Ambiguous Probabilistic Programs

Zhaolin Hu
School of Economics and Management, Tongji University, Shanghai 200092, China

L. Jeff Hong
Department of Industrial Engineering and Logistics Management
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

Anthony Man-Cho So
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, N. T., Hong Kong, China

Abstract

Probabilistic programs are widely used decision models. When implemented in practice, however, there often exists distributional ambiguity in these models. In this paper, we model the ambiguity using the likelihood ratio (LR) and use LR to construct various ambiguity sets. We consider ambiguous probabilistic programs which optimize under the worst case. Ambiguous probabilistic programs can be classified as ambiguous probability minimization problems (PM) and ambiguous chance constrained programs (CCP). We show that the ambiguous PM can be transformed to a pure PM under the nominal distribution, and that the ambiguous CCP can be transformed to a pure CCP with only the confidence level being rescaled from the original CCP. Our study indicates that ambiguous probabilistic programs with ambiguity modeled by LR essentially have the same complexity as the corresponding pure probabilistic programs and that risk and uncertainty have strong connections in probabilistic programs.

1 Introduction

Decision making often seeks stochastic programming models, where random outputs of a system of interested are mapped to some deterministic values using functionals and such deterministic values are then used for decision. Expectation is a frequently used functional. It represents the average of the output and is typically used if decision makers are risk-neutral. Probability is another functional of great interest. It represents the occurrence chance of some events that are of concern. It is often advocated by decision makers who are risk-averse to randomness. The use of probability functional in optimization models results in the so called probabilistic programs, which have been an important class of stochastic optimization problems and have been studied extensively in the literature. For a comprehensive review, readers are referred to Prékopa (2003).

In various situations, a decision maker hopes to optimize (say minimize) the probability of some events. Then she may formulate her problem as the following probabilistic program:

$$\underset{x\in X}{\text{minimize}} \ \Pr_{\sim P_0}\left\{H(x,\xi)>0\right\}, \tag{1}$$

where $X \subset \Re^d$ is the feasible region, x is the decision vector, ξ is the k-dimensional random vector, $H(\cdot,\cdot) : \Re^d \times \Re^k \to \Re$ is a deterministic function, which we call a loss function, and $\Pr_{\sim P_0}$ denotes that the probability is taken with respect to (w.r.t.) the distribution P_0. We refer to Problem (1) as a probability minimization model (PM) throughout this paper. Problem (1) is an important model and has been studied in many applications. For instance, in risk management, managers often want to minimize the probability of failure, ruin, or occurrence of certain undesirable events, whereas in goal driven optimization, which has deep root in bounded rationality of Simon (1955), decision makers often target to maximize the probability of attaining aspiration levels; see, e.g., Bordley and Pollock (2009), Chen and Sim (2009) and Brown and Sim (2009) for a thorough discussion.

In another class of situations, the decision maker aims to optimize an objective while require that the constraint with randomness be satisfied with a given probability. She may then formulate her problem as the following probabilistic program:

$$\begin{array}{ll} \underset{x \in X}{\text{minimize}} & h(x) \\ \text{subject to} & \Pr_{\sim P_0}\{H(x,\xi) \leq 0\} \geq 1-\beta, \end{array} \tag{2}$$

where h is the objective function to be optimized and β is the pre-specified confidence level. Problem (2) is a well known model called chance constrained program (CCP). CCP was first considered in Charnes et al. (1958), and has been studied extensively since then, see, e.g., Miller and Wagner (1965), Prékopa (1970), Hong et al. (2011), and Hu et al. (2013). CCPs are also often considered as an alternative formulation to the conventional robust optimization (RO) formulation. One can imagine that when $\beta = 0$, the CCP becomes a RO problem; see the discussion in Ben-Tal et al. (2009).

The probabilistic programs above assume that the random vector ξ follows a distribution P_0. To implement these probabilistic programs in real applications, a very first step is to specify the distribution P_0 for the random vector ξ, using whatever available information. However, it is a rare case that P_0 can be determined precisely. There often exist profound uncertainties for P_0. In this paper, we follow the convention of the economics literature (see, e.g., Ellsberg (1961) and Epstein (1999)) and use the notion "ambiguity" to describe the phenomenon that a distribution cannot be fully determined. One of the central issues for the decision makers is that the ambiguity of the random distribution may (severely) affect the decision. To handle such an issue, many proposals have been suggested, among which the distributionally robust optimization (DRO) has been a reasonable pursuit and has attracted increasing attention in recent years. DRO assumes that the distribution of the random vector is not precisely determined but is contained in a set, which is referred to as an ambiguity set. DRO then considers the worst case of the probability function when the distribution varies in the ambiguity set and proposes to optimize this worst case. Obviously, DRO is an ambiguity-averse approach.

To be a little more precise, we consider a distribution P of ξ and let $\mathbb{P}$ denote the ambiguity set of P. Applying the DRO approach to the PM we obtain the following problem:

$$\underset{x\in X}{\text{minimize}} \ \underset{P\in\mathbb{P}}{\text{maximize}} \ \Pr_{\sim P}\{H(x,\xi)>0\}. \tag{3}$$

Similarly, applying the DRO approach to the CCP, we obtain the following problem:

$$\begin{array}{ll}\underset{x\in X}{\text{minimize}} & h(x) \\ \text{subject to} & \Pr_{\sim P}\{H(x,\xi)\le 0\}\ge 1-\beta, \quad \forall\ P\in\mathbb{P}.\end{array} \tag{4}$$

Problem (3) is a very natural formulation. However, to the best of our knowledge, the study on this model is scarce in the literature. We will try to build some encouraging results for this model. The distributionally robust CCP (4) is also referred to as an ambiguous CCP in the literature. It was considered in Erdogan and Iyengar (2006) and Nemirovski and Shapiro (2006). We adopt the notion of Erdogan and Iyengar (2006) and call Problem (3) and Problem (4) ambiguous PM and ambiguous CCP, respectively. Clearly, when a risk-averse decision maker is also ambiguity-averse, she may want to formulate a probabilistic program into an ambiguous probabilistic program.

To implement the ambiguous probabilistic programs, a key is to specify the ambiguity set $\mathbb{P}$ for the underlying distribution. Significant amount of work has been devoted to the construction of the ambiguity set in DRO literature. Especially, there exist a bunch of studies that consider building ambiguity set using the moments of the distribution; see, e.g., Delage and Ye (2010) and Goh and Sim (2010) for detailed discussion. A number of situations have been identified where the ambiguity in probabilistic programs is tractable. For instance, El Ghaoui et al. (2003) studied worst-case value-at-risk in portfolio selection where the mean vector and covariance matrix of the underlying distribution are within bounded intervals and showed that the problems can be cast as semidefinite programs. Besides, many studies proposed tractable approximations for the ambiguous CCP. Chen et al. (2010) investigated conditional value-at-risk (CVaR) approximations to ambiguous CCPs and suggested a number of conservative approximations by incorporating distributional information such as the moments, the support, as well as the forward and backward deviations. For the ambiguity modeled using the first and second order moments, Zymler et al. (2013a) further showed that the approximations resulting from using CVaR to approximate chance constraint, i.e., the worst-case CVaR approximations, are indeed exact when the loss function is concave or quadratic in the random vector. These findings provide ways for Zymler et al. (2013a) to reformulate their ambiguous CCPs into semidefinite programs.

In many real situations, we have some data and/or information, using which we can often obtain, e.g., via statistical fitting or empirical justification, a nominal distribution P_0 of the random vector. Such a nominal distribution is often our best guess and contains valuable information about the stochastic nature of the parameters. Then, a natural approach to studying the effect

of ambiguity is to consider some level of perturbation or deviation of the nominal distribution. In this paper, we model the distributional ambiguity using the so called likelihood ratio (LR). Based on the data and information available, different ambiguity sets may be constructed via LR. We mainly consider two classes of constraints imposed on LR using convex functions: uniform constraints and expected constraints. The uniform constraints are somewhat straightforward. They actually define a uniform band of the LR which we call a band ambiguity set. Such ambiguity was studied in Shapiro and Ahmed (2004) under the context of minimax stochastic programs. The expected constraints naturally lead to the concept of distance of distributions. The approach of seeking some distance of distributions and then considering a neighborhood of the nominal distribution defined by the distance has been very popular for modeling distributional ambiguity. Many distances between probability distributions have been suggested. Particularly, imposing some minor regularity conditions on the convex function in the expected constraints, we obtain the neighborhood defined by a well-known class of distances called ϕ-divergence. The ϕ-divergence was introduced systematically in Pardo (2006) and Ben-Tal et al. (2013), and was used to construct ambiguity set in DRO framework. Yanıkoğlu and den Hertog (2012) further considered the ϕ-divergence in ambiguous CCPs and provided nice approximations for the ambiguous CCPs. ϕ-divergence contains many distances including the widely used Kullback-Leibler (KL) divergence, χ^2 distance, Hellinger distance, Variation distance, Burg entropy, and many others. In an earlier technical report Hu and Hong (2012), we studied the general DRO framework where the ambiguity set is defined by the KL divergence. We demonstrated the tractability of the ambiguous probabilistic programs (including both PM and CCP) in a unified way as for the general expectation based stochastic programs. Interestingly, we also realized the work of Jiang and Guan (2012) which studied a CCP model with affine loss functions and KL divergence defined ambiguity set and derived a similar tractability result for CCPs, as well as the work of El Ghaoui et al. (2003) which considered the KL divergence deviation from a Gaussian distribution and obtained the reformulation of the entropy-constrained value-at-risk. Our current paper has been greatly inspired by the work of Ben-Tal et al. (2013) and it updates and extends the framework of Hu and Hong (2012). One of the main contributions of our paper, as will be seen, is that we succeed to discover nice structures for the ambiguous probabilistic programs and disclose the fact that the complexity reduction of ambiguous probabilistic programs to pure probabilistic programs is indeed enjoyed by a large class of ambiguity sets constructed using LR. We expect our findings can complement the literature of DRO and our framework may be expanded for more plentiful ambiguity structures.

To provide a unified framework, the analysis of this paper is first conducted on an ambiguity set which combines the uniform constraints and expected constraints. We show the ambiguity set is general enough to model many real situations. We first study a worst-case probability function and show that it is the optimal value of a constrained functional optimization problem. Implementing

the Lagrangian duality, we show that the worst-case probability function is equal to the optimal value of an optimization problem with real decision variables. Based on this result, we then analyze ambiguous PMs and ambiguous CCPs separately. We show that solving an ambiguous PM can be reduced to solving a pure PM, where the underlying distribution is the nominal one. To the best of our knowledge, this is the first time that this kind of results is derived. We next consider an ambiguous CCP. We show that the ambiguous CCP can be converted to a pure CCP with only the confidence level being rescaled from the original CCP. The new confidence level is the optimal value of a nonlinear optimization problem and it can be derived within the convex framework. We further consider a number of specific instances of the ambiguity set. We discuss how to determine the sizes of these sets which admit some statistical meaning based on data available and show how to derive the new confidence levels for these sets.

