	1000		0.0424	0.0349	0.1053		0.0848	0.0418	0.0891
0.05	10000		0.0043	0.0745	0.0333		0.0085	0.0756	0.0318
0.1	100	$\beta = 0.05$	0.3998	6.3555e-006	0.6010	$\beta = 0.05$	0.7997	0.0028	0.2912
0.1	1000		0.0400	0.0103	0.2028		0.0800	0.0153	0.1731
0.1	10000		0.0040	0.0329	0.1104		0.0080	0.0338	0.1081
0.05	100		0.4241	3.9090e-006	0.6214		0.8482	0.0027	0.2935
0.05	1000		0.0424	0.0097	0.2072		0.0848	0.0148	0.1756
0.05	10000		0.0043	0.0323	0.1120		0.0085	0.0334	0.1091
0.1	100	Neyman	0.7997	0		Hellinger	0.1999	0	
0.1	1000	χ^2-distance	0.0800	0.0151	0.1741	distance	0.0200	0.0321	0.1126
0.1	10000	$\beta = 0.1$	0.0080	0.0732	0.0352	$\beta = 0.1$	0.0020	0.0748	0.0329
0.05	100		0.8482	0			0.2120	0	
0.05	1000		0.0848	0.0126	0.1878		0.0212	0.0307	0.1164
0.05	10000		0.0085	0.0723	0.0364		0.0021	0.0742	0.0338
0.1	100	$\beta = 0.05$	0.7997	0		$\beta = 0.05$	0.1999	0	
0.1	1000		0.0800	0			0.0200	0.0070	0.2304
0.1	10000		0.0080	0.0305	0.1170		0.0020	0.0323	0.1120
0.05	100		0.8482	0			0.2120	0	
0.05	1000		0.0848	0			0.0212	0.0064	0.2366
0.05	10000		0.0085	0.0299	0.1187		0.0021	0.0319	0.1131

methodology, a number of issues still need to be addressed. An important problem/issue is how to determine the ambiguity set based on the data available. We have discussed only the parametric case. For the more interesting and more practical nonparametric case, however, the problem is still open to us. If ideally, we can construct a confidence region for the unknown true distribution using a ϕ-divergence, then the RS approach can guarantee an exact confidence interval corresponding to the confidence region for the true performance measure. To resolve the problem, asymptotic statistical properties for the ϕ-divergence need to be built. However, these are significantly understudied in the literature and we will consider them in our future studies. Besides the specification of the ambiguity size, an important problem is the selection of ϕ-divergence. This is also a research question that needs answer. Another important problem is how to extend the RS approach to the cases of dynamic simulations, where multiple independent observations of the same uncertain input distributions are necessary in the simulation study, and still maintain the mathematical tractability. We are certainly interested in this problem and will study it in the future. Finally, in this paper, we have focused on the input error but have not taken into consideration the simulation error. In contrast to the input error, the simulation error can be controlled within any precision by setting the simulation effort. However, when implementing the RS approach in practice, the simulation error can not be ignored completely. Quantifying the simulation error in our RS approach will be a consecutive research problem.

Acknowledgments

A preliminary version of this paper (Hu and Hong 2015) was published in the *Proceedings of the 2015 Winter Simulation Conference*.

A Appendix

A.1 Proof of Theorem 1

Proof. To prove Theorem 1 and Proposition 1, we take a similar approach as that for proving Theorem 4.2 in Ben-Tal and Teboulle (2007). The general convex duality theory implies that the strong duality between (11) and (12) holds if the following constraint qualification holds (see, e.g., Corollary 4.8 of Borwein and Lewis 1992): $\exists\ L \in L^1$ such that $\mathrm{E}_{P_0}[L] = 1$, $\alpha < L < \beta$ a.e. Note that we assume $\alpha < 1 < \beta$. By setting $L = 1$, this constraint qualification holds. This concludes the proof. □

A.2 Proof of Proposition 1

Proof. Proposition 1 follows from Theorem 4.1 of Ben-Tal and Teboulle (2007) and the proof of Theorem 4.2 of Ben-Tal and Teboulle (2007). The difference is the integrand. In our problem, the integrand is also in normal convex. This justifies the result. □

