Phi-Divergence Constrained Ambiguous Stochastic Programs for Data-Driven Optimization

David K. Love
American Express, New York, NY, love.david.k@gmail.com

Güzin Bayraksan
Department of Integrated Systems Engineering, The Ohio State University, bayraksan.1@osu.edu

This paper investigates the use of ϕ-divergences in ambiguous (or distributionally robust) two-stage stochastic programs. Classical stochastic programming assumes the distribution of uncertain parameters are known. However, the true distribution is unknown in many applications. Especially in cases where there is little data or not much trust in the data, an ambiguity set of distributions can be used to hedge against the distributional uncertainty. ϕ-divergences (e.g., Kullback-Leibler divergence, χ^2 distance, etc.) provide a natural way to create an ambiguity set of distributions that are centered around a nominal distribution. The nominal distribution can be obtained by using observed data, expert opinions, simulations, and so forth. In this paper, we present a classification of ϕ-divergences to elucidate their use for models with different properties and sources of data. We illustrate our classification on ϕ-divergences that result in common risk optimization models. A condition for assessing the value of collecting additional data is derived, and we demonstrate that the ϕ-divergence-based ambiguous program behaves essentially the same as the associated non-ambiguous stochastic program as more data is collected. We present a decomposition-based solution algorithm to solve the resulting model. Finally, we demonstrate the behavior of ϕ-divergences in an optimization setting for a numerical example.

Key words: Ambiguous stochastic programming, distributionally robust optimization, phi-divergences, data-driven optimization

1. Introduction

Many optimization problems can be modeled by stochastic programs minimizing the expected value of an uncertain objective function. However, if the distribution of the uncertain parameters used in the model is incorrect, the stochastic program can give highly suboptimal results. This concern has led to the development of a modeling technique that replaces the probability distribution by a set of distributions. Then, the model optimizes the worst-case expectation with respect to the distributions in this set to hedge against the distributional uncertainty. This set of distributions is referred to as the *ambiguity set* of distributions, sometimes called the *uncertainty set*. This approach is not new, with early results dating back to Scarf (1958) and Dupačová as Žáčková (1966). This type of model has been referred to as an *ambiguous stochastic program*; see, for instance, Pflug and Wozabal (2007) and Erdoğan and Iyengar (2006). More recently, this approach has been called *distributionally robust optimization* (Delage and Ye 2010, Goh and Sim 2010, Mehrotra and Papp 2014, Hanasusanto et al. 2015).

There are different ways to form the ambiguity set of distributions. One recent approach that has been proposed by Ben-Tal et al. (2013) uses a set of distributions that are sufficiently close to a given "nominal" distribution according to a ϕ-divergence. ϕ-divergences quantify distances between probability distributions; we will shortly review them in Section 2. Of particular interest is the case when the nominal distribution takes the form of the empirical distribution determined by direct observation of data. However, this is not the only means of obtaining data. In addition to direct observation, data can come from simulations, forecasts, or expert opinions.

In this paper[1] we adapt the ϕ-divergence based ambiguity sets to two-stage stochastic linear programs with recourse. We call the resulting model two-stage ϕ-divergence constrained ambiguous stochastic linear program with recourse and denote it as ϕLP-2. While we focus on ϕLP-2, some of our results apply to a broad class of distributionally robust optimization problems using ϕ-divergences—we will point to these shortly. Data-driven models in the literature typically use empirical probabilities obtained through direct observations. In our modeling framework, we also allow unobserved data points (e.g., those given by expert opinions) to be represented in the model with zero nominal probabilities. We examine this case in more detail in the paper.

Stochastic programs with uncertain objective functions have long been studied by applying the minimax approach to an expected cost; see, e.g., Žáčková (1966) and Dupačová (1987). Two seminal papers by Shapiro and Kleywegt (2002) and Shapiro and Ahmed (2004) developed methods for converting stochastic minimax problems into equivalent stochastic programs with a certain distribution, laying the foundation for a commonly used reformulation technique.

In recent years, there has been a growing interest in distributionally robust methods. One common method for forming the ambiguity set is moment based, where all distributions that have the same moments (mean, variance, covariance, etc.) are admitted into the set. An early example comes from Scarf (1958), who provided a distributionally robust model for the newsvendor problem. More recent works using moment-based ambiguity sets include Delage and Ye (2010) and Wiesemann et al. (2014). Probability metrics, including the Kantorovich or Wasserstein metric (Pflug and Wozabal 2007, Esfahani and Kuhn 2015), Prokhorov metric (Erdoğan and Iyengar 2006), and ζ-structure metrics (Zhao and Guan 2015), have also been used. Hanasusanto et al. (2015) provide a comprehensive review of different types of ambiguity sets. We refer the readers to this paper and references therein for more details on different types of ambiguity sets.

As mentioned above, Ben-Tal et al. (2013) first systematically studied the ϕ-divergence based models and their computational tractability. Jiang and Guan (2015a) investigated ϕ-divergence based ambiguous chance-constrained programs, providing an exact approach to solve them; see also

Yanikoglu and den Hertog (2012). Specific ϕ-divergences—including the χ^2-distance (Klabjan et al. 2013), Kullback-Leibler divergence (Calafiore 2007, Hu and Hong 2013, Wang et al. 2015) and the variation distance (Jiang and Guan 2015b)—were also studied. Hu and Hong (2013) and Jiang and Guan (2015b) differ from this work and the others by considering continuous distributions. Close to our work, Bertsimas et al. (2014) study robust problems with ambiguity sets formed via goodness-of-fit test statistics. Some of their results include ϕ-divergences, but they consider other tests as well. Our work unites these papers with various ϕ-divergences in the finite support case, providing insight into conditions where each ϕ-divergence should be used. To the best of our knowledge, this is the first paper examining the behavior of different ϕ-divergences in an optimization setting.

The contributions of this paper, along with a motivation to study the corresponding research questions, are as follows:

(i) Given that there are many ϕ-divergences, a decision maker is left with the question of how each divergence behaves for his/her problem and which one to choose.

- In this paper we provide a classification of ϕ-divergences that begins to answer this open question. Our classification is based on the types of distributions admitted into the ambiguity set. This analysis provides insight into which class of ϕ-divergence may be most useful to which type of data and decision maker.
- Our main classification is a general feature of ϕ-divergences, and it applies to a broader class of ϕ-divergence constrained distributionally robust problems than the ones presented in this paper.

(ii) In a data-driven setting, several important questions arise. What happens as we add one more data? Will our solution change, and if so, will the overall cost decrease? Can we determine sampling from which scenarios result in a better (lower-cost) solution? Can we characterize the behavior of the problem as we add more data? We provide answers to these questions.

- First, we provide a simple condition to determine if sampling from a particular scenario will lower the cost, which again can be generalized beyond the ϕLP-2 setting. We refer to this as the *value of additional data.*
- Next, in a data-driven setting—where random data is collected to form the ambiguity set of distributions using ϕ-divergences—we show that asymptotically, ϕLP-2 behaves essentially the same as a stochastic program with the (unknown) true distribution.

(iii) Stochastic programs often become quite large, which raises questions of computational tractability. We devise a modified Bender's decomposition that can be used to solve ϕLP-2 efficiently by solving only linear problems.

(iv) Finally, we present examples of ϕ-divergences that result in commonly used risk models and illustrate our classification on these models. We also numerically illustrate our results on a small electricity generation example.

The rest of the paper is organized as follows. Section 2 introduces ϕ-divergences and lists several useful properties that are used throughout the paper. Section 3 presents the derivation of ϕ-divergence constrained ambiguous two-stage stochastic programs with recourse and discusses their basic properties. Data-driven properties of ϕLP-2 are explored in Section 4. Section 5 presents a classification of ϕ-divergences and illustrates different classes using risk models. Then, Section 6 discusses a decomposition method for solving the ϕLP-2. Section 7 numerically illustrates the results of the paper, and finally Section 8 concludes with a summary and future work.

2. ϕ-Divergences

In this section we define the concept of a ϕ-divergence and review several properties of ϕ-divergences that will be used throughout the paper. Pardo (2005) provides a good overview of much of the known properties of ϕ-divergences. We refer the readers to this book for further details. Many results in this section can be also found in Ben-Tal et al. (1991, 2013).

In the finite case, ϕ-divergences are used to measure the distance between two non-negative vectors $\mathbf{p} = (p_1, \ldots, p_n)^T$ and $\mathbf{q} = (q_1, \ldots, q_n)^T$. Specifically, when $\mathbf{p}$ and $\mathbf{q}$ are probability vectors (i.e., satisfying $\sum_{\omega=1}^n p_\omega = \sum_{\omega=1}^n q_\omega = 1$), ϕ-divergences are used to quantify the distance between two discrete distributions with finite support. The ϕ-divergence is defined by

$$I_\phi(\mathbf{p}, \mathbf{q}) = \sum_{\omega=1}^{n} q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right),$$

where $\phi(t)$, called the *ϕ-divergence function*, is a convex function on $t \geq 0$ with $\phi(1) = 0$. Additionally, it is defined that $0\phi(a/0) = a \lim_{t\to\infty} \frac{\phi(t)}{t}$ for $a > 0$ and $0\phi(0/0) = 0$. When both $\mathbf{p}$ and $\mathbf{q}$ are probability vectors—which is our setup throughout this paper—we can further assume that $\phi(t) \geq 0$ without loss of generality. Observe that the function $\phi(t)$ can be modified as $\psi(t) = \phi(t) + c(t-1)$ with an appropriately chosen constant c such that $\psi(t) \geq 0$ for all $t \geq 0$ and $I_\psi(\mathbf{p}, \mathbf{q}) = I_\phi(\mathbf{p}, \mathbf{q})$ for all probability vectors $\mathbf{p}, \mathbf{q}$. If $\phi(t)$ is differentiable at $t = 1$, this modification can be done by selecting $c = -\phi'(1)$. Throughout the remainder of this paper, we assume $\phi(t) \geq 0$ for $t \geq 0$, although we give an example of a ϕ-divergence that does not satisfy this condition in Table 1 below.

We extend $\phi(t)$ to the set of reals by setting $\phi(t) = +\infty$ for $t < 0$. We make a technical—but not restrictive—assumption that ϕ is a closed function because we will use ϕ-divergences in an optimization setting. For a proper convex function, closedness (i.e., its epigraph being closed) is the same as lower semi-continuity (Rockafellar 1970). (Recall that ϕ is a proper function because for at least one t $(t = 1)$, $\phi(t) = 0 < \infty$, and $\phi(t) > -\infty$, for all $t \in \mathbb{R}$.) Lower semi-continuity is a desirable property in our setting, and it is satisfied by common ϕ-divergences (see Table 1).

ϕ-divergences are not, in general, metrics. Most ϕ-divergences do not satisfy the triangle inequality and many are not symmetric in the sense that $I_\phi(\mathbf{p}, \mathbf{q}) \neq I_\phi(\mathbf{q}, \mathbf{p})$. One exception is the variation distance, which is equivalent to the L^1-distance between the vectors (see Table 1). A ϕ-divergence has an *adjoint*, defined by

$$\tilde{\phi}(t) = t\phi\left(\frac{1}{t}\right), \tag{1}$$

which satisfies all criteria for a ϕ-divergence (Ben-Tal et al. 1991) and has the property that $I_{\tilde{\phi}}(\mathbf{p},\mathbf{q}) = I_\phi(\mathbf{q},\mathbf{p})$. Divergences that are symmetric with respect to the input vectors are known as *self-adjoint*.

An important function related to the ϕ-divergence function is its *convex conjugate*, which is used, for instance, in the dual problem formulation (Section 3.1). We will also use the properties of the conjugate for our classification (Section 5). The conjugate $\phi^* : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is defined as

$$\phi^*(s) = \sup_{t \geq 0}\{st - \phi(t)\}. \tag{2}$$

It is a nondecreasing convex function, which may be undefined above some upper bound $\bar{s}$. Because ϕ is a proper closed convex function, $\phi^{**} = \phi$ and $t \in \partial\phi^*(s)$ if and only if $s \in \partial\phi(t)$ (Rockafellar 1970, Corollary 23.5.1). We will use the latter property in our analysis.

Table 1 lists some common examples of ϕ-divergences, along with their adjoints and conjugates. The value of the conjugate is listed only in its domain, i.e., $\{s : \phi^*(s) < \infty\}$. Most of these ϕ-divergences are widely used in statistics and information theory. Because many ϕ-divergences are commonly used in statistics—e.g., to conduct goodness-of-fit tests (Pardo 2005)—they provide natural ways to deal with data and distributions. Consequently, using ϕ-divergences can be more data driven. For instance, many ϕ-divergences use more distributional information than moments. Another advantage is that they form convex ambiguity sets. This opens up the tools of convex analysis and allows computationally tractable models. Finally, they encompass a fairly large class of problems, including some important risk-averse optimization problems. In Section 5.4, we present ϕ-divergences that assign a distance of either 0 or ∞, which result in commonly used risk models.

Table 1 lists a divergence, labeled “Likelihood,” that is somewhat different from the others. The Likelihood divergence is equivalent to the Burg entropy when comparing probability vectors, but it does not satisfy the normalizing condition $\phi(t) \geq 0$. This divergence is included because Wang et al. (2015) use it to formulate a distributionally robust problem so that the ambiguity set of distributions have a sufficiently high empirical likelihood. They refer to this method as the Likelihood Robust Optimization. We also note that Calafiore (2007) and Hu and Hong (2013) use a different naming convention than the one given here, referring to the Burg entropy as the Kullback-Leibler (KL) divergence—reversing the order of the arguments $\mathbf{p}$ and $\mathbf{q}$ relative to the notation presented here. In this paper, $\mathbf{q}$ denotes the nominal distribution.

