Kullback-Leibler Divergence Constrained Distributionally Robust Optimization

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

Abstract

In this paper we study distributionally robust optimization (DRO) problems where the ambiguity set of the probability distribution is defined by the Kullback-Leibler (KL) divergence. We consider DRO problems where the ambiguity is in the objective function, which takes a form of an expectation, and show that the resulted minimax DRO problems can be formulated as a one-layer convex minimization problem. We also consider DRO problems where the ambiguity is in the constraint. We show that ambiguous expectation-constrained programs may be reformulated as a one-layer convex optimization problem that takes the form of the Benstein approximation of Nemirovski and Shapiro (2006). We further consider distributionally robust probabilistic programs. We show that the optimal solution of a probability minimization problem is also optimal for the distributionally robust version of the same problem, and also show that the ambiguous chance-constrained programs (CCPs) may be reformulated as the original CCP with an adjusted confidence level. A number of examples and special cases are also discussed in the paper to show that the reformulated problems may take simple forms that can be solved easily. The main contribution of the paper is to show that the KL divergence constrained DRO problems are often of the same complexity as their original stochastic programming problems and, thus, KL divergence appears a good candidate in modeling distribution ambiguities in mathematical programming.

1 Introduction

Optimization models are often used in practice to guide decision makings. In many of these models there exist parameters that need to be specified or estimated. When these parameters appear in the objective function, the models can typically be formulated as

$$\underset{x\in X}{\text{minimize}}\ H(x,\xi), \tag{1}$$

where ξ denotes the vector of parameters, x is the vector of design (or decision) variables, and $X \subset \Re^d$ is the feasible region. Alternatively, when these parameters appear in the constraint function, the models can be formulated as

$$\begin{aligned} \underset{x\in X}{\text{minimize}} \quad & h(x) \\ \text{subject to} \quad & H(x,\xi) \le 0. \end{aligned} \tag{2}$$

Although it is possible to unify Problems (1) and (2) using, for instance, an epigraphical representation, we think it is necessary to study them separately due to their different structures and applications.

It is widely known that optimal solutions of optimization models, like (1) and (2), may depend heavily on the specification or estimation of their parameters. However, due to limited data/information availability on and possibly random nature of these parameters, it is often difficult to specify or estimate them precisely. Then, the optimal solutions of these models may turn out to be rather suboptimal or even infeasible for the true optimization problems. To address such an issue, a number of approaches have been suggested in the past decades. Robust optimization (see, for instance, Ben-Tal and Nemirovski (1998, 2000), Bertsimas and Sim (2004), and El Ghaoui et al. (1998)) targets to find an optimal solution that is, to some extent, immune to the ambiguity in the parameters. It typically models the ambiguity by restricting the parameters in a set, often called an uncertainty set, and optimizes under the worst case of the parameters in the set. For a comprehensive survey on robust optimization, readers are referred to Ben-Tal et al. (2009) and Bertsimas et al. (2011).

Modeling ambiguous parameters using an uncertainty set and allowing the parameters to take any value in the set sometimes ignore that the parameters may admit a stochastic nature (i.e., they are random variables) and, therefore, lead to solutions that may be excessively conservative, especially when the uncertainty set is large. To address this issue, distributionally robust optimization (DRO) was introduced. DRO considers a stochastic programming version of Problem (1) or (2) by, for instance, substituting $H(x,\xi)$ in Problems (1) and (2) by their expectations $\mathrm{E}_P\left[H(x,\xi)\right]$ with P denoting the distribution of ξ, models the ambiguity by restricting the distribution P in a set, often called an ambiguity set, and optimizes under the worst case of the distribution in the set. This approach is also in line with the economic literature that distinguishes between the randomness of the parameters (called the risk) and the ambiguity in specifying the randomness (called the uncertainty or ambiguity); see, e.g., Ellsberg (1961) and Epstein (1999). The literature on DRO is growing fast; see, e.g., the recent studies of Delage and Ye (2010) and Goh and Sim (2010).

An important question about DRO modeling is how to choose the ambiguity set. There exists a significant amount of work that constructs the ambiguity set by the moments of the distribution. For instance, Delage and Ye (2010) novelly constructed a confidence set for the mean vector and covariance matrix using historical data. Goh and Sim (2010) considered tractable conic representable sets for the mean vector coupled with information on directional deviations. We refer the readers to Delage and Ye (2010) and references therein on the literature on moment ambiguities. Note that in many practical situations, we may obtain an estimate of the distribution via statistical fitting, which is our best estimate (or best guess) of the distribution and often contains valuable information of the distribution. We call such a distribution *nominal distribution*. Then, a reasonable

approach is to construct the ambiguity set by requiring the distribution within a certain distance from the nominal distribution. There are different ways to define the distance between two probability distributions. In this paper we use the Kullback-Leibler (KL) divergence. The KL divergence originated in the field of information theory (Kullback and Leibler 1951), and it is now accepted widely as a good measure of distance between two distributions. KL divergence is also widely used in the area of operations research in recent years. For instance, in rare event simulation, the KL cross entropy is minimized in order to find a good importance sampling distribution that achieves variance reduction (e.g., Rubinstein 2002 and Homem-de-Mello 2007), whereas in simulation optimization, the KL divergence is minimized in order to obtain a good sampling distribution that guides the random search (e.g., Hu et al. 2007).

Ambiguity sets defined by distance measures have been investigated recently. Calafiore (2007) studied a portfolio selection problem where the ambiguity of the return distribution is described by KL divergence. Klabjan et al. (2012) considered the inventory management problem where the ambiguity of the demand distribution for a single item is depicted via the histogram and the χ^2-distance. Ben-Tal et al. (2012) considered the robust optimization problems where ambiguities are modeled using various ϕ-divergence measures. These studies all assume that the distribution of the random parameters is supported on a finite set of values and the ambiguity set is constructed for the discrete distribution. Prior to Ben-Tal et al. (2012), Ben-Tal et al. (2010) proposed a soft robust model under ambiguity and related the model to the theory of convex risk measures. Their work is not restricted to finite scenarios but requires a bounded support for the random function due to the bounded requirement for the dual representation of convex risk measures.

In this paper we study DRO problems where the underlying distributions are general, allowing them to be discrete or continuous and bounded or unbounded, and the ambiguity set consists of all probability distributions whose KL divergence from the nominal distribution is less than or equal to a positive constant. Such a constant is referred to as the *index of ambiguity*, since it controls the size of the ambiguity set. We first study the optimization models where the ambiguous parameters appear in the objective functions, and consider their minimax DRO problem. We implement a change-of-measure technique and reformulate the problem as a minimax problem with the inner problem maximizing over the likelihood ratio functional and the outer problem minimizing over x. To solve the inner functional optimization problem, we implement a functional optimization technique to solve its Lagrangian dual, obtain a closed-form expression of the optimal objective, and prove the strong duality. This closed-form expression allows us to convert the minimax DRO problem into a single minimization problem, which can be solved via either conventional deterministic optimization techniques or standard stochastic optimization techniques. Furthermore, as an interesting and important side result, we identify that having a light right tail is a sufficient and necessary condition for the worst-case expectation being finite in the DRO model, and this result

has profound implications in the modeling of ambiguities for distributions having heavy tails.

We next consider optimization models where random parameters appear in constraint functions, as in Problem (2). There are different approaches to handling these parameters. One approach is to requiring the expected value of the random function satisfy the constraint, i.e., substituting $H(x,\xi)$ in (2) by $\mathrm{E}_P[H(x,\xi)]$, and we call this problem an *expectation constrained program* (ECP). Another approach is to requiring the random constraint be satisfied with at least a given probability, i.e., substituting the constraint in (2) by $\Pr_{\sim P}\{H(x,\xi) \leq 0\} \geq 1-\beta$ for some $\beta \in (0,1)$, and we call this problem a *chance constrained program* (CCP). We first consider DRO formulations of ECPs, where we require the expectation constraints be satisfied for any P in an ambiguity set defined by KL divergence. We call such problems *ambiguous ECPs*. Implementing the functional approach developed for minimax DRO problems, we show that the ambiguous ECPs may be reformulated as single-layer optimization problems. More interestingly, we find that the formulations of ambiguous ECPs are equivalent to the famous Bernstein approximations of Nemirovski and Shapiro (2006), which are constructed to conservatively approximate CCPs. This result shows that the ECPs and CCPs are intrinsically interrelated via the KL divergence, and it allows us to understand the conservatism of the Bernstein approximations to CCPs from a robust optimization perspective.

When the performance measure "expectation function" in optimization models is substituted by a "probability function", the resulted stochastic programming model is often called a probabilistic program. Probabilistic programs are often studied separately from the expectation based stochastic programming in the literature due to their particularity (Prékopa 2003). Depending on where the probability function appears, probabilistic programs can be classified as the probability minimization problem and the CCP. We consider the DRO formulations for these probabilistic programs. DRO formulations of probability minimization problems propose to minimize the worst case of the probability function. We show that when ambiguity sets are defined by KL divergence, the minimax DRO for probability minimization is essentially the same as the original probability minimization problem. Therefore, to solve such DRO, it suffices to solve the original problem. DRO formulations of CCPs require the chance constraints be satisfied for all P in an ambiguity set. Note that such problems are also called *ambiguous CCPs* in the literature (e.g., Erdogan and Iyengar 2006 and Nemirovski and Shapiro 2006). We show that, when ambiguity sets are defined by KL divergence, the ambiguous CCPs may be reformulated as the original CCPs with only the confidence levels being rescaled to a more conservative level. This suggests that KL divergence-constrained ambiguous CCPs essentially have the same complexity as the original CCPs and can be solved using the same techniques that are used to solve the original CCPs. To generalize the results, we also study the distributionally robust value-at-risk (VaR) and conditional value-at-risk (CVaR) optimization problems, and show that they are also tractable when their ambiguity sets are defined by KL divergence.

Following the theoretical foundations developed in the paper, we consider a number of examples and special cases. For the affinely perturbed independent case, we show that the worst-case expectation can be written as a summation of convex functions that are generated by the logarithmic moment generating functions of the independent random parameters. For the linear case with a multivariate normal nominal distribution, we show the worst-case expectation has a second order cone representation. Moreover, the worst-case distribution is still a multivariate normal distribution with a shifted mean vector and the same covariance matrix. In the Appendix we re-derive this formulation by restricting ambiguous distribution to the family of multivariate normal distributions. Finally, we consider the broadly used exponential family of distributions and derive the general expressions of the worst-case distributions.

While this paper focuses mainly on solving various KL divergence constrained DRO problems, we are equally concerned about the modeling of ambiguity sets. In this paper we show that the use of KL divergence has some advantages. The first advantage is that KL divergence is a widely accepted measure of distances between distributions. The second (and perhaps more critical) advantage is its tractability in solving DRO problems. As shown in this paper, when the ambiguity set is defined by KL divergence, the worst-case expectation of a random performance may be derived analytically and the resulted DRO models, including minimax DRO, ambiguous ECP, ambiguous probabilistic programs (including probability minimization and CCP), and distributionally robust VaR/CVaR problems, can all be formulated into simple one-layer optimization problems that are readily solvable by standard optimization tools. Furthermore, these DRO models may be incorporated into more sophisticated models, such as multistage stochastic programming models and dynamic programming models (e.g., inventory systems, as considered in Klabjan et al. (2012)), to solve more complicated practical problems. In addition, we also show that KL divergence constrained DRO models may serve as analytically tractable conservative approximations to DRO models using many other distance measures.

Using KL divergence to model ambiguity sets, nevertheless, also has some limitations. First, there may not be any practical guidelines in determining the size of the ambiguity set (i.e., the index of ambiguity). When distributions are supported on a finite set of values, Ben-Tal et al. (2012) show that some confidence set may be derived using data. When distributions are continuous, however, we do not have such results. This is an important question for future research. Second, as shown in the paper, we find that KL divergence has difficulty in handling random functions that are heavy right tailed under the nominal distribution. In such cases, the worst-case expectation is infinite no matter how small the index of ambiguity is. Therefore, such a DRO model cannot be applied for stochastic optimization models with heavy tail random functions. It is worthwhile noting that similar modeling issues exist for many distance measures that can be bounded from above by KL divergence. Considering that heavy tail distributions are not uncommon in practical

situations (especially in financial risk management), it remains a very important problem to develop meaningful yet tractable DRO models for heavy tailed distributions.