Note that our analysis does not impose any structural assumptions on the loss function $H(x, \xi)$. Indeed, our approach can be used to handle the general probability function $\Pr_{\sim P}\{A(x, \xi)\}$ where $A(x, \xi)$ is an event depending on x and ξ. This naturally absorbs more complicated models such as joint CCPs (e.g., Miller and Wagner 1965) and CCPs with conic statements (e.g., Cheung et al. 2012). Of course, on the other hand, it should be admitted that for a general loss function, even the pure probabilistic programs may be difficult to solve. Our results show that the LR based ambiguity does not add much difficulty to the probabilistic programs. When we optimize a probability function, we often have simultaneously optimized a worst-case probability performance measure. In the CCP formulation, the ambiguity (uncertainty) and the randomness (risk) can often transform to each other. Therefore, we can often reduce some level of risk to take care of uncertainty. In this regard, the use of probability functional as performance measures in decision under uncertainty is quite appealing.

The rest of this paper is organized as follows. In Section 2, we consider the worst-case probability function, and discuss the ambiguous PM and ambiguous CCP. In Section 3 we consider the ambiguity set and discuss how to determine the size of the set and how to compute the new confidence level for ambiguous CCP. We conclude the paper in Section 4. Some lengthy proofs are provided in the Appendix.

2 Ambiguous Probabilistic Programs

Let $\mathbb{1}_{\{A\}}$ denote the indicator function, which is equal to 1 if A happens and 0 otherwise. Then the probability function $\Pr_{\sim P}\{H(x, \xi) > 0\}$ can be rewritten as $\mathrm{E}_P\left[\mathbb{1}_{\{H(x,\xi)>0\}}\right]$. Therefore, the ambiguous PM, i.e., Problem (3), can be reformulated as the following problem

$$\underset{x \in X}{\text{minimize}} \ \underset{P \in \mathbb{P}}{\text{maximize}} \ \mathrm{E}_P\left[\mathbb{1}_{\{H(x,\xi)>0\}}\right],$$

and the ambiguous CCP, i.e., Problem (4), can be rewritten as

$$\begin{array}{ll} \underset{x\in X}{\text{minimize}} & h(x) \\ \text{subject to} & \underset{P\in\mathbb{P}}{\text{maximize}}\ \mathrm{E}_P\left[\mathbb{1}_{\{H(x,\xi)>0\}}\right] \le \beta. \end{array}$$

Suppose we have obtained a nominal distribution P_0. Suppose the true but unknown distribution is P. We construct the ambiguity set by considering the difference between P_0 and P. Suppose the k-dimensional distributions P and P_0 have densities $p(z)$ and $p_0(z)$ on $\Xi \subset \Re^k$. Note that we do not differentiate P and $p(z)$ throughout this paper: The two notations denote the same distribution if no confusion is caused. Let $L = p/p_0$. Note that L is called a likelihood ratio (LR) in the literature. The definition of LR implicitly assumes that P is absolutely continuous w.r.t. P_0 (denoted as $P \ll P_0$), i.e., for every measurable set A, $P_0(A) = 0$ implies $P(A) = 0$. When P_0 is a discrete distribution, we understand $p_0(z)$ as the probability mass function. When P_0 follows a mixed distribution, $p_0(z)$ is the density at z if P_0 has zero mass at z, and is the probability mass function at z if P_0 has a positive mass at z. Clearly, LR is a good candidate for measuring the perturbation/deviation of the true distribution to the nominal one. As mentioned in Section 1, we use two different classes of constraints on the LR to model the ambiguity. The first is called uniform constraints. Specifically, we consider a convex function $\varphi : \Re \to \Re$, and construct the constraint

$$\varphi(L) \le \rho, \tag{5}$$

where ρ is a positive constant. To guarantee that the nominal distribution satisfies (5), we impose the regularity condition for φ that $\varphi(1) \le \rho$. Because φ is convex and finite valued, the constraint (5) defines a closed interval of L. Furthermore, a finite number of constraints taking the form of (5) still define an closed interval. Therefore, using the uniform constraints we are arriving at a set of p such that the LR falls in an interval, i.e., $a \le L \le b$ for some $0 \le a \le 1 \le b \le \infty$. Although L is itself a function of ξ, (5) requires the constraint be satisfied for all ξ. This is why we call (5) a uniform constraint.

The second class is called expected constraints. Specifically, consider a convex function ϕ on $\Re$ and construct the constraint $\mathrm{E}_{P_0}[\phi(L)] \le \eta$. Imposing some minor regularity conditions on ϕ, we are arriving at the famous ϕ-divergence in statistics. Seeking some distance of distributions and constructing the ambiguity set by requiring the distribution within a certain distance from the nominal distribution has been a natural approach to modeling ambiguity. The ϕ-divergence is often used to measure the distance of a distribution to another one. Therefore, imposing constraints on LR using ϕ-divergence admits a clear statistical and practical meaning. Following the notions 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

$$D_\phi(P\|P_0) = \int_\Xi p_0(z)\phi\left(\frac{p(z)}{p_0(z)}\right)\mathrm{d}z = \mathrm{E}_{P_0}\left[\phi\left(\frac{p(\xi)}{p_0(\xi)}\right)\right] = \mathrm{E}_{P_0}\left[\phi\left(L\right)\right]. \tag{6}$$

Similarly, we understand the integral in (6) 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. Using the ϕ-divergence, we can construct a neighborhood of P_0 defined by $D_\phi(P||P_0) \leq \eta$. As can be seen, instead of requiring L satisfy a constraint for all ξ in uniform constraints, in expected constraints one only requires L satisfy a constraint averagely.

We consider a unified form by combining the two classes of constraints, and construct the following ambiguity set of P:

$$\mathbb{P} = \left\{P \in \mathbb{D} : a \leq p/p_0 \leq b,\ D_{\phi_i}(P||P_0) \leq \eta_i, i = 1, \cdots, m\right\},$$

where $\mathbb{D}$ denotes the set of all probability distributions and $D_{\phi_i}(P||P_0)$ denotes the ϕ_i-divergence from P to P_0. In the ambiguity set $\mathbb{P}$, the constants a, b and $\eta_i, i = 1, 2, \cdots, m$ are indexes of ambiguity, which control the size of $\mathbb{P}$. In terms of L, $\mathbb{P}$ can also be represented as follows:

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

where we define $\mathbb{L}(a,b) = \{L : a \leq L \leq b \text{ a.s.}\}$. We defer the more detailed discussion on the ambiguity set in Section 3. In what follows, we discuss how to solve the ambiguous probabilistic programs with ambiguity set $\mathbb{L}$.

2.1 Worst-Case Probability Function

For simplicity of notation, we use $\mathbb{1}$ to denote $\mathbb{1}_{\{H(x,\xi)>0\}}$. In the ambiguous probabilistic programs, the critical is the worst-case probability function, that is, the optimal value of the following problem:

$$\underset{P\in\mathbb{P}}{\text{maximize}}\ \ \mathrm{E}_P\left[\mathbb{1}\right]. \tag{7}$$

Problem (7) is a rather abstract optimization problem. One of the major difficulties for solving the problem comes from that the randomness is embedded in the decision variable. A widely used technique that can separate them is the change-of-measure technique (e.g., Hu et al. (2012)). Applying the technique, we obtain that

$$\mathrm{E}_P\left[\mathbb{1}\right] = \int_\Xi \mathbb{1}p(z)\mathrm{d}z = \int_\Xi \mathbb{1}\frac{p(z)}{p_0(z)}p_0(z)\mathrm{d}z = \mathrm{E}_{P_0}\left[\mathbb{1}L(\xi)\right].$$

Recall the definition of the ambiguity set $\mathbb{P}$ and $\mathbb{L}$. Problem (7) can be stretched as

$$\begin{array}{ll} \text{maximize} & \mathrm{E}_{P_0}\left[\mathbb{1}L\right] \\ \text{subject to} & \mathrm{E}_{P_0}\left[\phi_i\left(L\right)\right] \leq \eta_i, i = 1, \cdots, m, \\ & \mathrm{E}_{P_0}\left[L\right] = 1, \\ & L \in \mathbb{L}(a, b). \end{array} \tag{8}$$

Problem (8) is a functional optimization problem with the decision variable L. It is not difficult to see that the problem is a convex optimization problem. One standard approach of handling such constrained functional optimization problem is to use the Lagrangian duality. We construct the Lagrangian functional associated with Problem (8):

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

Then Problem (8) is equivalent to

$$\underset{L \in \mathbb{L}(a,b)}{\text{maximize}} \ \underset{\lambda \in \Re, \alpha \geq 0}{\text{minimize}} \ \ell_0(\lambda, \alpha, L). \tag{9}$$

Interchanging the maximum and minimum in Problem (9), we obtain the Lagrangian dual of Problem (9):

$$\underset{\lambda \in \Re, \alpha \geq 0}{\text{minimize}} \ \underset{L \in \mathbb{L}(a,b)}{\text{maximize}} \ \ell_0(\lambda, \alpha, L). \tag{10}$$

The major concern about the above primal and dual problems are whether they have the same optimal value. Fortunately, the duality gap turns out to be zero without any extra qualification conditions. We summarize the result in the following theorem. The proof of the theorem can be found in the Appendix.

Theorem 1. *The optimal values of Problems (9) and (10) are the same. The optimal value of Problems (10) is attained at some $\lambda^* \in \Re$ and $\alpha^* \geq 0$.*

Theorem 1 guarantees that, to solve Problems (9) it suffices to solve Problem (10). Let $v(\lambda, \alpha)$ denote the optimal value of the inner maximization problem of Problem (10). We discuss first how to derive some simplified form for $v(\lambda, \alpha)$. We take an approach that was adopted in Ben-Tal and Teboulle (2007). This approach critically utilizes the following lemma, which can be found in Ben-Tal and Teboulle (2007), as well as in Rockafellar and Wets (1998).

