Phi-Divergence Constrained Ambiguous Stochastic Programs

David Love
University of Arizona, dlove@email.arizona.edu, http://math.arizona.edu/~dlove/

Güzin Bayraksan
The Ohio State University, bayraksan.1@osu.edu, http://www-iwse.eng.ohio-state.edu/biosketch_GBayraksan.cfm

This paper investigates the properties of ϕ-divergence constrained ambiguous stochastic programs. ϕ-divergences (e.g., Kullback-Leibler, χ^2 distance, etc.) provide a natural way to create an ambiguity set of distributions that are centered around a nominal distribution, which is often determined by collected observations. We present a classification of ϕ-divergences to elucidate their use for models with different properties and different sources of data. A condition of assessing the value of collecting additional data is derived and we demonstrate convergence of the ϕ-divergence-based ambiguous program to the associated non-ambiguous stochastic program. A decomposition-based solution algorithm to solve the resulting model is given.

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

1. Introduction and Motivation

In practice, 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. Such problems have led to the development of distributionally robust optimization, a modeling technique that replaces the probability distribution by a set of distributions, and optimizes the expected cost relative to the worst distribution in the uncertainty set. One approach that has been recently proposed by Ben-Tal et al. (2013) uses a set of distributions that have sufficiently small ϕ-divergence from a given "nominal" distribution (ϕ-divergences provide a measure of distance between two distributions).

Of particular interest is the case when the nominal distribution is determined by observation by making it an empirical distribution. In this paper, we adapt the ϕ-divergence method to the setting of a two-stage stochastic linear program with recourse and call this the two-stage ϕ-divergence constrained ambiguous stochastic linear program with recourse (ϕLP-2).

Using ϕ-divergences to model ambiguous probability distributions is an attractive data-driven approach because it uses the data directly—only those data points or scenarios of interest are used in the calculations. These scenarios can come from direct observation, results of simulation, or from expert opinion that the decision maker would especially like to be robust against. Because the ϕLP-2 depends only on these scenarios, the size of the problem is polynomial in the sample size, making the ϕLP-2 computationally tractable. Furthermore, many ϕ-divergences are commonly used in statistics (e.g, the χ^2 distance) and provide a natural way of modeling problems under uncertainty.

1.1. Related Literature

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

In recent years, there has been a growing interest in distributionally robust methods. Erdoğan and Iyengar (2006) study chance-constrained stochastic programs where the set of distributions considered is determined by the Prohorov metric. Calafiore and Campi (2005) develop a data-driven method for generating feasible solutions to chance constrained problems, and later Calafiore and Ghaoui (2006) develop a method for converting distributionally robust chance constraints into second-order cone constraints. Jiang and Guan

(2013) develop an exact approach to solving data-driven chance constrained programs. Pflug and Wozabal (2007) develop a data-driven method for solving a portfolio selection problem using the Kantorovich distance to define the set of distributions. Delage and Ye (2010) provide methods for modeling uncertain distributions of a specific form (e.g., Gaussian, exponential, etc.) or using moment-based constraints.

Three recent papers by Wang et al. (2010), Calafiore (2007), and Hu and Hong (2013) provide similar studies using a specific ϕ-divergence, described in Section 2, that is defined by the Kullback-Leibler divergence. Both Wang et al. (2010) and Hu and Hong (2013) produce dual problems similar to that presented in Ben-Tal et al. (2013) and used here. Hu and Hong (2013) differs from this work and the others by considering a continuous distributions, but doesn't relate the nominal distribution to observational data. Additionally, Klabjan et al. (2013) uses the χ^2 distance, another ϕ-divergence, to define an uncertain demand distribution for a stochastic lot-sizing problem using historical data. Our work unites these previous papers and provides insight into conditions where each ϕ-divergence should be used and the behavior of the optimization problem as more data is gathered.

1.2. Contributions

The contributions of this paper are as follows.

- One of the open problems identified by Ben-Tal et al. (2013) was to study the performance of different ϕ-divergences. 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. We provide a novel classification of ϕ-divergences dictated by the types of distributions that can be admitted to the set of distributions. This allows us to provide insight into which class of ϕ-divergence is most useful to which type of model. We note that this classification is a general feature of ϕ-divergences

and applies to a broader class of ϕ-divergence constrained problems than the ones presented in this paper.

- 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? In this paper, we provide answers to these questions. First, we provide a simple condition to determine if sampling from a particular scenario will rule out the current worst-case distribution, which can be generalized beyond the two-stage setting. Second, we show that ϕLP-2 converges to the stochastic program with the (unknown) true distribution.
- Stochastic programs often become quite large, which raises questions of computational tractability. We help to answer this problem by providing a modified Bender's decomposition that can be used to solve the ϕLP-2 efficiently by solving only linear programs.
- Finally, we provide some interesting examples of ϕ-divergences that result in commonly used risk models and illustrate our results numerically.

1.3. Organization

The rest of the paper is organized as follows. Section 2 introduces the ϕ-divergence and presents some useful properties. Section 3 presents the derivation of ϕ-divergence model for a two-stage stochastic program with recourse. Section 4 presents a classification of ϕ-divergences with some examples; Sections 5 describes the data-driven properties of ϕLP-2; Section 6 presents a decomposition method for solving the ϕLP-2 model; and in Section 7 we present numerical illustrations of our results. We end in Section 8 with conclusions and future work. All proofs are provided in the Appendix.

2. Introduction to ϕ-Divergences

In this section we define the concept of a ϕ-divergence, and describe some of the properties that will be used through the remainder of the paper. Pardo (2005) provides a good overview of much of the known properties of ϕ-divergences. Many results in this section can be also found in (Ben-Tal et al. 2013).

ϕ-divergences are used in statistics to measure the "distance" between two distributions. In the discrete case, ϕ-divergences can be used generally to measure the distance between two non-negative vectors $p = (p_1, \ldots, p_n)^T$ and $q = (q_1, \ldots, q_n)^T$, and specifically when p and q are probability vectors; i.e., satisfying $\sum_{\omega=1}^n p_\omega = \sum_{\omega=1}^n q_\omega = 1$. The ϕ-divergence is defined by

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

where $\phi(t)$, called the ϕ-divergence function, is a convex function on $t \geq 0$ such that $\phi(1) = 0$, and with the additional interpretations that $0\phi(a/0) = a \lim_{t\to\infty} \frac{\phi(t)}{t}$, and $0\phi(0/0) = 0$. If both p and q are probability vectors, as we assume throughout this paper, we can additionally assume without loss of generality that $\phi(t) \geq 0$. 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, and $I_\psi(p, q) = I_\phi(p, q)$ for all probability vectors p, q. If $\phi(t)$ is differentiable at $t = 1$ this can be done by selecting $c = -\phi'(1)$.