A.3 Proof of Proposition 2

Proof. For any fixed t, $st - \sum_{i=1}^{m} \alpha_i \phi_i(t)$ is linear, and thus convex, in (s, α). It is well known that the supremum preserves convexity. Thus $\Psi(s, \alpha)$ is convex in (s, α). Furthermore, $t \in \mathbb{L}(a, b)$ is always non-negative. Therefore, $\Psi(s, \alpha)$ is non-decreasing in s for any given $\alpha \geq 0$. Note that $1 \in \mathbb{L}(a, b)$ and $\phi_i(1) = 0$. We have $\Psi(s, \alpha) \geq s - \sum_{i=1}^{m} \alpha_i \phi_i(1) = s$. □

A.4 Proof of Corollary 1

Proof. We compute $\Psi(s,\alpha)$ using Lagrangian duality and properties of the conjugate:

$$\begin{aligned}
\Psi(s,\alpha) &= \sup_{t\geq 0} \inf_{\mu_1\geq 0,\mu_2\leq 0} \left\{ st - \sum_{i=1}^{m} \alpha_i \phi_i(t) + \mu_1(t-a) + \mu_2(t-b) \right\} \\
&= \inf_{\mu_1\geq 0,\mu_2\leq 0} \sup_{t\geq 0} \left\{ (s+\mu_1+\mu_2)\,t - \sum_{i=1}^{m} \alpha_i \phi_i(t) - a\mu_1 - b\mu_2 \right\} \\
&= \inf_{\mu_1\geq 0,\mu_2\leq 0} \left\{ \left(\sum_{i=1}^{m} \alpha_i \phi_i \right)^{*} (s+\mu_1+\mu_2) - a\mu_1 - b\mu_2 \right\} \\
&= \inf_{\mu_1\geq 0,\mu_2\leq 0} \inf_{\sum_{i=1}^{m} s_i = s+\mu_1+\mu_2} \left\{ \sum_{i=1}^{m} \alpha_i \phi_i^{*}\left(\frac{s_i}{\alpha_i}\right) - a\mu_1 - b\mu_2 \right\},
\end{aligned}$$

where the second equality follows from strong duality, the third one follows from the definition of the conjugate, and the last one follows from Proposition 1 of Ben-Tal et al. (2013) and the property that $(\alpha\phi)^{*}(s) = \alpha\phi^{*}(s/\alpha)$ (Ben-Tal et al. 2013). This concludes the proof. □

A.5 Proof of Theorem 5

Proof. We show (28) is equivalent to

$$\kappa(t) \leq \beta_l, \tag{36}$$

where β_l is defined by (30). From the definition of β_l, any $\kappa(t)$ satisfying (28) satisfies (36). Thus it suffices to show the opposite direction. Note that $\inf_{\lambda\in\Re,\alpha\geq 0} Z(\lambda,\alpha,\kappa)$ is a concave and nondecreasing in κ. Furthermore, it equals 0 when $\kappa = 0$ and equals 1 when $\kappa = 0$. Therefore, $\inf_{\lambda\in\Re,\alpha\geq 0} Z(\lambda,\alpha,\kappa)$ is continuous function of κ in $(0,1]$. This shows $\inf_{\lambda\in\Re,\alpha\geq 0} Z(\lambda,\alpha,\kappa(t)) \leq \beta$ when $\kappa(t) = \beta_l$. Therefore, any $\kappa(t)$ satisfying (36) satisfies (28).

The equivalence between (28) and (36) implies that Problem (26) can be transformed as

$$\begin{aligned}
&\underset{t\in\Re}{\text{minimize}} && t \\
&\text{subject to} && \Pr_{\sim P_0}\{H(\xi) - t \leq 0\} \geq 1 - \beta_l,
\end{aligned}$$

whose optimal value is (29) from the definition of the $(1-\beta_l)$-VaR. □

A.6 Proof of Theorem 6

Proof. The minimum V_l corresponds to the following optimization

$$\underset{P\in\mathbb{P}}{\text{minimize}}\ \mathrm{VaR}_{1-\beta,P}(H(\xi)). \tag{37}$$

From the definition of VaR, it is not difficult to verify Problem (37) can be rewritten as

$$\begin{aligned}
&\underset{t\in\Re}{\text{minimize}} && t \\
&\text{subject to} && \Pr_{\sim P}\{H(\xi) - t \leq 0\} \geq 1 - \beta, \quad \text{for some } P \in \mathbb{P}.
\end{aligned} \tag{38}$$

The constraint in (38) is equivalent to

$$\underset{P\in\mathbb{P}}{\text{maximize}}\ \mathrm{E}_P\left[\mathbb{1}_{\{H(\xi)-t\leq 0\}}\right]\geq 1-\beta. \tag{39}$$