Table 1 Definitions of some common ϕ-divergences, along with their adjoints $\tilde{\phi}(t)$ and conjugates $\phi^*(s)$

Divergence	$\phi(t)$	$\tilde{\phi}(t)$	$\phi(t), t \geq 0$	$I_\phi(p,q)$	$\phi^*(s)$
Kullback-Leibler	ϕ_{kl}	ϕ_b	$t\log t - t + 1$	$\sum p_\omega \log\left(\frac{p_\omega}{q_\omega}\right)$	$e^s - 1$
Burg Entropy	ϕ_b	ϕ_{kl}	$-\log t + t - 1$	$\sum q_\omega \log\left(\frac{q_\omega}{p_\omega}\right)$	$-\log(1-s),\ s < 1$
J-Divergence	ϕ_j	ϕ_j	$(t-1)\log t$	$\sum (p_\omega - q_\omega)\log\left(\frac{p_\omega}{q_\omega}\right)$	No closed form
Likelihood	ϕ_l	$t\log t$	$-\log t$	$\sum q_\omega \log\left(\frac{q_\omega}{p_\omega}\right)$	$-\log(-s) - 1,\ s < 0$
χ^2-Distance	ϕ_{χ^2}	$\phi_{m\chi^2}$	$\frac{1}{t}(t-1)^2$	$\sum \frac{(p_\omega - q_\omega)^2}{p_\omega}$	$2 - 2\sqrt{1-s},\ s \leq 1$
Modified χ^2-Dist.	$\phi_{m\chi^2}$	ϕ_{χ^2}	$(t-1)^2$	$\sum \frac{(p_\omega - q_\omega)^2}{q_\omega}$	$\begin{cases} -1 & s < -2 \\ s + \frac{s^2}{4} & s \geq -2 \end{cases}$
Variation Distance	ϕ_v	ϕ_v	$\lvert t-1 \rvert$	$\sum \lvert p_\omega - q_\omega \rvert$	$\begin{cases} -1 & s \leq -1 \\ s & -1 \leq s \leq 1 \end{cases}$
Hellinger Distance	ϕ_h	ϕ_h	$(\sqrt{t} - 1)^2$	$\sum (\sqrt{p_\omega} - \sqrt{q_\omega})^2$	$\frac{s}{1-s},\ s < 1$

3. ϕ-Divergence Constrained Ambiguous Stochastic Program

3.1. Formulation

We begin with a two-stage stochastic linear program with recourse (SLP-2). Let $\mathbf{x}$ be a vector of first-stage decision variables with cost vector $\mathbf{c}$, constraint matrix $\mathbf{A}$, and right-hand-side vector $\mathbf{b}$. We assume the random parameters of the second-stage problem have a finite distribution with realizations indexed by $\omega = 1, \ldots, n$ and probabilities denoted by q_ω, $\omega = 1, \ldots, n$. We refer to realizations interchangeably as scenarios. The SLP-2 is given by

$$\min_{\mathbf{x}} \left\{ \mathbf{c}\mathbf{x} + \sum_{\omega=1}^{n} q_\omega h_\omega(\mathbf{x}) :\ \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0 \right\}, \tag{3}$$

where

$$h_\omega(\mathbf{x}) = \min_{\mathbf{y}^\omega} \ \{\mathbf{g}^\omega \mathbf{y}^\omega :\ \mathbf{D}^\omega \mathbf{y}^\omega = \mathbf{B}^\omega \mathbf{x} + \mathbf{d}^\omega, \mathbf{y}^\omega \geq 0\}, \quad \omega = 1, \ldots, n. \tag{4}$$

For a given scenario ω, the second-stage decision variables $\mathbf{y}^\omega$ are optimized with respect to cost vector $\mathbf{g}^\omega$. The second-stage constraints with recourse matrix $\mathbf{D}^\omega$ depend on the first-stage variables $\mathbf{x}$ through the technology matrix $\mathbf{B}^\omega$, which appear on the right-hand side of constraints along with $\mathbf{d}^\omega$. Throughout the rest of the paper, we denote the first-stage feasible region as $X = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$. The first stage of a prototypical SLP-2 allocates capacities to different electricity generators before demand and reliability of the generators are known. Then, once the electricity demand and generator reliabilities become known, the second stage provides electricity to demand sites in a least-costly fashion. We will consider this problem for our numerical results in Section 7.

The SLP-2 formulation assumes that $\mathbf{q}$ is known. However, in many applications, the distribution is itself unknown, and there is no reliable way to obtain the probabilities of scenarios ω. By replacing the specific distribution in SLP-2 with a set of distributions sufficiently close to the nominal distribution $\mathbf{q}$ with respect to a ϕ-divergence, we create the ϕLP-2 model. In the ϕLP-2, the objective function is minimized with respect to the worst-case distribution selected from the ambiguity set of distributions. The resulting minimax formulation of ϕLP-2 is

$$\min_{\mathbf{x}\in X} \max_{\mathbf{p}\in\mathcal{P}} \left\{ \mathbf{cx} + \sum_{\omega=1}^{n} p_\omega h_\omega(\mathbf{x}) \right\}, \tag{5}$$

where the ambiguity set is

$$\mathcal{P} = \left\{ \begin{aligned} &\sum_{\omega=1}^{n} q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right) \le \rho, && (6)\\ &\sum_{\omega=1}^{n} p_\omega = 1, && (7)\\ &p_\omega \ge 0, \ \forall\omega \end{aligned} \right\}. \qquad (8)$$

We refer to (6) as the ϕ-divergence constraint, and (7) and (8) simply ensure a probability measure. We will discus how to determine ρ in (6) shortly in Section 3.3.

Taking the Lagrangian dual of the inner maximization problem, with dual variables λ and μ, of constraints (6) and (7), respectively, and combining the two minimizations gives ϕLP-2 in dual form

$$\begin{aligned} \min_{\mathbf{x},\lambda,\mu}\ & \mathbf{cx} + \mu + \rho\lambda + \lambda \sum_{\omega=1}^{n} q_\omega \phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{\lambda}\right) && (9)\\ \text{s.t. } & \mathbf{x}\in X \\ & h_\omega(\mathbf{x}) - \mu \le \left(\lim_{t\to\infty} \frac{\phi(t)}{t}\right)\lambda, \ \forall\omega && (10)\\ & \lambda \ge 0. \end{aligned}$$

When $\lambda = 0$, the last term in the objective function (9) has the following interpretations: $0\phi^*(b/0) = 0$ if $b \le 0$ and $0\phi^*(b/0) = +\infty$ if $b > 0$. When $\rho > 0$, $\mathbf{q}$ strictly satisfies the ϕ-divergence constraint $I_\phi(\mathbf{q},\mathbf{q}) = 0 < \rho$. So, the Slater condition holds, and we have strong duality.

Some ϕ-divergences, like the J-divergence, have no closed-form representation of ϕ^*. However, they can be expressed as the sum of other ϕ-divergences with closed-form conjugates. For example, sum of Burg Entropy and KL divergence gives the J-divergence. In this case, the dual can be formed similarly; see Ben-Tal et al. (2013) for details. Theorem 1 of Ben-Tal et al. (2013) contains a derivation of the dual problem, which is reprinted as part of the proof of Proposition 4.

The right-hand side of (10) contains a limit. This constraint results from a dual feasibility consideration. When this limit is finite, i.e., $\lim_{t\to\infty} \frac{\phi(t)}{t} = \bar{s} < \infty$, then for any $s > \bar{s}$, $\phi^*(s) = \infty$.

Therefore, ϕ-divergences with a finite limit (like the variation distance) induce this constraint. In (10), we moved λ to the right-hand side to allow for $\lambda = 0$. On the other hand, when this limit is ∞ (like the KL divergence), $\phi^*(s) < \infty$ for any finite value of s. In this case, constraint (10) can be removed from the formulation. Observe, in particular, that the dual formulation is accurate even for $q_\omega = 0$ for some ω. We will discuss this case in more detail in Section 5 (e.g., proof of Proposition 4).

3.2. Basic Properties of ϕLP-2

In this section we list some basic properties of ϕLP-2. Some of these properties have already been noted earlier (e.g., by Ben-Tal et al. (2013), Ben-Tal and Teboulle (2007) and by others for specific ϕ-divergences), but we list them for completeness. These properties help with our specialized solution method and our classification of ϕ-divergences, and we refer to them in later sections. Throughout the rest of the paper, we use the notation

$$s_\omega = \frac{h_\omega(\mathbf{x}) - \mu}{\lambda}. \tag{11}$$

Furthermore, $(\mathbf{x}^*, \mathbf{p}^*)$ denotes the optimal primal solution, and $(\mathbf{x}^*, \lambda^*, \mu^*)$ denotes the optimal dual solution. Optimal s^*_ω can then be found by plugging in the respective optimal solutions in (11). We assume $X \neq \emptyset$ and compact, and second-stage problems (4) are primal and dual feasible for all ω and all $\mathbf{x} \in X$. This ensures that $h_\omega(\mathbf{x})$ are (finite) real-valued, and both problems SLP-2 and ϕLP-2 have finite optimal solutions. The basic properties of ϕLP-2 are as follows.

Property 1. ϕLP-2 is a convex program.

Property 2. ϕLP-2 is equivalent to minimizing a coherent risk measure.

Property 3. ϕLP-2 preserves the time structure of SLP-2.

Property 4. (Primal-Dual Relation.) The (optimal) worst-case probabilities p^*_ω can be calculated with the equations

$$\frac{p^*_\omega}{q_\omega} \in \partial\phi^*\left(s^*_\omega\right), \qquad \sum_{\omega=1}^{n} q_\omega \phi\left(\frac{p^*_\omega}{q_\omega}\right) = \rho, \qquad \sum_{\omega=1}^{n} p^*_\omega = 1 \tag{12}$$

when $\lambda^* > 0$ and $q_\omega > 0$. With $\lambda^* > 0$ and $q_\omega = 0$, $p^*_\omega \in \partial\phi^*\left(s^*_\omega\right) q_\omega$ (i.e., $p^*_\omega = 0$) when $s^*_\omega < \bar{s}$; otherwise (i.e., when $s^*_\omega = \bar{s} < \infty$) use the last two equations in (12). With $\lambda^* = 0$, set $p^*_\omega = 0$ when $h_\omega(\mathbf{x}^*) - \mu^* < 0$; otherwise (when $h_\omega(\mathbf{x}^*) - \mu^* = 0$), use the last two equations in (12), but with the regular ϕ-divergence constraint $I_\phi(\mathbf{p}, \mathbf{q}) \le \rho$ instead of $I_\phi(\mathbf{p}, \mathbf{q}) = \rho$.

A coherent risk measure—first proposed by Artzner et al. (1999), and later refined by several authors including Rockafellar (2007), Shapiro et al. (2009)—has desired properties including convexity and monotonicity. These properties, combined with the facts that $h_\omega(\mathbf{x})$ is convex over X and X is a convex set, implies Property 1. See also Proposition 5 in Section 6 and Ben-Tal et al. (2013). Observe that Property 2 is valid even when we have $q_\omega = 0$ for some ω. The case of $q_\omega = 0$ plays an important role in the classification presented in Section 5. Therefore, we provide a proof of Property 2 in this case in Appendix A.

Property 3 helps with our decomposition-based solution method described in Section 6. The preservation of time structure can be seen in (9). Rewriting it slightly, we obtain

$$\min_{\mathbf{x}\in X,\lambda\geq 0,\mu} \left\{ \mathbf{cx} + \mu + \rho\lambda + \sum_{\omega=1}^{n} q_\omega h_\omega^\dagger(\mathbf{x},\lambda,\mu) : \;\; s_\omega \leq \bar{s} \right\}. \tag{13}$$

The above formulation preserves the two-stage structure of the SLP-2. The first-stage variables can now be viewed as $\mathbf{x}, \lambda$, and μ. The expectation is taken using the nominal probability vector $\mathbf{q}$. Finally, $h_\omega^\dagger(\mathbf{x},\lambda,\mu) = \lambda\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{\lambda}\right)$, where $h_\omega(\mathbf{x})$ are defined as before.

Property 4 lists the first order necessary conditions for optimality. The appearance of the conjugate ϕ^* in the objective of (9) gives a method for retrieving the worst-case distribution from the dual problem. It uses the fact that $\frac{p_\omega^*}{q_\omega} \in \partial\phi^*(s_\omega^*)$ if and only if $s_\omega^* \in \partial\phi\left(\frac{p_\omega^*}{q_\omega}\right)$. In many cases, the first equation in (12) is sufficient to calculate p_ω^*. In addition, ϕ^* is often differentiable, and so we have the relationship $p_\omega^* = q_\omega \phi^{*\prime}(s_\omega)$. Observe that because ϕ is a proper closed convex function, so is ϕ^* (Rockafellar 1970, Theorem 12.2). Hence, ϕ^* is subdifferentiable on the relative interior of its domain (Rockafellar 1970, Theorem 23.4). The boundary of the domain of ϕ^*—that is, at $s = \bar{s}$ when $\bar{s} < \infty$—might require special care, where we need the primal feasibility conditions (6) and (7). The ϕ-divergence constraint in (12) is written as an equality because with $\lambda^* > 0$, the complementary slackness dictates that this constraint must be active. With $\lambda^* = 0$, we have the regular ϕ-divergence constraint (6). While Property 4 summarizes how to obtain p_ω^* for the cases when $\lambda^* = 0$ or $q_\omega = 0$, we will discuss these special cases in more detail in Section 5.

3.3. The Level of Robustness

The literature on ϕ-divergences provides some insight on choosing a reasonable asymptotic value of ρ in the data-driven setting. In this setting, $\mathbf{q}$ is generated from observations, where scenario ω has been observed N_ω times with $N = \sum_{\omega=1}^{n} N_\omega$ total observations. So, the nominal probability of scenario ω is set to be $q_\omega = \frac{N_\omega}{N}$.

When ϕ is twice continuously differentiable around 1 with $\phi''(1) > 0$, Pardo (2005, Theorem 3.1) shows that the statistic

$$T_N^\phi(\mathbf{q}^N, \mathbf{q}^{\text{true}}) = \frac{2N}{\phi''(1)} I_\phi(\mathbf{q}^N, \mathbf{q}^{\text{true}})$$

converges in distribution to a χ^2-distribution with $n-1$ degrees of freedom, where $\mathbf{q}^N$ denotes the empirical distribution ($q_\omega^N = N_\omega/N$) and $\mathbf{q}^{\text{true}}$ denotes the underlying true distribution. Most ϕ-divergences in Table 1 satisfy this differentiability condition. Ben-Tal et al. (2013) then use this result to suggest the asymptotic value

$$\rho = \frac{\phi''(1)}{2N}\chi^2_{n-1,1-\alpha}, \tag{14}$$

where $\chi^2_{n-1,1-\alpha}$ is the $1-\alpha$ percentile of a χ^2_{n-1} distribution. This choice of ρ produces an approximate $1-\alpha$ confidence region on the true distribution. To correct for small sample sizes and for more details, we refer the readers to Pardo (2005) and Ben-Tal et al. (2013).