ϕ-divergences are not, in general, metrics. For example, most ϕ-divergences do not satisfy the triangle inequality and many are not symmetric in the sense that $I_\phi(p, q) \neq I_\phi(q, p)$. One exception is the Variation distance, which is equivalent to the L^1-distance between the vectors.

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 $I_{\tilde{\phi}}(p,q) = I_{\phi}(q,p)$. Divergences that are symmetric with respect to the input vectors are known as self-adjoint.

The problem formulation involves use of the conjugate $\phi^* : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, defined as

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

The conjugate ϕ^* is a nondecreasing convex function and may be undefined above some upper bound $\bar{s}$.

Table 1 lists some common examples of ϕ-divergences, along with their adjoints and conjugates. For all divergences, $\phi(t) = \infty$ for $t < 0$, and the value of the conjugate is listed only in its domain; i.e., $\{s : \phi^*(s) < \infty\}$. Most of these common divergences are widely used in statistics and information theory. In Section 4.5, we present other ϕ-divergences that assign a distance of either 0 or ∞, which result in commonly used risk models. Table 1 also 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 does not satisfy the normalizing condition $\phi(t) \geq 0$. This divergence is included because Wang et al. (2010) use it to formulate a distributionally robust newsvendor problem so that the ambiguity set of distributions have a sufficiently high empirical likelihood. They refer to this as the Likelihood Robust Optimization. We also note that Calafiore (2007), Hu and Hong (2013), and Wang et al. (2010) all use a different naming convention than the one given here, referring to the Likelihood divergence or Burg entropy as the "Kullback-Leibler (KL) divergence"—reversing the order of the arguments p and q relative to the notation presented here. In this paper, q denotes the nominal distribution.

3. ϕ-Divergence Constrained Ambiguous Stochastic Program

In this section we provide primal and dual formulations and basic properties of two-stage ambiguous stochastic linear programs constructed via ϕ-divergences.

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}	$\phi_{m\chi^2}$	$\frac{1}{t}(t-1)^2$	$\sum \frac{(p_\omega - q_\omega)^2}{p_\omega}$	$2-2\sqrt{1-s},\ s<1$
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	$(\sqrt{t}-1)^2$	$\sum(\sqrt{p_\omega}-\sqrt{q_\omega})^2$	$\frac{s}{1-s},\ s<1$

3.1. 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 $\mathbf{b}$. We assume a finite distribution given by q_ω with scenarios indexed by $\omega = 1,\ldots,n$. The SLP-2 is

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

where

$$h_\omega(\mathbf{x}) = \min_{\mathbf{y}^\omega} \left\{ \mathbf{k}^\omega \mathbf{y}^\omega : \mathbf{D}^\omega \mathbf{y}^\omega = \mathbf{B}^\omega \mathbf{x} + \mathbf{d}^\omega, \mathbf{y}^\omega \geq 0 \right\}. \tag{4}$$

We assume relatively complete recourse; i.e., the second-stage problems $h_\omega(\mathbf{x})$ are feasible for every feasible solution $\mathbf{x}$ of the first-stage problem; and that the second-stage problems $h_\omega(\mathbf{x})$ are dual feasible for every feasible solution $\mathbf{x}$ of the first-stage problem. For convenience, we denote $X = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$.

The SLP-2 formulation assumes that the distribution $\{q_\omega\}_{\omega=1}^n$ is known. However, in many applications the distribution is itself unknown. One technique to deal with this is to replace the known distribution with an *ambiguity set* of distributions; i.e., a set of

distributions that is believed to contain the true distribution. In this paper, we construct the ambiguity set by considering all distributions whose ϕ-divergence from the nominal distribution q is sufficiently small. Throughout portions of this paper, we focus on a data-driven setting and assume that q is generated from observations, where scenario ω has been observed N_ω times, with $N = \sum_{\omega=1}^{n} N_\omega$ total observations, although q can be obtained in other ways. In SLP-2, this data-driven setting would correspond to the probability of scenario ω to be $q_\omega = \frac{N_\omega}{N}$.

By replacing the specific distribution in SLP-2 with a set of distributions sufficiently close to the nominal distribution with respect to 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

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

where the ambiguity set is

$$\mathcal{P} = \left\{ \sum_{\omega=1}^{n} q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right) \le \rho, \tag{6}$$

$$\sum_{\omega=1}^{n} p_\omega = 1, \tag{7}$$

$$p_\omega \ge 0,\ \forall \omega \right\}. \tag{8}$$

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

Taking the 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

$$\min_{\mathbf{x},\lambda,\mu} \mathbf{cx} + \mu + \rho\lambda + \lambda \sum_{\omega=1}^{n} q_\omega \phi^*\left(\frac{h_\omega(\mathbf{x}) - \mu}{\lambda}\right) \tag{9}$$

$$\begin{aligned} \text{s.t. } & \mathbf{x} \in X \\ & \frac{h_\omega(\mathbf{x}) - \mu}{\lambda} \leq \lim_{t\to\infty} \frac{\phi(t)}{t}, \ \forall \omega \\ & \lambda \geq 0, \end{aligned} \tag{10}$$

where $h_\omega(\mathbf{x})$ and the second-stage problems are as given in (4), $0\phi^*(s/0) = 0$ if $s \leq 0$ and $0\phi^*(s/0) = +\infty$ if $s > 0$. Note that some ϕ, such as the J-Divergence, have no closed form representation of ϕ^*, but can be expressed as the sum of other divergences—Burg Entropy and KL divergence—which allows the dual to be formed; 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 1. Note in particular that the dual formulation is accurate even for $q_\omega = 0$ for some ω. Also note that the right-hand side of (10) contains a limit. For some ϕ-divergences, like the KL divergence, this limit is ∞, in which case (10) is redundant. However, other ϕ-divergences, like the Hellinger distance, have a finite limit, inducing this constraint. Throughout the paper, we use s_ω to denote

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

3.2. Basic Properties

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

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. The worst-case distribution can be calculated with the equations

$$\frac{p_\omega}{q_\omega} \in \partial\phi^*(s_\omega), \qquad \sum_{\omega=1}^{n} q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right) \leq \rho, \qquad \sum_{\omega=1}^{n} p_\omega = 1. \tag{12}$$

We now discuss these properties. A coherent risk measure is defined in (Rockafellar 2007), which shows that minimizing a coherent risk measure over a polyhedron implies that ϕLP-2 is a convex problem. The convexity of ϕLP-2 was also noted in (Ben-Tal et al. 2013). In Section 4.5, we present special ϕ-divergences that result in CVaR, or a convex combination of expectation with CVaR or the worst-case scenario. Properties 3 and 4 help with our decomposition-based solution method described in Section 6. The preservation of time structure, as can be seen in (9), allows us to decompose the problem and convert (sub-)derivatives of $h_\omega(\mathbf{x})$ to (sub-)derivatives of $\phi^*(s_\omega)$, aiding in our decomposition-based solution method. The appearance of the conjugate $\phi^*(s)$ in the objective of (9) gives a method for retrieving the worst-case distribution from the dual problem, as detailed in Property 4. In many cases, the first equation in (12) is sufficient to calculate $\{p_\omega\}_{\omega=1}^n$. In addition, ϕ^* is often differentiable and so we have the relationship $p_\omega = q_\omega \phi^{*\prime}(s_\omega)$. Special cases when $\lambda = 0$ or $q_\omega = 0$ for some ω are detailed in Section 4.