Let $\kappa(t)=\Pr_{\sim P_0}\{H(\xi)-t\leq 0\}$ and consider only the ambiguity set $\mathbb{L}$ defined in Equation (8). From Theorem 3 we have that (39) is equivalent to

$$\inf_{\lambda\in\Re,\alpha\geq 0} Z(\lambda,\alpha,\kappa(t))\geq 1-\beta \tag{40}$$

where $Z(\cdot,\cdot,\cdot)$ is defined by Equation (22). We show (40) is equivalent to

$$\kappa(t)\geq\beta_u \tag{41}$$

if $\beta_u>0$, and is equivalent to

$$\kappa(t)>\beta_u \tag{42}$$

if $\beta_u=0$, where β_u is defined by (32).

Note that $\inf_{\lambda\in\Re,\alpha\geq 0} Z(\lambda,\alpha,\kappa)$ is a concave and nondecreasing function of κ on $[0,1]$. It is strictly increasing before it reaches its maximal value 1. The only possible non-continuous point is $\kappa=0$. Therefore, when $\beta_u>0$, $\inf_{\lambda\in\Re,\alpha\geq 0} Z(\lambda,\alpha,\beta_u)=1-\beta$. Therefore, (40) is equivalent to (41). When $\beta_u=0$, we have $\inf_{\lambda\in\Re,\alpha\geq 0} Z(\lambda,\alpha,\kappa)>1-\beta$ for all $\kappa>0$ and $\inf_{\lambda\in\Re,\alpha\geq 0} Z(\lambda,\alpha,0)=0$. Thus (40) is equivalent to (42). Finally, since H is a continuous random variable, both cases indicates that the optimal value of Problem (38) is equal to that of

$$\begin{aligned}&\underset{t\in\Re}{\text{minimize}} && t\\ &\text{subject to} && \kappa(t)\geq\beta_u,\end{aligned}$$

whose optimal value is (31) from the definition of the β_u-VaR. □

A.7 Assumptions for Proposition 4

Let $(\Xi,\mathcal{F}_\Xi,P_\theta)_{\theta\in\Theta}$ be the statistical space and μ be the Lebesgue measure. We make the following assumptions.

(i1) For all $\theta_1\neq\theta_2\in\Theta\subset\Re^d$,

$$\mu\left(\{z\in\Xi: p_{\theta_1}(z)\neq p_{\theta_2}(z)\}\right)>0.$$

(i2) The set $S_\Xi=\{z\in\Xi: p_\theta(z)>0\}$ is independent of θ.

(i3) The first, second and third partial derivatives

$$\frac{\partial p_\theta(z)}{\partial\theta_i},\frac{\partial^2 p_\theta(z)}{\partial\theta_i\partial\theta_j},\frac{\partial^3 p_\theta(z)}{\partial\theta_i\partial\theta_j\partial\theta_k}, i,j,k=1,\cdots,d$$

exist everywhere.

(i4) The first, second and third partial derivatives of $p_\theta(z)$ with respect to θ are absolutely bounded by functions α, β and γ with finite integrals

$$\int_\Xi \alpha(z)d\mu(z) < \infty, \int_\Xi \beta(z)d\mu(z) < \infty, \int_\Xi \gamma(z)d\mu(z) < \infty.$$

(i5) For each $\theta \in \Theta$, the Fisher information matrix

$$\mathcal{I}_\mathcal{F}(\theta) = \left(\int_\Xi \frac{\partial \log p_\theta(z)}{\partial \theta_i} \frac{\partial \log p_\theta(z)}{\partial \theta_j} p_\theta(z) d\mu(z) \right)_{i,j=1,\cdots,d}$$

exists and is positive definite, with elements continuous in the variable θ.
(ii1) The function ϕ is twice continuously differentiable, with $\phi''(1) > 0$.
(ii2) For each $\theta_0 \in \Theta$ there exists an open neighborhood $N(\theta_0)$ such that for all $\theta \in N(\theta_0)$ and $1 \leq i, j \leq d$ it holds:

$$\frac{\partial}{\partial \theta_i} D_\phi(p_\theta, p_{\theta_0}) = \mathrm{E}_{P_0}\left[\frac{\partial}{\partial \theta_i} \phi\left(\frac{p_\theta(\xi)}{p_{\theta_0}(\xi)} \right) \right],$$
$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} D_\phi(p_\theta, p_{\theta_0}) = \mathrm{E}_{P_0}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \phi\left(\frac{p_\theta(\xi)}{p_{\theta_0}(\xi)} \right) \right],$$

and these expressions are continuous on $N(\theta_0)$.