The rest of this paper is organized as follows. In Section 2 we study minimax DRO problems and show that the worst-case expectation admits an analytical expression. In Section 3 we investigate DRO formulations of ECPs, and uncover some intrinsic relations between robust ECPs and CCPs. In Section 4 we analyze DRO formulations of probabilistic programs and their extensions to VaR/CVaR optimization problems. We consider a number of special cases in Section 5, followed by conclusions in Section 6. Some lengthy proofs are provided in the Appendix.

2 Minimax Distributionally Robust Optimization

We first analyze the case where the ambiguous random parameters are in the objective function. Consider the following minimax distributionally robust optimization problem:

$$\underset{x\in X}{\text{minimize}} \ \underset{P\in\mathbb{P}}{\text{maximize}} \ \mathrm{E}_P\left[H(x,\xi)\right] \tag{3}$$

where the feasible region X is assumed to be a convex compact subset of $\Re^d$, P denotes the distribution of ξ, $\mathrm{E}_P[\cdot]$ denotes that the expectation is taken with respect to a probability distribution P, $\mathbb{P}$ is an ambiguity set and the maximum is taken over all probability distributions contained in $\mathbb{P}$. As discussed in Section 1, the ambiguity set $\mathbb{P}$ may take different forms, depending on the available information and the modeler's belief. In this paper, we focus on the case where the distribution of ambiguous parameters is constrained by the Kullback-Leibler (KL) divergence. Specifically, we consider the ambiguity set

$$\mathbb{P} = \left\{P \in \mathbb{D} : D(P||P_0) \le \eta\right\}, \tag{4}$$

where $\mathbb{D}$ denotes the set of all probability distributions and $D(P\|P_0)$ denotes the KL divergence from distribution P to the nominal distribution P_0. The KL divergence $D(P\|P_0)$ implicitly assumes that P is absolutely continuous with respect to P_0 (denoted as $P << P_0$), i.e., for every measurable set A, $P_0(A) = 0$ implies $P(A) = 0$. 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. Then the KL divergence from P to P_0 is defined as

$$D(P\|P_0) = \int_{\Xi} p(z)\log\frac{p(z)}{p_0(z)}\,\mathrm{d}z. \tag{5}$$

When P_0 is a discrete distribution, we understand $p_0(z)$ in (5) as the probability mass function and the integral as the summation. 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, and the integral becomes a mixture of integral and summation. It can be shown that $D(P\|P_0) \ge 0$ and

the equality holds if and only if $p(z) = p_0(z)$ almost surely (a.s.) under P_0. As defined in Section 1, the constant η used in (4) is the index of ambiguity, which controls the size of the ambiguity set $\mathbb{P}$.

Problem (3) is a rather abstract optimization model, as the decision variable of the inner maximization problem is the probability distribution P, which does not explicitly appear in the objective function. This makes the problem difficult to handle. One step towards solving Problem (3) is to transform it into an explicit optimization problem via the so called change-of-measure technique (e.g., Hu et al. 2012 and Lam 2012). Note that $p_0(z)$ is the nominal distribution of the random vector ξ. Let $L(z) = p(z)/p_0(z)$. In the literature $L(z)$ is often called a likelihood ratio or a Radon-Nikodym derivative. It is easy to see that $L(z) \geq 0$ and $\mathrm{E}_{P_0}[L(\xi)] = 1$. When there is no confusion we suppress the variable z and just use L to denote $L(z)$. We denote by $\mathbb{L} = \{L \in \mathbb{L}(P_0) : \mathrm{E}_{P_0}[L] = 1, L \geq 0 \text{ a.s.}\}$ the set of likelihood ratios that are generated by all P such that $P << P_0$. By applying the change-of-measure technique, we obtain

$$D(P\|P_0) = \int_{\Xi} \frac{p(z)}{p_0(z)} \log \frac{p(z)}{p_0(z)} p_0(z)\,\mathrm{d}z = \mathrm{E}_{P_0}[L(\xi)\log L(\xi)].$$

Similarly, applying the change-of-measure technique to the objective function, we have

$$\mathrm{E}_P[H(x,\xi)] = \int_{\Xi} H(x,\xi)p(z)\mathrm{d}z = \int_{\Xi} H(x,\xi)\frac{p(z)}{p_0(z)}p_0(z)\mathrm{d}z = \mathrm{E}_{P_0}[H(x,\xi)L(\xi)].$$

Therefore, we can transform both the constraint function and the objective function into expectation forms where the expectation is taken with respect to the nominal distribution P_0. Then, the inner maximization problem in Problem (3) can be reformulated as

$$\begin{array}{ll} \text{maximize} & \mathrm{E}_{P_0}[H(x,\xi)L] \\ \text{subject to} & \mathrm{E}_{P_0}[L\log L] \leq \eta, \\ & L \in \mathbb{L}. \end{array} \tag{6}$$

Therefore, the change-of-measure technique converts an optimization problem on P (i.e., the inner maximization problem in Problem (3)) to an optimization problem on L (i.e., Problem (6)) which, we show in next subsection, can be solved analytically by a functional approach.

2.1 Solving the Inner Maximization Problem

A first yet critical observation is that, with L being the decision variable, Problem (6) is a convex optimization problem. To see this more clearly, let us consider any $\lambda \in [0,1]$ and any $L_i(\xi) \in \mathbb{L}, i = 1,2$. It can be verified that $L_\lambda(\xi) = \lambda L_1(\xi) + (1-\lambda)L_2(\xi) \in \mathbb{L}$. Furthermore, since $y\log y$ is a convex function of y on $\Re^+$, we have for every ξ, $L_\lambda(\xi)\log L_\lambda(\xi) \leq \lambda L_1(\xi)\log L_1(\xi) + (1-\lambda)L_2(\xi)\log L_2(\xi)$. It follows that $\mathrm{E}_{P_0}[L_\lambda(\xi)\log L_\lambda(\xi)] \leq \lambda \mathrm{E}_{P_0}[L_1(\xi)\log L_1(\xi)] + (1-\lambda)\mathrm{E}_{P_0}[L_2(\xi)\log L_2(\xi)]$. This

shows $\mathrm{E}_{P_0}[L \log L]$ is convex in L. Similarly, it can be shown that $\mathrm{E}_{P_0}[H(x,\xi)L]$ is convex in L. Thus, Problem (6) is a convex optimization problem.

For every $x \in X$, let $M_H(t) = \mathrm{E}_{P_0}\left[e^{tH(x,\xi)}\right]$ denote the moment generating function of $H(x,\xi)$ under P_0. Let $S = \{s \in \Re : s > 0, M_H(s) < +\infty\}$. Note that we suppress the dependence of $M_H(t)$ and S on x for notational simplicity. We make the following assumption on the original optimization problem.

Assumption 1. *For every $x \in X$, S is a nonempty set.*

Assumption 1 shows that under measure P_0, the moment generating function $M_H(t)$ of the random variable $H(x,\xi)$ is finite valued for some $s > 0$. Because $M_H(t)$ is convex in t, its effective domain $\mathrm{dom}M_H := \{t \in \Re : M_H(t) < +\infty\}$ is a convex set (Rockafellar 1970), which implies $[0, s] \subset \mathrm{dom}M_H$. Assumption 1 requires that the random variable $H(x,\xi)$ has a light right tail under P_0 for every $x \in X$. Note that $H(x,\xi)$ simply satisfies this assumption if it is supported on a finite set of values or if it is bounded a.s.

The basic idea of solving Problem (3) is to implement the duality theory of convex optimization, which is a key tool in robust optimization and has been used frequently in DRO; see, e.g., Delage and Ye (2010) and Goh and Sim (2010). To formulate the dual of Problem (6), we let

$$\ell_0(\alpha, L) = \mathrm{E}_{P_0}[H(x,\xi)L(\xi)] - \alpha\left(\mathrm{E}_{P_0}[L(\xi)\log L(\xi)] - \eta\right)$$

be the Lagrangian functional associated with Problem (6). Then, Problem (6) is equivalent to

$$\underset{L\in\mathbb{L}}{\text{maximize}} \ \underset{\alpha\geq 0}{\text{minimize}} \ \ell_0(\alpha, L). \tag{7}$$

Interchanging the order of the maximum and minimum operators, we obtain the Lagrangian dual of Problem (7), which is represented as

$$\underset{\alpha\geq 0}{\text{minimize}} \ \underset{L\in\mathbb{L}}{\text{maximize}} \ \ell_0(\alpha, L). \tag{8}$$

Conventionally, since Problem (6) is a convex optimization problem, the strong duality for Problems (7) and (8) should hold. We prove the strong duality later in this section (i.e., Theorem 2). The strong duality indicates that, to solve Problem (6), it suffices to solve Problem (8). In what follows we focus on Problem (8). Omitting the term $\alpha\eta$, the inner maximization problem in Problem (8) can be expressed as

$$\begin{array}{ll} \underset{L\in\mathbb{L}^0}{\text{maximize}} & \mathrm{E}_{P_0}[H(x,\xi)L(\xi) - \alpha L(\xi)\log L(\xi)] \\ \text{subject to} & \mathrm{E}_{P_0}[L(\xi)] = 1, \end{array} \tag{9}$$

where we define $\mathbb{L}^0 = \{L \in \mathbb{L}(P_0) : L \geq 0 \text{ a.s.}\}$.

Problem (9) is a convex functional optimization problem. Let $v(\alpha)$ denote the optimal objective value of Problem (9). To solve the problem, we consider three cases:

- Case 1, $\alpha = 0$;
- Case 2, $\alpha \neq 0$ and $1/\alpha \in S$;
- Case 3, $\alpha \neq 0$ and $1/\alpha \notin S$.

We first consider Case 1. Let $H_u(x)$ be the essential supremum of $H(x,\xi)$ under measure P_0, i.e.,

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

where $\Pr_{\sim P_0}$ denotes that the probability is taken with respect to P_0. Then, we can construct a sequence of distributions $P_j << P_0$ concentrating towards $H_u(x)$ and consequently construct $L_j \in \mathbb{L}^0$, such that $\mathrm{E}_{P_0}\left[H(x,\xi)L_j(\xi)\right]$ tends to $H_u(x)$ as $j \to +\infty$. Therefore, $v(\alpha) = H_u(x)$.

Next we consider Case 2. There are a number of potential approaches to solving relatively simple convex functional optimization problems. We find the method of Homem-de-Mello (2007) applicable. Homem-de-Mello (2007) investigated the rare event probability estimation using a "cross-entropy method". One critical step of his method is to find a density that minimizes the KL divergence to the nominal best but non-attainable density among a certain class. He formulated the problem into a convex functional optimization problem, and proposed an implementable approach to solving the problem. The basic idea of Homem-de-Mello (2007) is to compute the derivative of the functional with respect to the decision variable, which is a density function. In this paper, we implement a similar approach to solving Problem (9).