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. When ϕ is twice continuously differentiable around 1 with $\phi''(1) > 0$, Theorem 3.1 of Pardo (2005) shows that the statistic $T_N^\phi(q^N, q^{\text{true}}) = \frac{2N}{\phi''(1)} \sum_{\omega=1}^n q_\omega^{\text{true}} \phi\left(\frac{q_\omega^N}{q_\omega^{\text{true}}}\right)$ converges in distribution to a χ^2-distribution with $n-1$ degrees of freedom, where q^N denotes the empirical distribution ($q_\omega^N = N_\omega/N$), and q^{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

$$\rho = \frac{\phi''(1)}{2N} \chi^2_{n-1,1-\alpha}, \tag{13}$$

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

We are now ready to present the main contributions of this paper.

4. 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 $t \to 0$ and $t \to \infty$. 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 4.2. We also provide some special ϕ-divergences that result in common risk models used in the literature and discuss their behavior with respect to our classification in Section 4.5.

4.1. Suppressing and Popping of Scenarios

As motivation for our classification, consider a self-adjoint ϕ-divergence, which satisfies the relation

$$\frac{\phi(t)}{t} = \phi\left(\frac{1}{t}\right), \tag{14}$$

and consider $t \to \infty$. If both sides of (14) are finite in the limit, then we see a correspondence between the boundedness of $\phi(t)$ for $t < 1$ and linear growth of $\phi(t)$ for $t > 1$. On the other hand, infinite limits of (14) indicate a correspondence between superlinear growth of $\phi(t)$ for $t > 1$ and unboundedness of $\phi(t)$ for $t < 1$.

Recall the definition of the ambiguity set, in particular, the ϕ-divergence constraint (6). In the ϕLP-2, ϕ has arguments given by ratios of probabilities, $\frac{p_\omega}{q_\omega}$ and the limits $t \to 0$ and $t \to \infty$ correspond to the cases when $p_\omega = 0$ and $q_\omega = 0$, respectively. Consider each of these limiting cases:

- Case 1: $q_\omega > 0$ but $p_\omega = 0$. We call this the "**Suppress**" behavior because a scenario with a positive probability in the nominal distribution can take zero probability in the ambiguous problem. In this case we need to examine $\lim_{t \searrow 0} \phi(t)$:
  - If $\lim_{t \searrow 0} \phi(t) = \infty$, the ambiguity region will never contain distributions with $p_\omega = 0$ but $q_\omega > 0$.
  - On the other hand, if $\lim_{t \searrow 0} \phi(t) < \infty$, the ambiguity region could contain such a distribution, provided q_ω is sufficiently small or ρ is sufficiently large. We say that such a ϕ-divergence can *suppress* scenario ω.
- Case 2: $q_\omega = 0$ but $p_\omega > 0$. We call this the "**Pop**" behavior because a scenario with zero probability in the nominal distribution can have a positive probability (or, pop) in the ambiguous problem. In this case, we need to examine $\lim_{t \nearrow 0} \frac{\phi(t)}{t}$:
  - If $\lim_{t \nearrow 0} \frac{\phi(t)}{t} = \infty$, the ambiguity region can never contain distributions with $p_\omega > 0$ but $q_\omega = 0$.
  - On the other hand, if $\lim_{t \nearrow 0} \frac{\phi(t)}{t} < \infty$, the ambiguity region will admit sufficiently small p_ω. We say that these ϕ-divergences can *pop* scenario ω.
- Case 3: $p_\omega = 0$ but $q_\omega = 0$. Such a situation has no contribution to the divergence, since $0\phi\left(\frac{0}{0}\right) = 0$.

These two limiting cases describing suppressing and popping behavior in ϕ-divergences create four distinct categories. Examples of divergences in each category are given in Table 2. Note that ϕ can suppress scenarios if and only if its adjoint $\tilde{\phi}$ can pop scenarios, and vice versa. This means that self-adjoint ϕ-divergences are either capable of both popping and suppressing scenarios or capable of neither.

4.2. Modeling Considerations When Choosing a Divergence

We offer the following suggestions for choosing an appropriate ϕ-divergence classification for the data available. First, consider whether to choose a distribution that can suppress

Table 2 **Examples of ϕ-divergences fitting into each category. The number in parentheses under the "Can Suppress Scenarios" column denotes the subcategory detailed in Section 4.3.**

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

scenarios. If the problem scenarios come from high-quality observed data, one may wish to avoid divergences that can suppress scenarios. 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.

Next, consider whether to choose a distribution that allows for popping scenarios. If the problem scenarios come strictly from observation, with little theoretical understanding of the problem, we suggest choosing a divergence that cannot pop scenarios. However, if the problem 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, allowing the mathematical program to assign an appropriate probability to them.

4.3. Additional Details about Divergences that can Suppress

Recall the primal-dual variable relation, which specifies $\frac{p_\omega}{q_\omega} = \partial\phi^*(s_\omega)$, where $s_\omega = \frac{h_\omega(\mathbf{x})-\mu}{\lambda}$. Note that suppression ($p_\omega = 0$, $q_\omega > 0$) can occur only when $0 \in \partial\phi^*(s_\omega)$. For convenience, we assume ϕ^* is differentiable. We can examine suppressing in more detail by looking at the behavior of $\phi(t)$ as $t \searrow 0$. This analysis yields two subcategories within the ϕ-divergences that can suppress scenarios—one tends to suppress scenarios one at a time and the other simultaneously.

- SUBCATEGORY 1 ($\lim_{t\searrow 0} \phi'(t) > -\infty$). There are nonpositive constants $c, \underline{s}$ such that $\phi^*(s) = c$ 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 that satisfy the relation $\frac{h_\omega(\mathbf{x})-\mu}{\lambda} < \underline{s}$ are suppressed. As ρ increases, scenarios tend to be suppressed one at a time.