A.8 Proof Of Proposition 4

Proof. To build the asymptotics for $D_\phi(p_{\theta^*}, p_{\hat{\theta}})$, we construct the function $\varphi(t) := t\phi(t^{-1})$. It can be verified that φ is also a divergence function. Moreover, $\phi''(1) = \varphi''(1)$ and $D_\phi(p_{\theta^*}, p_{\hat{\theta}}) = D_\varphi(p_{\hat{\theta}}, p_{\theta^*})$. Thus it suffices to build asymptotics for $D_\varphi(p_{\hat{\theta}}, p_{\theta^*})$.

To express the divergence as a function of the parameter vectors, we use $D_\varphi(\theta, \theta^*)$ to denote $D_\varphi(p_\theta, p_{\theta^*})$. Note that $D_\varphi(\theta, \theta^*)$ is a function of θ. We can take a second order Taylor expansion for $D_\varphi(\theta, \theta^*)$ at θ^*, evaluated at $\theta = \hat{\theta}$. Following the proof (Pages 411-421) of Pardo (2006), we can obtain that

$$\frac{2n}{\varphi''(1)} D_\varphi(\hat{\theta}, \theta^*) \Rightarrow \chi^2_d.$$

Note that in the whole deduction, we shall interpret the integral as a multivariate integral. The proof is finished by noting $D_\phi(p_{\theta^*}, p_{\hat{\theta}}) = D_\varphi(p_{\hat{\theta}}, p_{\theta^*})$. □

References

Artzner, P., F. Delbaen, J.-M. Eber, D. Heath. 1999. Coherent measures of risk. *Mathematical Finance*, **9**(3) 203-228.

Barton, R. R. 2012. Tutorial: Input uncertainty in output analysis. *Proceedings of the 2003 Winter Simulation Conference*. Laroque C, Himmelspach J, Pasupathy R, Rose O, Uhrmacher AM, eds. IEEE Computer Society, Washington, DC, 67-78.

Barton. R. R., B. L. Nelson, W. Xie. 2014. Quantifying input uncertainty via simulation confidence intervals. *INFORMS Journal on Computing*, **26**(1) 74-87.

Barton, R. R., L. W. Schruben. 1993. Uniform and Bootstrap resampling of empirical distributions. *Proceedings of the 1993 Winter Simulation Conference*. G. W. Evans, M. Mollaghasemi, W. E. Biles, and E. C. Russell, eds. Piscataway, NJ, 503-508.

Barton, R. R., L. W. Schruben. 2001. Resampling methods for input modeling. *Proceedings of the 2001 Winter Simulation Conference*. B. A. Peters, J. S. Smith, D. J. Medeiros and M. W. Rohrer, eds. Piscataway, NJ, 372-378.

Ben-Tal, A., D. Bertsimas, D. Brown. 2010. A soft robust model for optimization under ambiguity. *Operations Research*, **58**(4) 1220-1234.

Ben-Tal, A., D. den Hertog, A. M. B. de Waegenaere, B. Melenberg, G. Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, **59**(2) 341-357.

Ben-Tal, A., M. Teboulle. 2007. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, **17**(3) 449-476.

Biller, B., C. Corlu. 2011. Accounting for parameter uncertainty in large-scale stochastic simulations with correlated inputs. *Operations Research*, **59** 661-673.

Borwein, J. M., A. S. Lewis. 1992. Partially finite convex programming. I. Quasi relative interiors and duality theory. *Mathematical Programming*, **57** 15-48.

Boyd, S., L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, Cambridge, U.K.

Casella, G., R. L. Berger. 2002. *Statistical Inference*, Second Edition. Duxbury Press.

Cheng, R. C. H., W. Holland. 1997. Sensitivity of computer simulation experiments to errors in input data. *Journal of Statistical Computation and Simulation*, **57** 219-241.

Chick, S. E. 1997. Bayesian analysis for simulation input and output. *Proceedings of the 1997 Winter Simulation Conference*. S. Andradottir, K. J. Healy, D. H. Withers, and B. L. Nelson, eds. Piscataway, NJ, 253-260.

Chick, S. E. 2001. Input distribution selection for simulation experiments: Accounting for input uncertainty. *Operations Research*, **49**(5) 744-758.

Delage, E., Y. Ye. 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, **58** 595-612.