Define the functionals

$$\mathcal{J}(L(\xi)) := \mathrm{E}_{P_0}\left[H(x,\xi)L(\xi) - \alpha L(\xi)\log L(\xi)\right],$$
$$\mathcal{J}_c(L(\xi)) := \mathrm{E}_{P_0}\left[L(\xi)\right] - 1.$$

Note that $\mathcal{J}(L(\xi))$ is convex in L and $\mathcal{J}_c(L(\xi))$ is linear in L. This allows us to calculate the derivative of the functionals (see, e.g., Shapiro et al. (2009) for the definition of a derivative). Let $D\mathcal{J}(L(\xi))$ denote the derivative of $\mathcal{J}(L(\xi))$. Then, for any feasible direction $V(\xi)$ at $L(\xi)$,

$$\begin{aligned} D\mathcal{J}(L(\xi))V &= \lim_{t\to 0}\frac{\mathcal{J}(L(\xi)+tV(\xi)) - \mathcal{J}(L(\xi))}{t} \\ &= \lim_{t\to 0}\frac{\mathrm{E}_{P_0}\left[H(x,\xi)\left(L+tV\right) - \alpha\left(L+tV\right)\log L\right] - \mathrm{E}_{P_0}\left[H(x,\xi)L - \alpha L\log L\right]}{t} \\ &= \mathrm{E}_{P_0}\left[H(x,\xi)V\right] - \alpha\lim_{t\to 0}\mathrm{E}_{P_0}\left[\frac{\left(L+tV\right)\log\left(L+tV\right) - L\log L}{t}\right]. \end{aligned} \tag{10}$$

Note that the function $y\log y$ is convex in y on $\Re^+$. It follows that for any y and v feasible, the function $\left[(y+tv)\log(y+tv) - y\log y\right]/t$ is monotone in t. Therefore, by the monotone convergence theorem (Durrett 2005), we can interchange the order of the operators $\lim_{t\to 0}$ and E_{P_0} in (10). It

follows that

$$\begin{aligned} D\mathcal{J}(L(\xi))V &= \mathrm{E}_{P_0}\left[H(x,\xi)V\right] - \alpha \mathrm{E}_{P_0}\left[\lim_{t\to 0}\frac{(L+tV)\log(L+tV) - L\log L}{t}\right] \\ &= \mathrm{E}_{P_0}\left[H(x,\xi)V\right] - \alpha \mathrm{E}_{P_0}\left[(\log L + 1)V\right] \\ &= \mathrm{E}_{P_0}\left[(H(x,\xi) - \alpha(\log L + 1))V\right]. \end{aligned}$$

Similarly, it is also straightforward to obtain that

$$D\mathcal{J}_c(L(\xi))V = \mathrm{E}_{P_0}\left[V(\xi)\right].$$

We have obtained the derivatives of the functionals. As in Homem-de-Mello (2007), we now construct the Lagrangian functional associated with Problem (9) as follows:

$$\begin{aligned} \ell(L,\lambda) &= \mathrm{E}_{P_0}\left[H(x,\xi)L(\xi) - \alpha L(\xi)\log L(\xi)\right] + \lambda\left(\mathrm{E}_{P_0}\left[L(\xi)\right] - 1\right) \\ &= \mathrm{E}_{P_0}\left[H(x,\xi)L - \alpha L\log L + \lambda L\right] - \lambda. \end{aligned}$$

Recall that Proposition 3.3 of Bonnans and Shapiro (2000) shows, if there exists a pair $(L^*(\xi), \lambda^*)$ such that $L^*(\xi) \in \mathbb{L}^0$, $\mathcal{J}_c(L^*(\xi)) = 0$ and

$$L^*(\xi) \in \arg\max_{L\in\mathbb{L}^0} \ell(L,\lambda^*), \tag{11}$$

$L^*(\xi)$ is an optimal solution of Problem (9). The problem thus simplifies to solving Problem (11).

Note that Problem (11) is a convex optimization problem with essentially no constraints. Thus, its optimal solution should be a stationary point of a certain sense, i.e., it should enforce the derivative of $\ell(L,\lambda)$ to be zero in some sense. From the expressions of the derivatives and the linearity of the derivative operator, we immediately obtain

$$D\ell(L,\lambda)V = \mathrm{E}_{P_0}\left[(H(x,\xi) - \alpha(\log L + 1) + \lambda)V\right].$$

Then, we have the following proposition whose strict proof is provided in the Appendix.

Proposition 1. *Suppose $L = L^*(\xi,\lambda)$ satisfies $H(x,\xi) - \alpha(\log L + 1) + \lambda = 0$, which means $D\ell(L,\lambda) = 0$ (i.e., $D\ell(L,\lambda)$ is the zero linear operator). Then, $\ell(L^*(\xi,\lambda),\lambda) < +\infty$ and $L^*(\xi,\lambda) \in arg\max_{L\in\mathbb{L}^0}\ell(L,\lambda)$.*

Proposition 1 shows that the optimal objective value of the functional optimization problem (11) is finite. Moreover, it shows the optimal solution takes the following form

$$L^*(\xi,\lambda) = e^{(\lambda-\alpha)/\alpha}\cdot e^{H(x,\xi)/\alpha}.$$

Setting $\lambda^* = -\alpha\log\mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right] + \alpha$, we have $\mathcal{J}_c(L^*(\xi,\lambda^*)) = 0$. Therefore,

$$L^*(\xi) = L^*(\xi,\lambda^*) = \frac{e^{H(x,\xi)/\alpha}}{\mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right]} \tag{12}$$

and λ^* form a pair that satisfies the conditions in Proposition 3.3 of Bonnans and Shapiro (2000). This shows that $L^*(\xi)$ solves Problem (9). Plugging $L^*(\xi)$ into Problem (9) we obtain the optimal objective value of Problem (9):

$$v(\alpha) = \alpha \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right] + \alpha\eta. \tag{13}$$

Finally, we consider Case 3. In this case, we must have $H_u(x) = +\infty$. Now we consider a positive real sequence $\{R_j\}$ such that $\lim_{j\to+\infty} R_j = +\infty$. Let $\mathbb{1}_{\{A\}}$ denote the indicator function which is equal to 1 if the event A happens and 0 otherwise. We use $H(x,\xi)\mathbb{1}_{\{H(x,\xi)\le R_j\}}$ to replace $H(x,\xi)$ in Problem (9) and denote the resulted problem as Problem (R_j). Denote the optimal objective value of Problem (R_j) as $v_j(\alpha)$. Because $H(x,\xi)\mathbb{1}_{\{H(x,\xi)\le R_j\}}$ is bounded by R_j from above, its moment generating function exists for all $s \ge 0$. Therefore, we can solve Problem (R_j) using the functional approach in Case 2 and obtain the optimal objective value $v_j(\alpha)$. It follows that

$$v_j(\alpha) = \alpha \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)\mathbb{1}_{\{H(x,\xi)\le R_j\}}/\alpha}\right] + \alpha\eta.$$

Because $\alpha \notin S$, we have $v_j(\alpha) \to +\infty$ as $j \to +\infty$. Note that the objective function of Problem (R_j) is always a lower bound of the objective function of Problem (9). Moreover, the feasible regions of the two problems are the same. Thus, we have $v_j(\alpha) \le v(\alpha)$ for any $j > 0$. This implies $v(\alpha) = +\infty$.

Because Assumption 1 is satisfied, for any $x \in X$, there always exists $\alpha > 0$ such that (13) is finite. This shows the optimal objective value of Problem (8) is finite. Note further that in the Appendix we show (43) holds. Therefore, we can incorporate Case 1 into Case 2. Combining the three cases, we obtain the following theorem.

Theorem 1. *Suppose that Assumption 1 is satisfied. Problem (8) is equivalent to the following one-layer optimization problem*

$$\underset{\alpha\ge 0}{\textit{minimize}}\ \ h_x(\alpha) := \alpha \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right] + \alpha\eta. \tag{14}$$

Remark 1. *Recently, Lam (2012) proposed a one dimensional analog of Problem (6). His goal is to study the robustness of the random system outputs to the simulation input distributions, as what is suggested and investigated in Hu et al. (2012). The theme of model misspecification was also considered in robust control (Hansen and Sargent 2008). Hansen and Sargent (2008) modeled distribution perturbations of the shock process that enters the transition equation of a control problem, and proposed Problem (8) to penalize the misspecification. Both Hansen and Sargent (2008) and Lam (2012) seeked the expression of the optimal solution of Problem (8) by a heuristic approach and then verified the optimality by using the Jensen's inequality. In this paper, the decision models and source of randomness are drastically different from that of control theory.*

Moreover, the variable α which is allowed to take values on $\Re^+$ becomes a decision variable adjoint with x. Different cases are considered and the problem is solved by a more systematic functional optimization approach. This solution approach provides us more insights about the optimal solution (as shown in Section 2.2) and may be used to solve more general functional optimization problems that may arise in DRO.

Let $\alpha^*(x)$ be an optimal solution of Problem (14). Let $\kappa_u = \Pr_{\sim P_0}\{H(x,\xi) = H_u(x)\}$, i.e., κ_u is the mass of the distribution P_0 on its essential supremum. We have the following proposition. The proof of the proposition is provided in the Appendix.

Proposition 2. *Suppose Assumption 1 is satisfied. Then $\alpha^*(x) = 0$ or $1/\alpha^*(x) \in S$. Moreover, $\alpha^*(x) = 0$ if and only if $H_u(x) < +\infty$, $\kappa_u > 0$ and $\log \kappa_u + \eta \geq 0$.*

Proposition 2 shows that the optimal solution of Problem (14) is finite. It also provides the equivalent conditions for that the optimality is attained at 0. This will be used in analyzing the complementary slackness between Problems (7) and (8). Now we can state and prove strictly the following theorem regarding the strong duality.

Theorem 2. *Suppose that Assumption 1 is satisfied. Then, the optimal objective values of Problems (7) and (8) are equal.*

Proof. Consider the Lagrangian functional $\ell_0(\alpha, L)$. We first show that if there exists a saddle point $(\tilde{\alpha}, \tilde{L})$ for $\ell_0(\alpha, L)$, i.e., for any $\alpha \geq 0$ and $L \in \mathbb{L}$,

$$\ell_0(\tilde{\alpha}, L) \leq \ell_0(\tilde{\alpha}, \tilde{L}) \leq \ell_0(\alpha, \tilde{L}), \tag{15}$$

then the strong duality holds.

Let v_p and v_d denote the optimal objective values of Problem (7) and Problem (8) respectively. By weak duality, we immediately obtain $v_p \leq v_d$. By (15), we have $\ell_0(\tilde{\alpha}, \tilde{L}) \leq \inf_{\alpha \geq 0} \ell_0(\alpha, \tilde{L})$. It follows that

$$\ell_0(\tilde{\alpha}, \tilde{L}) \leq \inf_{\alpha \geq 0} \ell_0(\alpha, \tilde{L}) \leq \sup_{L \in \mathbb{L}} \inf_{\alpha \geq 0} \ell_0(\alpha, L) = v_p.$$

On the other hand, by (15), we have $\sup_{L \in \mathbb{L}} \ell_0(\tilde{\alpha}, L) \leq \ell_0(\tilde{\alpha}, \tilde{L})$. It follows that

$$v_d = \inf_{\alpha \geq 0} \sup_{L \in \mathbb{L}} \ell_0(\alpha, L) \leq \sup_{L \in \mathbb{L}} \ell_0(\tilde{\alpha}, L) \leq \ell_0(\tilde{\alpha}, \tilde{L}).$$

Therefore, we obtain $v_p = v_d = \ell_0(\tilde{\alpha}, \tilde{L})$.

We next show the existence of the saddle point. Let $\tilde{\alpha} = \alpha^*(x)$. We consider two cases: Case A, $\tilde{\alpha} \neq 0$; Case B, $\tilde{\alpha} = 0$. For Case A, let

$$\tilde{L} = \frac{e^{H(x,\xi)/\tilde{\alpha}}}{\mathrm{E}_{P_0}\left[e^{H(x,\xi)/\tilde{\alpha}}\right]}.$$

We show that $(\tilde{\alpha}, \tilde{L})$ is a saddle point. Because $\tilde{L}$ solves Problem (9) as $\alpha = \tilde{\alpha}$, we have $\ell_0(\tilde{\alpha}, L) \leq \ell_0(\tilde{\alpha}, \tilde{L})$. Now we prove the second inequality of (15). We show that it is actually an equality. Note that $\tilde{\alpha} = \alpha^*(x)$ is an optimal solution of Problem (14). Furthermore, from Proposition 2 we have $0 < \alpha^*(x) < +\infty$. Therefore,

$$0 = \nabla_\alpha \left[\alpha \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right] + \alpha\eta\right]\Big|_{\alpha=\tilde{\alpha}} = -\frac{\mathrm{E}_{P_0}\left[e^{H(x,\xi)/\tilde{\alpha}} H(x,\xi)/\tilde{\alpha}\right]}{\mathrm{E}_{P_0}\left[e^{H(x,\xi)/\tilde{\alpha}}\right]} + \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\tilde{\alpha}}\right] + \eta.$$

It follows that

$$-\mathrm{E}_{P_0}\left[\tilde{L}(\xi)\log\tilde{L}(\xi)\right] + \eta = -\frac{\mathrm{E}_{P_0}\left[e^{H(x,\xi)/\tilde{\alpha}} H(x,\xi)/\tilde{\alpha}\right]}{\mathrm{E}_{P_0}\left[e^{H(x,\xi)/\tilde{\alpha}}\right]} + \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\tilde{\alpha}}\right] + \eta = 0.$$

Therefore,

$$\begin{aligned} \ell_0(\alpha, \tilde{L}) &= \mathrm{E}_{P_0}\left[H(x,\xi)\tilde{L}(\xi)\right] - \alpha\left(\mathrm{E}_{P_0}\left[\tilde{L}(\xi)\log\tilde{L}(\xi)\right] - \eta\right) \\ &= \mathrm{E}_{P_0}\left[H(x,\xi)\tilde{L}(\xi)\right] \;=\; \ell_0(\tilde{\alpha}, \tilde{L}). \end{aligned}$$

Consider now Case B. By Proposition 2, we have $H_u(x) < +\infty$, $\kappa_u > 0$, and $\log\kappa_u + \eta \geq 0$. We let P_{H_u} denote the probability distribution of ξ such that $H(x,\xi)$ is concentrated on the single point $H_u(x)$, and $\tilde{L}$ denote the corresponding likelihood ratio. Note that $\tilde{L}$ is well defined since $\kappa_u > 0$. We now show that $(\tilde{\alpha}, \tilde{L})$ is still a saddle point. The first inequality in (15) is straightforward. We only need to verify the second one. It suffices to show $\mathrm{E}_{P_0}\left[\tilde{L}\log\tilde{L}\right] - \eta \leq 0$. The result then follows from that $\mathrm{E}_{P_0}\left[\tilde{L}\log\tilde{L}\right] - \eta = \log(1/\kappa_u) - \eta \leq 0$. □

Theorems 1 and 2 are important results of this paper. They together show that, when the random function has a light right tail, the worst-case expectation admits an analytical expression. The light right tail we identify includes the bounded case and numerous other interesting cases in practical applications (perhaps a simplest example is the normal distribution; see Section 5). Such a property guarantees the tractability of KL divergence in modeling ambiguity.