- SUBCATEGORY 2 ($\lim_{t\searrow 0}\phi'(t) = -\infty$). In this case, $\phi^*(s) \searrow c$ as $s \to -\infty$ asymptotically, but never reaches the bound. As a result, scenarios can only be suppressed if $s_\omega = -\infty$, which can only occur if $\lambda = 0$ and $h_\omega(\mathbf{x}) < \mu$. Consequently, all solutions 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 p is a probability distribution. This means that all but the most expensive scenario(s) will vanish simultaneously. Divergences of this type can be difficult to deal with numerically when suppression occurs given the denominator of $\lambda = 0$ (see Section 6.4 for details).

Table 2 lists ϕ-divergences that belong to these subcategories. We present the one-by-one and simultaneous suppressing behavior numerically for the Modified χ^2-Distance and KL Divergence, respectively, in Section 7.

4.4. Additional Details about 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 $\phi^*(s)$. The primal-dual variable relation specifies $\frac{p_\omega}{q_\omega} = \partial\phi^*(s_\omega)$, but the left-hand side is undefined when $q_\omega = 0$. Intuitively, we can think of $\frac{p_\omega}{0} = \infty$ if $p_\omega > 0$, and thus popping a scenario can only occur when the right-hand side subdifferential also includes ∞. This, in turn, occurs only when $s_\omega = \bar{s}$. The next proposition makes this statement rigorous.

PROPOSITION 1. *Suppose there is a finite* $\bar{s} = \lim_{t\to\infty}\frac{\phi(t)}{t}$. *A scenario* ω *for which* $q_\omega = 0$ *can only be popped if* $s_\omega = \bar{s}$.

REMARK 1. Because $s_\omega \le \bar{s}$ for all ω and $s_\omega = \bar{s}$ for any popped scenarios, only the most expensive scenario could be popped.

REMARK 2. Finding the probability of the popped scenario cannot be done by differentiating ϕ^* as with other scenarios, thus the probability must be calculated with $\sum_\omega p_\omega = 1$.

4.5. Some Special ϕ-Divergences

The class of ϕ-divergence constrained problems includes some interesting special cases, which we document here, followed by a discussion of their suppressing and popping behavior.

EXAMPLE 1. (CVaR). The coherent risk measure Conditional Value-at-Risk (CVaR) is well studied in financial applications. Minimizing

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

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

$$\phi(t) = \begin{cases} 0 & 0 \le t \le \frac{1}{1-\beta} \\ \infty & \text{otherwise,} \end{cases}$$

for $0 < \beta < 1$. We see that $\phi(0) = 0$, indicating that CVaR will suppress some scenarios. This appears in the definition of CVaR as the positive part in the expected value, $\mathbb{E}\,[[h(\mathbf{x}) - \mu]^+]$. Scenarios cannot be popped because the expectation is taken with respect to the nominal distribution.

The CVaR ϕ-divergence is bounded above, which leads to the question of what happens when a divergence is bounded below.

EXAMPLE 2 ("REVERSE" CVAR). The ϕ-divergence constrained problem with

$$\phi(t) = \begin{cases} 0 & t \ge 1-\beta \\ \infty & t < 1-\beta, \end{cases}$$

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

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

over $\mathbf{x} \in X$, where the expectation is taken with respect to the nominal distribution q. Note that $\lim_{t\to\infty} \frac{\phi(t)}{t} = 0$, indicating that this divergence will pop scenarios. This behavior appears in the term $\sup_\omega h_\omega(\mathbf{x})$. However, $\phi(0) = \infty$ indicates that scenarios will not be suppressed, which is demonstrated by the expectation term $\mathbb{E}\left[h(\mathbf{x})\right]$, which 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.

Example 3 (Conbination CVaR and Expectation). The ϕ-divergence constrained problem with

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

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

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

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. Data-Driven Considerations

In this section we assume the nominal distribution q is the empirical distribution ($q_\omega = \frac{N_\omega}{N}$) and provide insight into how the ϕLP-2 changes as data is added: first how it might change

with a single additional observation in Section 5.1, then as more and more data is gathered with asymptotic results in Section 5.2. This analysis must consider how ρ changes as additional samples are taken; therefore, we use ρ_N to emphasize the dependence on sample size in this section. To be consistent with the known ϕ-divergence results stated in Section 3.3, we assume $\rho_N = \frac{\rho_0}{N}$.

5.1. The Value of Data

With a data-driven formulation such as ϕLP-2, it is natural to ask how the optimal value and solution changes as more data is gathered. In particular, 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 because new information removes the current worst-case distribution from the ambiguity set. Below, we present a simple method of determining if taking an additional sample will eliminate the old worst-case distribution and allow for better optimization; i.e., a lower-cost solution.

THEOREM 1. *An additional sample of scenario $\hat{\omega}$ will result in a decrease in the worst-case expected cost of the ϕLP-2 if the following condition is satisfied*

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

where $s^*_\omega = \dfrac{h_\omega(\mathbf{x}^*_N) - \mu^*_N}{\lambda^*_N}$ and $(\mathbf{x}^*_N, \mu^*_N, \lambda^*_N)$ solve the N-sample problem with $q_\omega = \frac{N_\omega}{N}$.

We can interpret (15) as follows. If an additional sample is taken from the unknown distribution and the resulting observed scenario $\hat{\omega}$ satisfies (15), then the $(N+1)$-sample problem will have a lower cost than the N-sample problem that was already solved. This

is equivalent to saying that an additional observation of $\hat{\omega}$ will rule out the computed worst-case distribution given by $\{p_\omega\}_{\omega=1}^n$ in (12).

It is possible to simplify the condition in (15) for some ϕ-divergences and we detail this in the corollary below.

COROLLARY 1. *An additional sample of scenario $\hat{\omega}$ will result in a decrease in the worst-case expected cost of the ϕLP-2 if the following condition is satisfied for:*

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

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

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

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

The simple conditions in Theorem 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 and therefore can be considered "good" 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

$$\min_r \left\{ \sum_{\omega \in L} r_\omega : \ r \in \mathcal{P} \right\}. \tag{16}$$

That is, we find the minimum probability of L within the ambiguity set defining ϕLP-2 (since we do not know the true distribution). We solve (16) by taking its dual.

5.2. Asymptotic Analysis

We now wish to show that ϕLP-2 behaves essentially the same as the corresponding SLP-2 with the (unknown) true distribution q^{true} as $N \to \infty$. This requires that the sequence

of nominal distributions converge to the true distribution q^{true} in L^∞, a situation that is satisfied by the assumed empirical distribution. To emphasize the dependence of N in the nominal distribution used, we prefer to use q^N in this section.