Lemma 1. *Let Ω be a σ-finite measure space, and let $\mathcal{X} := L^p(\Omega, \mathcal{F}, P), p \in [1, +\infty]$. Let $g : \Re \times \Omega \to (-\infty, +\infty]$ be a normal integrand, and define on $\mathcal{X}$ the integral functional $I_g(x) := \int_\Omega g(x(\omega), \omega) dP(\omega)$. Then*

$$\inf_{x \in \mathcal{X}} \int_\Omega g(x(\omega), \omega) dP(\omega) = \int_\Omega \inf_{s \in \Re} g(s, \omega) dP(\omega)$$

provided the left-hand side is finite. Moreover,

$$\bar{x} \in arg\min_{x \in \mathcal{X}} I_g(x) \Longleftrightarrow \bar{x}(\omega) \in arg\min_{s \in \Re} g(s, \omega), a.e.\ \omega \in \Omega.$$

Lemma 1 guarantees that we can put the supremum in the expectation in the expression of $v(\lambda, \alpha)$. Therefore,

$$v(\lambda, \alpha) = \mathrm{E}_{P_0}\left[\sup_{L \in \mathbb{L}(a,b)} \left\{(\mathbb{1} + \lambda) L - \sum_{i=1}^m \alpha_i \phi_i(L)\right\}\right] + \sum_{i=1}^m \alpha_i \eta_i - \lambda. \tag{11}$$

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

$$\Psi(s, \alpha) = \sup_{t \in \mathbb{L}(a,b)} \left\{st - \sum_{i=1}^m \alpha_i \phi_i(t)\right\}. \tag{12}$$

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

Proposition 1. *$\Psi(s, \alpha)$ is convex in (s, α), is non-decreasing in s, and satisfies $\Psi(s, \alpha) \geq s$.*

Proposition 1 summarizes important properties of $\Psi(s, \alpha)$. We will frequently refer to this proposition in the analysis followed. To further simplify the notation, we let $\kappa(x) = \Pr_{\sim P_0}\{H(x, \xi) > 0\}$. Then it follows from (11) and (12) that

$$\begin{aligned} v(\lambda, \alpha) &= \mathrm{E}_{P_0}\left[\Psi(\mathbb{1} + \lambda, \alpha)\right] + \sum_{i=1}^m \alpha_i \eta_i - \lambda \\ &= \Psi(1 + \lambda, \alpha)\kappa(x) + \Psi(\lambda, \alpha)(1 - \kappa(x)) + \sum_{i=1}^m \alpha_i \eta_i - \lambda \\ &= \left[\Psi(1 + \lambda, \alpha) - \Psi(\lambda, \alpha)\right]\kappa(x) + \Psi(\lambda, \alpha) + \sum_{i=1}^m \alpha_i \eta_i - \lambda, \end{aligned} \tag{13}$$

where the second equality follows from the definition of the random variable $\mathbb{1}$. Then, we have the following result on the expression of the worst-case probability function.

Theorem 2. *Suppose that the ambiguity set is $\mathbb{L}$. Then the optimal value of Problem (7) is equal to $\inf_{\lambda \in \Re, \alpha \geq 0}\ v(\lambda, \alpha)$ where $v(\lambda, \alpha)$ is given by (13).*

Theorem 2 builds that the worst-case probability function is equal to the optimal value of an optimization problem with real decision variables. In the following subsections, we derive results for both ambiguous PM and ambiguous CCP based on Theorem 2.

2.2 Ambiguous Probability Minimization

Consider first the ambiguous PM. We have the following result.

Theorem 3. *Suppose that the ambiguity set is* $\mathbb{L}$*. Then any optimal solution of Problem (1) solves Problem (3).*

Proof. Suppose that the ambiguity set is $\mathbb{L}$. Then it follows from Theorem 2 that

$$\inf_{x\in X}\sup_{P\in\mathbb{P}} \mathrm{E}_P\left[\mathbb{1}_{\{H(x,\xi)>0\}}\right] = \inf_{x\in X}\inf_{\lambda\in\Re,\alpha\geq 0}\ v(\lambda,\alpha,x)$$

where $v(\lambda,\alpha,x) = v(\lambda,\alpha)$ which is defined by (13). Suppose x^* is an optimal solution of Problem (1). Then for any $x \in X$, $\kappa(x^*) \leq \kappa(x)$. From Proposition 1 we have $\Psi(s,\alpha)$ is non-decreasing in s. Therefore $\Psi(1+\lambda,\alpha) - \Psi(\lambda,\alpha) \geq 0$. It follows that $v(\lambda,\alpha,x^*) \leq v(\lambda,\alpha,x)$ for all (λ,α). Thus $\inf_{\lambda\in\Re,\alpha\geq 0}\ v(\lambda,\alpha,x^*) \leq \inf_{\lambda\in\Re,\alpha\geq 0}\ v(\lambda,\alpha,x)$. This indicates that x^* solves Problem (3). □

Theorem 3 shows that, for the ambiguity set $\mathbb{L}$, a solution that minimizes the original probability function simultaneously minimizes the worst-case probability function, no matter what divergences are used and what values the indexes of ambiguity a, b, η_i take. Therefore, to solve Problem (3), it suffices to solve Problem (1). It reflects that the probability minimization model has already taken care of the ambiguity (at least the ambiguity defined in this paper) of the distribution of the random parameters. This result suggests in probabilistic programs, risk and ambiguity are interrelated. It seems that in the ambiguous PM, risk and ambiguity are the two sides of the same coin. If we take care of one, we may have already taken care of the other. In the literature, it is often criticized that one serious issue of the model of optimizing probability risk measure is that it is difficult to precisely determine a distribution for the random parameters. Theorem 3 suggests that the model of optimizing a probability functional (measure) can be quite robust to the distribution ambiguity.

2.3 Ambiguous Chance Constrained Program

We now turn to the ambiguous CCP. In contrast to ambiguous PM, there have been encouraging results for the ambiguous CCP in the literature. The use of distribution distance in ambiguous CCP was pioneered by Erdogan and Iyengar (2006), who considered ambiguous CCPs in which the ambiguity set is

$$\{P \in \mathbb{D} : D_{PV}(P||P_0) \leq \eta\}$$

where D_{PV} denotes the Prohorov metric. They studied the scenario approach, and proposed a robust sampled problem where the sample is simulated from the nominal distribution P_0, to approximate the ambiguous CCP, and built a lower bound for the sample size which ensures that the optimal solution of the robust sampled problem is included in the feasible region of the ambiguous CCP with a given probability. Other studies on ambiguous CCPs include El Ghaoui et al. (2003),

Nemirovski and Shapiro (2006), Chen et al. (2010), Jiang and Guan (2012), Hu and Hong (2012), and Zymler et al. (2013a,b). Most closely to our setting, Yanıkoğlu and den Hertog (2012) considered the ϕ-divergence in ambiguous CCP. They proposed safe approximations to the ambiguous CCPs.

In this subsection, we study ambiguous CCPs with ambiguity set $\mathbb{L}$. Different from Yanıkoğlu and den Hertog (2012), we derive exact reformulation for the ambiguous CCPs. Our key results for ambiguous CCPs are summarized in the following theorem.

Theorem 4. *Suppose that the ambiguity set is $\mathbb{L}$. Then Problem (4) is equivalent to the following CCP*

$$\begin{array}{ll} \underset{x\in X}{\textit{minimize}} & h(x) \\ \textit{subject to} & \Pr_{\sim P_0}\left\{H(x,\xi)\leq 0\right\}\geq 1-\bar{\beta}, \end{array}$$

where

$$\bar{\beta}=\sup_{\lambda\in\Re,\alpha\geq 0}\frac{\beta-(\Psi(\lambda,\alpha)+\sum_{i=1}^{m}\alpha_i\eta_i-\lambda)}{\Psi(1+\lambda,\alpha)-\Psi(\lambda,\alpha)}. \tag{14}$$

Proof. From Theorem 2 we have that $\underset{P\in\mathbb{P}}{\text{maximize}}\ \mathrm{E}_P\left[\mathbb{1}_{\{H(x,\xi)>0\}}\right]\leq\beta$ is equivalent to

$$\inf_{\lambda\in\Re,\alpha\geq 0} v(\lambda,\alpha)\leq\beta \tag{15}$$

where $v(\lambda,\alpha)$ is defined by (13). From Theorem 1, the infimum in (15) is attained at some $\lambda^*\in\Re$ and $\alpha^*\geq 0$, Thus (15) is equivalent to

$$\exists\ \lambda\in\Re,\alpha\geq 0,\ \ \text{such that}\ \ \left[\Psi(1+\lambda,\alpha)-\Psi(\lambda,\alpha)\right]\kappa(x)+\Psi(\lambda,\alpha)+\sum_{i=1}^{m}\alpha_i\eta_i-\lambda\leq\beta. \tag{16}$$

Suppose that A is the set of (x,λ,α) such that (x,λ,α) satisfies (16). We show that for any $(x,\lambda,\alpha)\in A$,

$$\Psi(1+\lambda,\alpha)-\Psi(\lambda,\alpha)\neq 0.$$

To see this, we note that if $\Psi(1+\lambda,\alpha)-\Psi(\lambda,\alpha)=0$, then from (16) we have $\Psi(\lambda,\alpha)+\sum_{i=1}^{m}\alpha_i\eta_i-\lambda\leq\beta$. However, from Proposition 1, we have $\Psi(\lambda,\alpha)=\Psi(1+\lambda,\alpha)\geq 1+\lambda$. This further implies that $\Psi(\lambda,\alpha)+\sum_{i=1}^{m}\alpha_i\eta_i-\lambda\geq 1+\sum_{i=1}^{m}\alpha_i\eta_i\geq 1$. We obtained a contradiction since $\beta<1$.

The analysis shows that (16) is equivalent to the following

$$\exists\ \lambda\in\Re,\alpha\geq 0,\ \ \text{such that}\ \ \kappa(x)\leq\frac{\beta-(\Psi(\lambda,\alpha)+\sum_{i=1}^{m}\alpha_i\eta_i-\lambda)}{\Psi(1+\lambda,\alpha)-\Psi(\lambda,\alpha)}. \tag{17}$$

It now suffices to show that (17) can be equivalently strengthened as

$$\kappa(x)\leq\bar{\beta}, \tag{18}$$

where $\bar{\beta}$ is defined by (14). We briefly justify the equivalence as follows. Let A and B denote the set of x satisfying (17) and (18) respectively. Then it is obvious that $A \subset B$. We now prove the opposite inclusion. Consider any $x^* \in B$. If the supremum in (14) for x^* is attained at some finite $\lambda^* \in \Re$ and $\alpha^* \geq 0$, then $x^* \in A$. Suppose the supremum is attained only when some $\alpha_i \to +\infty$. Then by Proposition 1, $\Psi(\lambda, \alpha) + \sum_{i=1}^m \alpha_i \eta_i - \lambda \geq \alpha_i \eta_i \to +\infty$. Note that $\Psi(1+\lambda, \alpha) - \Psi(\lambda, \alpha) \geq 0$. We have $\bar{\beta} \leq 0$. This implies $\kappa(x^*) \leq 0$. Thus x^* must satisfy the original worst-case chance constraint and consequently $x^* \in A$. Suppose now the supremum is attained at finite α_i but only when $\lambda \to +\infty$. Consider the situation where $b > 1$ (In the degenerate case where $b = 1$, it is easy to see the ambiguity set has only one element, i.e., the nominal distribution. It is easy to show that $\bar{\beta} = \beta$). We select a finite $b_0 \in (1, b]$. Then $\Psi(\lambda, \alpha) + \sum_{i=1}^m \alpha_i \eta_i - \lambda \geq \Psi(\lambda, \alpha) - \lambda \geq b_0 \lambda - \sum_{i=1}^m \alpha_i \phi_i(b_0) - \lambda \to +\infty$. We also have $\bar{\beta} \leq 0$. This also implies $x^* \in A$. Similarly, the result $\bar{\beta} \leq 0$ holds when the supremum is attained only when $\lambda \to -\infty$. This concludes the proof of the theorem. □

Theorem 4 shows that the ambiguous CCP can be equivalently formulated as the original CCP with only the confidence level being adjusted from the original one. This suggests that the ambiguous CCP can be solved by using standard techniques developed for pure CCPs. Furthermore, it is not difficult to verify that $\bar{\beta} \leq \beta$. This shows that, to compensate the distributional robustness of the CCP, a certain amount of allowed error probability needs to be given up. Similar to the discussions followed Theorem 3, again, we see that risk and ambiguity are interrelated. Theorem 4 shows that, in the LR constrained ambiguous CCP, the ambiguity averseness is equivalent to an increase of risk averseness in the original CCP. From a modeling perspective, we can take care of the ambiguity by reducing the confidence level, i.e., by considering more conservative confidence level.