2.2 Modeling Difficulty for Heavy Tail

The results shown in Theorems 1 and 2 require the assumption that the random function has a light right tail. We now investigate what happens if the random function has a heavy right tail. Suppose that S is empty for x. Then, $H_u(x) = +\infty$ and we can find a positive real sequence $\{R_j\}$ tending to $+\infty$ such that the sequence of probability masses of $H(x,\xi)\mathbb{1}_{\{H(x,\xi)\leq R_j\}}$ at corresponding essential supremums diminishes to 0. Let $\alpha_j^*(x)$ denote the optimal solution of Problem (14) where the function $H(x,\xi)$ is replaced with $H(x,\xi)\mathbb{1}_{\{H(x,\xi)\leq R_j\}}$. Then, from Proposition 2, $0 < \alpha_j^*(x) < +\infty$ starting from sufficiently large j. Construct the sequence

$$L_j = \frac{e^{H(x,\xi)\mathbb{1}_{\left\{H(x,\xi)\leq R_j\right\}}/\alpha_j^*(x)}}{\mathrm{E}_{P_0}\left[e^{H(x,\xi)\mathbb{1}_{\left\{H(x,\xi)\leq R_j\right\}}/\alpha_j^*(x)}\right]},\ j = 1, 2, \cdots.$$

Then, following the analysis in the proof of Theorem 2, we have $\{L_j\}$ is a sequence of feasible solutions of Problem (6). Furthermore, the sequence of objective values of $\{L_j\}$ tends to $+\infty$. This shows the optimal objective value of Problem (6) is positive infinite.

It is now clear that the light right tail of the random function is the sufficient and necessary condition for Problem (6) to have a finite optimal value. The result also shows, when the random function has a heavy tail distribution, the worst-case expectation is positive infinite no matter how small the ambiguity set is. In such a case, the DRO formulation becomes meaningless and can no longer be applied.

The difficulty of modeling ambiguous heavy tail distributions does not only exist for KL divergence. Other distance measures may suffer from the same difficulty. To see this, we consider a general distance measure D_M and the KL divergence D. For any two functions $B_1(y)$ and $B_2(y)$, we say $B_1(D_M) \leq B_2(D)$, if $B_1(D_M(P_1||P_2)) \leq B_2(D(P_1||P_2))$ holds for any distributions P_1 and P_2. Then, we have the following theorem whose proof is provided in the Appendix.

Theorem 3. *Suppose that there exists a nonnegative increasing function $B(y)$ on $\Re^+$ such that $B(y) > 0$ if $y > 0$ and $B(D_M) \leq D$. Then for any $\eta > 0$,*

$$\mathbb{P}_M := \{P \in \mathbb{D} : D_M(P||P_0) \leq \eta\} \supset \{P \in \mathbb{D} : D(P||P_0) \leq B(\eta)\}.$$

Furthermore, suppose that S is empty for x. Then, $\sup_{P\in\mathbb{P}_M} \mathrm{E}_P[H(x,\xi)] = +\infty$.

Theorem 3 shows that if we can find some function $B(y)$ for D_M, such that D_M can be bounded from above by the KL divergence together with $B(y)$, the worst-case expectation for the ambiguity set $\mathbb{P}_M$ is also infinite given that $H(x,\xi)$ is heavy tailed under P_0. This shows the distance measure D_M cannot be used in modeling ambiguous heavy tail distributions as well.

For many distance measures, it is easy to find the function $B(y)$. Gibbs and Su (2002) studied a number of distances of distributions. They showed that the Discrepancy, Hellinger distance, Kolmogorov (or Uniform) metric, Lévy metric, Prokhorov metric, and Total variation distance, when well defined on an underlying space, can all be bounded from above by KL divergence together with some functions. This means that we can find $B(y)$ for all these distances provided they are well defined on the considered distribution space. Take Hellinger distance D_H, Total variation distance D_{TV} and Prokhorov metric D_{PV} as examples. From Gibbs and Su (2002), we have $D_H^2 \leq D$, $2D_{TV}^2 \leq D$ and $2D_{PV}^2 \leq D$. Therefore, we can set $B(y) = y^2$ for Hellinger distance, $B(y) = 2y^2$ for Total variation distance, and $B(y) = 2y^2$ for Prokhorov metric on $y \geq 0$. Theorem 3 shows that, on the other hand, if we want to use some distance measure to model ambiguous heavy tailed distributions, we have to look for distance measures that cannot be bounded by KL divergence.

Nevertheless, heavy tailed distributions appear frequently in practical applications, especially in financial risk management. Therefore, it is an important question to investigate how to modify the

KL divergence constrained ambiguity set $\mathbb{P}$, maybe by incorporating some additional constraints, such that the new set is meaningful for heavy tailed distributions and, at the same time, keeps the tractability of the original set. Here we consider adding a perturbation constraint

$$L_l \leq L \leq L_u,$$

where L_l and L_u are some nonnegative functions of z and the inequalities hold for all $z \in \Xi$. The functional approach developed in Section 2.1 allows us to look into the specific structures of the problems. Therefore, it may be applicable to handle these more sophisticated ambiguity sets. Our preliminary study via using the functional approach shows a Monte Carlo approach may be necessary to estimate a worst-case performance in this case. The basic idea is that L is now restricted and cannot take values freely on $\Re^+$, and therefore we need to compare the values L_l, L_u and the value of $L(z)$ that enforces the gradient to 0. We will investigate such an extension in our future research.

2.3 Solving the Minimax Problem

From Theorem 2 we have the following theorem. The proof of the theorem is straightforward following the analysis above, and thus is omitted here.

Theorem 4. *Suppose Assumption 1 is satisfied. Then, Problem (3) is equivalent to*

$$\underset{x\in X,\ \alpha\geq 0}{\textit{minimize}}\ h(x,\alpha) := \alpha \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right] + \alpha\eta. \tag{16}$$

In Theorem 4, in order to emphasize (x, α) is the joint decision vector, we use $h(x, \alpha)$ rather than $h_x(\alpha)$ to denote the objective function. Suppose that $H(x, \xi)$ is convex in x for every ξ. The objective function $h(x, \alpha)$ is a convex function of (x, α). Indeed, the convexity follows from the fact that the functional $\ell_0(\alpha, L)$ is convex in (x, α), and $h(x, \alpha)$ is obtained by maximizing $\ell_0(\alpha, L)$ over $L \in \mathbb{L}$. Therefore, Problem (16) is a $d+1$-dimensional convex optimization problem. Note that the first term of $h(x, \alpha)$ is exactly the logarithmic moment generating function of $H(x, \xi)$ under the probability measure P_0. In some cases, the logarithmic moment generating function has a closed-form expression. Then, Problem (16) can be transformed to a deterministic convex optimization problem that can be solved by standard optimization algorithms; see examples in Section 5.

When the closed-form expression of the logarithmic moment generating function is not available, Problem (16) is a typical stochastic optimization problem with a fixed probability distribution P_0. We can then use standard stochastic optimization techniques, such as sample average approximation (SAA) and stochastic approximation (SA) to solve the problem (Shapiro et al. 2009). For instance, to apply the SAA, we first generate an independent and identically distributed (i.i.d.) sample $\xi_j, j = 1, \cdots, N$ from the distribution P_0, and then use the following optimization problem to

approximate (16):

$$\underset{x\in X,\ \alpha\geq 0}{\text{minimize}}\ \hat{h}_N(x,\alpha) := \alpha \log\left(\frac{1}{N}\sum_{j=1}^{N} e^{H(x,\xi_j)/\alpha}\right) + \alpha\eta. \tag{17}$$

By the strong law of large numbers, we have $\frac{1}{N}\sum_{j=1}^{N} e^{H(x,\xi_j)/\alpha}$ converges to $\mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right]$ with probability one (w.p.1) as N goes to infinity for every $x \in X$ and $\alpha > 0$. Then, by the continuous mapping theorem, $\alpha \log\left(\frac{1}{N}\sum_{j=1}^{N} e^{H(x,\xi_j)/\alpha}\right)$ converges to $\alpha \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right]$ w.p.1 as N goes to infinity for every $x \in X$ and $\alpha > 0$. Because $h(x,\alpha)$ is jointly convex in x and α, by Theorem 7.50 of Shapiro et al. (2009), we have that $\hat{h}_N(x,\alpha)$ converges to $h(x,\alpha)$ w.p.1 uniformly on $X \times \Re^+$. Therefore, the convergence of the optimal value and the set of optimal solutions of the SAA, i.e., Problem (17), to those of the true problem, i.e., Problem (16), can be guaranteed; see, e.g., Theorem 5.3 of Shapiro et al. (2009) and the followed discussions.

Before ending this section, we briefly discuss the structure of the probability distribution that achieves the worst-case performance. Suppose $\alpha^*(x) \neq 0$. Let $p^*(z,\alpha)$ denote the probability distribution that achieves the maximal value of $\ell(L,\lambda^*)$. Then,

$$p^*(z,\alpha) = p_0(z)L^*(z) = \frac{p_0(z)e^{H(x,z)/\alpha}}{\mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right]}.$$

It follows that the probability measure

$$p^*(z,\alpha^*(x)) = \frac{p_0(z)e^{H(x,z)/\alpha^*(x)}}{\mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha^*(x)}\right]} \tag{18}$$

is the optimal distribution that achieves the worst-case expectation in the inner maximization problem of Problem (3). This structure shows that the optimal distribution is proportional to the nominal distribution composite with the exponential term $e^{H(x,z)/\alpha^*(x)}$. When $p_0(z)$ is a density function, $p^*(z,\alpha^*(x))$ is also a density function, and it has the same support as $p_0(z)$. This is different from many results in the robust optimization literature, where optimal distributions are often atomic (i.e., they allocate positive probabilities on a finite set of values). For many parametric families of distributions, we find that the optimal distribution and the nominal one are in the same family. We discuss this further in Section 5.

3 Ambiguous Expectation Constrained Programs

The minimax DRO is a natural formulation when the ambiguous random parameters appear in the objective function of an optimization model. In many practical models, these parameters may appear in the constraints of the optimization models, like Problem (2). When a decision maker is risk-neutral to the randomness, he or she may only require the constraint be satisfied "averagely".

Then, we have the following formulation of an ECP:

$$\begin{array}{ll} \underset{x\in X}{\text{minimize}} & h(x) \\ \text{subject to} & \mathrm{E}_{P_0}\left[H(x,\xi)\right] \leq 0. \end{array} \tag{19}$$

In this section we consider a robust version of Problem (19), which requires the constraint be satisfied for all the distributions in the ambiguity set $\mathbb{P}$, where $\mathbb{P}$ is defined by (4). That is, we are interested in solving

$$\begin{array}{ll} \underset{x\in X}{\text{minimize}} & h(x) \\ \text{subject to} & \underset{P\in\mathbb{P}}{\text{maximize}}\ \mathrm{E}_{P}\left[H(x,\xi)\right] \leq 0. \end{array} \tag{20}$$

We call Problem (20) an ambiguous ECP. Following the functional approach developed in Section 2, we obtain the following theorem.