We begin by showing that the worst-case distribution obtained by solving the ϕLP-2 converges weakly to the true distribution as $N \to \infty$. Let $(\Xi, \mathcal{F}, \mathbb{P}^\infty)$ be the probability space associated with taking infinitely many random samples from the distribution q^{true}. Let $\Xi' \subset \Xi$ be a measure 1 set such that $\|q^N(\xi) - q^{\text{true}}\|_\infty \to 0$.

PROPOSITION 2. *Suppose $\phi(t) \geq 0$ has a unique root at $t = 1$. For all $\epsilon > 0$ and $\xi \in \Xi'$, there exists N' such that $\forall N \geq N'$, $I_\phi(p, q^N(\xi)) \leq \frac{\rho_0}{N}$ implies $\max_\omega |p_\omega - q_\omega^{true}| \leq \epsilon$.*

The requirements on ϕ in Proposition 2 are satisfied by every divergence in Table 1 except Likelihood (which, however, can be rewritten as Burg Entropy). The unique root requirement, however, is violated for the special cases introduced in Section 4.5. Proposition 2 implies that the worst-case distributions of (5) converge weakly to q^{true}, which is used to show the desired result below.

THEOREM 2. *Assume X is compact and $h_\omega(\mathbf{x})$ are primal and dual feasible for every $\mathbf{x} \in X$. Then, the optimal value of ϕLP-2 (9) converges to that of SLP-2 (3) with distribution q^{true} and all limit points of the solutions of ϕLP-2 solve SLP-2 with distribution q^{true}.*

6. A Decomposition-Based Solution Method

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 method for solving ϕLP-2. The algorithm removes feasibility constraint (10) and exchanges it with a series of feasibility cuts in the first-stage problem. The master problem is given by

$$\min_{\mathbf{x},\lambda,\mu} \mathbf{cx} + \mu + \rho\lambda + \theta \tag{17}$$

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

$$\theta \geq T_j(\mathbf{x}, \mu, \lambda)^T + t_j, \quad j \in J \tag{18}$$

$$\mu + \bar{s}\lambda \geq M_k \mathbf{x} + m_k, \quad k \in K \tag{19}$$

where (18) are the objective cuts, (19) are the feasibility cuts on constraint (10) if $\bar{s} = \lim_{t\to\infty} \frac{\phi(t)}{t} < \infty$, and J and K are the sets of objective and feasibility cuts, respectively.

The proposed algorithm is shown in Algorithm 1. The modified Bender's decomposition presented here has the following features: *(i)* it solves the original linear second stage problems, rather than nonlinear subproblems, and uses them to quickly generate subgradients of the nonlinear term $\lambda\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{\lambda}\right)$, *(ii)* exchanges the polyhedral ϕ-divergence constraints $\frac{h_\omega(\mathbf{x})-\mu}{\lambda} \leq \bar{s}$ for a (potentially much) smaller set of easily generated feasibility cuts, and thus *(iii)* maintains linear master and subproblems.

Algorithm 1. Decomposition algorithm for solving ϕLP-2

Initialize $z_l = -\infty, z_u = \infty$
Solve master problem (17) with $\theta = 0$ to generate $\mathbf{x}$
Solve all second-stage scenario subproblems (4) to obtain $h_\omega(\mathbf{x})$
Initialize $\lambda \leftarrow 1$ and μ so that $\frac{h_\omega(\mathbf{x})-\mu}{\lambda} < \bar{s}$
Generate initial objective cut
while $z_u - z_l \geq \texttt{TOL}\min\{|z_u|, |z_l|\}$ **do**
  Solve master problem (17), get $\mathbf{x}$,λ,μ,θ_M
  Solve subproblems (4) to obtain $h_\omega(\mathbf{x})$
  $\theta_{\text{true}} \leftarrow \sum_{\omega=1}^{n} q_\omega h_\omega(\mathbf{x}, \lambda, \mu)$
  if $\frac{h_\omega(\mathbf{x})-\mu}{\lambda} > \bar{s}$ **then**
    Generate feasibility cut
    Find μ so that $\frac{h_\omega(\mathbf{x})-\mu}{\lambda} < \bar{s}$
  else
    $z_l \leftarrow$ master optimal cost $\mathbf{cx} + \mu + \bar{N}\lambda + \theta_{\text{true}}$
  end if
  Generate objective cut
  if $\mathbf{cx} + \mu + \bar{N}\lambda + \theta_{\text{true}} < z_u$ **then**
    $z_u \leftarrow \mathbf{cx} + \mu + \bar{N}\lambda + \theta_{\text{true}}$
    $\mathbf{x}_{\text{best}} \leftarrow \mathbf{x}, \lambda_{\text{best}} \leftarrow \lambda, \mu_{\text{best}} \leftarrow \mu$
    $p_\omega \leftarrow \phi^{*\prime}(\frac{h_\omega(\mathbf{x})-\mu}{\lambda})$ for $\omega = 1, \ldots, n$
  end if
end while

6.1. Objective Cuts

Let $(\hat{\mathbf{x}}, \hat{\mu}, \hat{\lambda})$ be the candidate solution from the master problem (17), $\hat{s}_\omega = \frac{h(\hat{\mathbf{x}})-\mu}{\hat{\lambda}}$, and $h^\dagger_\omega(s_\omega) = \lambda\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{\lambda}\right)$ be the nonlinear portion of the objective function, which will be

used to generate the objective cuts. An objective cut can be computed by solving the SLP-2 subproblems $h_\omega(\hat{\mathbf{x}})$ to obtain the optimal dual solutions $\pi^{*,\omega}$ to each second-stage problem. Using these to compute the partial (sub)derivatives of the ϕLP-2 subproblems

$$T_j^\omega = \left(\phi^{*\prime}(\hat{s}_\omega)\pi^{*,\omega}B^\omega,\, -\phi^{*\prime}(\hat{s}_\omega),\, \phi^*(\hat{s}_\omega) - \phi^{*\prime}(\hat{s}_\omega)\hat{s}_\omega \right),$$
$$t_j^\omega = \hat{\lambda}\phi^{*\prime}(\hat{s}_\omega)\left[\hat{s}_\omega - \frac{\pi^{*,\omega}B^\omega\hat{\mathbf{x}} - \hat{\mu}}{\hat{\lambda}}\right].$$

For the single-cut master problem proposed, $T_j = \sum_\omega q_\omega T_j^\omega$ and $t_j = \sum_\omega q_\omega t_j^\omega$ and multi-cut versions are also possible.