4. Data-Driven Properties

In this section we assume the nominal probabilities $\mathbf{q}$ are the empirical probabilities $\mathbf{q}^N$—i.e., $q_\omega = q_\omega^N = N_\omega/N$—and provide insight into how the ϕLP-2 changes as data is added. First, we investigate how ϕLP-2 might change with a single additional observation in Section 4.1. Next, we examine what happens as more and more data is gathered with asymptotic results in Section 4.2. This analysis must consider how the level of robustness ρ changes as additional observations are obtained. Therefore, in this section, we use the notation ρ_N to emphasize the dependence of ρ on the number observations. To be consistent with the known ϕ-divergence results stated in Section 3.3, we set $\rho_N = \frac{\rho_0}{N}$. Observe that $\rho_0 = \frac{\phi''(1)}{2}\chi^2_{n-1,1-\alpha}$ in (14).

4.1. Value of Additional Data

With a data-driven formulation such as ϕLP-2, one might be concerned about being overly conservative in the problem formulation, and thus, missing the opportunity to find a better solution to the true distribution. For ϕLP-2, this means that the initial model is likely to be more conservative in an effort to be robust, while the new information could make the model less conservative. This would happen, for instance, when the new information removes the current worst-case distribution from the ambiguity set. Below, we present a simple method of determining if taking an additional observation will allow for a lower-cost solution.

Our main goal is to come up with simple conditions by using the current solution. In particular, we would like to use only the current optimal worst-case probabilities p_ω^*, nominal probabilities q_ω, and the number of observations N. One could, of course, solve the problem with an additional observation of ω, see if the optimal value is lowered, and check this for every scenario ω. We would like to avoid resolving the problem. Toward this end, we first provide a general result. Using this general result, we then provide simpler conditions for a subset of the ϕ-divergences by using only $\mathbf{q}$, $\mathbf{p}^*$ and N in Corollary 1.

PROPOSITION 1. *Let $(\mathbf{x}_N^*, \mu_N^*, \lambda_N^*)$ solve the N-observation (dual) problem with $q_\omega = \frac{N_\omega}{N}$. Suppose $s_\omega^* = \dfrac{h_\omega(\mathbf{x}_N^*) - \mu_N^*}{\lambda_N^*}$ is finite and ϕ^* is subdifferentiable at s_ω^* for all $\omega = 1, \ldots, n$. An additional observation of scenario $\hat{\omega}$ will result in a decrease in the worst-case expected cost of ϕLP-2 if the following condition is satisfied*

$$\sum_{\omega=1}^{n} q_\omega \phi^{*\prime}\left(\frac{N}{N+1}s_\omega^*\right)\left(\frac{N}{N+1}s_\omega^*\right) > \phi^*\left(\frac{N}{N+1}s_{\hat{\omega}}^*\right). \tag{15}$$

In (15), $\phi^{*\prime}(s)$ denotes the derivative of ϕ^* at s if it is differentiable, and it denotes a subgradient of ϕ^* at s otherwise.

PROOF OF PROPOSITION 1. For ease of exposition, assume ϕ^* is differentiable. We begin the proof with the change of variables $\kappa = \frac{\lambda}{N}$ and note that $N\rho_N = (N+1)\rho_{N+1} = \rho_0$ is constant. With this change of variables, the objective function of the N-observation problem is given by

$$f_N(\mathbf{x}, \mu, \kappa) = c\mathbf{x} + \mu + \rho_0\kappa + \sum_{\omega=1}^{n} N_\omega\left[\kappa\phi^*\left(\frac{h_\omega(\mathbf{x}) - \mu}{\kappa N}\right)\right].$$

Let z_N^* be the optimal value and $(\mathbf{x}_N^*, \mu_N^*, \kappa_N^*)$ be the optimal solution of the N-observation dual problem (9) with the change of variables. We wish to find a simple estimate of the decrease in the optimal cost $z_N^* - z_{N+1}^*$ associated with taking an additional observation of a specific scenario $\hat{\omega}$. In particular, we look for a condition under which $z_N^* - z_{N+1}^* > 0$. Notice that $(\mathbf{x}_N^*, \mu_N^*, \kappa_N^*)$ is a feasible but not necessarily an optimal solution to the $(N+1)$-observation problem. Then, $z_N^* - f_{N+1}(\mathbf{x}_N^*, \mu_N^*, \kappa_N^*)$ provides a lower bound on the decrease in optimal cost $z_N^* - z_{N+1}^*$. We will find scenarios $\hat{\omega}$ such that $z_N^* - f_{N+1}(\mathbf{x}_N^*, \mu_N^*, \kappa_N^*) > 0$.

The objective of the $(N+1)$-observation problem for a given $(\mathbf{x}, \mu, \kappa)$ is $f_{N+1}(\mathbf{x}, \mu, \kappa) = \mathbf{c}\mathbf{x} + \mu + \rho_0\kappa + \sum_{\omega=1}^{n} N'_\omega\left[\kappa\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{\kappa(N+1)}\right)\right]$, where N'_ω is the number of observations of scenario ω after $N+1$ total observations (e.g., $N_{\hat{\omega}} = N_{\hat{\omega}} + 1$ and $N'_\omega = N_\omega$ for others). Then, $z_N^* - z_{N+1}^*$ is bounded by $\kappa\sum_{\omega=1}^{n}\left[N_\omega\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{\kappa N}\right) - N'_\omega\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{\kappa(N+1)}\right)\right]$, which must be positive to guarantee a drop in optimal cost. If $\hat{\omega}$ is the next scenario observed, we can rewrite this condition as

$$\kappa\sum_{\omega=1}^{n} N_\omega\left[\phi^*\left(\frac{h_\omega(\mathbf{x}) - \mu}{N\kappa}\right) - \phi^*\left(\frac{h_\omega(\mathbf{x}) - \mu}{(N+1)\kappa}\right)\right] - \kappa\phi^*\left(\frac{h_{\hat{\omega}}(x) - \mu}{(N+1)\kappa}\right) > 0. \tag{16}$$

Let $s_\omega^N = \frac{h_\omega(\mathbf{x})-\mu}{\kappa N}$ and $s_\omega^{N+1} = \frac{h_\omega(\mathbf{x})-\mu}{\kappa(N+1)}$, and note that $s_\omega^{N+1} = \frac{N}{N+1}s_\omega^N$. The difference $\phi^*(s_\omega^N) - \phi^*(s_\omega^{N+1})$ will be approximated by the derivative. Because $\phi^*(s)$ is convex, the gradient inequality gives $\phi^*(s_\omega^N) - \phi^*(s_\omega^{N+1}) \geq \frac{1}{N}\phi^{*\prime}(s_\omega^{N+1})s_\omega^{N+1}$. Using this inequality, we can guarantee (16) is satisfied with the condition $\kappa\sum_{\omega=1}^{n}\frac{N_\omega}{N}\phi^{*\prime}(s_\omega^{N+1})s_\omega^{N+1} - \kappa\phi^*\left(\frac{h_{\hat{\omega}}(\mathbf{x})-\mu}{(N+1)\kappa}\right) > 0$. Rearranging and dividing by $\kappa > 0$

$$\sum_{\omega=1}^{n}\frac{N_\omega}{N}\phi^{*\prime}(s_\omega^{N+1})s_\omega^{N+1} > \phi^*(s_{\hat{\omega}}^{N+1}).$$

Finally, we return to the original variables with the substitution $s_\omega^{N+1} = \frac{N}{N+1}s_\omega^*, \forall\omega$. □

We can interpret (15) as follows. If an additional observation is taken from the unknown distribution and the resulting observed scenario $\hat{\omega}$ satisfies (15) with the current solution, then the $(N+1)$-observation problem will have a lower cost than the N-observation problem that was already solved. Observe that the statement of Proposition 1 eliminates the case $\lambda_N^* = 0$. A closer look at (16) and the version of (16) after the use of the gradient inequality, and recalling that $\kappa = \lambda/N$, reveals that the condition (15) is not satisfied at $\lambda_N^* = 0$.

It is possible to simplify condition (15) for some ϕ-divergences, and we detail this in the corollary below. Condition (15) uses the dual formulation and (sub)gradient $\phi^{*\prime}$. This allows us to utilize the primal-dual relationship ($p_{\hat{\omega}}^* = \phi^{*\prime}(s_{\hat{\omega}}^*) q_{\hat{\omega}}$) to provide simplified conditions using $\mathbf{q}, \mathbf{p}^*$ and N.

COROLLARY 1. *Let* $\mathbf{p}^* = (p_1^*, p_2^*, \ldots, p_n^*)$ *solve the* N*-observation (primal) problem with* $q_\omega = \frac{N_\omega}{N}$. *An additional observation of scenario* $\hat{\omega}$ *will result in a decrease in the worst-case expected cost of* ϕ*LP-2 if the following condition is satisfied for:*

Burg entropy: $\frac{p_{\hat{\omega}}^*}{q_{\hat{\omega}}} < \frac{N}{N+1}$,

χ^2*-distance:* $\sum_\omega \frac{(q_\omega)^2}{p_\omega^*} + \sqrt{\frac{N+1}{N}} < 2\frac{q_{\hat{\omega}}}{p_{\hat{\omega}}^*}$,

Hellinger: $\sum_\omega q_\omega \sqrt{\frac{p_\omega^*}{q_\omega}} + \sqrt{\frac{p_{\hat{\omega}}^*}{q_{\hat{\omega}}}} < 2\frac{N}{N+1}$,

Modified χ^2*:* $2\sum_\omega \frac{(p_\omega^*)^2}{q_\omega} > \left(\frac{p_{\hat{\omega}}^*}{q_{\hat{\omega}}}\right)^2 + \left(\frac{N+1}{N}\right)^2$.

PROOF OF COROLLARY 1. For any real number c, we can define $\phi_c(t) = \phi(t) + c(t-1)$, which satisfies $I_{\phi_c}(\mathbf{p}, \mathbf{q}) = I_\phi(\mathbf{p}, \mathbf{q})$ for probability vectors $\mathbf{p}$ and $\mathbf{q}$. This changes the conjugate as $\phi_c^*(s) = \phi^*(s-c) + c$. For some ϕ, we can choose c such that $\phi_c^{*\prime}(s)$ is separable, i.e., $\phi_c^{*\prime}(as) = f(a)\phi_c^{*\prime}(s)$ for some function f. Using this separability, we can simplify (15) for some ϕ by choosing:

Burg entropy: $c = -1$, so $\phi_{b,c}^{*\prime}(s) = -\frac{1}{s}$, $s < 0$,

χ^2*-distance:* $c = -1$, so $\phi_{\chi^2,c}^{*\prime}(s) = \frac{1}{\sqrt{-s}}$, $s < 0$,

Hellinger: $c = -1$ so $\phi_{h,c}^{*\prime}(s) = \frac{1}{s^2}$, $s < 0$,

Modified χ^2*:* $c = 2$, so

$$\phi_{m\chi^2,c}^{*\prime}(s) = \begin{cases} 0 & s < 0 \\ \frac{1}{2}s & s \geq 0. \end{cases}$$

We illustrate the rest of the steps using Burg entropy. Because $I_{\phi_c}(\mathbf{p}, \mathbf{q}) = I_\phi(\mathbf{p}, \mathbf{q})$, we can equivalently solve ϕLP-2 by using ϕ_c instead. Applying (15) to $\phi_{b,c}$, after some algebra, we obtain the simplified condition $-\left(\frac{N}{N+1}\right) s_{\hat{\omega}}^* > 1$ for Burg entropy. Property 4 yields the relationship $\frac{p_{\hat{\omega}}^*}{q_{\hat{\omega}}} = \frac{-1}{s_{\hat{\omega}}^*}$. Substituting this into the simplified condition gives the desired result. Modified χ^2 must consider the split at 0. This can be easily handled. The left-hand side of (15) contains the term $\phi_{m\chi^2,c}^{*\prime}(s_\omega^*) s_\omega^*$, which is equal to $2\left(\frac{p_\omega^*}{q_\omega}\right)^2$. Both terms equal to 0 if $s_\omega^* < 0$ and $(s_\omega^*)^2/2$ otherwise. The right-hand side can be handled similarly, resulting in the final condition presented above. □

Let us take a closer look into Corollary 1's condition for Burg entropy. Recall that using Burg entropy, we obtain the likelihood robust optimization of Wang et al. (2015). Furthermore, it has been used by some authors as the KL divergence because of the change between $\mathbf{p}$ and $\mathbf{q}$. In this case, we have the condition $p_{\hat{\omega}}^* < \frac{N}{N+1} q_{\hat{\omega}}$. This means that the ϕLP-2 has assigned a worst-case probability $p^*_{\hat{\omega}}$ that is less than the slightly adjusted observed frequency $q_{\hat{\omega}} = \frac{N_{\hat{\omega}}}{N}$ of scenario $\hat{\omega}$. The ϕLP-2 focuses on the worst-case cost within the ambiguity set. Therefore, it tends to assign higher probabilities p^*_ω to costly scenarios. Because $p^*_{\hat{\omega}} < \frac{N}{N+1} q_{\hat{\omega}}$, the condition in Corollary 1 suggests that $\hat{\omega}$ might not be a very costly scenario. If we observe one more from this scenario, we believe it is more likely in the nominal (or true) distribution. Consequently, a lower-cost scenario being more likely would decrease the optimal cost.

The simple conditions in Proposition 1 and Corollary 1 provide insight into different scenarios for a decision maker. Let $L = \left\{ \hat{\omega} : \sum_{\omega=1}^{n} q_\omega \phi^{*\prime}\left(\frac{N}{N+1} s^*_\omega\right)\left(\frac{N}{N+1} s^*_\omega\right) > \phi^*\left(\frac{N}{N+1} s^*_{\hat{\omega}}\right)\right\}$. Set L divides the scenarios into two—the ones in L guarantee a drop in the overall cost if sampled one more. Therefore, these scenarios can be considered 'good' or 'optimistic' scenarios. Note that scenarios not in L can also result in the cost decrease. The numerical experiments in Section 7 suggest that L is an adequate indicator of 'good' scenarios for our test problem.

Finally, one might be interested in obtaining a lower bound on the probability that the next sample will decrease the optimal cost. An approximate lower bound on the probability of selecting a sample in L can be found by solving

$$\min_r \left\{ \sum_{\omega \in L} r_\omega : \; r \in \mathcal{P} \right\}. \tag{17}$$

Because we do not know the true distribution, we find the minimum probability of the scenarios in L within the ambiguity set defining ϕLP-2. One can solve (17) by taking its dual.