Theorem 5. *Suppose that Assumption 1 is satisfied. Then, Problem (20) is equivalent to*

$$\begin{array}{ll} \underset{x\in X}{\textit{minimize}} & h(x) \\ \textit{subject to} & \inf\limits_{\alpha\geq 0}\ \alpha\log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right] + \alpha\eta \leq 0. \end{array} \tag{21}$$

Theorem 5 shows that the ambiguous ECP can be simplified as a one-layer optimization problem, which is convex if $h(x)$ is convex in x and $H(x,\xi)$ is convex in x for every ξ. Therefore, it may be solved efficiently using standard optimization techniques.

3.1 Relation to Chance Constrained Programs

A different, often more natural, approach to modeling the randomness in the decision problem (2) is to require that the constraint be satisfied with at least a given probability. Such an approach leads to the following optimization problem:

$$\begin{array}{ll} \underset{x\in X}{\text{minimize}} & h(x) \\ \text{subject to} & \Pr_{\sim P_0}\left\{H(x,\xi)\leq 0\right\} \geq 1-\beta, \end{array} \tag{22}$$

where $1-\beta \in (0,1)$ is called the confidence level of the probability constraint. Problem (22) is often called a CCP; see, e.g., Charnes et al. (1958), Prékopa (2003), Nemirovski and Shapiro (2006), and Hong et al. (2011) for more details about CCPs. Compared to the ECP formulation, the CCP formulation is in general (but not necessarily) a more conservative approach, which may be advocated by decision makers who are risk-averse to the randomness in ξ. Because CCPs are generally nonconvex optimization problems and are often difficult to solve, a convex conservative approximation approach is often used to tackle them (see, e.g., Ben-Tal and Nemirovski 2000, Nemirovski and Shapiro 2006, and Chen et al. 2010).

The Bernstein approximation of Nemirovski and Shapiro (2006) is a famous example of such an approach. It takes the following form:

$$\begin{aligned} \underset{x\in X}{\text{minimize}} \quad & h(x) \\ \text{subject to} \quad & \inf_{\alpha>0} \left[\alpha \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right] - \alpha \log \beta\right] \leq 0. \end{aligned} \tag{23}$$

Nemirovski and Shapiro (2006) showed that Problem (23) is a convex conservative approximation of Problem (22). Using Jensen's inequality, we have

$$\alpha \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right] \geq \alpha \mathrm{E}_{P_0}\left[\log\left(e^{H(x,\xi)/\alpha}\right)\right] = \mathrm{E}_{P_0}\left[H(x,\xi)\right].$$

It follows that

$$\inf_{\alpha>0} \left[\alpha \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right] - \alpha \log \beta\right] \geq \mathrm{E}_{P_0}\left[H(x,\xi)\right] + \inf_{\alpha>0} \left\{-\alpha \log \beta\right\} = \mathrm{E}_{P_0}\left[H(x,\xi)\right].$$

Therefore, the Bernstein approximation, i.e., Problem (23), is also a convex conservative approximation of the ECP, i.e., Problem (19).

Comparing Problems (21) and (23) we have the following theorem that reveals the links between ambiguous ECPs and Bernstein approximations.

Theorem 6. *If $\eta = \log(\beta^{-1})$, or equivalently $\beta = e^{-\eta}$, Problems (21) and (23) are the same.*

Theorem 6, we think, is an interesting result. Note that the formulation of CCP reflects a decision maker's risk averseness and the formulation of ambiguous ECP reflects a decision maker's ambiguity averseness. Even though we often treat risk and ambiguity differently (see, for instance, Ellsberg (1961) and Epstein (1999)), Theorem 6 shows that they are interrelated via the KL divergence. By solving the Bernstein approximation, we obtain a solution that not only approximates the solution of the corresponding CCP, but is also optimal under an ambiguous ECP with an appropriately determined index of ambiguity; and vice versa.

Table 1: Relation between Confidence Level and Index of Ambiguity

confidence level β	index of ambiguity η	index of ambiguity η	confidence level β
0.1	2.3026	0.5	0.6065
0.05	2.9957	1	0.3679
0.01	4.6052	1.5	0.2231

Theorem 6 also provides valuable information on the selection of the index of ambiguity in DRO models. From Theorem 6 we immediately see that, the confidence level $\beta = 0.05$ corresponds to the index of ambiguity $\eta = \log(\beta^{-1}) \approx 3.0$, while the index of ambiguity $\eta = 0.5$ corresponds to the confidence level $\beta = e^{-\eta} \approx 0.6$. Some more correspondences between the confidence level and the index of ambiguity are shown in Table 1, to help obtain a sense of their relationships.

4 Distributionally Robust Probabilistic Programs

In Sections 2 and 3 we focus mainly on performance measures that are defined as expectations. In many situations, however, decision makers who are risk-averse to randomness may prefer using probabilities as performance measures. Then, they may consider a probabilistic program. Probabilistic programming is an important area within stochastic programming and it has been studied extensively in the literature; see Prékopa (2003) for a comprehensive review. Depending on whether the probability function appears in the objective or in the constraint, they can be roughly classified into the problems of optimizing a probability function and the CCPs that are discussed in Section 3.1. When a decision maker is both risk averse and ambiguity averse, he or she may want to formulate a probabilistic program into a distributionally robust probabilistic program, which we study in this section.

4.1 Minimax Probability Optimization

Consider the following problem of minimizing a probability performance measure,

$$\underset{x\in X}{\text{minimize}}\ \ \Pr_{\sim P_0}\left\{H(x,\xi)>0\right\}. \tag{24}$$

This model has 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, decision makers often target to maximize the probability of attaining aspiration levels; see, e.g., Bordley and Pollock (2009) and Chen and Sim (2009). In this subsection we are interested in finding how this model is affected by the ambiguity in the distribution of ξ. Suppose that the ambiguity set $\mathbb{P}$ is defined by (4). We then have the following formulation of the minimax DRO for Problem (24):

$$\underset{x\in X}{\text{minimize}}\ \ \underset{P\in\mathbb{P}}{\text{maximize}}\ \ \Pr_{\sim P}\left\{H(x,\xi)>0\right\}, \tag{25}$$

which can also be written as

$$\underset{x\in X}{\text{minimize}}\ \ \underset{P\in\mathbb{P}}{\text{maximize}}\ \ \mathrm{E}_P\left[\mathbb{1}_{\{H(x,\xi)>0\}}\right], \tag{26}$$

where $\mathbb{1}_{\{A\}}$ is the indicator function. Therefore, Problem (26) may be considered as a special instance of the minimax DRO model (3). Let v denote the optimal objective value of Problem (24). Then, based on Theorem 4, we have the following theorem.

Theorem 7. (a) *Any optimal solution of Problem (24) is also an optimal solution of the outer minimization problem of Problem (25).*

(b) *If* $\log v+\eta<0$, *any optimal solution of the outer minimization problem of Problem (25) is also an optimal solution of Problem (24); if* $\log v+\eta\geq 0$, *the objective value of the inner*

maximization problem of Problem (25) equals 1 *for all* $x \in X$*, and all* $x \in X$ *are optimal solutions of the outer minimization problem of Problem (25).*

Proof. For simplicity of notation, we let $\kappa(x) = \Pr_{\sim P_0}\{H(x,\xi) > 0\}$. Note that $\mathbb{1}_{\{H(x,\xi)>0\}}$ only takes two values $0, 1$. Therefore Assumption 1 is satisfied for $\mathbb{1}_{\{H(x,\xi)>0\}}$. Using Theorem 4 by setting $H(x,\xi)$ in Theorem 4 as $\mathbb{1}_{\{H(x,\xi)>0\}}$, we obtain that

$$\begin{aligned}
\inf_{x\in X}\sup_{P\in\mathbb{P}} \mathrm{E}_P\left[\mathbb{1}_{\{H(x,\xi)>0\}}\right] &= \inf_{x\in X}\inf_{\alpha\geq 0}\ \alpha\log \mathrm{E}_{P_0}\left[e^{\mathbb{1}_{\{H(x,\xi)>0\}}/\alpha}\right] + \alpha\eta \qquad (27)\\
&= \inf_{x\in X}\inf_{\alpha\geq 0}\ \alpha\log\left[\kappa(x)e^{1/\alpha} + (1-\kappa(x))\right] + \alpha\eta\\
&= \inf_{x\in X}\inf_{\alpha\geq 0}\ \alpha\log\left[\kappa(x)\left(e^{1/\alpha}-1\right)+1\right] + \alpha\eta\\
&= \inf_{\alpha\geq 0}\inf_{x\in X}\ \alpha\log\left[\kappa(x)\left(e^{1/\alpha}-1\right)+1\right] + \alpha\eta. \qquad (28)
\end{aligned}$$

Because $e^{1/\alpha} - 1 > 0$ for all $\alpha \geq 0$ and $\log(\cdot)$ is a strictly increasing function, we have if $\bar{x}$ is an optimal solution of Problem (24), it attains the inner infimum in (28) and thus is an optimal solution of the outer minimization problem of Problem (25). Therefore (a) holds.

We next show (b). Consider first the case that $\log v + \eta < 0$. Suppose that $\bar{x}$ is an optimal solution of Problem (24). Then $v = \kappa(\bar{x})$. For $x = \bar{x}$, from Proposition 2, the inner infimum of (27) is attained at $\bar{\alpha} > 0$ and the objective value of the inner maximization problem of Problem (25) is less than 1. Consider any optimal solution $\hat{x}$ of the outer minimization problem of Problem (25). If $\log\kappa(\hat{x}) + \eta \geq 0$, then for $x = \hat{x}$, the inner infimum of (27) is attained at $\hat{\alpha} = 0$ and the objective value of the inner maximization problem of Problem (25) equals 1. This contradicts with the optimality of $\hat{x}$. Therefore we have $\log\kappa(\hat{x}) + \eta < 0$. Similarly, for $x = \hat{x}$, Proposition 2 implies that the inner infimum of (27) is attained at $\hat{\alpha} > 0$. Suppose $\hat{x}$ is not an optimal solution of Problem (24). Then $\kappa(\hat{x}) > \kappa(\bar{x})$. It follows that

$$\begin{aligned}
\inf_{x\in X}\sup_{P\in\mathbb{P}} \mathrm{E}_P\left[\mathbb{1}_{\{H(x,\xi)>0\}}\right] &= \hat{\alpha}\log\left[\kappa(\hat{x})\left(e^{1/\hat{\alpha}}-1\right)+1\right] + \hat{\alpha}\eta\\
&> \hat{\alpha}\log\left[\kappa(\bar{x})\left(e^{1/\hat{\alpha}}-1\right)+1\right] + \hat{\alpha}\eta\\
&\geq \inf_{\alpha\geq 0}\ \alpha\log\left[\kappa(\bar{x})\left(e^{1/\alpha}-1\right)+1\right] + \alpha\eta\\
&\geq \inf_{x\in X}\inf_{\alpha\geq 0}\ \alpha\log\left[\kappa(x)\left(e^{1/\alpha}-1\right)+1\right] + \alpha\eta\\
&= \inf_{x\in X}\sup_{P\in\mathbb{P}} \mathrm{E}_P\left[\mathbb{1}_{\{H(x,\xi)>0\}}\right].
\end{aligned}$$

This is a contradiction. Therefore, $\hat{x}$ is an optimal solution of Problem (24).

Consider now the case that $\log v + \eta \geq 0$. Since $v \leq \kappa(x)$ for all $x \in X$, we have $\log\kappa(x) + \eta \geq 0$ for all $x \in X$. From Proposition 2, for any $x \in X$, the inner infimum of (27) is attained at $\alpha = 0$ and the objective value of the inner maximization problem of Problem (25) equals 1. Therefore all $x \in X$ solve Problem (25). This concludes the proof of the theorem. □

Theorem 7 shows that when the ambiguity set is defined by the KL divergence, a solution that optimizes the original probability function simultaneously optimizes the worst-case probability function, no matter what value the index of ambiguity η takes. Theorem 7 suggests that, to solve Problem (25), it suffices to solve Problem (24). In many practical situations, the optimal objective value v of Problem (24) is small (e.g., ≤ 0.05) and the index of ambiguity η is not very large (see also the discussions in Section 4.2). Thus the case that $\log v + \eta \geq 0$ is not very likely to happen and is often of no interest. In such situations, the original probability optimization problem and its DRO are actually the same problem. This result again suggests that risk and ambiguity are interrelated via the KL divergence. It seems that in the KL divergence-constrained distributionally robust probability optimization problems, 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.