6.2. Feasibility Cuts

After the subproblems $h_\omega(\hat{\mathbf{x}})$ are solved, it may be the case that $\hat{s}_\omega < \bar{s}$ for some ω, rendering $\hat{\mu}$ and $\hat{\lambda}$ infeasible. This is corrected using the feasibility problem

$$U_\omega(\mathbf{x},\mu,\lambda) = \min_{y^\omega \geq 0, z \geq 0} \left\{ z \,:\; z + \bar{s}\lambda + \mu - k^\omega y^\omega \geq 0,\; D^\omega y^\omega = d^\omega + B^\omega x \right\},$$

which is solved by $z = h_\omega(\mathbf{x}) - \bar{s}\lambda - \mu$. The subdifferentials can be easily found as $\frac{\partial z^*}{\partial \mathbf{x}} = \pi^{*,\omega}B^\omega$, $\frac{\partial z^*}{\partial \mu} = -1$, and $\frac{\partial z^*}{\partial \lambda} = -\bar{s}$. Then for infeasible candidate solution $(\hat{\mathbf{x}}, \hat{\lambda}, \hat{\mu})$ we get the inequality $U_\omega(\mathbf{x},\mu,\lambda) \geq \pi^{*,\omega}B^\omega(\mathbf{x} - \hat{\mathbf{x}}) - (\mu - \hat{\mu}) - \bar{s}(\lambda - \hat{\lambda}) + (h_\omega(\hat{\mathbf{x}}) - \hat{\mu} - \bar{s}\hat{\lambda})$, and setting $U_\omega(\mathbf{x},\mu,\lambda) = 0$ to find a feasible solution gives the feasibility cut $\mu + \bar{s}\lambda \geq \pi^{*,\omega}B^\omega\mathbf{x} + (h_\omega(\hat{\mathbf{x}}) - \pi^{*,\omega}B^\omega\hat{\mathbf{x}})$.

6.3. Computational Enhancements

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 in Nocedal and Wright (1999).

6.4. Implementation Notes on Different ϕ

First, note that when (10) is not present, there is no need to use feasibility cuts. Let's now look at implementation issues with respect to suppression and popping.

Divergences that can suppress in Subcategory 2 (see §4.3) 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. We recommend forcing λ to be nonzero and checking optimality condition at $\lambda = 0$ separately.

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 $\phi^{*\prime}(\bar{s}_{\text{comp}}) \geq \frac{1}{\min_\omega q_\omega}$. 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

To illustrate the techniques discussed in this paper, we applied the decomposition method from Section 6 to a small electricity generation problem. On this problem, we demonstrate the popping and suppressing behavior described in Section 4 for several ϕ-divergences and show how the worst-case cost decrease condition (15) compares to the actual cost decrease when an additional sample of each scenario is taken.

We modified an SLP-2 test problem, denoted APL1P, which is a power expansion problem with 5 independent random variables and 1280 realizations (Infanger 1992). To clearly

Table 3 Numerical results of ϕAPL1P for various divergences.

ϕ	ρ	Cost	Worst-Case Distribution					
$\phi_{m\chi^2}$	1.845	30735	0.1353	0.6354	0.2293	0	0	0
ϕ_{kl}	0.9225	30921	0.1050	0.7208	0.1507	0.0052	0.0108	0.0075
ϕ_b	0.9225	30714	0.0636	0.7751	0.0768	0.0253	0.0308	0.0285
ϕ_b	1.107	29775	0.1065	0.6273	0.1311	0.0402	0.0494	0.0455

demonstrate how the worst-case distribution changes with ρ and especially to demonstrate suppressing and popping behavior, we took 6 unique samples from APL1P to form the ϕLP-2. We denote the resulting problem as ϕAPL1P. The first-stage determines the capacity to be built for two generators and the generators are operated under uncertain demands and availability of the generators in the second stage.

In this problem, the second scenario (displayed in green in Figures 1 and 2) is the most costly, thus the only candidate for popping. Furthermore, the second scenario will eventually have unit probability in suppressing divergences.

7.1. Numerical Results

Table 3 shows a few select results for the ϕAPL1P using different ϕ-divergences. Behavior is shown assuming one observation per scenario for divergences above the line, and with the most costly scenario unobserved below the line to demonstrate popping. The value of ρ is chosen in accordance with an asymptotic 95% confidence region in (13). 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. Notice that, 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. Illustrations of Popping and Suppressing

We begin with an examination of two divergences that can suppress scenarios. Figure 1 shows how the worst-case distribution changes with ρ for both the Modified χ^2-Distance

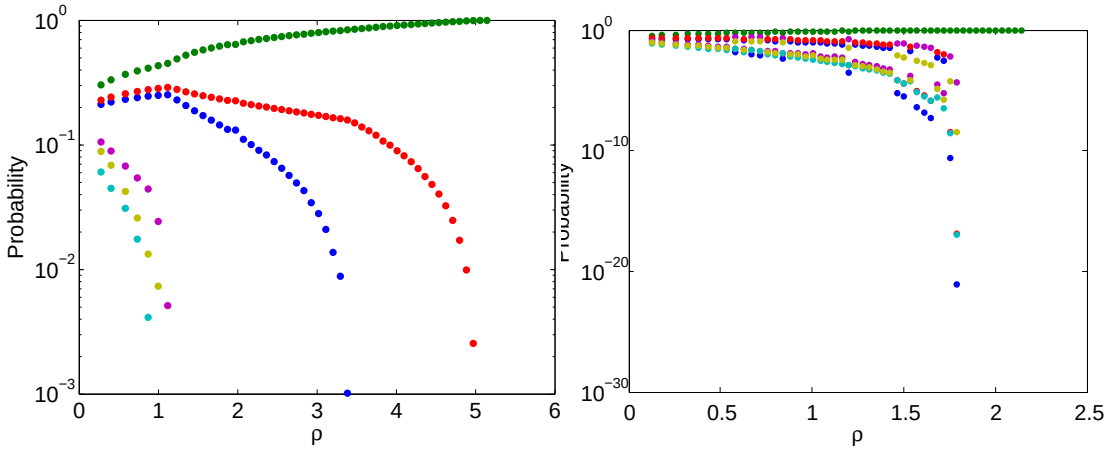

Figure 1 **Examples of distributions that can suppress: Modified χ^2 distance (left; one-at-a-time suppression) and KL Divergence (right; simultaneous suppression).**

(left) and the KL divergence (right). As shown in Section 4.3, the Modified χ^2-Distance suppresses scenarios one at a time, starting with the least expensive; while the KL divergence will suppress all (simultaneously) but the most costly scenario.

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.