4.2. Asymptotic Analysis

We now show that ϕLP-2 behaves essentially the same as the corresponding SLP-2 with the (unknown) true distribution $\mathbf{q}^{\text{true}}$ as $N \to \infty$. This requires that the sequence of nominal probabilities $\mathbf{q}$ converge to $\mathbf{q}^{\text{true}}$ with probability one (w.p.1) uniformly in ω—a situation that is satisfied by the assumed empirical probabilities under mild conditions. To emphasize the nominal distribution's dependence on N, we use $\mathbf{q}^N$ in this section. We begin by showing that the worst-case probabilities $\mathbf{p}^*$ obtained by solving ϕLP-2 have a similar asymptotic behavior as $\mathbf{q}^N$.

Proposition 2. *Suppose $\phi(t) \geq 0$ has a unique root at $t = 1$ and the observations are independent and identically distributed from a distribution with probability mass function $\mathbf{q}^{true}$. Then, w.p.1, $\sup_\omega |p^*_\omega - q^{true}_\omega| \to 0$ as $N \to \infty$.*

Proof of Proposition 2. Because $q^N_\omega = \frac{N_\omega}{N}$ obeys the strong law of large numbers and $\omega = 1, ..., n$ is a finite set, we have uniform convergence over ω. That is, $\sup_\omega |q^N_\omega - q^{\text{true}}_\omega| \to 0$ as $N \to \infty$, w.p.1. Now, on a sample path that this convergence occurs, we will show that for all $\epsilon > 0$, there exists N' (depending on the sample path) such that $\forall N \geq N'$ (on that path), $I_\phi(\mathbf{p}^*, \mathbf{q}^N) \leq \frac{\rho_0}{N}$

implies $\sup_\omega |p^*_\omega - q^{\text{true}}_\omega| \le \epsilon$. Because such sample paths have measure 1, our desired result will occur w.p.1. Below, for simplicity of notation we skip the dependence on a particular sample path.

First, note that $\sup_\omega |p^*_\omega - q^{\text{true}}_\omega| \le \sup_\omega |p^*_\omega - q^N_\omega| + \sup_\omega |q^N_\omega - q^{\text{true}}_\omega|$. Assume, again for simplicity, ϵ is chosen so that $\min_\omega q^{\text{true}}_\omega > \frac{\epsilon}{2}$. Let N'' be such that $\sup_\omega |q^N_\omega - q^{\text{true}}_\omega| \le \frac{\epsilon}{2}$ for all $N \ge N''$. This implies $q^N_\omega > 0$ for all $N \ge N''$ for all ω. Now suppose $\sup_\omega |p^*_\omega - q^N_\omega| > \frac{\epsilon}{2}$. Then, for at least one ω—let's denote it $\bar\omega$—we have either $p^*_{\bar\omega} > q^N_{\bar\omega} + \frac{\epsilon}{2}$ or $p^*_{\bar\omega} < q^N_{\bar\omega} - \frac{\epsilon}{2}$. In either case, because $\phi(t) \ge 0$ is a convex function with a root at $t = 1$, we can say for $\bar\omega$

$$\phi\left(\frac{p^*_{\bar\omega}}{q^N_{\bar\omega}}\right) \ge \min\left\{\phi\left(\frac{q^N_{\bar\omega} + \frac{\epsilon}{2}}{q^N_{\bar\omega}}\right), \phi\left(\frac{q^N_{\bar\omega} - \frac{\epsilon}{2}}{q^N_{\bar\omega}}\right)\right\} \ge \min\left\{\phi\left(1 + \frac{\epsilon}{2}\right), \phi\left(1 - \frac{\epsilon}{2}\right)\right\}.$$

The last inequality follows from the fact that $\frac{a+\frac{\epsilon}{2}}{a} \ge 1 + \frac{\epsilon}{2}$ and $\frac{a-\frac{\epsilon}{2}}{a} \le 1 - \frac{\epsilon}{2}$ for $0 < a \le 1$ and again the properties of ϕ. Putting this all together,

$$\begin{aligned} I_\phi(\mathbf{p}^*, \mathbf{q}^N) &= \sum_{\omega=1}^n q^N_\omega \phi\left(\frac{p^*_\omega}{q^N_\omega}\right) \\ &\ge \min_\omega\{q^N_\omega\} \cdot \phi\left(\frac{p^*_{\bar\omega}}{q^N_{\bar\omega}}\right) \\ &\ge \min_\omega\left\{q^{\text{true}}_\omega - \frac{\epsilon}{2}\right\} \cdot \min\left\{\phi\left(1 + \frac{\epsilon}{2}\right), \phi\left(1 - \frac{\epsilon}{2}\right)\right\}. \end{aligned} \tag{18}$$

The right-hand side of (18) is positive because ϕ has a unique root at $t = 1$. By choosing $N' \ge N''$ to satisfy $\min_\omega \left\{q^{\text{true}}_\omega - \frac{\epsilon}{2}\right\} \cdot \min\left\{\phi\left(1 + \frac{\epsilon}{2}\right), \phi\left(1 - \frac{\epsilon}{2}\right)\right\} \ge \frac{\rho_0}{N'}$, we see that $\sup_\omega |p^*_\omega - q^N_\omega| > \frac{\epsilon}{2}$ implies $I_\phi(\mathbf{p}^*, \mathbf{q}) > \frac{\rho_0}{N}$ for all $N \ge N'$. Because the ϕ-divergence constraint ensures $I_\phi(\mathbf{p}^*, \mathbf{q}) \le \frac{\rho_0}{N}$, we must have $\sup_\omega |p^*_\omega - q^N_\omega| \le \frac{\epsilon}{2}$ for all $N \ge N'$, and the desired result follows. □

We are now ready to present the main result on the asymptotic behavior of ϕLP-2.

Theorem 1. *Suppose $\phi(t) \ge 0$ has a unique root at $t = 1$ and the observations are independent and identically distributed from a distribution with probability mass function $\mathbf{q}^{true}$. Then, the optimal value of ϕLP-2 given in (9) converges to that of SLP-2 given in (3) with $\mathbf{q}^{true}$, and all limit points of the solutions $\mathbf{x}^*$ of ϕLP-2 solve SLP-2 with $\mathbf{q}^{true}$ as $N \to \infty$, w.p.1.*

Proof of Theorem 1. Assumptions on problems (4) and set X stated in Section 3.2 ensure $\sup_{\omega, \mathbf{x}\in X} |h_\omega(\mathbf{x})| < C$ for some constant $C < \infty$ and that SLP-2 with $\mathbf{q}^{\text{true}}$ and ϕLP-2 w.p.1 have finite optimal solutions. Let $f(\mathbf{x}, \omega) = \mathbf{cx} + h_\omega(\mathbf{x})$. View the objective of ϕLP-2 as $\mathbb{E}_{\mathbf{p}^*}[f(\mathbf{x}, \omega)] = \sum_{\omega=1}^n p^*_{\omega,N}(\mathbf{x}) f(\mathbf{x}, \omega)$ and the objective of SLP-2 with $\mathbf{q}^{\text{true}}$ as $\mathbb{E}_{\mathbf{q}^{\text{true}}}[f(\mathbf{x}, \omega)] = \sum_{\omega=1}^n q^{\text{true}}_\omega f(\mathbf{x}, \omega)$. Here, $\mathbf{p}^*$ depends on the number of observations N, the actual observations collected, and also $\mathbf{x}$. So, we use the longer notation $p^*_{\omega,N}(\mathbf{x})$ inside the summation for clarity. Following the same arguments as in the proof Proposition 2, we have for each $\mathbf{x} \in X$, $\sup_\omega |p^*_{\omega,N}(\mathbf{x}) - q^{\text{true}}_\omega| \to 0$ as $N \to \infty$, w.p.1. Using this result, we can obtain pointwise strong law of large numbers (SLLN)—that is, for each $\mathbf{x} \in X$, $|\mathbb{E}_{\mathbf{p}^*}[f(\mathbf{x}, \omega)] - \mathbb{E}_{\mathbf{q}^{\text{true}}}[f(\mathbf{x}, \omega)]| \to 0$ as $N \to \infty$ w.p.1. Because X is convex and compact, $f(\mathbf{x},\omega)$ is convex and continuous on X for all ω, and $\mathbb{E}_{\mathbf{q}^{\text{true}}}[f(\mathbf{x},\omega)]$ is finite valued and continuous on X, the desired result follows (see, e.g., Theorem 4 of Shapiro (2003)). □

The non-negativity condition on ϕ in Proposition 2 and Theorem 1 are satisfied by every divergence in Table 1 except Likelihood (which, however, can be rewritten as Burg Entropy). The unique root requirement, on the other hand, is violated for the special cases to be introduced in Section 5.4.

5. A Classification of ϕ-Divergences

Given that there are many ϕ-divergences to choose from, it is important to study how ϕ-divergences act within an ambiguous (or, distributionally robust) stochastic optimization model. We present a classification of ϕ-divergences into four types, resulting from an examination of the limiting behavior of $\phi(t)$ as $t \searrow 0$ and $t \nearrow \infty$. We begin in Section 5.1 by defining two behaviors—suppressing and popping of scenarios—that result in our main classification. Additional details on these behaviors along with a subclassification follow in Sections 5.2 and 5.3. Different classifications may be suitable to different problem types and desired qualities in the ambiguous model. We discuss modeling considerations with respect to our classification in Section 5.5, and we demonstrate the classification on risk models in Section 5.4. Throughout this section, we assume $0 < \rho < \infty$, unless otherwise stated.

5.1. Suppressing and Popping of Scenarios: A Main Classification

Recall the definition of the ambiguity set $\mathcal{P}$, in particular, the ϕ-divergence constraint

$$\sum_{\omega=1}^{n} q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right) \le \rho.$$

Above, ϕ has arguments given by ratios of probabilities, $\frac{p_\omega}{q_\omega}$. The case $q_\omega > 0$, $p_\omega = 0$ means that a scenario with a positive nominal probability has been assigned zero probability by the ambiguous counterpart problem. This case corresponds to the limit of ϕ as $t \searrow 0$. On the other hand, the case $q_\omega = 0$, $p_\omega > 0$ could mean that scenario ω has never been observed before—so it has a zero probability in the nominal distribution—but the ambiguous counterpart problem has assigned it a positive probability. Recall that, by definition, $0\phi(a/0) = a\lim_{t\to\infty}\frac{\phi(t)}{t}$, and so in the latter case, we need to examine $\lim_{t\nearrow\infty}\frac{\phi(t)}{t}$.

Consider each of the limiting cases in more detail:

- Case 1: ($q_\omega > 0$ but $p_\omega = 0$) We call this the "**Suppressing**" behavior because a scenario with a positive probability in the nominal distribution can take zero probability in the ambiguous problem. In other words, such a scenario can be suppressed.

Table 2 Examples of ϕ-divergences fitting into each class. The number in parentheses under the "Can Suppress Scenarios" column denotes the subclass detailed in Section 5.2.

	Can Suppress Scenarios	Cannot Suppress Scenarios
Can Pop Scenarios	Variation Distance (1) Hellinger Distance (2)	χ^2-Distance Burg Entropy
Cannot Pop Scenarios	Modified χ^2-Distance (1) Kullback-Leibler Divergence (2)	J-Divergence

- If $\lim_{t\searrow 0}\phi(t)=\infty$, the ambiguity region will never contain distributions with $p_\omega=0$ but $q_\omega>0$. We say that such a ϕ-divergence *cannot suppress* scenarios.
- If $\lim_{t\searrow 0}\phi(t)<\infty$, the ambiguity region could contain such a distribution, provided q_ω is sufficiently small and/or ρ is sufficiently large. We say that such a ϕ-divergence *can suppress* scenarios.

- Case 2: ($q_\omega=0$ but $p_\omega>0$) We call this the "**Popping**" behavior because a scenario with zero probability in the nominal distribution can have a positive probability (or, pop) in the ambiguous problem.
  - If $\lim_{t\nearrow\infty}\frac{\phi(t)}{t}=\infty$, the ambiguity region will never contain distributions with $p_\omega>0$ but $q_\omega=0$. We say that such a ϕ-divergence *cannot pop* scenarios.
  - If $\lim_{t\nearrow\infty}\frac{\phi(t)}{t}<\infty$, the ambiguity region can admit sufficiently small p_ω. We say that these ϕ-divergences *can pop* scenarios.
- Case 3: ($p_\omega=0$ but $q_\omega=0$) Such a situation has no contribution to the divergence because $0\phi\left(\frac{0}{0}\right)=0$.

The two limiting cases describing suppression and popping behavior create four distinct classes of ϕ-divergences. Table 2 categorizes the ϕ-divergences listed in Table 1 (except for Likelihood). We will shortly provide a subclassification for the ϕ-divergences that can suppress. We end this section with a simple proposition that relates the behavior of a ϕ-divergence with its adjoint.

Proposition 3. *A ϕ-divergence can suppress scenarios if and only if its adjoint $\tilde{\phi}$ can pop scenarios.*

Proof of Proposition 3. Suppose ϕ can suppress scenarios. Then, $\lim_{t\searrow 0}\phi(t)<\infty$. Rearranging the adjoint relationship (1) as $\frac{\tilde{\phi}(t)}{t}=\phi\left(\frac{1}{t}\right)$, and taking limits as $t\nearrow\infty$, we see that $\lim_{t\nearrow\infty}\frac{\tilde{\phi}(t)}{t}<\infty$. This means that the adjoint $\tilde{\phi}$ can pop scenarios. The reverse is also true by the adjoint relationship. □

Proposition 3 implies that a self-adjoint ϕ-divergence can only appear in the top left (can suppress and can pop) or bottom right (cannot suppress and cannot pop) corners of Table 2, as can be seen by the Variation and Hellinger distances and the J-divergence. Proposition 3 also implies that a ϕ-divergence can suppress but cannot pop scenarios if and only if its adjoint $\tilde{\phi}$ can pop but cannot suppress scenarios. Examples of such adjoints can be seen in the opposite corners of Table 2 by the pairs (Modified χ^2-Distance, χ^2-Distance) and (Kullback-Leibler Divergence, Burg Entropy).