4.2 Ambiguous Chance Constrained Programs

We next consider an ambiguous CCP that requires the chance (or probability) constraint be satisfied for all distributions in an ambiguity set. This problem has been considered in the literature. Erdogan and Iyengar (2006) 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 (Gibbs and Su 2002). 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 feasible region of the robust sampled problem is contained in the feasible region of the ambiguous CCP with a given probability. Besides proposing the Bernstein approximations, Nemirovski and Shapiro (2006) also considered ambiguous CCPs. They built Bernstein-type approximations to ambiguous CCPs where the ambiguity set is comprised of some product distributions. In this subsection, we study ambiguous CCPs where the ambiguity set is defined by the KL divergence. Suppose that the ambiguity set $\mathbb{P}$ is defined by (4). We then have the following formulation of an ambiguous CCP:

$$\begin{array}{ll} \underset{x \in X}{\text{minimize}} & h(x) \\ \text{subject to} & \Pr_{\sim P}\{H(x,\xi) \leq 0\} \geq 1-\beta, \quad \forall\ P \in \mathbb{P}. \end{array} \tag{29}$$

Similar to Problem (25), Problem (29) can be written as

$$\begin{array}{ll} \underset{x \in X}{\text{minimize}} & h(x) \\ \text{subject to} & \underset{P \in \mathbb{P}}{\text{maximize}}\ \mathrm{E}_P\left[\mathbb{1}_{\{H(x,\xi)>0\}}\right] \leq \beta. \end{array} \tag{30}$$

Therefore, Problem (29) may be considered as a special instance of ambiguous ECPs. Then, based on Theorem 5, we have the following theorem on the equivalent form of an ambiguous CCP.

Theorem 8. *Problem (29) 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

$$\bar{\beta}=\sup_{t>0}\frac{e^{-\eta}(t+1)^{\beta}-1}{t}. \tag{31}$$

Proof. Using Theorem 5 by setting $H(x,\xi)$ in Theorem 5 as $\mathbb{1}_{\{H(x,\xi)>0\}}$, and following the analysis in Theorem 7, we obtain constraint (30) is equivalent to

$$\inf_{\alpha\ge 0}\ \alpha\log\left[\kappa(x)\left(e^{1/\alpha}-1\right)+1\right]+\alpha\eta\le\beta. \tag{32}$$

Let A denote the set defined by (32). We now show that A is equal to a set B which is defined by the following constraint

$$\exists\ \alpha>0,\ \alpha\log\left[\kappa(x)\left(e^{1/\alpha}-1\right)+1\right]+\alpha\eta\le\beta. \tag{33}$$

It is obvious that $B\subset A$. Thus it suffices to show $A\subset B$. Consider any $x\in A$. If $\kappa(x)=0$, then x also satisfies (33) by setting, e.g., $\alpha=\beta/(2\eta)$. Suppose $\kappa(x)>0$. Note that the left hand side of (32) tends to 1 as $\alpha\to 0$, and $+\infty$ as $\alpha\to+\infty$. Therefore, the infimum in (32) cannot be attained at $\alpha=0,+\infty$ and has to be attained at a positive and finite α. This shows $x\in B$. Therefore $A=B$.

Elementary algebra shows that constraint (33) can be simplified as

$$\exists\ \alpha>0,\ \kappa(x)\le\frac{e^{\frac{\beta}{\alpha}-\eta}-1}{e^{\frac{1}{\alpha}}-1}.$$

which can further be transformed as the following constraint via a one-to-one transformation $t=e^{\frac{1}{\alpha}}-1$:

$$\exists\ t>0,\ \kappa(x)\le\frac{e^{-\eta}(t+1)^{\beta}-1}{t}. \tag{34}$$

Because $\left(e^{-\eta}(t+1)^{\beta}-1\right)/t$ tends to $-\infty$ as $t\to 0$ and tends to 0 as $t\to+\infty$, and it is strictly larger than 0 when $t>e^{\eta/\beta}-1$, it attains its maximum over $t>0$ at some positive and finite t. Therefore, constraint (34) can be strengthened as $\kappa(x)\le\bar{\beta}$ where $\bar{\beta}$ is defined by (31). This concludes the proof of the theorem. □

Remark: *A similar result as Theorem 8 for ambiguous CCPs was also derived by Jiang and Guan (2012) using a different approach.*

Theorem 8 shows that the ambiguous CCP can be equivalently formulated as the original CCP with only the confidence level being adjusted. This suggests that it can be solved by using standard

CCP algorithms. Furthermore, note that $\left[(t+1)^{\beta}-1\right]/t \leq \beta$. Thus, $\bar{\beta} \leq \beta$. 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 Theorems 6 and 7, again, we see that risk and ambiguity are interrelated via the KL divergence. Theorem 8 shows that, in the KL divergence-constrained ambiguous CCP, the ambiguity averseness is equivalent to an increase of risk averseness in the original CCP.

To determine the new confidence level $\bar{\beta}$, we need to solve a one dimensional optimization problem. The problem 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.

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

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

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

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

If $\Phi(t^*(\tilde{\beta})) = 0$, stop.

We compute the adjusted 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 2.

In Erdogan and Iyengar (2006), the index of ambiguity η cannot be larger than the confidence level β of the original CCP. In our formulation, we do not have this restriction. For any $\eta > 0$, the adjusted confidence level $\bar{\beta}$ is larger than 0. However, from Table 2, it is clear that $\bar{\beta}$ may be very small (leading to extreme conservativeness) if η is significantly larger than β.

Table 2: Relation between Rescaled Confidence Level and Index of Ambiguity

	index of ambiguity η	rescaled 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

4.3 Distributionally Robust Optimization for Other Performance Measures

The results derived for DRO problems in preceding sections may be extended to other performance measures. Here we discuss two important risk measures, value-at-risk (VaR) and conditional value-at-risk (CVaR), which are widely used in financial risk management. We briefly show how to derive the DRO reformulations of VaR and CVaR related stochastic programs. Consider the following DRO formulation of a VaR optimization problem:

$$\underset{x\in X}{\text{minimize}} \ \underset{P\in\mathbb{P}}{\text{maximize}} \ \text{VaR}_{1-\beta,P}(H(x,\xi)) \tag{35}$$

where the subscript P denotes the distribution of ξ and $\mathbb{P}$ is the KL divergence constrained ambiguity set defined in (4). Problem (35) suggests to minimize the worst-case VaR. We then have the following proposition.

Proposition 3. *Problem (35) is equivalent to*

$$\underset{x\in X}{\textit{minimize}} \ \ \textit{VaR}_{1-\bar{\beta},P_0}(H(x,\xi)), \tag{36}$$

where $\bar{\beta}$ is defined by (31).

Proof. From the definition of VaR (e.g., Trindade et al. 2007), it is not difficult to verify Problem (35) can be rewritten as

$$\begin{array}{ll} \underset{x\in X, t\in\Re}{\text{minimize}} & t \\ \text{subject to} & \Pr_{\sim P}\{H(x,\xi) - t \le 0\} \ge 1-\beta, \quad \forall\ P \in \mathbb{P}. \end{array} \tag{37}$$

Using Theorem 8, Problem (37) can be transformed as

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

which is equivalent to Problem (36) from the definition of VaR. □

Consider next the following distributionally robust VaR constrained program:

$$\begin{array}{ll} \underset{x\in X}{\text{minimize}} & h(x) \\ \text{subject to} & \text{VaR}_{1-\beta,P}(H(x,\xi)) \le 0, \quad \forall\ P \in \mathbb{P}. \end{array} \tag{38}$$

Problem (38) requires the worst-case VaR satisfy the non-positive constraint. We have the following proposition which can be proven following the argument in Proposition 3.

Proposition 4. *Problem (38) is equivalent to*

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

where $\bar{\beta}$ is defined by (31).

Propositions 3 and 4 show that DRO formulations of VaR optimization problems can be converted to VaR optimization problems of different confidence levels. Therefore, we can implement standard VaR optimization algorithms to solve these DROs.

We further consider a distributionally robust CVaR constrained program, which may be formulated as

$$\begin{array}{ll} \underset{x\in X}{\text{minimize}} & h(x) \\ \text{subject to} & \text{CVaR}_{1-\beta,P}(H(x,\xi)) \le 0, \quad \forall\ P \in \mathbb{P}. \end{array} \tag{39}$$

Problem (39) requires the worst-case CVaR satisfy the non-positive constraint. We have the following proposition.

Proposition 5. *Suppose that Assumption 1 is satisfied. Problem (39) is equivalent to*

$$\begin{array}{ll} \underset{x\in X, t\in\Re, \alpha\ge 0}{\textit{minimize}} & h(x) \\ \textit{subject to} & \alpha \log \text{E}_{P_0}\left[e^{[H(x,\xi)+t]^+/\alpha}\right] + \alpha\eta - \beta t \le 0. \end{array} \tag{40}$$

Proof. By the stochastic program representation of CVaR (Rockafellar and Uryasev 2000), it is clear that Problem (39) is equivalent to

$$\begin{array}{ll} \underset{x\in X, t\in\Re}{\text{minimize}} & h(x) \\ \text{subject to} & \text{E}_P\left[[H(x,\xi)+t]^+\right] - \beta t \le 0, \quad \forall\ P \in \mathbb{P}. \end{array}$$

Assumption 1 guarantees that for any t, $\text{E}_{P_0}\left[e^{s[H(x,\xi)+t]^+}\right]$ is finite for some $s > 0$. Applying Theorem 5, the optimization problem above is equivalent to Problem (40). □

Note that Problem (40) is a typical convex stochastic program that may be solved by applying a sample-average approximation (see, e.g., Shapiro et al. (2009)). Then, by Proposition 5, Problem (39) is also solvable by the same approach.

4.4 Conservative Approximation to Other Distance Measures

In Section 2.2 we showed that many distance measures can be bounded from above by the KL divergence. Therefore, the inability of the KL divergence in handling heavy tailed distributions also implies the inabilities of those distance measures. There are also many distance measures that can bound the KL divergence from above (Gibbs and Su 2002). For those distance measures, the KL divergence can serve as a conservative approximation. We summarize the result in the following theorem. Since the result is straightforward, we omit the proof.

Theorem 9. *Suppose there exists an increasing function $B(y)$ on $\Re^+$ such that $D \leq B(D_M)$. Then, for any $\eta > 0$,*

$$\mathbb{P}_M := \{P \in \mathbb{D} : D_M(P||P_0) \leq \eta\} \subset \{P \in \mathbb{D} : D(P||P_0) \leq B(\eta)\} := \mathbb{P}(B(\eta)).$$

Consequently, $\sup_{P\in\mathbb{P}_M} \mathrm{E}_P[H(x,\xi)] \leq \sup_{P\in\mathbb{P}(B(\eta))} \mathrm{E}_P[H(x,\xi)]$.

Suppose that D_M and $\mathbb{P}_M$ are used in defining the ambiguity set in an ambiguous ECP or ambiguous CCP, but the distance measure D_M may not have the mathematical tractability as the KL divergence. In such cases we may use the KL divergence to construct a new ambiguity set $\mathbb{P}(B(\eta))$. By Theorem 9, we know that the new ambiguous ECP or CCP is a tractable conservative approximation to the original problem.

To demonstrate how to specify the function $B(y)$, we consider two examples. The first is the χ^2-distance D_{χ^2}, which is used by Klabjan et al. (2012) to study multi-period inventory management problems. It follows from Gibbs and Su (2002) that $D \leq \log\left(1 + D_{\chi^2}\right)$. Then, we can set $B(y) = \log(1+y)$. The second example is the J-divergence D_J, which belongs to the ϕ-divergence class (Ben-Tal et al. 2012). Following the definition in Ben-Tal et al. (2012), we have $D_J = D + D_B$ where D_B denotes the Burg entropy. This implies $D \leq D_J$ and thus we can simply set $B(y) = y$.