Dellino, G., Kleijnen, J. P., Meloni, C. 2012. Robust optimization in simulation: Taguchi and Krige Combined. *INFORMS Journal on Computing*, **24** 471-484.

Durrett, R. 2005. *Probability: Theory and Examples*, Third Edition. Duxbury Press, Belmont.

Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, **75** 643-669.

Epstein, L. G. 1999. A definition of uncertainty aversion. *Review of Economic Studies*, **66** 579-608.

Fan, W. W., L. J. Hong, X. W. Zhang. 2013. Robust selection of the best. *Proceedings of the 2013 Winter Simulation Conference*. R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds. IEEE Press Piscataway, NJ, 868-876.

Freimer, M., L. W. Schruben. 2002. Collecting data and estimating parameters for input distributions. *Proceedings of the 2002 Winter Simulation Conference*. E. Yücesan, C. Chen, J. L. Snowdon, and J. M. Charnes, eds. Piscataway, NJ, 392-399.

Ghadimi, S., G. Lan. 2015. Stochastic approximation methods and their finite-time convergence properties. Chapter 7, *Handbook of Simulation Optimization*, Fu, M. C. Editor, Springer.

Grant, M., S. Boyd. 2013. CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx.

Henderson, S. G. 2003. Input model uncertainty: Why do we care and what should we do about it? *Proceedings of the 2003 Winter Simulation Conference*. S. Chick, P. J. Sanchez, D. Ferrin, and D. J. Morrice, eds. IEEE, Piscataway NJ, 90-100.

Hong, L. J., Z. Hu., G. Liu. 2014a. Monte Carlo methods for value-at-risk and conditional value-at-risk: A review. *ACM Transactions on Modeling and Computer Simulation*, **24**(4) Article 22.

Hong, L. J., Z. Hu., L. Zhang. 2014b. Conditional value-at-risk approximation to value-at-risk constrained programs: A remedy via Monte Carlo. *INFORMS Journal on Computing*, **26**(2) 385-400.

Hu, Z., J. Cao, L. J. Hong. 2012. Robust simulation of global warming policies using the DICE model. *Management Science*, **58**(12) 2190-2206.

Hu, Z., L. J. Hong. 2015. Robust simulation of stochastic systems with input uncertainties modeled by statistical divergences. *Proceedings of the 2015 Winter Simulation Conference*, forthcoming.

Jorion, P. 2006. *Value at Risk*, Second Edition. New York: McGraw-Hill, Inc.

Lam, H. 2012. Robust sensitivity analysis for stochastic systems. Working paper.

Law, A. M., Kelton, W. D. 2000. *Simulation Modeling and Analysis*, Third Edition. McGraw Hill.

Maxwell, M. S., E. C. Ni, C. Tong, S. R. Hunter, S. G. Henderson, and H. Topaloglu. 2014. A bound on the performance of an optimal ambulance redeployment policy". *Operations Research*, **62**(5): 1014-1027.

Nemirovski, A., A. Juditsky, G. Lan, A. Shapiro. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* **19**(4) 1574-1609.

Ng, S. H., S. E. Chick. 2006. Reducing parameter uncertainty for stochastic systems. *ACM Transactions on Modeling and Computer Simulation* **16** 26-51.

Pardo, L. 2006. *Statistical Inference Based on Divergence Measures*, Chapman & Hall/CRC, Boca Raton, Florida.

Pierreval, H., Durieux-Paris, S. 2007. Robust simulation with a base environmental scenario. *European Journal of Operational Research*, **182** 783-793.

Sanchez, S. M. 2000. Robust design: seeking the best of all possible worlds. *Proceedings of the 2000 Winter Simulation Conference*. J. A. Joines, R. R. Barton, K. Kang, P. A. Fishwick, eds. Piscataway, NJ, 69-76.

Shapiro, A., D. Dentcheva, A. Ruszczyński. 2014. *Lectures on Stochastic Programming: Modeling and Theory*, Second Edition. SIAM, Philadelphia.

Wang, R., L. Peng, J. Yang. 2013. Bounds for the sum of dependent risks and worst Value-at-Risk with monotone marginal densities. *Finance and Stochastics*, **17**(2) 395-417.

Xie, W., B. L. Nelson, R. R. Barton. 2014. A Bayesian framework for quantifying uncertainty in stochastic simulation. *Operations Research*, **62**(6) 1439-1452.

Zouaoui, F., J. R. Wilson. 2004. Accounting for input-model and input-parameter uncertainties in simulation. *IIE Transactions*, **36** 1135-1151.