2.3.1 Computing New Confidence Level

To obtain the new CCP, we still need to derive the new confidence level $\bar{\beta}$ which is defined by (14). Because the corresponding optimization problem is typically non-convex, it may be difficult to obtain $\bar{\beta}$ by directly solving (14). To this end, we go back to the definition of $v(\lambda, \alpha)$. We show that $\bar{\beta}$ can be obtained via solving a sequence of convex optimization problems. Note that $\bar{\beta} \in [0, 1]$. Therefore, we only need to seek $\bar{\beta}$ from $[0, 1]$ and $\bar{\beta}$ is equal to the optimal value of the following optimization problem:

$$\begin{aligned} \underset{0 \leq y \leq 1, \lambda \in \Re, \alpha \geq 0}{\text{maximize}} \quad & y \\ \text{subject to} \quad & y \leq \frac{\beta - (\Psi(\lambda, \alpha) + \sum_{i=1}^m \alpha_i \eta_i - \lambda)}{\Psi(1+\lambda, \alpha) - \Psi(\lambda, \alpha)}, \end{aligned}$$

which can be reformulated as

$$\begin{aligned} \underset{0\le y\le 1,\lambda\in\Re,\alpha\ge 0}{\text{maximize}} \quad & y \\ \text{subject to} \quad & y\Psi(1+\lambda,\alpha)+(1-y)\Psi(\lambda,\alpha)+\sum_{i=1}^{m}\alpha_i\eta_i-\lambda\le\beta. \end{aligned} \tag{19}$$

From Proposition 1, $\Psi(s,\alpha)$ is convex in (s,α). This suggests for any given $y\in[0,1]$, the constraint function in Problem (19) is convex in (λ,α). Furthermore, the constraint function is nondecreasing in y. The nice structures allow for the following bisection procedure to solve Problem (19).

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

$$\underset{\lambda\in\Re,\alpha\ge 0}{\text{minimize}} \quad y_i\Psi(1+\lambda,\alpha)+(1-y_i)\Psi(\lambda,\alpha)+\sum_{i=1}^{m}\alpha_i\eta_i-\lambda$$

to obtain its optimal value v.

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

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

It is not difficult to see that the sequence $\{y_i\}$ generated by the Bisection Search procedure converges to the optimal value of Problem (19), i.e., $\bar{\beta}$, and the convergence rate 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 function $\Psi(\lambda,\alpha)$ is itself defined as a supremum, we suggest obtaining the dual of the supremum and building the corresponding strong duality. This may help obtain an equivalent reformulation of the problem in Step i and the reformulation may have a closed form. In next section, we discuss how to obtain a reformulation for the ambiguity set $\mathbb{L}$. It is worthwhile noting that at each iteration of the Bisection Search, we do not really want to solve the problem, but to identify whether its optimal value is less than β. Therefore, when actually solving the optimization problem using some nonlinear optimization algorithm, e.g., an interior point method, if we have obtained an objective value that is already less than or equal to β, we do not need to proceed the algorithm any more and we can go directly to the next iteration of Bisection Search. This may save some computational effort.

2.3.2 Extending to Value-at-Risk

The results derived for ambiguous CCP may be extended to the important risk measure, value-at-risk (VaR), which is widely used in financial risk management. Distributionally robust VaR

optimization has recently been discussed in Zymler et al. (2013b). Zymler et al. (2013b) studied the ambiguity of nonlinear portfolio optimization modeled using the first two moments. Different from Zymler et al. (2013b), we focus on in this section the ambiguity defined by the LR in VaR optimization. We briefly show how to derive the DRO reformulations of VaR related stochastic programs. Consider the following ambiguous VaR minimization problem:

$$\underset{x\in X}{\text{minimize}} \ \underset{P\in\mathbb{P}}{\text{maximize}} \ \text{VaR}_{1-\beta,P}(H(x,\xi)) \tag{20}$$

where the subscript P denotes that the VaR is calculated when ξ follows P. Problem (20) suggests to minimize the worst-case VaR. We then have the following corollary.

Corollary 1. *Suppose that the ambiguity set is $\mathbb{L}$. Then Problem (20) is equivalent to*

$$\underset{x\in X}{\textit{minimize}} \ \ \textit{VaR}_{1-\bar{\beta},P_0}(H(x,\xi)), \tag{21}$$

where $\bar{\beta}$ is defined by (14).

Proof. From the definition of VaR, it is not difficult to verify Problem (20) can be rewritten as

$$\begin{array}{ll} \underset{x\in X,t\in\Re}{\text{minimize}} & t \\ \text{subject to} & \text{Pr}_{\sim P}\left\{H(x,\xi)-t\leq 0\right\}\geq 1-\beta, \quad \forall\ P\in\mathbb{P}. \end{array} \tag{22}$$

Using Theorem 4, Problem (22) can be transformed as

$$\begin{array}{ll} \underset{x\in X,t\in\Re}{\text{minimize}} & t \\ \text{subject to} & \text{Pr}_{\sim P_0}\left\{H(x,\xi)-t\leq 0\right\}\geq 1-\bar{\beta}, \end{array}$$

which is equivalent to Problem (21) from the definition of VaR. □

Consider next the following ambiguous VaR constrained program:

$$\begin{array}{ll} \underset{x\in X}{\text{minimize}} & h(x) \\ \text{subject to} & \text{VaR}_{1-\beta,P}(H(x,\xi))\leq 0, \quad \forall\ P\in\mathbb{P}. \end{array} \tag{23}$$

Problem (23) requires the worst-case VaR satisfy the non-positive constraint. We have the following corollary which can be proven following the argument in Corollary 1.

Corollary 2. *Suppose that the ambiguity is $\mathbb{L}$. Then Problem (23) is equivalent to*

$$\begin{array}{ll} \underset{x\in X}{\textit{minimize}} & h(x) \\ \textit{subject to} & \textit{VaR}_{1-\bar{\beta},P_0}(H(x,\xi))\leq 0, \end{array}$$

where $\bar{\beta}$ is defined by (14).

Corollaries 1 and 2 show that ambiguous VaR optimization problems can be converted to pure VaR optimization problems of different confidence levels. Therefore, we can implement standard VaR optimization algorithms to solve these ambiguous VaR optimization problems.

3 Ambiguity Set

In preceding sections we have obtained the main results for a general ambiguity set for the ambiguous probabilistic programs. In this section, we consider a number of cases that may be used in practical applications and discuss how to construct the sets based on data. Furthermore, although the results derived show that in solving the ambiguous PM, we do not need to differentiate the ambiguity set, the new confidence level for ambiguous CCP depends on the ambiguity set. In this section, we demonstrate how to derive the new confidence level for CCPs for the various ambiguity sets.

3.1 Band Ambiguity Set

Consider first an instance of $\mathbb{L}$ in which there are only uniform constraints. It simply takes the following form

$$\mathbb{L}_{a,b} := \{L \in \mathbb{L}(a,b) : \mathrm{E}_{P_0}[L] = 1\}.$$

We call $\mathbb{L}_{a,b}$ a *band ambiguity set*. The band ambiguity was studied in Shapiro and Ahmed (2004). This is a very intuitive ambiguity set, particularly suitable for sensitivity analysis, e.g., we can perturb the probability distribution by a percentage and see the worst case of the probability performance measure. When the distribution is discrete, the band ambiguity set constrains the probability mass at each scenario within an interval.

The following corollary shows the new confidence level can be derived analytically for the ambiguous CCP with $\mathbb{L}_{a,b}$.

Corollary 3. *Suppose that the ambiguity set is $\mathbb{L}_{a,b}$. Then Problem (4) is equivalent to the following CCP:*

$$\begin{array}{ll} \underset{x\in X}{\textit{minimize}} & h(x) \\ \textit{subject to} & \Pr_{\sim P_0}\{H(x,\xi) \le 0\} \ge 1-\bar{\beta}, \end{array}$$

where

$$\bar{\beta} = \begin{cases} \frac{\beta+a-1}{a} & \text{if } \beta > \frac{1-a}{1-a/b} \\ \frac{\beta}{b} & \text{if } \beta \le \frac{1-a}{1-a/b} \end{cases}.$$

Proof. Consider the function $\Psi(s,\alpha)$ defined by (12). Because for the ambiguity set $\mathbb{L}_{a,b}$, we do not have constraint defined by ϕ_i, the function $\Psi(s,\alpha)$ is only a function of s. For simplicity, we denote it by $\Psi(s)$. It can be verified that

$$\Psi(s) = \sup_{t\in\mathbb{L}(a,b)} \{st\} = bs^+ - a[-s]^+ = bs^+ - a\left(s^+ - s\right) = as + (b-a)\,s^+,$$

where $y^+ = \max\{y, 0\}$. It follows from Theorem 4 that

$$\begin{aligned}
\bar{\beta} &= \sup_{\lambda \in \Re} \frac{\beta - (a\lambda + (b-a)\lambda^+ - \lambda)}{[a(1+\lambda) + (b-a)[1+\lambda]^+] - [a\lambda + (b-a)[\lambda]^+]} \\
&= \sup_{\lambda \in \Re} \frac{\beta - (a\lambda + (b-a)\lambda^+ - \lambda)}{a + (b-a)\left[[1+\lambda]^+ - [\lambda]^+\right]} \\
&= \begin{cases} \frac{\beta+a-1}{a} & \text{if } \beta \geq \frac{1-a}{1-a/b} \\ \frac{\beta}{b} & \text{if } \beta \leq \frac{1-a}{1-a/b} \end{cases},
\end{aligned} \tag{24}$$

where the last equality follows from the fact that if $\beta \geq (1-a)/(1-a/b)$, then the supremum in (24) is attained at $\lambda = 0$, and if $\beta \leq (1-a)/(1-a/b)$, then the supremum in (24) is attained at $\lambda = 1$. This concludes the proof of the corollary. □

Table 1: Relation between New Confidence Level and Bounds of Band

	a	b	new confidence level β
$\beta = 0.1$	0.9	1.1	0.0909
	0.5	1.5	0.0667
	0.95	10	0.0526
	0.01	100	0.0010
$\beta = 0.05$	0.9	1.1	0.0455
	0.5	1.5	0.0333
	0.95	10	0.0050
	0.01	100	0.0005

We compute the new confidence levels for some combinations of a and b and report the results in Table 1. From the table, we see that for the band ambiguity set, the new confidence level decreases approximately in a linear fashion with respect to the increase of bounds of the band.