5.2. Additional Details about ϕ-Divergences that can Suppress: A Subclassification

ϕ-divergences that can suppress scenarios (i.e., $\lim_{t\searrow 0}\phi(t) = \underline{c}$ for some $0 \le \underline{c} < \infty$) induce a finite limit on $\phi^*(s)$ as $s \searrow -\infty$ (that is, $\lim_{s\searrow -\infty}\phi^*(s) = -\underline{c}$). This can be seen from the fact that ϕ^* is a monotone non-decreasing function that is bounded below by $-\phi(0) = -\lim_{t\searrow 0}\phi(t)$. However, this limit on ϕ^* as $s \searrow -\infty$ can be achieved in different ways. Consider, for instance, Modified χ^2-distance and KL divergence. For both divergences, $\lim_{t\searrow 0}\phi(t) = 1$ and $\lim_{s\searrow -\infty}\phi^*(s) = -1$. However, these two divergences are inherently different. For the Modified χ^2-distance, $\phi^*_{m\chi^2}(s) = -1$ for all $s < -2 = \lim_{t\searrow 0}\phi'_{m\chi^2}(t)$. In contrast, the same limit is reached only asymptotically as $s \searrow -\infty = \lim_{t\searrow 0}\phi'_{kl}(t)$ for the KL divergence. This, in turn, relates to the way the scenarios are suppressed by these divergences.

To see how the above discussion relates to suppression, recall the primal-dual variable relation (12), which specifies $\frac{p^*_\omega}{q_\omega} \in \partial\phi^*(s^*_\omega)$, where $s^*_\omega = \frac{h_\omega(\mathbf{x}^*)-\mu^*}{\lambda^*}$. Note that suppression ($p^*_\omega = 0$, $q_\omega > 0$) can occur only when $0 \in \partial\phi^*(s^*_\omega)$. Recall also that, by the convex conjugate's properties, we have $0 \in \partial\phi^*(s^*_\omega)$ if and only if $s^*_\omega \in \partial\phi(0)$. Assume ϕ and ϕ^* are differentiable for convenience. Examining $\lim_{t\searrow 0}\phi'(t)$ reveals when $0 = \phi^{*\prime}(s^*_\omega)$ by the aforementioned property of conjugates. This analysis yields two subclasses within the ϕ-divergences that can suppress scenarios—one tends to suppress scenarios one at a time, and the other suppresses all but the most costly scenario(s) simultaneously.

- Subclass 1 ($\lim_{t\searrow 0}\phi'(t) > -\infty$) In this case, there are two constants $\underline{c} \ge 0$, and $\underline{s} \le 0$ such that (i) $\lim_{t\searrow 0}\phi'(t) = \underline{s}$ and (ii) $\phi^*(s) = -\underline{c} = -\lim_{t\searrow 0}\phi(t)$ for all $s < \underline{s}$. Thus, $\phi^{*\prime}(s^*_\omega) = 0$ when $s^*_\omega < \underline{s}$, suppressing all such scenarios. In other words, all scenarios ω that satisfy $\frac{h_\omega(\mathbf{x}^*)-\mu^*}{\lambda^*} < \underline{s}$ are suppressed. As ρ increases, scenarios tend to be suppressed one at a time as their s^*_ω reaches $\underline{s}$. Modified χ^2-distance belongs to this subclass.

- Subclass 2 ($\lim_{t\searrow 0}\phi'(t) = -\infty$) In this case, there is a constant $\underline{c} \ge 0$ such that $\phi^*(s) \searrow -\underline{c} = -\lim_{t\searrow 0}\phi(t)$ as $s \to -\infty$ asymptotically, but never reaches the bound. As a result, scenarios can only be suppressed if $s^*_\omega = -\infty$. Intuitively, this can only occur if $\lambda^* = 0$ and $h_\omega(\mathbf{x}^*) < \mu^*$. Consequently, all scenarios ω with $h_\omega(\mathbf{x}^*) < \mu^*$ have $p^*_\omega = 0$, and we must have $\mu^* = \max_\omega h_\omega(\mathbf{x}^*)$ to ensure that scenarios $\omega \in \arg\max h_\omega(\mathbf{x}^*)$ are given positive probability so that $\mathbf{p}^*$ is a probability vector. This means that all but the most expensive scenario(s) will vanish simultaneously. KL divergence belongs to this subclass.

Table 2 lists ϕ-divergences that belong to these subclasses with the number in parentheses. Divergences in the second subclass can be difficult to deal with numerically when suppression occurs because of the $\lambda^* = 0$ in the denominator (see Section 6.3.2 for details). In addition to numerical illustration of this subclassification in Section 7.2.2, we will shortly provide examples of ϕ-divergences to illuminate the different behaviors in Section 5.4.

5.3. Additional Details about ϕ-Divergences that can Pop

Divergences that can pop a scenario have $\phi(t)$ grow linearly as $t \to \infty$, which causes the existence of an upper bound $\bar{s} = \lim_{t\to\infty} \frac{\phi(t)}{t}$ on the domain of $\phi^*(s)$. (Recall the discussion in Section 3.1 regarding ϕLP-2 in dual form, in particular regarding constraint (10).) The primal-dual variable relation (12) specifies $\frac{p^*_\omega}{q_\omega} \in \partial\phi^*(s^*_\omega)$, but the left-hand side is undefined when $q_\omega = 0$. Intuitively, we can think of $\frac{p^*_\omega}{0} = \infty$ if $p^*_\omega > 0$. Thus popping a scenario can be thought to occur when the right-hand side subdifferential also includes ∞. This, in turn, can only happen at the boundary of the domain, $s^*_\omega = \bar{s}$, or when scenario ω has the highest cost. Therefore, only the most expensive scenarios can be popped. The next proposition makes this statement rigorous.

Proposition 4. *Suppose there is a finite $\bar{s} = \lim_{t\to\infty} \frac{\phi(t)}{t}$. A scenario $\bar{\omega}$ for which $q_{\bar{\omega}} = 0$ can only be popped if it has the highest cost, i.e., $\bar{\omega} \in \arg\max_\omega h_\omega(\mathbf{x}^*)$.*

Proof of Proposition 4. We present here an abridged derivation of the dual problem (9), which can be found in full in Ben-Tal et al. (2013) and additionally consider the case where $q_\omega = 0$. For this proof, we assume for simplicity that the first-stage cost vector is $\mathbf{c} = \mathbf{0}$. We begin with the Lagrangian of (5), $\mathcal{L}(\mathbf{p}, \mu, \lambda) = \sum_{\omega=1}^n p_\omega h_\omega(\mathbf{x}) + (1 - \sum_{\omega=1}^n p_\omega)\mu + \left(\rho - \sum_{\omega=1}^n q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right)\right)\lambda$. We then generate the dual of the inner problem as

$$\begin{aligned}
&\min_{\lambda\geq 0,\mu} \max_{\mathbf{p}\geq 0} \sum_{\omega=1}^n p_\omega h_\omega(\mathbf{x}) + \left(1 - \sum_{\omega=1}^n p_\omega\right)\mu + \left(\rho - \sum_{\omega=1}^n q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right)\right)\lambda \\
&= \min_{\lambda\geq 0,\mu} \mu + \rho\lambda + \sum_{\omega=1}^n \max_{p_\omega \geq 0} \left\{ p_\omega (h_\omega(\mathbf{x}) - \mu) - \lambda q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right)\right\} \qquad (19) \\
&= \min_{\lambda\geq 0,\mu} \mu + \rho\lambda + \lambda \sum_{\omega=1}^n q_\omega \max_{t_\omega\geq 0} \{ s_\omega t_\omega - \phi(t_\omega)\} \qquad (20) \\
&= \min_{\lambda\geq 0,\mu} \mu + \rho\lambda + \lambda \sum_{\omega=1}^n q_\omega \phi^*(s_\omega),
\end{aligned}$$

where in (19), we used the fact that the problem is separable for each scenario ω, and in (20), we assumed $q_\omega > 0$ with $t_\omega = \frac{p_\omega}{q_\omega}$. As mentioned earlier, the above formulation is valid when $\lambda = 0$ with the following interpretations: $0\phi^*(b/0) = 0$ if $b \leq 0$ and $0\phi^*(b/0) = +\infty$ if $b > 0$.

To account for the possibility that $q_\omega = 0$, and demonstrate popping behavior, equality (20) must be slightly modified. Consider a term inside the summation in (19) for scenario $\bar{\omega}$ with $q_{\bar{\omega}} = 0$:

$$\max_{p_{\bar{\omega}}\geq 0} \left\{ p_{\bar{\omega}}(h_{\bar{\omega}}(\mathbf{x}) - \mu) - \lambda q_{\bar{\omega}} \phi\left(\frac{p_{\bar{\omega}}}{q_{\bar{\omega}}}\right)\right\} = \max_{p_{\bar{\omega}}\geq 0} \left\{ p_{\bar{\omega}}(h_{\bar{\omega}}(\mathbf{x}) - \mu) - \lambda 0 \phi\left(\frac{p_{\bar{\omega}}}{0}\right)\right\}$$

$$= \max_{p_{\bar\omega}\geq 0} \left\{ p_{\bar\omega}\left(h_{\bar\omega}(\mathbf{x}) - \mu - \lambda\bar{s}\right)\right\}. \qquad (21)$$

The behavior of (21) depends on the sign of $(h_{\bar\omega}(\mathbf{x}) - \mu - \lambda\bar{s})$ (or equivalently, relation between $s_{\bar\omega}$ and $\bar{s}$). There are three cases:

Case 1: $h_{\bar\omega}(\mathbf{x}) - \mu > \lambda\bar{s}$ selects $p_{\bar\omega} = \infty$, making (21) unbounded. Because we are trying to minimize the overall objective, this case will be eliminated. In other words, this case induces the constraint $h_{\bar\omega}(\mathbf{x}) - \mu \leq \bar{s}\lambda$ given in (10).

Case 2: $h_{\bar\omega}(\mathbf{x}) - \mu < \lambda\bar{s}$ selects $p_{\bar\omega} = 0$.

Case 3: $h_{\bar\omega}(\mathbf{x}) - \mu = \lambda\bar{s}$ places no restrictions on the value of $p_{\bar\omega}$ because (21) is identically zero, and hence allows for $p_{\bar\omega} > 0$ (popping).

Because $h_{\bar\omega}(\mathbf{x}) - \mu \leq \lambda\bar{s}$ for all $\bar\omega$ by the above discussion, and $h_{\bar\omega}(\mathbf{x}) - \mu = \lambda\bar{s}$ for any popped scenarios, only the most expensive scenario can be popped. □

Observe that the dual formulation (9) is still valid when $\lambda = 0$ and $q_{\bar\omega} = 0$ with the aforementioned interpretations. When $\lambda > 0$, we set $q_{\bar\omega}\phi^*(\bar{s}) = 0 \cdot \infty = 0$ if $\phi^*(\bar{s}) = \infty$ when popping occurs by using the typical extended arithmetic rules (Rockafellar 1970). Finding the probability of the popped scenario cannot be done by using the subgradients of ϕ^* as with other scenarios. Thus, the probability must be calculated with the primal feasibility conditions as detailed in Property 4.

5.4. Illustration of Suppression and Popping via Risk Models

The class of ϕ-divergence constrained problems includes some special cases that result in common risk models, which we document here. We follow each example with a discussion of their suppressing and popping behavior. Derivations of these examples are provided in the appendix. Below and in the appendix, we use the notation $[a]^+ = \max\{a, 0\}$, and we view $\mathbf{h}(\mathbf{x})$ as a random variable that takes on values $h_\omega(\mathbf{x})$ according to the nominal probabilities $\mathbf{q}$ when writing expectations, etc.

Example 1 (CVaR). The coherent risk measure Conditional Value-at-Risk (CVaR) is one of the most widely used risk measure in the literature. Minimizing

$$\mathbf{cx} + \mathrm{CVaR}_\beta(\mathbf{h}(\mathbf{x})) = \mathbf{cx} + \min_{\mu\in\mathbb{R}} \left\{ \mu + \frac{1}{1-\beta}\mathbb{E}\left[[\mathbf{h}(\mathbf{x}) - \mu]^+ \right] \right\}$$

over $\mathbf{x} \in X$ is equivalent to the ϕ-divergence constrained ϕLP-2 with

$$\phi_{\mathrm{CVaR}}(t) = \begin{cases} 0 & 0 \leq t \leq \frac{1}{1-\beta} \\ \infty & \text{otherwise,} \end{cases}$$

for $0 < \beta < 1$. We see that $\phi_{\mathrm{CVaR}}(0) = 0$, indicating that CVaR will suppress some scenarios. This appears in the definition of CVaR as the positive part in the expected value, $\mathbb{E}\left[[\mathbf{h}(\mathbf{x}) - \mu]^+\right]$. Furthermore, suppression will occur one at a time. Scenarios cannot be popped because the expectation is taken with respect to the nominal distribution. This can also been seen by the limit $\lim_{t\nearrow\infty} \frac{\phi_{\mathrm{CVaR}}(t)}{t} = \infty$. □

The CVaR ϕ-divergence is bounded above, which leads to the question of what happens when a ϕ-divergence is bounded below. The "reverse" CVaR, turns out, is equivalent to minimizing a convex combination of expectation and worst-case. We discuss this example below.

Example 2 (Convex Combination of Expectation and Worst-Case). The ϕ-divergence constrained ϕLP-2 with

$$\phi_{\text{EW}}(t) = \begin{cases} \infty & t < 1-\beta \\ 0 & t \geq 1-\beta, \end{cases}$$

for $0 < \beta < 1$ is equivalent to minimizing the convex combination of expectation and worst-case

$$\mathbf{cx} + \beta \sup_{\omega} h_\omega(\mathbf{x}) + (1-\beta)\mathbb{E}\left[\mathbf{h}(\mathbf{x})\right],$$

over $\mathbf{x} \in X$, where the expectation is taken with respect to the nominal probability vector $\mathbf{q}$. Note that $\lim_{t\to\infty} \frac{\phi_{\text{EW}}(t)}{t} = 0$, indicating that this divergence will pop scenarios. This behavior appears in the term $\sup_\omega h_\omega(\mathbf{x})$. However, $\phi_{\text{EW}}(0) = \infty$ indicates that scenarios will not be suppressed, which is demonstrated by the expectation term $\mathbb{E}\left[\mathbf{h}(\mathbf{x})\right]$. Recall that the expectation term takes into account every scenario with positive nominal probability. □

An objective function taking a weighted sum of expected value and CVaR often comes up in practice. The next example shows how to generate a convex combination of expectation and CVaR using ϕ-divergences.