5 Special Cases and Illustrations

In this section, we consider some special cases for the DRO formulations, and show that the DRO problems can often be transformed to simple optimization problems in these cases.

5.1 Affinely Perturbed Independent Case

Suppose that $H(x,\xi) = \sum_{i=1}^k h_i(x)\xi_i$, where $\xi_i, i = 1,\cdots,k$ are independent of each other, the functions $h_i(x), i = 1,\cdots,k$ are convex in X, and for those $i \geq 1$ such that the support of ξ_i is not a subset of $\Re^+$, the function $h_i(x)$ is affine. These conditions are also used by Nemirovski

and Shapiro (2006) to analyze their Bernstein approximation. Denote the logarithmic moment generating function of $\xi_i, i = 1, \cdots, k$ by

$$\Lambda_i(s_i) = \log \mathrm{E}_{P_{0i}} \left[e^{s_i \xi_i} \right], \ i = 1, \cdots, k,$$

where $\mathrm{E}_{P_{0i}}$ denotes that the expectation is taken with respect to the marginal distribution of ξ_i under the nominal probability distribution P_0. Then,

$$\alpha \log \mathrm{E}_{P_0} \left[e^{H(x,\xi)/\alpha} \right] = \alpha \sum_{i=1}^{k} \Lambda_i(\alpha^{-1} h_i(x)).$$

Consequently, the worst-case expectation can be expressed as

$$\inf_{\alpha \geq 0} \ \alpha \sum_{i=1}^{k} \Lambda_i(\alpha^{-1} h_i(x)) + \alpha\eta.$$

Note that, in the affinely perturbed independent case, $\mathrm{E}_P [H(x,\xi)] = (h_1(x), \cdots, h_k(x))^T \mu$, where $\mu = \mathrm{E}_P(\xi)$. This makes us wonder whether we can transform a KL divergence constrained ambiguity set to a mean constrained ambiguity set, which is a simple case of a moments constrained ambiguity set (see, for instance, Delage and Ye (2010)). However, we find it difficult to make such a transformation. This may indicate that the ambiguity defined by the moments and the ambiguity defined by the entire distribution may be quite different, revealing that the DRO approach with moment ambiguities and the DRO approach taking into consideration the whole distribution may be quite different modeling approaches.

5.2 Linear Case with Normal Nominal Distribution

Suppose that $H(x,\xi) = \xi^T x$, and that the nominal distribution of ξ is a multivariate normal distribution $N(\mu_0, \Sigma_0)$ with mean μ_0 and covariance matrix Σ_0. Then, $H(x,\xi)$ follows a normal distribution $N(\mu_0^T x, x^T \Sigma_0 x)$ under the probability measure $N(\mu_0, \Sigma_0)$, and

$$\begin{aligned} \inf_{\alpha \geq 0} \ \alpha \log \mathrm{E}_{P_0} \left[e^{H(x,\xi)/\alpha} \right] + \alpha\eta &= \inf_{\alpha \geq 0} \ \alpha \log \left(e^{\frac{\mu_0^T x}{\alpha} + \frac{x^T \Sigma_0 x}{2\alpha^2}} \right) + \alpha\eta \\ &= \mu_0^T x + \inf_{\alpha \geq 0} \ \frac{x^T \Sigma_0 x}{2\alpha} + \alpha\eta \\ &= \mu_0^T x + \sqrt{2\eta} \sqrt{x^T \Sigma_0 x}. \end{aligned} \tag{41}$$

This shows the minimax DRO problem and the ambiguous ECP can be transformed to second order cone programs that can be solved easily.

We now derive the optimal distribution that achieves the worst-case expectation in the non-degenerate case ($x \neq 0$). Note that $\alpha^*(x) = \sqrt{x^T \Sigma_0 x / 2\eta}$. It follows from (18) that the optimal distribution is

$$p^*(z, \alpha^*(x)) = \frac{(2\pi)^{\frac{k}{2}} e^{-\frac{1}{2}(z-\mu_0)^T \Sigma_0^{-1} (z-\mu_0) - \frac{1}{2} \log \det \Sigma_0} e^{z^T x / \sqrt{x^T \Sigma_0 x / 2\eta}}}{\mathrm{E}_{P_0} \left[e^{\xi^T x / \sqrt{x^T \Sigma_0 x / 2\eta}} \right]},$$

where $\det A$ denotes the determinant of a matrix A. Some simple algebra shows that

$$p^*(z, \alpha^*(x)) = (2\pi)^{\frac{k}{2}} e^{-\frac{1}{2}(z-\mu^*)^T \Sigma^{*-1}(z-\mu^*) - \frac{1}{2}\log\det\Sigma^*},$$

where

$$\mu^* = \mu_0 + \frac{\Sigma_0 x}{\sqrt{x^T \Sigma_0 x/(2\eta)}} \tag{42}$$

and $\Sigma^* = \Sigma_0$. Therefore, $p^*(z, \alpha^*(x))$ is the density of the multivariate normal distribution $N(\mu^*, \Sigma^*)$. This result shows that, in the linear case, if the nominal distribution is a multivariate normal distribution, the optimal distribution that achieves the worst-case expectation is still a multivariate normal distribution with the same covariance matrix. Only the mean vector is changed. In Appendix A.4 we show that the same optimal distribution can also be derived by restricting the candidate distributions to the family of multivariate normal distributions. Furthermore, it is worthwhile noting that this linear case may be generalized to $H(x, \xi) = \sum_{i=1}^k h_i(x)\xi_i$, where $h_i(x), i = 1, \cdots, k$ are affine functions of x.

5.3 Exponential Families

In Section 5.2 we show that, in the linear case with a multivariate normal nominal distribution, the worst-case distribution and the nominal distribution belong to the same distribution family. In this subsection we show that this property can be extended to exponential families of distributions. Exponential families include many useful families of distributions. Brown (1986) pointed out "many if not most of the successful mathematical formulations of statistical questions involve specific exponential families of distributions." It is well known that an exponential family associated with its sufficient statistics $\phi = (\phi_1, \phi_2, \cdots, \phi_{k'})^{\mathrm{T}}$ consists of the following parameterized collection of probability density (mass) functions

$$p(z, \theta) = e^{\phi(z)^{\mathrm{T}}\theta - A(\theta)},$$

taken with respect to some underlying measure $d\nu$ (Wainwright and Jordan 2008), where $d\nu$ is not necessarily the Lebesgue measure. The vector θ is often called the natural parameter, and the quantity A, known as the log partition function or cumulant function, is defined by the integral

$$A(\theta) = \log \int_Z \exp\{\phi(z)^T \theta\} \nu(dz),$$

where Z is the support that is independent of the natural parameter θ. A key structure of such a canonical form of exponential family is that the log partition function $A(\theta)$ is convex (Wainwright and Jordan 2008).

Suppose that the nominal distribution has a density (mass) function $p(z, \theta_0)$. Then, from (18), we have that the optimal distribution takes the following form

$$p^*(z, \alpha^*(x)) = e^{\phi(z)^{\mathrm{T}}\theta_0 - A(\theta_0) + H(x,z)/\alpha^*(x)},$$

with some new measure $d\tilde{\nu}$. If $\phi(z) = z$ and $H(x,z) = x^T z$, then $p^*(z, \alpha^*(x)) = e^{\phi(z)^{\mathrm{T}}\theta_1 - A(\theta_1 - x/\alpha^*(x))}$ where $\theta_1 = \theta_0 + x/\alpha^*(x)$. In such a case, the optimal distribution $p^*(z, \alpha^*(x))$ belongs to the same exponential family.

6 Conclusions

In this paper, we have studied various DRO problems with KL divergence constrained ambiguity sets. We have shown that the resulted minimax DRO problems, the ambiguous ECPs and the distributionally robust probabilistic programs are all quite tractable. We have also considered other optimization models and other distance measures in modeling ambiguity sets, and have also uncovered some interesting relations between different optimization models. Furthermore, it is worthwhile noting that the probability distributions considered in this paper are quite general. They can be either continuous, discrete, or mixed. The only condition is they have to have a light right tail.

In this paper we have discussed the advantages and disadvantages of using the KL divergence in modeling ambiguities in distributions, and have linked it to various other distance measures. In conclusion, we find that the properties of KL divergence make it a very special and tractable distance measure and a good candidate for modeling ambiguities.

A Appendix

A.1 Proof of Proposition 1

Proof. Because $H(x,\xi)L - \alpha L \log L + \lambda L$ is concave in L, we have for every ξ and $L(\xi)$,

$$\begin{aligned} H(x,\xi)L - \alpha L \log L + \lambda L &\leq H(x,\xi)L^*(\xi,\lambda) - \alpha L^*(\xi,\lambda)\log L^*(\xi,\lambda) + \lambda L^*(\xi,\lambda) \\ &= \alpha e^{(\lambda-\alpha)/\alpha} \cdot e^{H(x,\xi)/\alpha}. \end{aligned}$$

Because $1/\alpha \in S$, we have for every $L \in \mathbb{L}^0$,

$$\ell(L,\lambda) \leq \ell(L^*(\xi,\lambda),\lambda) < +\infty.$$

The inequality above further shows $L^*(\xi,\lambda) \in \arg\max_{L\in\mathbb{L}^0} \ell(L,\lambda)$. □

A.2 Proof of Proposition 2

Proof. We first show that $\alpha^*(x) < +\infty$. For every $\alpha > 0$, let $\beta = 1/\alpha$. Then

$$\begin{aligned} \lim_{\alpha\to+\infty} \alpha \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right] &= \lim_{\beta\to 0} \frac{\log \mathrm{E}_{P_0}\left[e^{H(x,\xi)\beta}\right]}{\beta} = \lim_{\beta\to 0} \nabla_\beta \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)\beta}\right] \\ &= \lim_{\beta\to 0} \frac{\mathrm{E}_{P_0}\left[e^{H(x,\xi)\beta} H(x,\xi)\right]}{\mathrm{E}_{P_0}\left[e^{H(x,\xi)\beta}\right]} = \mathrm{E}_{P_0}\left[H(x,\xi)\right] \end{aligned}$$

where the second equality follows from L'Hospital rule (Rudin 1976), the third equality follows from changing the order of ∇_β and E_{P_0}, which can be ensured by the Dominated Convergence Theorem, and the fourth equality follows from changing the order of $\lim_{\beta\to 0}$ and E_{P_0}, which is also ensured by the Dominated Convergence Theorem. Therefore, the objective function $h_x(\alpha)$ will tend to $+\infty$ as $\alpha \to +\infty$. Because S is not empty, there exists $\alpha > 0$ such that $h_x(\alpha)$ is finite. Thus $\alpha^*(x) < +\infty$ and we further have $\alpha^*(x) = 0$ or $1/\alpha^*(x) \in S$.

Recall that $H_u(x)$ is the essential supremum of $H(x,\xi)$ under measure P_0. We first show that

$$\lim_{\alpha\to 0} h_x(\alpha) = H_u(x). \tag{43}$$

If $H_u(x) < +\infty$, then $h_x(\alpha)$ is finite valued for all $\alpha > 0$ and $h_x(\alpha) \le H_u(x) + \alpha\eta$. Therefore $\lim_{\alpha\to 0} h_x(\alpha) \le H_u(x)$. By the definition of $H_u(x)$, for any given $M < H_u(x)$, we have

$$\kappa_M := \Pr_{\sim P_0}\{H(x,\xi) \ge M\} > 0.$$

It follows that

$$\lim_{\alpha\to 0} \alpha \log \mathrm{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right] \ge \lim_{\alpha\to 0} \alpha \log(\kappa_M e^{M/\alpha}) = M. \tag{44}$$

Therefore we have (43) holds. If $H_u(x) = +\infty$, for any given $M < H_u(x)$, we have $\kappa_M > 0$, and thus (44) holds. Therefore we also have (43) holds.