3.2 ϕ-divergence Constrained Ambiguity Set

Next we consider another special case of $\mathbb{L}$, a neighborhood of the nominal distribution which is defined by the ϕ-divergence. As has been mentioned, the ϕ-divergence was suggested in Ben-Tal et al. (2013) and Yanıkoğlu and den Hertog (2012) to model the uncertainty in optimization problems. Ben-Tal et al. (2013) considered DRO problems. They assumed that the underlying distribution in the optimization problem is discrete, and constructed a confidence region using a set of realizations of the random vector. They showed the tractability of the resulting DRO for various ϕ-divergences. Yanıkoğlu and den Hertog (2012) further considered the ϕ-divergence in ambiguous CCP. They also focused on discrete distribution setting. They developed safe approximation to the ambiguous CCP and showed that the resulted safe approximation using ϕ-divergence could be less conservative than some existing safe approximation methods.

Of particular importance of the ϕ-divergence is its conjugate, which is defined as

$$\phi^*(s) = \sup_{t \geq 0} \{st - \phi(t)\}.$$

Table 2 extracted from Ben-Tal et al. (2013) summarized information of various ϕ-divergence measures. In Table 2 the second column shows various ϕ-divergence functions, whereas the third

Table 2: Some ϕ-Divergence Functions and Their Conjugates

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

column shows the corresponding conjugates. Let

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

We call $\mathbb{L}_\phi$ a *ϕ-divergence constrained ambiguity set*. We have the following result for the ambiguous CCP with $\mathbb{L}_\phi$ and the proof is provided in the Appendix.

Corollary 4. *Suppose that the ambiguity set is $\mathbb{L}_\phi$. Suppose that $\phi^*(\cdot)$ is the conjugate of $\phi(\cdot)$. Then Problem (4) is equivalent to the following CCP*

$$\begin{array}{ll} \underset{x \in X}{\textit{minimize}} & h(x) \\ \textit{subject to} & \Pr_{\sim P_0}\left\{H(x,\xi) \leq 0\right\} \geq 1 - \bar{\beta}, \end{array}$$

where

$$\bar{\beta} = \sup_{\lambda \in \Re, \alpha \geq 0} \frac{\beta + \lambda - \eta\alpha - \alpha\phi^*\left(\frac{\lambda}{\alpha}\right)}{\alpha\left[\phi^*\left(\frac{1+\lambda}{\alpha}\right) - \phi^*\left(\frac{\lambda}{\alpha}\right)\right]}. \tag{25}$$

Corollary 4 shows the expression of the new confidence level for $\mathbb{L}_\phi$. Similarly, we can use the Bisection Search procedure proposed in Section 2.3 to compute $\bar{\beta}$. In that procedure, the sequence of optimization problems we want to solve become the following

$$\underset{\lambda \in \Re, \alpha \geq 0}{\text{minimize}} \quad y_i \alpha \phi^*\left(\frac{1+\lambda}{\alpha}\right) + (1 - y_i)\alpha\phi^*\left(\frac{\lambda}{\alpha}\right) + \alpha\eta - \lambda. \tag{26}$$

For the ϕ-divergence (except J-divergence) in Table 2, we have the closed form of the conjugate ϕ^*, which indicates Problem (26) has an explicit expression. In what follows, we consider further the various ϕ-divergences and discuss the tractability of Problem (26).

3.2.1 Tractability of Divergences

Problem (26) has a similar structure as Problem (13) of Ben-Tal et al. (2013). Ben-Tal et al. (2013) showed that Problem (13) therein can be reformulated into simple optimization problems such as conic quadratic program (CQP) and linear program (LP) for various ϕ-divergences. Similarly, Problem (26) can also be reformulated as simple optimization problems. Below we first build reformulations of Problem (26) for χ^2-distance, Modified χ^2-distance, Hellinger distance, χ-distance of order $\theta > 1$, and Cressie-Read.

χ^2-distance.

$$\begin{array}{ll} \text{minimize} & (2+\eta)\alpha - \lambda + y_i\mu_1 + (1-y_i)\mu_2 \\ \text{subject to} & \sqrt{\mu_1^2 + (1+\lambda)^2} \leq 2\alpha - \lambda - 1, \\ & \sqrt{\mu_2^2 + \lambda^2} \leq 2\alpha - \lambda, \\ & \lambda - \alpha + 1 \leq 0, \alpha \geq 0, \lambda, \mu_1, \mu_2 \in \Re. \end{array}$$

Modified χ^2-distance.

$$\begin{array}{ll} \text{minimize} & (\eta-1)\alpha - \lambda + \dfrac{1}{4}y_i\mu_3 + \dfrac{1}{4}(1-y_i)\mu_4 \\ \text{subject to} & \sqrt{\mu_1^2 + \dfrac{1}{4}\left(\alpha - \mu_3\right)^2} \leq \dfrac{1}{2}\left(\alpha + \mu_3\right), \\ & \sqrt{\mu_2^2 + \dfrac{1}{4}\left(\alpha - \mu_4\right)^2} \leq \dfrac{1}{2}\left(\alpha + \mu_4\right), \\ & \mu_1 \geq 1 + \lambda + 2\alpha, \mu_2 \geq \lambda + 2\alpha, \alpha \geq 0, \mu_1 \geq 0, \mu_2 \geq 0, \lambda, \mu_3, \mu_4 \in \Re. \end{array}$$

Hellinger Distance.

$$\begin{array}{ll} \text{minimize} & (\eta-1)\alpha - \lambda + y_i\mu_1 + (1-y_i)\mu_2 \\ \text{subject to} & \sqrt{\alpha^2 + \dfrac{1}{4}\left[\alpha - (1+\lambda) - \mu_1\right]^2} \leq \dfrac{1}{2}\left[\alpha - (1+\lambda) + \mu_1\right], \\ & \sqrt{\alpha^2 + \dfrac{1}{4}\left(\alpha - \lambda - \mu_2\right)^2} \leq \dfrac{1}{2}\left(\alpha - \lambda + \mu_2\right), \\ & \alpha - (1+\lambda) \geq 0, \alpha \geq 0, \lambda, \mu_1, \mu_2 \in \Re. \end{array}$$

χ-distance of Order $\theta > 1$.

$$\begin{array}{ll} \text{minimize} & \eta\alpha + y_i(\theta-1)\alpha\left(\dfrac{\mu_1}{\theta\alpha}\right)^{\frac{\theta}{\theta-1}} + (1-y_i)(\theta-1)\alpha\left(\dfrac{\mu_2}{\theta\alpha}\right)^{\frac{\theta}{\theta-1}} + y_i \\ \text{subject to} & 1 + \lambda \leq \mu_1, 1 + \lambda \geq -\mu_1, \\ & \lambda \leq \mu_2, \lambda \geq -\mu_2, \\ & \alpha \geq 0, \lambda, \mu_1, \mu_2 \in \Re. \end{array}$$

Cressie-Read.

$$\begin{array}{ll} \text{minimize} & \left(\eta-\theta^{-1}\right)\alpha-\lambda+y_i\theta^{-1}\alpha\left(\frac{\mu_1}{\alpha}\right)^{\theta/(\theta-1)}+(1-y_i)\theta^{-1}\alpha\left(\frac{\mu_2}{\alpha}\right)^{\theta/(\theta-1)} \\ \text{subject to} & \alpha-(1-\theta)(1+\lambda)=\mu_1, \\ & \alpha-(1-\theta)\lambda=\mu_2, \\ & \alpha\geq 0, \lambda, \mu_1, \mu_2\in\Re. \end{array}$$

The reformulations above are all CQPs, which can be solved easily. Next we discuss Variation distance.

Variation Distance. It is not difficult to show that for the Variation distance, Problem (26) can be reformulated as a LP. However, the following corollary shows that the new confidence level for Variation distance has a closed expression. The proof of the corollary can be finished by a standard calculation and thus is omitted.

Corollary 5. *Suppose that the ambiguity set is $\mathbb{L}_\phi$ and that the ϕ-divergence is the Variation distance. Then the new confidence level defined by (25) is $\bar{\beta}=\max\left\{\beta-\frac{\eta}{2},0\right\}$.*

Corollary 5 reveals that if the index of ambiguity η is greater than or equal to 2β, the new confidence level degenerates to 0. This means it is not possible to require the worst-case probability satisfy the constraint when the constraint $H(x,\xi)\leq 0$ does not hold almost surely under P_0. Similar phenomenon for the Prohorov metric was observed in Erdogan and Iyengar (2006). In Erdogan and Iyengar (2006), the index of ambiguity η cannot be larger than the confidence level β of the original CCP.

For Burg entropy and KL divergence, Problem (26) admits a self-concordant barrier and thus can be solved readily. Readers can refer to Ben-Tal et al. (2013) for the analysis on self-concordant barrier. Below we discuss more on the KL divergence.

Kullback-Leibler Divergence. Among the ϕ-divergence class, the KL divergence has received the most attention, in different fields including information theory, communication, and operations research. We show that for the KL divergence, the new confidence level can be derived by analytically solving a sequence of optimization problems. Consider the ambiguity set $\mathbb{L}_\phi$ where ϕ is the KL divergence. It follows from Corollary 4 that

$$\begin{aligned} \bar{\beta} &= \sup_{\lambda\in\Re,\alpha\geq 0}\frac{\beta+\lambda-\eta\alpha-\alpha e^{\frac{\lambda}{\alpha}-1}}{\alpha\left[e^{\frac{1+\lambda}{\alpha}-1}-e^{\frac{\lambda}{\alpha}-1}\right]}=\sup_{\lambda\in\Re,\alpha\geq 0}\frac{\left(\frac{\beta}{\alpha}-\eta+\frac{\lambda}{\alpha}\right)e^{-\frac{\lambda}{\alpha}+1}-1}{e^{\frac{1}{\alpha}}-1} \\ &= \sup_{\alpha\geq 0}\sup_{t>0}\frac{\left(\frac{\beta}{\alpha}-\eta+1\right)t-t\log t-1}{e^{\frac{1}{\alpha}}-1}=\sup_{\alpha\geq 0}\frac{e^{\frac{\beta}{\alpha}-\eta}-1}{e^{\frac{1}{\alpha}}-1}=\sup_{t>0}\frac{e^{-\eta}(t+1)^\beta-1}{t}. \end{aligned}$$

where the first equality follows by plugging the conjugate of the KL divergence into (25), the second equality follows from dividing both numerator and denominator by $\alpha e^{\frac{\lambda}{\alpha}-1}$ and from the

careful analysis indicating that the supremum cannot be attained at $\alpha = 0$, the third equality follows from making transformation $t := e^{-\frac{\lambda}{\alpha}+1}$, the fourth equality follows from directly solving the problem over $t > 0$, and the last equality follows from making transformation $t := e^{\frac{1}{\alpha}} - 1$.