Example 3 (Convex Combination of Expectation and CVaR). The ϕ-divergence constrained ϕLP-2 with

$$\phi_{\text{EC}}(t) = \begin{cases} 0 & 1-\alpha \leq t \leq \frac{1}{1-\beta} \\ \infty & \text{otherwise}, \end{cases}$$

for $\alpha, \beta \in (0,1)$ is equivalent to minimizing, over $\mathbf{x} \in X$,

$$\mathbf{cx} + (1-\alpha)\mathbb{E}\left[\mathbf{h}(\mathbf{x})\right] + \alpha \text{CVaR}_{\frac{\beta}{\alpha(1-\beta)+\beta}}\left[\mathbf{h}(\mathbf{x})\right].$$

This divergence will neither pop (because both the expectation and CVaR term are taken with respect to the nominal distribution) nor suppress (because the expectation term includes every scenario). □

5.5. Modeling Considerations When Choosing a ϕ-Divergence

We offer the following suggestions for choosing an appropriate ϕ-divergence class for the data available and the preferences of a decision maker.

First, consider whether to choose a ϕ-divergence that can suppress scenarios. If the scenarios come from high-quality observed data, one may wish to avoid ϕ-divergences that can suppress scenarios. Alternatively, if the decision maker wishes to consider every observed scenario with a positive probability in the final model, then ϕ-divergences that cannot suppress would be preferable. However, if the data is poorly sampled or comes from opinion rather than observation or simulation, the option of suppressing scenarios may result in a solution with better robustness properties. Suppression one at a time may be preferred by decision makers who wish to see the effect of robustness level (or risk level) on the solutions.

Next, consider whether to choose a ϕ-divergence that allows for popping scenarios. If the scenarios strictly come from observation, with little theoretical understanding of the problem, we suggest choosing a ϕ-divergence that cannot pop scenarios. However, if the scenarios come from a mix of observed/simulated data and expert opinion about scenarios of interest, then divergences that can pop present an interesting modeling choice. This allows for including interesting but unobserved scenarios and letting the mathematical program to assign an appropriate probability to them.

A risk-averse decision maker that is flexible might wish to consider ϕ-divergences that can both pop and suppress. This way, the model assigns an appropriate probability in a risk-averse manner to each scenario. Suppression one at a time may again be preferred to see the effect of robustness level on the individual scenarios.

6. A Decomposition-Based Solution Method

6.1. Preliminaries

As the model gets larger, a direct solution of ϕLP-2 becomes computationally expensive. Decomposition-based methods could significantly reduce the solution time and allow larger problems to be solved efficiently. We propose a Bender's decomposition-based algorithm for solving ϕLP-2 using the dual formulation. The heart of the algorithm lies in the fact that dual ϕLP-2 (9) is a convex program (Property 1) that can be decomposed by scenario (Property 3). We begin with a proposition to make the first point (convexity) clear. The proof is provided in the appendix.

Proposition 5. *$h^\dagger_\omega(\mathbf{x},\lambda,\mu) = \lambda\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{\lambda}\right)$ is a convex function over $\lambda \geq 0$, μ, and $\mathbf{x} \in X$.*

By Proposition 5, the expectation of this function with respect to the nominal distribution $\sum_{\omega=1}^{n} q_\omega h^\dagger_\omega(\mathbf{x},\lambda,\mu)$ is also convex. The algorithm replaces this convex function with a number of affine cutting planes, forming a lower approximation. It is possible to use a single-cut or multicut version of the algorithm. The single-cut version replaces the overall convex function with a number of affine cutting planes, whereas the multicut version creates affine cutting planes for each individual function $h^\dagger_\omega(\mathbf{x},\lambda,\mu)$, $\omega = 1,\ldots,n$. The algorithm also removes the dual feasibility constraint (10)—when it is present in the formulation—and exchanges it with a series of feasibility cuts. Recall that feasibility constraint (10) is present only for ϕ-divergences that can pop scenarios.

6.2. Algorithm

Algorithm 1 presents a basic implementation of the proposed method. The modified Bender's decomposition presented here has the following features: *(i)* it solves the original linear second-stage subproblems (4) rather than nonlinear subproblems in (13), and it uses them to quickly generate subgradients of the nonlinear term $h_\omega^\dagger(\mathbf{x}, \lambda, \mu)$; *(ii)* when $\lim_{t\to\infty} \frac{\phi(t)}{t} = \bar{s} < \infty$, it exchanges the nonlinear constraints $h_\omega(\mathbf{x}) - \mu \le \bar{s}\lambda$ for a (potentially much) smaller set of easily generated affine feasibility cuts; and thus *(iii)* it maintains a linear master problem and linear subproblems.

The single-cut master problem is given by

$$\min_{\mathbf{x},\lambda,\mu} \ \mathbf{cx} + \mu + \rho\lambda + \theta \tag{22}$$

$$\text{s.t. } \mathbf{x} \in X, \quad \lambda \ge 0$$

$$\theta \ge \mathbf{T}_j \cdot (\mathbf{x}\ \lambda\ \mu)^T + t_j, \quad j \in J \tag{23}$$

$$\mu + \bar{s}\lambda \ge \mathbf{M}_k \cdot \mathbf{x} + m_k, \quad k \in K \tag{24}$$

where constraints (23) are the objective cuts, constraints (24) are the feasibility cuts replacing constraint (10) if $\lim_{t\to\infty} \frac{\phi(t)}{t} = \bar{s} < \infty$, and J and K are the index sets for objective and feasibility cuts, respectively. In the algorithmic statement and discussions below, any variable with a 'hat' over it (e.g., $\hat{\mathbf{x}}, \hat{\mu}$) indicates a current solution. In contrast, regular variables (e.g., $\mathbf{x}, \mu$) appear in the definition of master problem, cutting planes (23) and (24), etc.

Algorithm 1. Decomposition algorithm for solving ϕLP-2

```
Initialize $z_l = -\infty, z_u = +\infty$; $J \leftarrow \emptyset, K \leftarrow \emptyset$; select $\texttt{TOL} \ge 0$
Initialize $\hat{\lambda} \leftarrow 1$, $\hat{\mu} \leftarrow 0$, and $\hat{\theta} \leftarrow 0$
while $z_u - z_l \ge \texttt{TOL} \cdot \min\{|z_u|, |z_l|\}$ do
    if at first iteration then
        Solve master problem (22) with $\theta = \mu = 0$ to generate $\hat{\mathbf{x}}$
    else
        Solve master problem (22) to obtain $\hat{\mathbf{x}}$, $\hat{\lambda}$, $\hat{\mu}$, and $\hat{\theta}$
    end if
    Solve subproblems (4) to obtain $h_\omega(\hat{\mathbf{x}})$ and duals $\hat{\pi}_\omega$, $\omega = 1, \ldots, n$
    if $\bar{s} < \infty$ and $\sup_\omega h_\omega(\hat{\mathbf{x}}) - \hat{\mu} > \bar{s}\hat{\lambda}$ then
        Generate feasibility cut and add to master problem; update $K$
        Find $\hat{\mu}$ so that $\sup_\omega h_\omega(\hat{\mathbf{x}}) - \hat{\mu} < \bar{s}\hat{\lambda}$
    else
        Set $z_l \leftarrow$ master optimal cost $\mathbf{c}\hat{\mathbf{x}} + \hat{\mu} + \rho\hat{\lambda} + \hat{\theta}$
    end if
    Generate objective cut and add to master problem; update $J$
    Set $\theta_{\text{true}} \leftarrow \sum_{\omega=1}^{n} q_\omega h_\omega^\dagger(\hat{\mathbf{x}}, \hat{\lambda}, \hat{\mu})$
    if $\mathbf{c}\hat{\mathbf{x}} + \hat{\mu} + \rho\hat{\lambda} + \theta_{\text{true}} < z_u$ then
        $z_u \leftarrow \mathbf{c}\hat{\mathbf{x}} + \hat{\mu} + \rho\hat{\lambda} + \theta_{\text{true}}$
        $\mathbf{x}_{\text{best}} \leftarrow \hat{\mathbf{x}}, \lambda_{\text{best}} \leftarrow \hat{\lambda}, \mu_{\text{best}} \leftarrow \hat{\mu}$
        $p_\omega \leftarrow \phi^{*\prime}(\frac{h_\omega(\mathbf{x}) - \mu}{\lambda}) q_\omega$ for $\omega = 1, \ldots, n$
    end if
end while
```

In line 2 of Algorithm 1, we initialize $\hat{\lambda}=1$ somewhat arbitrarily. However, by setting $\hat{\theta}=\hat{\mu}=0$, we obtain a candidate solution $\hat{\mathbf{x}}$ in the first iteration by essentially solving $\min\{\mathbf{cx}:\mathbf{x}\in X\}$. It is possible to obtain an initial $\hat{\mathbf{x}}$ in other ways. If we detect that current $\hat{\mu}$ is infeasible, we do not update the lower bound z_l so that the master is solved again with the feasibility cuts. It is easy to obtain a new $\hat{\mu}$ in line 12 of the Algorithm. For example, when $\bar{s}\hat{\lambda}>0$, simply setting $\hat{\mu}=\sup_\omega h_\omega(\hat{\mathbf{x}})-\bar{s}\hat{\lambda}(1-10^3)$ gives us a feasible—but not necessarily optimal—solution. We can then update the upper bound and current solution in lines 18–21 as necessary. We now go through in more detail how to obtain the objective and feasibility cuts.

6.2.1. Objective Cuts The chain rule translates the (sub)gradients of $h_\omega(\mathbf{x})$ to (sub)gradients of $h^\dagger_\omega(\mathbf{x},\lambda,\mu)$. Recall notation $s_\omega=\frac{h(\mathbf{x})-\mu}{\lambda}$ and let $h^\dagger_\omega(s_\omega)=\lambda\phi^*(s_\omega)$ denote the non-linear portion of the objective function. Consider for simplicity $\hat{\lambda}>0$ and ϕ^* is differentiable. The cut coefficients of $\mathbf{x}$ are formed through $\partial h^\dagger/\partial\mathbf{x}=(\partial h^\dagger_\omega/\partial s_\omega)\cdot(\partial s_\omega/\partial\mathbf{x})$. This gives $\phi^{*\prime}(\hat{s}_\omega)\cdot(\hat{\pi}_\omega\mathbf{B}^\omega)$ as the coefficients of $\mathbf{x}$, where $\mathbf{B}^\omega$ is the technology matrix of subproblem (4). The cut coefficients of λ and μ are found in a similar way. Using the (sub)gradient inequality, we can also obtain the intercept term t^ω_j. This gives us:

$$\begin{aligned}\mathbf{T}^\omega_j&=\begin{pmatrix}\phi^{*\prime}(\hat{s}_\omega)\cdot(\hat{\pi}_\omega\mathbf{B}^\omega) & \phi^*(\hat{s}_\omega)-\phi^{*\prime}(\hat{s}_\omega)\hat{s}_\omega & -\phi^{*\prime}(\hat{s}_\omega)\end{pmatrix},\\ t^\omega_j&=\phi^{*\prime}(\hat{s}_\omega)\left[h_\omega(\hat{\mathbf{x}})-\hat{\pi}_\omega\mathbf{B}^\omega\hat{\mathbf{x}}\right].\end{aligned}$$

For the single-cut master problem proposed, the vector of cut coefficients and intercept in constraints (23) can be found by $\mathbf{T}_j=\sum_\omega q_\omega\mathbf{T}^\omega_j$ and $t_j=\sum_\omega q_\omega t^\omega_j$. The multi-cut version replaces the last term in the objective of (22) with $\sum_\omega q_\omega\theta_\omega$ and uses the individual cuts $\theta_\omega\geq\mathbf{T}^\omega_j\cdot(\mathbf{x}\ \lambda\ \mu)^T+t^\omega_j$ for each scenario $\omega=1,\ldots,n$.

6.2.2. Feasibility Cuts After the linear subproblems (4) are solved, it may be the case that $h_\omega(\hat{\mathbf{x}})-\hat{\mu}>\bar{s}\hat{\lambda}$ for some scenario ω, rendering $\hat{\mu}$ and $\hat{\lambda}$ infeasible. We need $h_\omega(\mathbf{x})-\mu-\bar{s}\lambda\leq 0$ for feasibility. This is a convex function over $\lambda\geq 0$, μ and $\mathbf{x}\in X$, and we can use the lower approximation of this function to generate a feasibility cut. The cut coefficients are obtained through the subgradients $(\hat{\pi}_\omega\mathbf{B}^\omega\ \ -1\ \ -\bar{s})$ for $(\mathbf{x}\ \ \mu\ \ \lambda)$. This leads to the feasibility constraints (24) with $\mathbf{M}_k=\hat{\pi}_\omega\mathbf{B}^\omega$ and $m_k=\hat{\pi}_\omega\mathbf{d}^\omega$, where $\mathbf{B}^\omega$ and $\mathbf{d}^\omega$ are from the subproblem (4).

6.3. Computational Considerations

6.3.1. Trust Regions In order to enhance the performance of the above decomposition algorithm, we included an L^∞-norm trust region which is scaled up (by a factor of 3) or down (by a factor of $\frac{1}{4}$) when the trust region inhibits finding the optimal solution or when the polyhedral lower approximation is far from the second-stage expected cost, respectively. The trust region is an implementation of Algorithm 4.1 of Nocedal and Wright (1999). This speeds up the algorithm considerably.

6.3.2. Implementation Notes on Different ϕ-Divergences First, when constraint (10) is not present, there is no need to use feasibility cuts.

We recommend forcing λ to be nonzero and checking optimality condition at $\lambda = 0$ separately. Especially, divergences that can simultaneously suppress all but the most expensive scenarios in Subclass 2 (see Section 5.2) can be computationally difficult to work with because $\lambda = 0$ could occur. Floating point finite tolerance can alleviate this somewhat for the KL divergence, for which $\phi^*(s) = e^s$, because $e^{-800} = 0$ to machine precision.

Divergences that can pop require a check for any $s_\omega = \bar{s}$. The probability of a popped scenario can be determined by enforcing $\sum_\omega p_\omega = 1$ after determining the probability of the other scenarios. For divergences that cannot pop, it can be useful to add a computational upper bound on s, $\bar{s}_{\text{comp}}$. Such an upper bound can be computed easily by bounding the ratio $\frac{p_\omega}{q_\omega} \leq \frac{1}{\min_\omega q_\omega}$. The computational upper bound can then be selected so that $\phi^{*\prime}(\bar{s}_{\text{comp}}) \geq \frac{1}{\min_\omega q_\omega}$ according to machine precision. Note, however, that an artificial upper bound will induce artificial popping behavior if the nominal distribution contains impossible scenarios. This technique is especially useful for the KL divergence because e^s overflows on double-precision machines for $s \geq 710$.

7. Numerical Illustration

7.1. Experimental Setup

To illustrate the techniques discussed in this paper, we applied the specialized Bender's decomposition algorithm from Section 6 to a small electricity generation problem. The algorithm was implemented in MATLAB using the CPLEX linear program solver.