Now we prove the "only if" direction. Suppose $\alpha^*(x) = 0$. Because Assumption 1 is satisfied, from (43) we have $H_u(x) < +\infty$. We first show that $\kappa_u > 0$. Suppose not. By the definition of $H_u(x)$, we can find $H_l(x) < H_u(x)$ such that

$$0 < \kappa_l := \Pr_{\sim P_0}\{H_l(x) \le H(x,\xi) \le H_u(x)\} \le e^{-2\eta}.$$

Let $\varepsilon = H_u(x) - H_l(x)$ and $q = 1 - \kappa_l$. Then $\varepsilon > 0$, $0 < q < 1$ and

$$\begin{aligned} h_x(\alpha) &\le \alpha \log\left(q e^{H_l(x)/\alpha} + \kappa_l e^{H_u(x)/\alpha}\right) + \alpha\eta \\ &= H_u(x) + \alpha \log\left(q e^{-\varepsilon/\alpha} + \kappa_l\right) + \alpha\eta. \end{aligned}$$

Consider $\alpha \log\left(q e^{-\varepsilon/\alpha} + \kappa_l\right) + \alpha\eta$. We have $\lim_{\alpha\to 0} \alpha \log\left(q e^{-\varepsilon/\alpha} + \kappa_l\right) + \alpha\eta = 0$. Moreover, simple calculation shows that $\lim_{\alpha\to 0} \nabla_\alpha \left[\alpha \log\left(q e^{-\varepsilon/\alpha} + \kappa_l\right) + \alpha\eta\right] = \log\kappa_l + \eta \le -\eta < 0$. This shows that there exists $\bar\alpha > 0$ such that $\bar\alpha \log\left(q e^{-\varepsilon/\bar\alpha} + p\right) + \bar\alpha\eta < 0$. Thus $h_x(\bar\alpha) < H_u(x)$. This contradicts with that $\alpha^*(x) = 0$ is an optimal solution. Therefore $\kappa_u > 0$.

We then show $\log\kappa_u + \eta \ge 0$. Note that $h_x(\alpha)$ is differentiable at every $\alpha > 0$.

$$\begin{aligned} \nabla_\alpha h_x(\alpha) &= \nabla_\alpha \left[\alpha \log\left(\kappa_u + \mathrm{E}_{P_0}\left[e^{(H(x,\xi)-H_u(x))/\alpha}\mathbb{1}_{\{H(x,\xi)<H_u(x)\}}\right]\right) + \alpha\eta\right] \\ &= \log\left(\kappa_u + \mathrm{E}_{P_0}\left[e^{(H(x,\xi)-H_u(x))/\alpha}\mathbb{1}_{\{H(x,\xi)<H_u(x)\}}\right]\right) \\ &\quad + \frac{\mathrm{E}_{P_0}\left[e^{(H(x,\xi)-H_u(x))/\alpha}\mathbb{1}_{\{H(x,\xi)<H_u(x)\}}\left(H_u(x) - H(x,\xi)\right)/\alpha\right]}{\kappa_u + \mathrm{E}_{P_0}\left[e^{(H(x,\xi)-H_u(x))/\alpha}\mathbb{1}_{\{H(x,\xi)<H_u(x)\}}\right]} + \eta. \end{aligned}$$

By using Dominated Convergence Theorem, it can be verified that $\lim_{\alpha\to+\infty}\nabla_\alpha h_x(\alpha)=\eta$. Let $h_1(\alpha):=\mathrm{E}_{P_0}\left[e^{(H(x,\xi)-H_u(x))/\alpha}\mathbb{1}_{\{H(x,\xi)<H_u(x)\}}\right]$. Then $h_1(\alpha)\geq 0$. Moreover, using Dominated Convergence Theorem, we have $\lim_{\alpha\to 0}h_1(\alpha)=0$. Let

$$h_2(\alpha):=\frac{\mathrm{E}_{P_0}\left[e^{(H(x,\xi)-H_u(x))/\alpha}\mathbb{1}_{\{H(x,\xi)<H_u(x)\}}\left(H_u(x)-H(x,\xi)\right)/\alpha\right]}{\kappa_u+\mathrm{E}_{P_0}\left[e^{(H(x,\xi)-H_u(x))/\alpha}\mathbb{1}_{\{H(x,\xi)<H_u(x)\}}\right]}.$$

Similarly, we have $h_2(\alpha)\geq 0$ and $\lim_{\alpha\to 0}h_2(\alpha)=0$. It follows that $\lim_{\alpha\to 0}\nabla_\alpha h_x(\alpha)=\log\kappa_u+\eta$. If $\log\kappa_u+\eta<0$, then we can find $\bar{\alpha}>0$, such that $\nabla_\alpha h_x(\bar{\alpha})=0$. This again contradicts with that $\alpha^*(x)=0$ is an optimal solution. Thus we have proven the "only if" direction.

Finally we briefly verify the "if" direction. For any $\alpha>0$, $\nabla_\alpha h_x(\alpha)=\log\left(\kappa_u+h_1(\alpha)\right)+h_2(\alpha)+\eta$. If $\kappa_u=1$, then $h_1(\alpha)=h_2(\alpha)=0$ and $\nabla_\alpha h_x(\alpha)=\eta>0$ for all $\alpha>0$. If $0<\kappa_u<1$, then $h_1(\alpha)>0$ and $h_2(\alpha)>0$ for all $\alpha>0$. Because $\log\kappa_u+\eta\geq 0$, we have $\nabla_\alpha h_x(\alpha)>0$ for all $\alpha>0$. This shows when $H_u(x)<+\infty, \kappa_u>0$ and $\log\kappa_u+\eta\geq 0$, $h_x(\alpha)$ is differentiable and $\nabla_\alpha h_x(\alpha)>0$ for all $\alpha>0$. Note that $h_x(\alpha)$ is convex in α. We have $\alpha^*(x)=0$. This finishes the proof of the proposition. □

A.3 Proof of Theorem 3

Proof. Consider any $P\in\mathbb{D}$ satisfying $D(P||P_0)\leq B(\eta)$. Since $B(D_M)\leq D$ and $B(\cdot)$ is increasing, $D_M(P||P_0)\leq B^{-1}(D(P||P_0))\leq\eta$ where $B^{-1}(\cdot)$ is the inverse function of $B(\cdot)$. Therefore $P\in\mathbb{P}_M$. Suppose that S is empty for x. Since $B(\eta)>0$, following the analysis in Section 2.2, we have $\sup_{P\in\{P\in\mathbb{D}:D(P||P_0)\leq B(\eta)\}}\mathrm{E}_P[H(x,\xi)]=+\infty$. Therefore, $\sup_{P\in\mathbb{P}_M}\mathrm{E}_P[H(x,\xi)]=+\infty$. □

A.4 Alternative Formulation for Linear Normal Case

The KL divergence between two multivariate normal distributions can be expressed using their mean vectors and covariance matrices. Specifically, consider P_0 with distribution $N(\mu_0,\Sigma_0)$ and P with distribution $N(\mu,\Sigma)$. We have

$$D(P||P_0)=\frac{1}{2}\left(\mathrm{tr}\left(\Sigma_0^{-1}\Sigma\right)+(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)-\log\left(\frac{\det\Sigma}{\det\Sigma_0}\right)-k\right),$$

where $\mathrm{tr}A$ denotes the trace of a matrix A, and k is the dimension of ξ. Note that $\mathrm{E}_P[H(x,\xi)]=x^T\mu$, where μ is the expectation of ξ under distribution P. By restricting the candidate distributions to the family of multivariate normal distributions, the worst-case expectation in the DRO problem is equal to the optimal objective value of the following semi-definite optimization problem with the mean vector μ and the covariance matrix Σ being the decision variables.

$$\begin{array}{ll}\underset{\mu,\ \Sigma\succ 0}{\text{maximize}} & x^T\mu \\ \text{subject to} & \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_0^{-1}\Sigma\right)+(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)-\log\left(\frac{\det\Sigma}{\det\Sigma_0}\right)-k\right)\leq\eta.\end{array} \tag{45}$$

It is well known that the logarithm determinant function $\log\det\Sigma$ is a concave function of Σ (see, e.g., Hu et al. (2012)). Thus $-\log\det\Sigma$ is convex in Σ. Note that $\mathrm{tr}\left(\Sigma_0^{-1}\Sigma\right)$ is a linear function of Σ. Therefore, Problem (45) is a convex optimization problem of μ and Σ. Observe further that the objective function of Problem (45) does not include Σ. Therefore, we can take the infimum for the constraint function over $\Sigma \succ 0$ to eliminate the decision variable Σ. Let ∇_Σ denote the derivative of a function with respect to the matrix Σ. Note that (Hu et al. 2012)

$$\nabla_\Sigma \left[\mathrm{tr}\left(\Sigma_0^{-1}\Sigma\right) - \log\det\Sigma\right] = \Sigma_0^{-1} - \Sigma^{-1}.$$

Therefore, the infimum of the constraint function in Problem (45) is attained at $\Sigma = \Sigma_0$. Plugging $\Sigma = \Sigma_0$ into Problem (45) and noting that $\mathrm{tr}\left(\Sigma_0^{-1}\Sigma_0\right) = k$, we obtain that Problem (45) is equivalent to the following optimization problem

$$\begin{array}{ll} \text{maximize} & x^T\mu \\ \text{subject to} & \dfrac{1}{2}\left(\mu-\mu_0\right)^T\Sigma_0^{-1}(\mu-\mu_0) \leq \eta. \end{array} \tag{46}$$

Problem (46) is a convex quadratic program of μ. Using the Lagrangian duality, we can solve Problem (46) analytically. We find that the optimal solution μ^* of Problem (46) is exactly given by (42). Furthermore, this solution yields the optimal objective value $\mu_0^T x + \sqrt{2\eta}\sqrt{x^T\Sigma_0 x}$. Therefore, we again obtain the second order cone representation (41). Meanwhile, we obtain the worst-case normal distribution is $N(\mu^*, \Sigma^*)$ where μ^* is given by (42) and $\Sigma^* = \Sigma_0$. This is the same as what has been derived using the functional approach in Section 5.2.

References

Ben-Tal, A., A. Nemirovski. 1998. Robust convex optimization. *Mathematics of Operations Research*, **23** 769-805.

Ben-Tal, A., A. Nemirovski. 2000. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, **88** 411-424.

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. 2012. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, forthcoming.

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

Bertsimas, D., D. B. Brown, C. Caramanis. 2011. Theory and applications of robust optimization. *SIAM Review*, **53**(3) 464-501.

Bertsimas, D., M. Sim. 2004. Price of robustness. *Operations Research*, **52**(1) 35-53.

Bonnans, J. F., A. Shapiro. 2000. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research, Springer-Verlag, New York.

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

Brown, L. D. 1986. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. Inst. of Math. Statist., Hayward, California.

Calafiore, G. C. 2007. Ambiguous risk measures and optimal robust portfolios. *SIAM Journal on Optimization*, **18** 853-877.

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.

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

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

El Ghaoui, L., F. Oustry, H. Lebret. 1998. Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization*, **9**(1) 33-52.

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.

Hansen, L. P., T. J. Sargent. 2008. *Robustness*. Princeton University Press.

Hong, L. J., Y. Yang, L. Zhang. 2011. Sequential convex approximations to joint chance constrained programs: A Monte Carlo approach. *Operations Research*, **59**(3) 617-630.

Homem-de-Mello, T. 2007. A study on the cross-entropy method for rare-event probability estimation. *INFORMS Journal on Computing*, **19**(3) 381-394.

Hu, J., M. C. Fu and S. I. Marcus. 2007. A model reference method for global optimization. *Operations Research*, **55**(3) 549-568.

Hu, Z., J. Cao, L. J. Hong. 2012. Robust simulation of global warming policies using the DICE

model. *Management Science*, forthcoming.

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

Klabjan, D., D. Simchi-Levi, M. Song. 2012. Robust stochastic lot-sizing by means of histograms. *Production and Operations Management*, forthcoming.

Kullback, S., R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, **22**(1) 79-86.

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

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

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

Rockafellar, R. T. 1970. *Convex Analysis*. Princeton University Press, Princeton, NJ.

Rockafellar, R. T., S. Uryasev. 2000. Optimization of conditional value-at-risk. *The Journal of Risk*, **2** 21-41.

Rubinstein, R. Y. 2002. Cross-entropy and rare events for maximal cut and partition problems. *ACM Transactions on Modeling and Computer Simulation*, **12**(1) 27-53.

Rudin, W. 1976. *Principles of Mathematical Analysis*, Third Edition. McGraw-Hill.

Shapiro, A., D. Dentcheva, A. Ruszczyński. 2009. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelphia.

Trindade, A. A., S. Uryasev, A. Shapiro, G. Zrazhevsky. 2007. Financial prediction with constrained tail risk. *Journal of Banking and Finance*, **31** 3524-3538.

Wainwright, M. J., M. I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, **1** 1-305.