The last supremum in the equation above has a nice structure that allows us to design a bisection search algorithm to solve it. The basic idea is to check whether the set

$$T_{\tilde{\beta}} = \left\{ t : t > 0, \frac{e^{-\eta}(t+1)^{\beta} - 1}{t} > \tilde{\beta} \right\}$$

is empty for a given $\tilde{\beta} > 0$. If $T_{\tilde{\beta}}$ is non-empty, then $\bar{\beta} > \tilde{\beta}$ and we should search $\bar{\beta}$ in $(\tilde{\beta}, \beta]$. Otherwise, we should search $\bar{\beta}$ in $(0, \tilde{\beta}]$. Checking the non-emptiness of $T_{\tilde{\beta}}$ can be transformed to checking whether the maximum of

$$\Phi(t) = e^{-\eta}(t+1)^{\beta} - 1 - \tilde{\beta}t$$

over $t \geq 0$ is larger than 0. Note that $\Phi(t)$ is a concave function of t on $[0, +\infty)$, and its maximum over $t \geq 0$ is attained at

$$t^*(\tilde{\beta}) = \max \left\{ 0, \left(\frac{\tilde{\beta} e^{\eta}}{\beta} \right)^{\frac{1}{\beta - 1}} - 1 \right\}.$$

When $\Phi(t^*(\tilde{\beta})) > 0$, we have $t^*(\tilde{\beta}) > 0$ and $\left(e^{-\eta}(t^*(\tilde{\beta}) + 1)^{\beta} - 1 \right) / t^*(\tilde{\beta}) > \tilde{\beta}$. This shows $T_{\tilde{\beta}}$ is non-empty. Similarly, some careful analysis shows when $\Phi(t^*(\tilde{\beta})) < 0$, we have $T_{\tilde{\beta}}$ is empty and $\bar{\beta} < \tilde{\beta}$, and when $\Phi(t^*(\tilde{\beta})) = 0$, we have $\bar{\beta} = \tilde{\beta}$. Therefore, the following bisection search algorithm can be used to solve the one dimensional problem and obtain a solution with arbitrary accuracy.

Bisection Search for Kullback-Leibler Divergence

Step 0. Set $i = 0$. Set $\beta_l := 0$ and $\beta_u := \beta$

Step i. Set $\beta_i = \frac{\beta_l + \beta_u}{2}$ and compute $\Phi(t^*(\beta_i))$.

If $\Phi(t^*(\beta_i)) > 0$, update $\beta_l =: \beta_i$. Set $i = i + 1$.

If $\Phi(t^*(\beta_i)) < 0$, update $\beta_u =: \beta_i$. Set $i = i + 1$.

If $\Phi(t^*(\beta_i)) = 0$, stop.

We compute the new confidence levels for some η values using the Bisection Search (stop if $\beta_u - \beta_l \leq 10^{-12}$) and report the results in Table 3.

Different from the Variation distance, for the KL divergence we do not have the restriction on the value of η. For any $\eta > 0$, the adjusted confidence level $\bar{\beta}$ is larger than 0. However, from Table 3, it is clear that $\bar{\beta}$ may be very small (leading to extreme conservativeness) if η is significantly larger than β. In contrast to the linear fashion for the band ambiguity set, the new confidence level

Table 3: Relation between New Confidence Level and Index of Ambiguity

	index of ambiguity η	new confidence level $\bar{\beta}$
$\beta = 0.1$	1	1.7589e-006
	0.1	0.0166
	0.05	0.0313
	0.01	0.0629
$\beta = 0.05$	1	3.8563e-011
	0.1	0.0027
	0.05	0.0081
	0.01	0.0250

could decrease very rapidly w.r.t. the increase of the index of ambiguity for the KL divergence constrained ambiguity set.

J-divergence. The conjugate of J-divergence is not analytically available. However, the J-divergence can be expressed as the sum of the KL divergence and the Burg entropy. Furthermore, the conjugate of the sum of two functions can be expressed as the infimum of the conjugates of the two functions; see Proposition 1 of Ben-Tal et al. (2013). Therefore, Problem (26) for J-divergence is equivalent to the following problem:

$$\begin{array}{ll} \text{minimize} & y_i\alpha\left[\phi_{KL}^*\left(\frac{t_1}{\alpha}\right)+\phi_B^*\left(\frac{t_2}{\alpha}\right)\right]+(1-y_i)\alpha\left[\phi_{KL}^*\left(\frac{t_3}{\alpha}\right)+\phi_B^*\left(\frac{t_4}{\alpha}\right)\right]+\alpha\eta-t_3-t_4 \\ \text{subject to} & t_1+t_2=1+t_3+t_4 \\ & \alpha\geq 0, t_i\in\Re, i=1,2,3,4, \end{array}$$

where ϕ_{KL}^* and ϕ_B^* are the conjugates of the KL divergence function and the Burg entropy function, respectively. Clearly, this technique may be used to handle the ϕ-divergence which can be expressed as the sum of multiple ϕ-divergences.

Ambiguity Set $\mathbb{L}$. Using the convolution technique, we can now derive an analytical expression for the general ambiguity set $\mathbb{L}$. More specifically, we have the following corollary. The proof of the corollary can be found in the Appendix.

Corollary 6. *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 6 essentially generalizes Corollary 4 of Ben-Tal et al. (2013). From Corollary 6, we immediately have that the optimization problem in Step i of the Bisection Search is equivalent to

the following problem:

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

Note that the problem above can again be reformulated into simple optimization problems as that for a single ϕ-divergence. Thus, the Bisection Search procedure can be implemented readily.

3.3 Specifying the Ambiguity Set

In the ambiguous probabilistic programs, the size of the ambiguity set is also critical to the model. In this section we discuss how to specify the band ambiguity set and the ϕ-divergence constrained ambiguity set when the distribution is discrete. We leave the continuous case unresolved due to the challenge of the problem. But we propose some idea of handling continuous case and we try to solve the problem in a future study.

Suppose that the distribution of ξ is discrete and is supported on a finite number of scenarios $\{\xi_1, \xi_2, \cdots, \xi_s\}$ with probabilities $\{p_1, p_2, \cdots, p_s\}$ where $\sum_{i=1}^{s} p_i = 1$. We first discuss how to determine bounds a and b for $\mathbb{L}_{a,b}$ in practical applications. Suppose we obtained a sample of the random vector ξ with sample size n. Let $\hat{p} = (\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_s)$ denote the empirical estimate of $p = (p_1, p_2, \cdots, p_s)$. From Wasserman (2004) we have

$$\sqrt{n} \left(\frac{\hat{p}_1}{p_1} - 1, \frac{\hat{p}_2}{p_2} - 1, \cdots, \frac{\hat{p}_s}{p_s} - 1 \right) \Rightarrow Y,$$

where "$\Rightarrow$" denotes convergence in distribution, and $Y = (Y_1, Y_2, \cdots, Y_s)$ is a random vector following a multivariate normal distribution $N(0, \Sigma)$ with mean 0 and covariance matrix

$$\Sigma = \begin{bmatrix} p_1(1-p_1) & -p_1 p_2 & \cdots & -p_1 p_s \\ -p_1 p_2 & p_2(1-p_2) & \cdots & -p_2 p_s \\ \vdots & \vdots & \vdots & \vdots \\ -p_1 p_s & -p_2 p_s & \cdots & p_s(1-p_s) \end{bmatrix}.$$

We further normalize the random vector Y by letting

$$Z = (Z_1, Z_2, \cdots, Z_s) = \left(\frac{Y_1}{\sqrt{p_1(1-p_1)}}, \frac{Y_2}{\sqrt{p_2(1-p_2)}}, \cdots, \frac{Y_s}{\sqrt{p_s(1-p_s)}} \right).$$

Then Z follows a multivariate normal distribution with standard normal marginals and a correlation matrix determined by Σ. Because the summation of all column vectors of Σ equals 0, Σ is a singular matrix. Therefore, Y and Z follow degenerate multivariate normal distributions. This

makes building confidence region potentially difficult. In this paper, we use a famous result in Šidák (1967) to build a conservative confidence region. The result states that for a multivariate normal distribution with any covariance structure, the joint probability function can be bounded from below by the product of the marginal probability functions; see Theorem 1 of Šidák (1967). It follows from Theorem 1 of Šidák (1967) that

$$\Pr\left\{|Z_1| \leq c_1, |Z_2| \leq c_2, \cdots, |Z_s| \leq c_s\right\} \geq \Pr\left\{|Z_1| \leq c_1\right\} \Pr\left\{|Z_2| \leq c_2\right\} \cdots \Pr\left\{|Z_s| \leq c_s\right\}. \quad (27)$$

Therefore, for any $\alpha \in [0,1]$, by selecting $c_i, i = 1, 2, \cdots, s$ such that

$$\Pr\left\{|Z_1| \leq c_1\right\} \Pr\left\{|Z_2| \leq c_2\right\} \cdots \Pr\left\{|Z_s| \leq c_s\right\} \geq 1 - \alpha, \quad (28)$$

we can ensure that the left hand side of (27) is also larger than or equal to $1-\alpha$. This suggests we can construct an approximate confidence region for the LR based on the limiting distribution, i.e., the distribution of Z. The basic idea is as follows. We first find some $c_i, i = 1, 2, \cdots, s$ satisfying (28). We use $\sigma_i = \sqrt{\hat{p}_i\left(1-\hat{p}_i\right)}$ to estimate $\sqrt{p_i\left(1-p_i\right)}$. Then we can construct the following approximate $1-\alpha$ confidence region:

$$-c_i \leq \frac{\sqrt{n}}{\sigma_i}\left(\frac{\hat{p}_2}{p_2} - 1\right) \leq c_i, i = 1, 2, \cdots, s,$$

which can be equivalently transformed as

$$\frac{1}{1 + \frac{c_i \sigma_i}{\sqrt{n}}} \leq \frac{p_i}{\hat{p}_i} \leq \frac{1}{1 - \frac{c_i \sigma_i}{\sqrt{n}}}, i = 1, 2, \cdots, s.$$

Define

$$a = \min_{i=1,\cdots,s}\left\{\frac{1}{1 + \frac{c_i \sigma_i}{\sqrt{n}}}\right\}, \quad b = \max_{i=1,\cdots,s}\left\{\frac{1}{1 - \frac{c_i \sigma_i}{\sqrt{n}}}\right\}.$$

Then $\mathbb{L}_{a,b}$ is an approximate $1-\alpha$ confidence band for L. Now we discuss how to find the best $c_i^*, i = 1, 2, \cdots, s$ to achieve the largest lower bound a and smallest upper bound b. We show that $c_i^*, i = 1, 2, \cdots, s$ are determined by the following conditions:

$$c_1\sigma_1 = c_2\sigma_2 = \cdots = c_s\sigma_s \quad (29)$$

$$\Pr\left\{|Z_1| \leq c_1\right\} \Pr\left\{|Z_2| \leq \frac{\sigma_1}{\sigma_2} c_1\right\} \cdots \Pr\left\{|Z_s| \leq \frac{\sigma_1}{\sigma_s} c_1\right\} = 1 - \alpha, \quad (30)$$

and the optimal bounds are

$$a^* = \frac{1}{1 + \frac{c_1^* \sigma_1}{\sqrt{n}}}, \quad b^* = \frac{1}{1 - \frac{c_1^* \sigma_1}{\sqrt{n}}}.$$

We justify the results via contradiction. Consider any $c_i, i = 1, 2, \cdots, s$ satisfying (28) but violating (29). Without loss of generality we assume $c_1\sigma_1$ and $c_2\sigma_2$ are the largest and the smallest values of $c_i\sigma_i, i = 1, 2, \cdots, s$. Then we must have $c_1\sigma_1 > c_1^*\sigma_1$. The reason is as follows. Suppose

$c_1\sigma_1 \leq c_1^*\sigma_1$. Then $c_i\sigma_i \leq c_1^*\sigma_1$ for all $i = 1, 2, \cdots, s$, and the strict inequality holds for at least some i. Consequently, $c_i \leq \frac{\sigma_1}{\sigma_i}c_1^*$ for all $i = 1, 2, \cdots, s$, and the strict inequality holds for at least some i. In this case, (28) cannot hold, which leads to a contradiction. Similarly, we can show that $c_2\sigma_2 < c_1^*\sigma_1$. It follows that $a < a^*$ and $b > b^*$. This justifies that $c_i^*, i = 1, 2, \cdots, s$ are the best choice.