7.3. Illustration of Value of Data

A comparison of the worst-case expected cost decrease condition (15) with the actual decrease in expected cost resulting from an additional observation is shown in Figure 3 for the Modified χ^2-Distance (left) and the Burg Entropy (right). The solid lines indicate regions where (15) is satisfied and dotted regions indicate that (15) is not satisfied. Note that (15) is a sufficient but not necessary condition. Both plots were generated using the reformulations to (15) provided in Corollary 1. The most costly scenario, shown in green

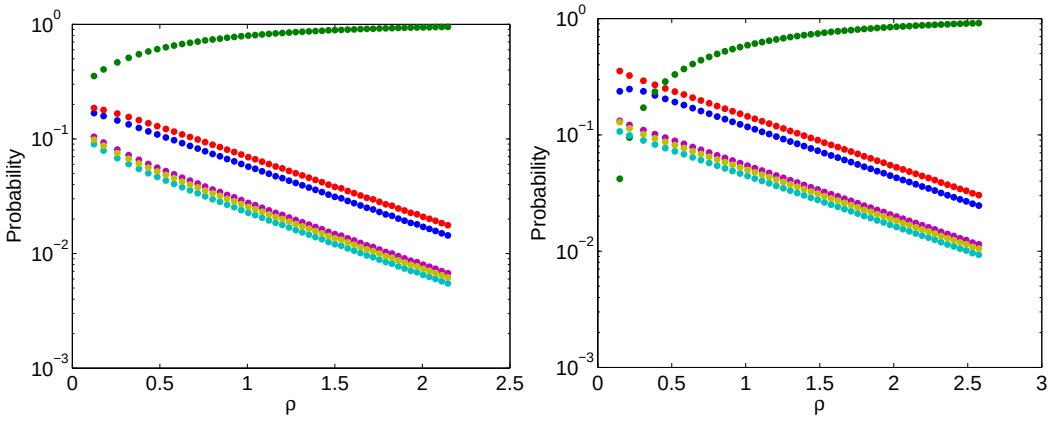

Figure 2 **Example of a distribution that can pop—the Burg entropy: all scenarios have a single observation (left); the most costly scenario having no observation (right).**

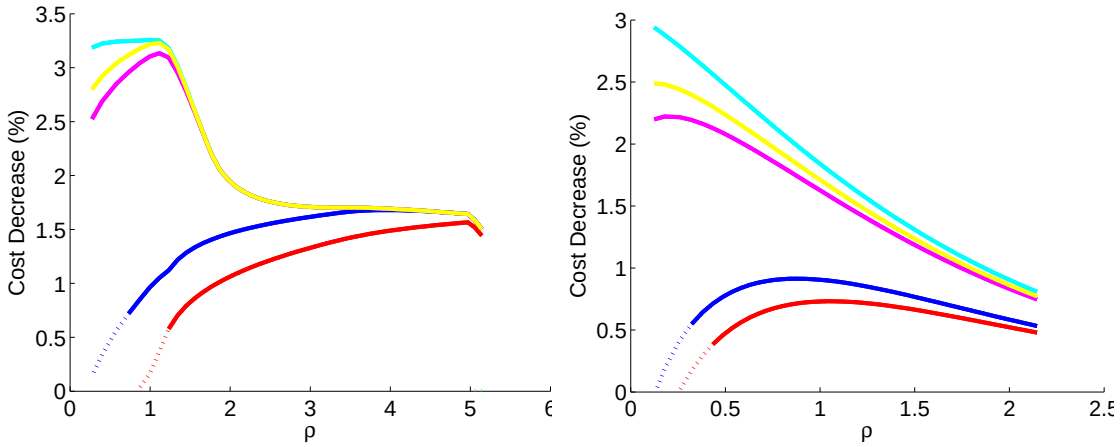

Figure 3 **Decrease in the worst-case expected cost from an additional observation for Modified χ^2-Distance (left) and Burg Entropy (right). Solid lines indicate when (15) is satisfied, while dotted lines show when an additional observation decrease the worst-case expected cost although (15) is not satisfied.**

in all plots, is not visible in Figure 3 because an additional observation will increase the worst-case expected cost. Condition (15) was never satisfied for this scenario.

8. Summary and Future Work

We proposed to use ϕ-divergences to define an ambiguity set of probability distributions, possibly using observed data, 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. One is the multi-stage extensions of the work provided here. Ways to handle continuous distributions in the ϕLP-2 also merits further research. There are other divergences, probability metrics, and statistical ways to measure the distance between two distributions. Generalizations of the results presented in this paper to other distance measures is another area of future research. Finally, it would be useful to study applications to real-world problems.

Appendix. Proofs

In this appendix, we provide proofs of all propositions, theorems, and the corollary in the order they appear in the paper.

Proof of Proposition 1. 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 $\mathbf{c} = 0$. We begin with the Lagrangian of (5), $\mathcal{L}(p,\mu,\lambda) = \sum_{\omega=1}^n p_\omega h_\omega(\mathbf{x}) + (1 - \sum_{\omega=1}^n p_\omega)\,\mu + \left(\rho - \sum_{\omega=1}^n q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right)\right)\lambda$, for which we generate the dual problem as

$$\begin{aligned}
&\min_{\lambda\geq 0,\mu}\max_{p\geq 0} \sum_{\omega=1}^n p_\omega h_\omega(\mathbf{x}) + \left(1 - \sum_{\omega=1}^n p_\omega\right)\mu + \left(\rho - \sum_{\omega=1}^n q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right)\right)\lambda \\
&= \min_{\lambda\geq 0,\mu} \mu + \rho\lambda + \sum_{\omega=1}^n \max_{p_\omega \geq 0}\left\{p_\omega(h_\omega(\mathbf{x}) - \mu) - \lambda q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right)\right\} \qquad (20)\\
&= \min_{\lambda\geq 0,\mu} \mu + \rho\lambda + \lambda\sum_{\omega=1}^n q_\omega \max_{t_\omega \geq 0}\{s_\omega t_\omega - \phi(t_\omega)\} \qquad (21)\\
&= \min_{\lambda\geq 0,\mu} \mu + \rho\lambda + \lambda\sum_{\omega=1}^n q_\omega \phi^*(s_\omega),
\end{aligned}$$

where $t_\omega = \frac{p_\omega}{q_\omega}$.

To account for the possibility that $q_\omega = 0$ and demonstrate popping behavior, equality (21) must be modified slightly. Consider a term in the summation in (20) for which $q_\omega = 0$:

$$\begin{aligned}\max_{p_\omega \geq 0}\left\{p_\omega(h_\omega(\mathbf{x})-\mu) - \lambda q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right)\right\} &= \max_{p_\omega \geq 0}\left\{p_\omega(h_\omega(\mathbf{x})-\mu) - \lambda 0 \phi\left(\frac{p_\omega}{0}\right)\right\} \\ &= \max_{p_\omega \geq 0}\left\{p_\omega\left(h_\omega(\mathbf{x})-\mu-\lambda\bar{s}\right)\right\}. \qquad (22)\end{aligned}$$

The behavior of (22) depends on the sign of $(h_\omega(\mathbf{x})-\mu-\lambda\bar{s})$, or equivalently, relation between s_ω and $\bar{s}$. There are three cases:

Case 1: $s_\omega > \bar{s}$ selects $p_\omega = \infty$, which induces the constraint $\frac{h_\omega(\mathbf{x})-\mu}{\lambda} \leq \bar{s}$ for scenarios with $q_\omega = 0$.