We modified an SLP-2 test problem, denoted APL1P, which has 5 independent random variables and 1280 realizations (Infanger 1992). The first-stage determines the capacity to be built for two electricity generators. The generators are operated under uncertain demands and reliability of the generators in the second stage. To clearly demonstrate how the worst-case distribution changes with ρ—and especially to demonstrate suppressing and popping behavior and value of data—we took 6 unique scenarios from APL1P. We denote the resulting problem as ϕAPL1P.

To make our presentation clear, we ordered the scenarios of ϕAPL1P from most costly to least costly. So, `scen1` is the most costly scenario (displayed in dark green in Figures 1 and 2 that appears typically at the top), and `scen6` is the least costly scenario (displayed in light green in Figures 1 and 2 that appears typically at the very bottom). We assume each scenario is equally likely in the nominal distribution. In the case where we wish to investigate the popping behavior, we set the nominal probability of the most expensive scenario to $q_{\texttt{scen1}} = 0$ and again have the other scenarios be equally likely.

Table 3 Numerical results of ϕAPL1P for various divergences.

			p^*_ω					
ϕ	ρ	Opt. Cost	`scen1`	`scen2`	`scen3`	`scen4`	`scen5`	`scen6`
$\phi_{m\chi^2}$	1.845	30735	0.6354	0.2293	0.1353	0	0	0
ϕ_{kl}	0.9225	30921	0.7208	0.1507	0.1050	0.0108	0.0075	0.0052
ϕ_b	0.9225	30714	0.7751	0.0768	0.0636	0.0308	0.0285	0.0253
ϕ_b	1.107	29775	0.6273	0.1311	0.1065	0.0494	0.0455	0.0402

For our numerical experiments, we used the following ϕ-divergences: (i) Modified χ^2-distance, (ii) KL divergence and (iii) Burg entropy. Modified χ^2-distance and KL divergence are used to demonstrate the one-at-a-time and simultaneous suppression, respectively. Burg entropy, on the other hand, is used to demonstrate the popping behavior.

7.2. Results

7.2.1. Optimal Value and Solution Table 3 shows the optimal value and solution of ϕAPL1P using different ϕ-divergences. As mentioned above, we assume one observation per scenario for the first three divergences in Table 3 and the most costly scenario unobserved in the last row to demonstrate popping. The value of ρ is chosen in accordance with an asymptotic 95% confidence region in (14). For the popping example, while $n = 6$ is the same, N is one less, resulting in a different ρ.

All divergences put the highest probability in the most costly scenario. At this level of robustness, the Modified χ^2-distance has suppressed three scenarios, while the KL divergence has not yet suppressed any. In the next section, as ρ increases, the KL divergence will suppress all but the most costly scenario simultaneously. The total costs are similar except for the Burg entropy with popping, which is slightly lower.

7.2.2. Numerical Illustrations of Suppressing and Popping Figure 1 shows how p^*_ω changes with ρ for both the Modified χ^2-distance (left) and the KL divergence (right). As shown in Section 5.2, the Modified χ^2-Distance suppresses scenarios one at a time, starting with the least expensive; while the KL divergence suppresses all scenarios but the most costly scenario simultaneously at a high enough value of ρ.

An example of a ϕ-divergence that can pop, the Burg entropy, is given in Figure 2. The left plot in Figure 2 demonstrates the worst-case distribution assuming that all scenarios have a single observation. The right plot shows the worst-case distribution when all scenarios but the most costly have a single observation, which is unobserved. Notice, in particular, that the probability of the most costly scenario becomes small as ρ decreases. Other divergences that can pop but not suppress look qualitatively similar.

Figure 1 Examples of distributions that can suppress: Modified χ^2 distance (left; one-at-a-time suppression) and KL Divergence (right; simultaneous suppression). Note the x-axis scale difference in the plots.

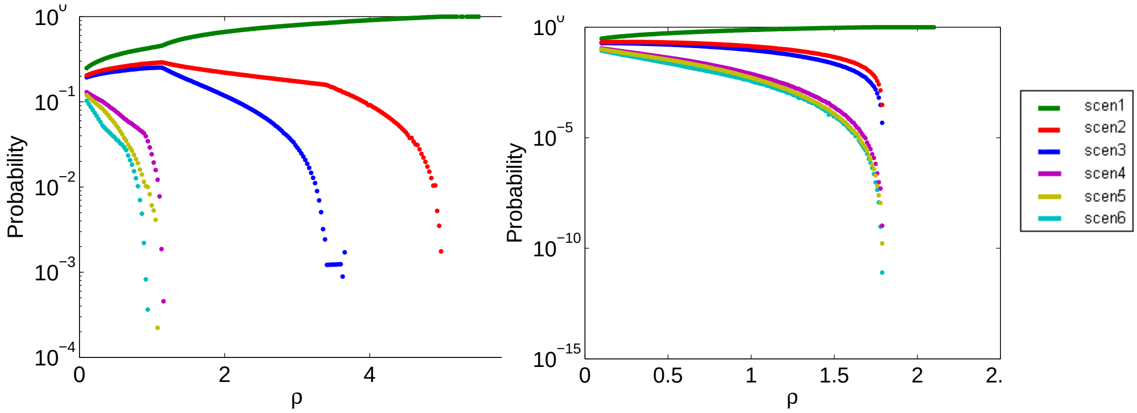
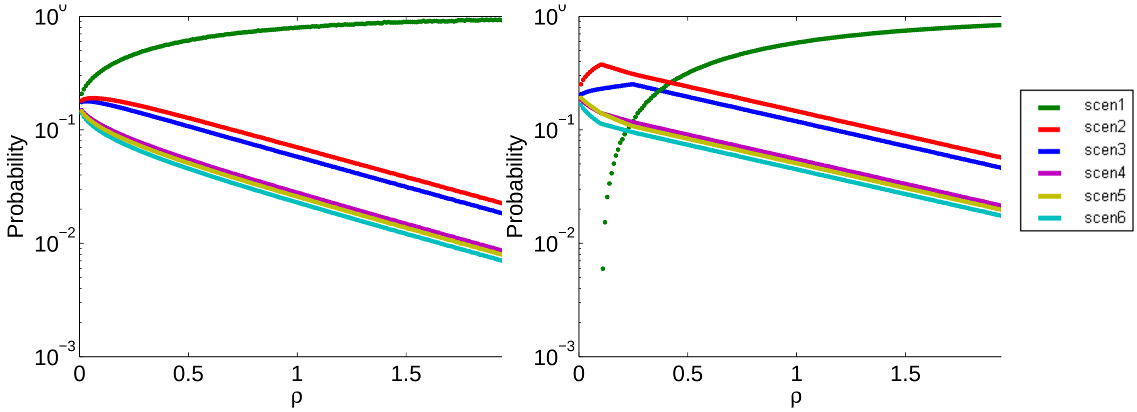


Figure 2 Example of a distribution that can pop—the Burg entropy: all scenarios have a single observation (left); the most costly scenario having no observation (right).

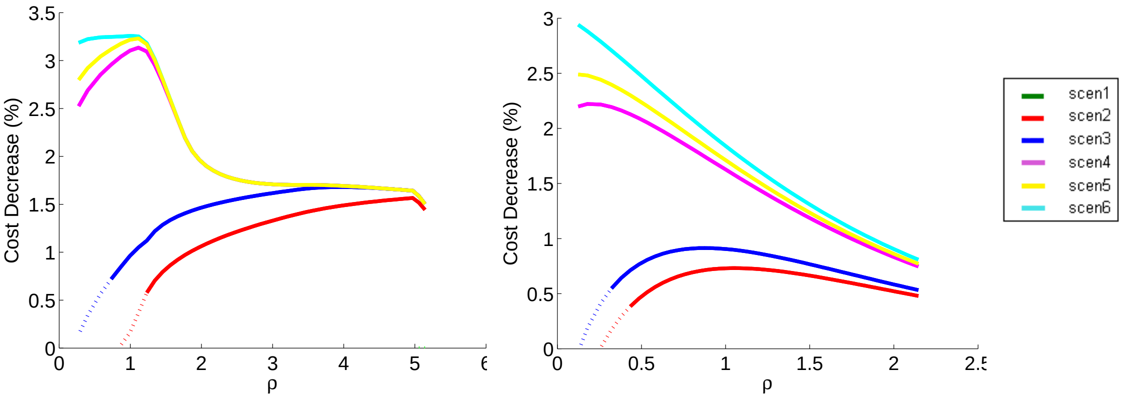


7.2.3. Numerical Illustration of Value of Data We tested the value of data (VoD) condition from Corollary 1 for the Modified χ^2-distance and the Burg entropy for various values of ρ. Note that VoD condition in (15) and the simplified conditions in Corollary 1 are sufficient but not necessary conditions.

Figure 3 compares the VoD condition from Corollary 1 to the actual cost decrease resulting from an additional observation for the Modified χ^2-distance (left) and Burg entropy (right). The solid lines indicate when the VoD condition is satisfied, while dotted lines show when an additional observation decreased the optimal value although VoD condition was not satisfied. VoD conditions detect the actual cost decrease in majority of the cases for this example.

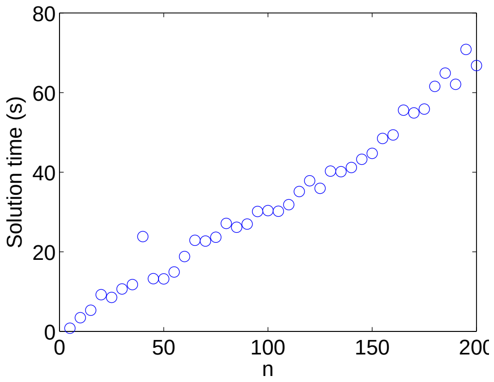
We see a reversal of the scenarios in Figure 3 compared to Figures 1 and 2. The least costly scenarios (`scen6`, `scen5`, and `scen4`) cause the largest decrease in the optimal cost if sampled one more observation. This is expected because having one more observation from the low-cost scenarios increases our belief (or, nominal probability) that future can be less costly. Higher cost

Figure 3 Percentage decrease in the worst-case expected cost from an additional observation for Modified χ^2-distance (left) and Burg entropy (right). The solid lines indicate regions where the condition is satisfied, and dotted regions indicate that the condition is not satisfied.



scenarios (`scen3` and `scen2`) cause a smaller decrease in the optimal value, and the VoD condition is less effective in detecting these scenarios, especially at lower values of ρ. The most costly scenario (`scen1`) is not visible in Figure 3 because an additional observation increases the optimal value. VoD conditions are never satisfied for this scenario.

Figure 4 Solution time of ϕAPL1P with the proposed decomposition algorithm against the number of scenarios for the Modified χ^2-distance.



7.2.4. Running Time of the Decomposition Algorithm Finally, to illustrate the linear dependence of the problem complexity on the number of scenarios, we tracked the time to solve ϕAPL1P (in seconds) for an increasing number of scenarios, from 6 to 200. Figure 4 depicts the linear increase in the running time with the number of scenarios in ϕLP-2 formed using the Modified χ^2-distance. Other ϕ-divergences have similar behavior.

8. Summary and Future Work

We proposed to use ϕ-divergences to define an ambiguity set of probability distributions—possibly using observed data, simulated data, forecasts, expert opinions, etc.—and optimize the worst-case expected cost with respect to this ambiguity set in a two-stage setting. We provided a new classification of ϕ-divergences that can be used in determining which ϕ-divergence is most appropriate in practice for different model types and decision makers. A computationally simple method is established to determine if an additional sample will result in a lower-cost solution. We have shown that as more data is gathered, the optimal value and solution of ϕLP-2 converge to those of SLP-2. We have also provided a Bender's decomposition-based solution algorithm to solve ϕLP-2 efficiently and used it to illustrate some of the properties of the ϕLP-2.

There are many interesting avenues for future work. Extensions to multistage problems and ways to handle continuous distributions ϕLP-2 merit further research, and some work has already started to appear in these areas; e.g., Hu and Hong (2013), Jiang and Guan (2015b). There are other divergences, probability metrics, and statistical ways to measure the distance between two distributions. Generalizations of the classification presented in this paper to other distance measures is another area of future research. Further refinement of the classification would be valuable to deepen our understanding of why each ambiguity set might be preferred. Finally, while some applications have appeared in the literature—some for specific phi-divergences, e.g., Calafiore (2007), Klabjan et al. (2013) and some recent work for general phi-divergences, e.g., Love and Bayraksan (2015)—further applications of this class of problems to real-world problems would be beneficial.

Endnotes

1. This work is largely derived from the dissertation Love (2013). An earlier version of Section 5 has appeared in the conference paper Love and Bayraksan (2014), and select results have been summarized in the tutorial Bayraksan and Love (2015). This paper contains more results, full derivations and proofs, and further numerical illustration of results. Parts of Bayraksan and Love (2015) have been included by permission.

Acknowledgments

This work has been partially supported by the National Science Foundation through grant CMMI-1345626. We also thank Hamed Rahimian for comments on an earlier version of the paper.

Appendix A: Proof of Property 2 and Proposition 5

Proof of Property 2 A coherent risk measure has a dual representation that can be viewed as worst-case expectation from a set of probability measures. Specifically, $\mathcal{R}$ is a real-valued coherent risk measure if and only if there exists a convex, bounded, and closed set $\mathcal{U}$ such that $\mathcal{R}(Y) = \sup_{\zeta \in \mathcal{U}} \sum_{\omega=1}^{n} r_\omega \zeta_\omega Y_\omega$. Here, random variables Y are seen as functions in an $\mathcal{L}^p$ space with sample space Ω and a reference probability measure $\mathbf{r}$ (see, e.g., Rockafellar (2007), Theorem 6.6 of Shapiro et al. (2009) for details). In our context, $Y \equiv \mathbf{h}(\mathbf{x})$, taking values $h_\omega(\mathbf{x})$.