To obtain a^* and b^*, we need to compute c_1^*, which is the unique root of (30), or equivalently, the unique root of the following equation:

$$[2\Phi(c_1) - 1]\left[2\Phi\left(\frac{\sigma_1}{\sigma_2}c_1\right) - 1\right]\cdots\left[2\Phi\left(\frac{\sigma_1}{\sigma_2}c_1\right) - 1\right] = 1 - \alpha, \tag{31}$$

where Φ is the cumulative distribution function of a standard normal distribution. Because the left hand side of (31) is monotone in c_1, we can compute c_1 via a bisection search procedure. The procedure only involves evaluating Φ for different points which is computationally very easy.

Specifying the size of the ambiguity set $\mathbb{L}_\phi$ for discrete case was studied by Ben-Tal et al. (2013), who built an approximate confidence region for probability mass p based on asymptotic results of the ϕ-divergence. Consider the discrete distribution p and the empirical distribution $\hat{p}$. The idea of Ben-Tal et al. (2013) is to use that

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

where χ^2_{s-1} is a χ^2-distribution with $s - 1$ degrees of freedom. Then based on a finite number of observations, an approximate $1 - \alpha$ confidence region for p can be built:

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

where $\eta = \frac{\phi''(1)}{2n}\chi^2_{s-1,1-\alpha}$. The case that the distribution is supported on s scenarios is called base case in Ben-Tal et al. (2013). Ben-Tal et al. (2013) also discussed more general cases and studied how to improve the approximate confidence region using correction parameters.

Specifying the ambiguity set for continuous distribution turns out to be much more difficult. We briefly discuss it in what follows. Suppose we have n observations of ξ, denoted as $\xi_1, \xi_2, \cdots, \xi_n$. The first step should be to fit a continuous distribution based on the data. There are several ways for constructing the distribution. A simple approach is to construct the density histogram (Scott 1992). Suppose the support Ξ of ξ is a bounded hyper-rectangle and suppose it is partitioned into sub-rectangles of size $h_1 \times h_2 \times \cdots \times h_k$. Consider one sub-rectangular bin labeled B_j, which contains ν_j observations of the sample. Then a density histogram takes the following form:

$$\hat{p}(z) = \frac{\nu_j}{nh_1h_2\cdots h_k} \quad \text{for} \quad z \in B_j.$$

A more popular approach is to use the kernel density estimation (Scott 1992). Suppose $K : \Re^k \to \Re$ is a kernel function and B is a $k \times k$ nonsingular matrix. Let $|B|$ denote the determinant of B. Then the following probability density

$$\hat{p}(z) = \frac{1}{n|B|} \sum_{i=1}^{n} K\left(B^{-1}\left(z - \xi_i\right)\right)$$

is a typical kernel density estimator.

Once we obtain a density estimator, the natural idea is then to replace the probability mass in the discrete case with the estimated probability density and obtain the corresponding ambiguity set for the true density p. The question is whether the constructed ambiguity set (the band ambiguity set or the ϕ-divergence constrained ambiguity set) is still a confidence region with the specified confidence level. The answer is no. For the histogram estimation and kernel density estimation, the estimator is itself biased and there often exists a tradeoff between bias and variance for the estimator. Moreover, the convergence rate will be affected by the bin width or the bandwidth and will be certainly slower than the rate n^{-1} which is the convergence rate of the discrete setting. The remaining question is how to specify the index of ambiguity η to obtain a confidence region. This turns out to be a difficult statistical problem. The methodology that determines the index for discrete setting does not naturally carry over to the continuous setting. Below we briefly introduce some idea of tackling the problem when we use kernel density estimation. Suppose $\hat{p}$ is some kernel estimator of the true density p. We want to build the asymptotics for $D_\phi(p, \hat{p})$. For this we construct the function $\varphi(t) = t\phi(t^{-1})$. Then φ is also a divergence function. Moreover, $\phi''(1) = \varphi''(1)$ and $D_\phi(p, \hat{p}) = D_\varphi(\hat{p}, p)$. Thus it suffices to build asymptotics for $D_\varphi(\hat{p}, p)$.

Define a functional $g(f) = \int_\Xi p\varphi\left(\frac{f}{p}\right) dz$. Then $g(\hat{p}) = D_\varphi(\hat{p}, p)$ and $g(p) = 0$. Assume that some regularity conditions are satisfied. The second order Taylor expansion for $g(\hat{p})$ at p yields

$$\begin{aligned} g(\hat{p}) &= g(p) + \int_\Xi \varphi'(1)\left(\hat{p} - p\right) dz + \int_\Xi \varphi''(1) \frac{\left(\hat{p} - p\right)^2}{p} dz + o\left(\|\hat{p} - p\|^2\right) \\ &= \phi''(1) \int_\Xi \frac{\left(\hat{p} - p\right)^2}{p} dz + o\left(\|\hat{p} - p\|^2\right). \end{aligned}$$

Therefore, it suffices to study the asymptotics of the statistic $\int_\Xi \left(\hat{p} - p\right)^2 p^{-1} dz$. For the one-dimensional case, Bickel and Rosenblatt (1973) built some asymptotic results for this statistic. Especially, they showed $\int_\Xi \left(\hat{p} - p\right)^2 p^{-1} dz$ multiplied by a term (which goes to ∞ as $n \to \infty$) will converge in distribution to a normal distribution. Based on the convergence result, we can then build a confidence region accordingly. To the best of our knowledge, there is no result for the multi-dimensional case. But Rosenblatt (1976) did some work that generalizes some asymptotic results for another statistic $\max_{z \in \Xi} \left\{\left|\hat{p}(z) - p(z)\right| p(z)^{-1/2}\right\}$ to the multi-dimensional case. We expect to generalize their approach to show that the statistic $\int_\Xi \left(\hat{p} - p\right)^2 p^{-1} dz$ converges in distribution to some normal distribution. The analysis is quite involved. Furthermore, building asymptotics for the

ϕ-divergence in continuous setting is itself an important statistical problem and is of independent interest from this paper. We will try to carry out these ideas in a future work.

When it is difficult to construct the ambiguity set, or there is no data for specifying the set, we can absorb some expert opinion to the ambiguity set. Alternatively, the decision maker can take a sensitivity analysis viewpoint, and can derive the optimal solutions for a number of indexes of ambiguity to see the effect of the distribution ambiguity.

3.4 Bounding Relationship between Distances

We have discussed a number of ϕ-divergences. Besides, there are many other distances that can be used to model ambiguity. It turns out that there may exists certain bounding relationship between two distances, i.e., one distance can bound the other one, or be bounded by the other one; see, e.g., fruitful results in Gibbs and Su (2002). In DRO we can often use the relationship to generate tractable approximations for the distances that may be difficult to handle. In particular, for two distances D_1 and D_2, we have the following property.

Proposition 2. *Suppose there exists an increasing function $B(y)$ on $\Re^+$ such that $D_1 \leq B(D_2)$. Then, for any $\eta > 0$,*

$$\mathbb{P}_{D_2} := \{P \in \mathbb{D} : D_2(P||P_0) \leq \eta\} \subset \{P \in \mathbb{D} : D_1(P||P_0) \leq B(\eta)\} := \mathbb{P}_{D_1}.$$

Consequently, $\sup_{P\in\mathbb{P}_{D_2}} \mathrm{E}_P\left[\mathbb{1}_{\{H(x,\xi)>0\}}\right] \leq \sup_{P\in\mathbb{P}_{D_1}} \mathrm{E}_P\left[\mathbb{1}_{\{H(x,\xi)>0\}}\right]$.

Suppose that $\mathbb{P}_{D_2}$ is used as the ambiguity set in an ambiguous CCP, but the distance measure D_2 may be mathematically less tractable than the tractable distance measure D_1. Suppose D_1 and D_2 satisfy conditions in Proposition 2. Then we can use D_1 to construct a new ambiguity set $\mathbb{P}_{D_1}$ and the corresponding new ambiguous CCP is a tractable conservative approximation of the original ambiguous CCP.

To demonstrate how to specify the function $B(y)$, we consider some simple examples. From Gibbs and Su (2002), we have $D_V \leq D_H$, $D_V \leq \sqrt{D_{\chi^2}}/2$, $D_V \leq \sqrt{D_{KL}/2}$, and $D_V \leq D_S$, where D_V, D_H, D_{χ^2}, D_{KL} and D_S denote the Variation distance, Hellinger distance, χ^2-distance, KL divergence and Separation distance, respectively. Therefore, for D_V versus D_H, D_{χ^2}, D_{KL} and D_S, we can set $B(y) = y$, $B(y) = \sqrt{y}/2$, $B(y) = \sqrt{y/2}$, and $B(y) = y$ respectively. We know that the new confidence level for Variation distance has analytical expression. Then we can build conservative approximations for the other four distances using Variation distance, and the new confidence levels for the conservative approximations can be computed analytically.

4 Conclusions

In this paper, we have studied ambiguous probabilistic programs. We have considered different ambiguity sets constructed using the likelihood ratio of the random distribution. The main con-

tribution of the paper is that we show the ambiguous probabilistic programs with a certain class of ambiguity sets essentially have the same complexity as their pure probabilistic program counterparts. Therefore, the probability functional often provides a reasonable performance measure in decision under uncertainty. One main managerial insight of the paper is that we do not need to fear that much the ambiguity in probabilistic program based decision models. The results derived in this paper complemented the DRO literature.

A Appendix

A.1 Proof of Theorem 1

Proof. We prove this theorem in a measure-theoretic framework. Let $(\Omega, \mathcal{F})$ be a measurable space, where Ω is non-empty and $\mathcal{F}$ is a σ-algebra on Ω. Let $\mathcal{X}$ be the linear space of real-valued measurable functions on $(\Omega, \mathcal{F})$. Let M^+ be the set of probability measures on $(\Omega, \mathcal{F})$. Now, consider a fixed probability measure $P_0 \in M^+$. Let $P \in M^+$ be absolutely continuous w.r.t. P_0 (denoted by $P \ll P_0$). Then, by the Radon-Nikodym theorem, there exists a measurable function $f : (\Omega, \mathcal{F}) \to \Re_+$, called the *Radon-Nikodym derivative* of P w.r.t. P_0 and denoted by $f = dP/dP_0$, such that

$$P(A) = \int_A f \, dP_0 \qquad \text{for all } A \in \mathcal{F}.$$

Given a divergence function $\phi : \Re \to \Re$, the ϕ-divergence from P to P_0 is defined as

$$D_\phi(P \| P_0) = \int_\Omega \phi\left(\frac{dP}{dP_0}\right) dP_0.$$

Furthermore, define $\mathcal{C} \subset \mathcal{X}$ to be the set

$$\mathcal{C} = \{f \in \mathcal{X} : a \le f(\omega) \le b \ \ P_0\text{-almost surely}\},$$

where $0 \le a < 1 < b$. Note that $\mathcal{C}$ is convex. Roughly speaking, the set $\mathcal{C}$ contains all the Radon-Nikodym derivatives of measures in M^+ w.r.t. P_0 that are bounded between $[a, b]$ P_0-almost surely.

Let $A \in \mathcal{F}$ and $\eta_1, \ldots, \eta_m > 0$ be given. (For instance, one can take $A = \{\omega : H(x, \omega) > 0\}$, where $H(x, \cdot) \in \mathcal{X}$ for each x.) Furthermore, let $\phi_1, \ldots, \phi_m$ be given divergence functions. Consider the following problem:

$$\begin{aligned} \sup_{f \in \mathcal{C}} \quad & \int_A f \, dP_0 \\ \text{such that} \quad & \int_\Omega \phi_i(f) \, dP_0 \le \eta_i \quad \text{for } i = 1, \ldots, m, \\ & \int_\Omega f \, dP_0 = 1. \end{aligned} \tag{32}$$

Note that Problem (32) can be written in the form

$$\begin{array}{rcll} v_p^* & = & \sup\limits_{f\in\mathcal{C}} & \displaystyle\int_A f\,dP_0 \\ & & \text{such that} & \Phi(f)\in Q, \end{array} \tag{33}$$

where $\Phi : \mathcal{X} \to \Re^{m+1}$ is given by

$$\Phi(f) = \left(\int_\Omega \phi_1(f)\,dP_0 - \eta_1, \cdots, \int_\Omega \phi_m(f)\,dP_0 - \eta_m, \int_\Omega f\,dP_0 - 1 \right),$$

and $Q \subset \Re^{m+1}$ is the convex cone given by $Q = \Re^m_- \times \{0\}$. Note that Φ is convex on $\mathcal{C} \subset \mathcal{X}$. We remark that Problem (33) is an instance of the so-called *nonlinear programming with generalized constraints* (Rockafellar (1974), p. 26, Example 4').

Strong Duality

To analyze the dual of Problem (33), let $\pi : \Re^{m+1} \to [-\infty, +\infty]$ be the *optimal value function* defined by

$$\pi(u) = \inf\left\{ -\int_A f\,dP_0 : f\in\mathcal{C},\ \Phi(f) - u \in Q \right\} = \inf_{f\in\mathcal{X}} F(f,u),$$

where

$$F(f,u) = \begin{cases} -\displaystyle\int_A f\,dP_0 & \text{if } f\in\mathcal{C},\ \Phi(f)-u\in Q, \\ +\infty & \text{otherwise.} \end{cases}$$

Note that F is convex in (f,u) and closed in u (see Rockafellar (1974), p. 26, Example 4'). The domain of π is

$$\mathrm{dom}(\pi) = \left\{ u\in\Re^{m+1} : \exists\, f\in\mathcal{C} \text{ such that } \Phi(f)-u\in Q \right\}.$$

The *Lagrangian function* $K : \mathcal{X}\times\Re^{m+1} \to [-\infty,+\infty]$ is given by (cf. Section 4 and Example 4' of Rockafellar (1974))

$$\begin{array}{rcl} K(f,y) & = & \inf\limits_{u\in\Re^{m+1}} \left\{ F(f,u) + u^T y \right\} \\ & = & \begin{cases} \inf\limits_{v\in Q} \left\{ -\displaystyle\int_A f\,dP_0 + (\Phi(f)-v)^T y \right\} & \text{if } f\in\mathcal{C}, \\ +\infty & \text{if } f\notin\mathcal{C}, \end{cases} \\ & = & \begin{cases} -\displaystyle\int_A f\,dP_0 + \Phi(f)^T y & \text{if } f\in\mathcal{C}, y\in\Re^m_+\times\Re, \\ -\infty & \text{if } f\in\mathcal{C}, y\notin\Re^m_+\times\Re, \\ +\infty & \text{if } f\notin\mathcal{C}. \end{cases} \end{array}$$

The dual of Problem (33) can then be given by

$$v_d^* = \sup_{y\in\Re^m_+\times\Re} \inf_{f\in\mathcal{C}} \left\{ -\int_A f\,dP_0 + \Phi(f)^T y \right\}.$$

By Theorem 17 of Rockafellar (1974), strong duality $v_p^* = v_d^*$ holds if we can show that π is bounded above in a neighborhood of 0. By Theorem 18(b) of Rockafellar (1974), it suffices to show that $0 \in \text{core dom}(\pi)$, i.e.,

$$\forall u \in \Re^{m+1}, \exists \epsilon > 0 \text{ such that } \lambda u \in \text{dom}(\pi) \ \ \forall \lambda \in [-\epsilon, \epsilon].$$

Since $\text{dom}(\pi)$ is finite-dimensional, the above condition is equivalent to $0 \in \text{int dom}(\pi)$; see Section 6 of Rockafellar (1974). To establish the latter, we simply note that if $f = 1 + \kappa$ with $|\kappa| > 0$ sufficiently small, then $f \in \mathcal{C}$, and $|\Phi(f) - \Phi(1)|$ is small due the the continuity of $\phi_1, \cdots, \phi_m$. □

A.2 Proof of Proposition 1

Proof. For any fixed z, $sz - \sum_{i=1}^m \alpha_i \phi_i(z)$ is linear, and thus convex, in (s, α). Furthermore, it is well known that the supremum preserves convexity. Thus $\Psi(s, \alpha)$ is convex in (s, α). Furthermore, $z \in \mathbb{L}(a, b)$ 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.3 Proof of Corollary 4

Proof. For the ambiguity set $\mathbb{L}_\phi$, we have

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

Then it follows from Theorem 4 that the result of Corollary 4 holds. □

A.4 Proof of Corollary 6

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. □

References

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., L. El-Ghaoui, A. Nemirovski. 2009. *Robust Optimization*. Princeton Series in Applied Mathematics.

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

Bickel, P. J., M. Rosenblatt. 1973. On some global measures of the deviations of density function estimates. *The Annals of Statistics*, **1**(6) 1071-1095.

Bordley, R. F., S. M. Pollock. 2009. A decision-analytic approach to reliability-based design optimization. *Operations Research*, **57**(5) 1262-1270.

Brown, D., M. Sim. 2009. Satisficing measures for analysis of risky positions. *Management Science*, **55**(1) 71-84.

Charnes, A., W. W. Cooper, G. H. Symonds. 1958. Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil. *Management Science*, **4** 235-263.

Chen, W., M. Sim. 2009. Goal-driven optimization. *Operations Research*, **57**(2) 342-357.

Chen, W., M. Sim, J. Sun, C-P Teo. 2010. From CVaR to uncertainty set: Implications in joint chance constrained optimization. *Operations Research*, **58** 470-485.

Cheung, S., A. M. So, K. Wang. 2012. Linear matrix inequalities with stochastically dependent perturbations and applications to chance-constrained semidefinite optimization. *SIAM Journal on Optimization*, **22**(4) 1394-1430.

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

El Ghaoui L., M. Oks, F. Oustry. 2003. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, **51**(4) 543-556.

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.

Erdogan, E., G. Iyengar. 2006. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, **107** 37-61.

Gibbs, A. L., F. E. Su. 2002. On choosing and bounding probability metrics. *International Statistical Review*, **7**(3) 419-435.

Goh, J., M. Sim. 2010. Distributionally robust optimization and its tractable approximations. *Operations Research*, **58**(4) 902-917.

Hong, L. J., Y. Yang, L. Zhang. 2011. Sequential convex approximations to joint chance constrained programs: A Monte Carlo approach. *Operations Research*, **59** 617-630.

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. 2012. Kullback-Leibler divergence constrained distributionally robust optimization. Technical report. http://www.optimization-online.org/DB_FILE/2012/11/3677.pdf.

Hu, Z., L. J. Hong, L. Zhang. 2013. A smooth Monte Carlo approach to joint chance constrained program. *IIE Transactions*, **45**(7) 716-735.

Jiang, R., Y. Guan. 2012. Data-driven chance constrained stochastic program. http://www.optimization-online.org/DB_FILE/2012/07/3525.pdf.

Miller, L. B., H. Wagner. 1965. Chance-constrained programming with joint constraints. *Operations Research*, **13** 930-945.

Nemirovski, A., A. Shapiro. 2006. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, **17** 969-996.

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

Prékopa, A. 1970. On probabilistic constrained programming. *Proceedings of the Princeton Symposium on Mathematical Programming*, 113-138.

Prékopa, A. 2003. Probabilistic programming. In *Stochastic Programming, Handbooks in OR&MS*. Vol. 10, A. Ruszczynski and A. Shapiro, eds., Elsevier.

Rockafellar, R. T. 1974. *Conjugate Duality and Optimization*, volume 16 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1974.

Rockafellar, R. T., R. J.-B. Wets. 1998. *Variational Analysis*, Springer-Verlag, New York.

Rosenblatt, M. 1976. On the maximal deviation of k-dimensional density estimates. *The Annuals of Probability*, **4**(6) 1009-1015.

Scott, D. W. 1992. *Multivariate Density Estimation*, New York, Wiley.

Shapiro, A., S. Ahmed. 2004. On a class of minimax stochastic programs. *SIAM Journal on Optimization*, **14** 1237-1249.

Šidák, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, **62** 626-633.

Simon, H. A. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics*, **69** 99-118.

Wasserman, L. 2004. *All of Statistics: A Concise Course in Statistical Inference*.

Yanıkoğlu, İ., D. den Hertog. 2012. Safe approximations of ambiguous chance constraints using historical data. *INFORMS Journal on Computing*, forthcoming.

Zymler, S., Kuhn, D., B. Rustem. 2013a. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, **137** 167-198.

Zymler, S., Kuhn, D., B. Rustem. 2013b. Worst-case value at risk of nonlinear portfolios. *Management Science*, **59**(1) 172-188.