Observe that with our assumptions, $\sup_{\omega,\mathbf{x}\in X}|h_\omega(\mathbf{x})|<C$ for some $C<\infty$. Because we have finitely many elements of ω, the choice of norm or $\mathcal{L}^p$ space does not matter; so, $\mathcal{U}\subset\{\zeta\in\mathbb{R}^n\colon\ \sum_{\omega=1}^n r_\omega\zeta_\omega=1,\quad \zeta\geq 0\}$. To handle the case $q_\omega=0$ for some ω, we use the discrete uniform distribution as the reference distribution; $r_\omega=\frac{1}{n},\forall\omega$. Then, we rewrite ϕLP-2 in primal form (5) with the change of variables $\tilde{p}_\omega=\frac{p_\omega}{1/n}$ and $\tilde{q}_\omega=\frac{q_\omega}{1/n}$. By setting $\zeta_\omega=\tilde{p}_\omega$, we obtain $\mathcal{U}=\left\{\tilde{\mathbf{p}}\in\mathbb{R}^n\colon\ \sum_{\omega=1}^n\frac{1}{n}\tilde{q}_\omega\phi\left(\frac{\tilde{p}_\omega}{\tilde{q}_\omega}\right)\leq\rho,\ \sum_{\omega=1}^n\tilde{p}_\omega=1,\ \ \tilde{\mathbf{p}}\geq 0\right\}$—a nonempty convex compact set—again with the interpretations $0\phi\left(\frac{a}{0}\right)=a\lim_{t\to\infty}\frac{\phi(t)}{t}$ for $a>0$ and $0\phi\left(\frac{0}{0}\right)=0$ for $a=0$. Finally, we can rewrite the inner maximization of (5) as $\mathcal{R}(\mathbf{h}(\mathbf{x}))=\sup_{\tilde{\mathbf{p}}\in\mathcal{U}}\sum_{\omega=1}^n\frac{1}{n}\tilde{p}_\omega h_\omega(\mathbf{x})=\max_{\mathbf{p}\in\mathcal{P}}p_\omega h_\omega(\mathbf{x})$. Thus we see that ϕLP-2 minimizes a coherent risk measure. □

Remark 1. The above proof can be simplified by using $\tilde{p}_\omega=\frac{p_\omega}{q_\omega}$ if $q_\omega>0$ for all ω. However, the case of $q_\omega=0$ plays an important role in the classification presented in Section 5.

Proof of Proposition 5. The conjugate ϕ^* is a nondecreasing convex function, and $h_\omega(\mathbf{x})-\mu$ is a convex function over $\mathbf{x}\in X$ and μ. Therefore, their composition $\phi^*\left(h_\omega(\mathbf{x})-\mu\right)$ is convex. We consider $\lambda>0$ and $\lambda=0$ cases separately. Because the perspective of a convex function is convex, $\lambda\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{\lambda}\right)$ is convex over $\lambda>0$, $\mathbf{x}\in X$ and μ. When $\lambda=0$, by definition, $0\phi^*\left(\frac{b}{0}\right)=0$ for $b\leq 0$, $0\phi^*\left(\frac{b}{0}\right)=+\infty$ for $b>0$. This is a nondecreasing convex function; so by the same arguments, we obtain that $0\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{0}\right)$ is convex over $\lambda=0,\mu$ and $\mathbf{x}\in X$. Consider $\mathbf{x}_1\in X$, $\mathbf{x}_2\in X$, μ_1,μ_2, and cases (i) $\lambda_1=\lambda_2=0$, (ii) $\lambda_1=0$, $\lambda_2>0$, (iii) $\lambda_1>0$, $\lambda_2=0$, and (iv) both $\lambda_1,\lambda_2>0$. Let $a\in[0,1]$ and define $\bar{\mathbf{x}}=a\mathbf{x}_1+(1-a)\mathbf{x}_2$, $\bar{\mu}=a\mu_1+(1-a)\mu_2$, and $\bar{\lambda}=a\lambda_1+(1-a)\lambda_2$. Then, using the above definitions and convexity arguments, one can show that for all $a\in[0,1]$, $\bar{\lambda}\phi^*\left(\frac{h_\omega(\bar{\mathbf{x}})-\bar{\mu}}{\bar{\lambda}}\right)\leq a\lambda_1\phi^*\left(\frac{h_\omega(\mathbf{x}_1)-\mu_1}{\lambda_1}\right)+(1-a)\lambda_2\phi^*\left(\frac{h_\omega(\mathbf{x}_2)-\mu_2}{\lambda_2}\right)$. □

Appendix B: Derivations of ϕ-Divergences in Section 5.4

Proof of Example 1 (CVaR). Note that ϕ_{CVaR} only admits two distance values: 0 or ∞. Thus any choice of $\rho<\infty$ is equivalent to $\rho=0$. The conjugate of ϕ_{CVaR} is

$$\phi^*_{\text{CVaR}}(s)=\begin{cases}0 & s<0\\ \frac{1}{1-\beta}s & s\geq 0,\end{cases}$$

or equivalently $\phi^*_{\text{CVaR}}(s)=\max\left\{0,\frac{1}{1-\beta}s\right\}$. This results in the dual problem

$$\begin{aligned}\min_{\lambda\geq 0,\mu}\mu+\rho\lambda+\lambda\sum_\omega q_\omega\phi^*_{\text{CVaR}}\left(\frac{h_\omega(\mathbf{x})-\mu}{\lambda}\right)&=\min_{\lambda\geq 0,\mu}\mu+0\lambda+\lambda\sum_\omega q_\omega\max\left\{0,\frac{1}{1-\beta}\frac{h_\omega(\mathbf{x})-\mu}{\lambda}\right\}\\&=\min_\mu\mu+\frac{1}{1-\beta}\sum_\omega q_\omega\max\left\{0,h_\omega(\mathbf{x})-\mu\right\}\\&=\min_\mu\mu+\frac{1}{1-\beta}\mathbb{E}\left[[\mathbf{h}(\mathbf{x})-\mu]^+\right],\end{aligned}$$

which is one definition of $\text{CVaR}_\beta(\mathbf{h}(\mathbf{x}))$. Observe that selecting $\lambda=0$ in the above problem results in the objective $\sup_\omega h_\omega(\mathbf{x})$. Because $\text{CVaR}_\beta(h(\mathbf{x}))\leq\sup_\omega h_\omega(\mathbf{x})$, it is optimal to pick $\lambda>0$. □

Proof of Example 2 (Convex Combination of Expectation and Worst-Case). Without loss of generality, let $\rho = 0$. The conjugate of ϕ_{EW} is

$$\phi^*_{\mathrm{EW}}(s) = \begin{cases} (1-\beta)s & s \le 0 \\ \infty & s > 0. \end{cases}$$

This gives the dual problem

$$\begin{aligned}\min_{\lambda\ge 0,\mu} \mu + \rho\lambda + \lambda \sum_\omega q_\omega \phi^*_{\mathrm{EW}}\left(\frac{h_\omega(\mathbf{x}) - \mu}{\lambda}\right) &= \min_{\mu \ge \sup_\omega h_\omega(\mathbf{x})} \mu + \sum_\omega q_\omega (1-\beta)(h_\omega(\mathbf{x}) - \mu) \\ &= \min_{\mu \ge \sup_\omega h_\omega(\mathbf{x})} \beta\mu + (1-\beta)\sum_\omega q_\omega h_\omega(\mathbf{x}) \\ &= \beta \sup_\omega h_\omega(\mathbf{x}) + (1-\beta)\mathbb{E}\left[\mathbf{h}(\mathbf{x})\right]. \quad \square\end{aligned}$$

Proof of Example 3 (Convex Combination of Expectation and CVaR). Without loss of generality, let $\rho = 0$. The conjugate of ϕ_{EC} is

$$\phi^*_{\mathrm{EC}}(s) = \begin{cases} (1-\alpha)s & s < 0 \\ \frac{1}{1-\beta}s & s \ge 0. \end{cases}$$

Noting that $0 < 1-\alpha < 1 < \frac{1}{1-\beta}$, we can rewrite the conjugate as $\phi^*_{\mathrm{EC}}(s) = \max\left\{(1-\alpha)s, \frac{1}{1-\beta}s\right\} = (1-\alpha)s + \left(\frac{1}{1-\beta} - (1-\alpha)\right)[s]^+$. This gives the dual problem

$$\begin{aligned}&\min_{\lambda\ge 0,\mu} \mu + \rho\lambda + \lambda\sum_\omega q_\omega \phi^*\left(\frac{h_\omega(\mathbf{x}) - \mu}{\lambda}\right) \\ =&\min_\mu \mu + \sum_\omega q_\omega \max\left\{(1-\alpha)(h_\omega(\mathbf{x}) - \mu), \frac{(h_\omega(\mathbf{x}) - \mu)}{1-\beta}\right\} \\ =&\min_\mu \mu + (1-\alpha)\mathbb{E}\left[\mathbf{h}(\mathbf{x}) - \mu\right] + \left(\frac{1}{1-\beta} - (1-\alpha)\right)\mathbb{E}\left[[\mathbf{h}(\mathbf{x}) - \mu]^+\right] \\ =&(1-\alpha)\mathbb{E}\left[\mathbf{h}(\mathbf{x})\right] + \alpha \min_\mu \left\{\mu + \left(1 - \frac{\beta}{\alpha(1-\beta)+\beta}\right)^{-1} \mathbb{E}\left[[\mathbf{h}(\mathbf{x}) - \mu]^+\right]\right\}.\end{aligned}$$

Then, using the definition of CVaR, we obtain $(1-\alpha)\mathbb{E}\left[\mathbf{h}(\mathbf{x})\right] + \alpha\mathrm{CVaR}_{\frac{\beta}{\alpha(1-\beta)+\beta}}[\mathbf{h}(\mathbf{x})]$. □

References

Artzner, Philippe, Freddy Delbaen, Jean-Marc Eber, David Heath. 1999. Coherent measures of risk. *Mathematical Finance* **9**(3) 203–228.

Bayraksan, G., D. K. Love. 2015. Data-driven stochastic programming using phi-divergences. *INFORMS Tutorials in Operations Research*. the Institute for Operations Research and the Management Sciences, 5521 Research Park Drive, Suite 200, Catonsville, Maryland 21228, USA, 1–19. Published online: 26 Oct 2015; 1-19. http://dx.doi.org/10.1287/educ.2015.0134.

Ben-Tal, A., A. Ben-Israel, M. Teboulle. 1991. Certainty equivalents and information measures: duality and extremal principles. *Journal of Mathematical Analysis and Applications* **157**(1) 211–236.

Ben-Tal, A., D. Den Hertog, A. De Waegenaere, B. Melenberg, G. Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* **59** 341–357.

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

Bertsimas, D., V. Gupta, N. Kallus. 2014. Data-driven robust optimization. Tech. rep., Massachusetts Institute of Technology. Available at arXiv: 1401.0212v2.

Calafiore, G.C. 2007. Ambiguous risk measures and optimal robust portfolios. *SIAM Journal on Optimization* **18**(3) 853–877.

Delage, E., Y. Ye. 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* **58**(3) 595–612.

Dupačová, J. 1987. The minimax approach to stochastic programming and an illustrative application. *Stochastics* **20** 73–88.

Erdoğan, E., G. Iyengar. 2006. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming* **107**(1-2) 37–61.

Esfahani, P.M., D. Kuhn. 2015. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. Tech. rep., Available at Optimization Online.

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

Hanasusanto, G. A., V. Roitch, D. Kuhn, W. Wiesemann. 2015. A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming* doi: 10.1007/s10107-015-0896-z. Published online before print.

Hu, Z., L. J. Hong. 2013. Kullback-Leibler divergence constrained distributionally robust optimization. Tech. rep., The Hong Kong University of Science and Technology. Available at: Optimization Online `www.optimization-online.org`.

Infanger, G. 1992. Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs. *Annals of Operations Research* **39**(1) 69–95.

Jiang, R., Y. Guan. 2015a. Data-driven chance constrained stochastic program. *Mathematical Programming* 1–37Published online before print.

Jiang, R., Y. Guan. 2015b. Risk-averse two-stage stochastic program with distributional ambiguity. Tech. rep., Available on Optimization Online.

Klabjan, D., D. Simchi-Levi, M. Song. 2013. Robust stochastic lot-sizing by means of histograms. *Production and Operations Management* **22**(3) 691–710.

Love, D., G. Bayraksan. 2014. A Classification of Phi-Divergences for Data-Driven Stochastic Optimization. *Proceedings of the 2014 Industrial and Systems Engineering Research Conference*.

Love, D., G. Bayraksan. 2015. A data-driven method for robust water allocation under uncertainty. Tech. rep., The Ohio State University.

Love, David. 2013. Data-driven methods for optimization under uncertainty with application to water allocation. Ph.D. Dissertation, University of Arizona.

Mehrotra, S., D. Papp. 2014. A cutting surface algorithm for semi-infinite convex programming, with an application to distributionally robust optimization. *SIAM Journal on Optimization* **24**(4) 1670–1697.

Nocedal, J., S. Wright. 1999. *Numerical Optimization*. Springer Verlag, New York, NY.

Pardo, L. 2005. *Statistical Inference Based On Divergence Measures*, vol. 185. Chapman and Hall/CRC.

Pflug, G., D. Wozabal. 2007. Ambiguity in portfolio selection. *Quantitative Finance* **7**(4) 435–442.

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

Rockafellar, R.T. 2007. Coherent approaches to risk in optimization under uncertainty. T. Klastorin, ed., *Tutorials in Operations Research*, vol. 3. INFORMS, Hanover, MD, 38–61.

Scarf, H. 1958. A min-max solution of an inventory problem. K.J. Arrow, S. Karlin, H. Scarf, eds., *Studies in the Mathematical Theory of Inventory and Production*. Stanford University Press, Stanford, CA, 201–209.

Shapiro, A. 2003. Monte Carlo sampling methods. A. Ruszczyński, A. Shapiro, eds., *Handbooks in Operations Research and Management Science, Volume 10: Stochastic Programming*. Elsevier, 353–425.

Shapiro, A., S. Ahmed. 2004. On a class of minimax stochastic programs. *SIAM Journal on Optimization* **14**(4) 1237–1249.

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

Shapiro, A., A. Kleywegt. 2002. Minimax analysis of stochastic problems. *Optimization Methods and Software* **17**(3) 523–542.

Wang, Z., P.W. Glynn, Y. Ye. 2015. Likelihood robust optimization for data-driven problems. *Computational Management Science* 1–21doi:10.1007/s10287-015-0240-3. Published online before print.

Wiesemann, W., D. Kuhn, M. Sim. 2014. Distributionally robust convex optimization. *Operations Research* **62**(6) 1358–1376.

Yanikoglu, Ihsan, Dick den Hertog. 2012. Safe approximations of ambiguous chance constraints using historical data. *INFORMS Journal on Computing* **25**(4) 666–681.

Žáčková, J. 1966. On minimax solutions of stochastic linear programming problems. *Časopis pro Pěstování Matematiky* **91**(4) 423–430.

Zhao, C., Y. Guan. 2015. Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics. Tech. rep., Available on Optimization Online.