Case 2: $s_\omega < \bar{s}$ selects $p_\omega = 0$.

Case 3: $s_\omega = \bar{s}$ places no restrictions on the value of p_ω, since (22) is identically zero, and hence allows for $p_\omega > 0$ (popping). □

Proof of Theorem 1. For ease of exposition, we assume ϕ^* is differentiable, although the proof works without this assumption with little modification. We begin this proof with the change of variables $\kappa = \frac{\lambda}{N}$, and note that $N\rho_N = \rho_0$ is constant. With this change of variables, the objective function is given by

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

Let $z_N = \min_{\mathbf{x},\mu,\kappa} f_N(\mathbf{x},\mu,\kappa)$. We wish to find a simple estimate of the decrease in the optimal cost, $z_N - z_{N+1}$, associated with taking an additional sample of, say, $\hat{\omega}$, looking in particular for a condition under which $z_N - z_{N+1} > 0$. Let $(\mathbf{x}_N^*, \mu_N^*, \kappa_N^*)$ minimize f_N. Then $z_N - f_{N+1}(\mathbf{x}_N^*, \mu_N^*, \kappa_N^*)$ is 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 $N+1$ sample problem for a given $(\mathbf{x},\mu,\kappa)$ is $\mathbf{cx}+\mu+\rho_0\kappa+\sum_{\omega=1}^n N'_\omega\left[\kappa\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{(N+1)\kappa}\right)\right]$, where N'_ω is the number of observations of ω after $N+1$ total observations. Then the difference between the two optimal costs is $\kappa\sum_{\omega=1}^n\left[N_\omega\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{N\kappa}\right) - N'_\omega\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{(N+1)\kappa}\right)\right]$, which must be positive to guarantee a drop in optimal cost. Let $\hat{\omega}$ be the scenario observed on the next observation, then we can rewrite the condition as

$$\kappa\sum_{\omega=1}^{n} N_\omega\left[\phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{N\kappa}\right) - \phi^*\left(\frac{h_\omega(\mathbf{x})-\mu}{(N+1)\kappa}\right)\right] - \kappa\phi^*\left(\frac{h_{\hat{\omega}}(x)-\mu}{(N+1)\kappa}\right) > 0. \qquad (23)$$

Let $s_\omega^N = \frac{h_\omega(x)-\mu}{N\kappa}$ and $s_\omega^{N+1} = \frac{h_\omega(x)-\mu}{(N+1)\kappa}$, 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. Note that $\phi^*(s)$ is convex, so, using the gradient inequality we

have $\phi^*(s_\omega^N) - \phi^*(s_\omega^{N+1}) \geq \frac{1}{N}\phi^{*\prime}(s_\omega^{N+1})s_\omega^{N+1}$. Using this, we can guarantee (23) is satisfied with the condition $\kappa\sum_{\omega=1}^n \frac{N_\omega}{N}\phi^{*\prime}(s_\omega^{N+1})s_\omega^{N+1} - \kappa\phi^*\left(\frac{h_{\hat{\omega}}(x)-\mu}{(N+1)\kappa}\right) > 0$, or, rearranging and dividing by $\kappa > 0$,

$$\sum_{\omega=1}^n \frac{N_\omega}{N}\phi^{*\prime}(s_\omega^{N+1})s_\omega^{N+1} > \phi^*(s_\omega^{N+1}). \tag{24}$$

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

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

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

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

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

Modified χ^2*:* $c = 2$, so

$$\phi^{*\prime}(s) = \begin{cases} \frac{1}{2}s & s \geq 0 \\ 0 & s < 0. \end{cases} \quad \square$$

Proof of Proposition 2. Let $Z = \{\omega : q_\omega^{\text{true}} = 0\}$ be the set of impossible scenarios. For simplicity, we assume ϵ is chosen so that $\max_{\omega\notin Z} q_\omega^{\text{true}} > \frac{\epsilon}{2}$ and drop the dependence on $\xi \in \Xi'$. First, note that $\max_\omega |p_\omega - q_\omega^{\text{true}}| \leq \max_\omega |p_\omega - q_\omega^N| + \max_\omega |q_\omega^N - q_\omega^{\text{true}}|$. Let N'' be such that $\max_\omega |q_\omega^N - q_\omega^{\text{true}}| \leq \frac{\epsilon}{2}$ for all $N \geq N''$. To complete the proof, we will show that one can choose $N' \geq N''$ such that $\forall N \geq N'$, $\max_\omega |p_\omega - q_\omega^N| > \frac{\epsilon}{2} \Rightarrow I_\phi(p,q) > \frac{\rho_0}{N}$. First, bound the divergence by

$$\begin{aligned}
I_\phi(p, q^N) &= \sum_{\omega=1}^n q_\omega^N \phi\left(\frac{p_\omega}{q_\omega^N}\right) \\
&= \bar{s}\mathbb{I}_{\bar{s}<\infty}\sum_{\omega\in Z} p_\omega + \sum_{\omega\notin Z} q_\omega^N \phi\left(\frac{p_\omega}{q_\omega^N}\right) \\
&\geq \bar{s}\mathbb{I}_{\bar{s}<\infty}\sum_{\omega\in Z} p_\omega + \min_{\omega\notin Z}\{q_\omega^N\}\cdot\max_{\omega\notin Z}\left\{\phi\left(\frac{p_\omega}{q_\omega^N}\right)\right\} \\
&\geq \bar{s}\mathbb{I}_{\bar{s}<\infty}\sum_{\omega\in Z} p_\omega + \min_{\omega\notin Z}\{q_\omega^N\}\cdot\min\left\{\phi\left(1+\frac{\epsilon}{2}\right), \phi\left(1-\frac{\epsilon}{2}\right)\right\} \qquad (25) \\
&\geq \bar{s}\mathbb{I}_{\bar{s}<\infty}\sum_{\omega\in Z} p_\omega + \min_{\omega\notin Z}\left\{q_\omega^{\text{true}} - \frac{\epsilon}{2}\right\}\cdot\min\left\{\phi\left(1+\frac{\epsilon}{2}\right), \phi\left(1-\frac{\epsilon}{2}\right)\right\},
\end{aligned}$$

where $\bar{s}\mathbb{I}_{\bar{s}<\infty}$ is the indicator function taking value $\bar{s}$ if $\bar{s} < \infty$ (i.e., if ϕ can pop scenarios—please see Section 4.1 for details), and zero otherwise. Inequality (25) is true because $\phi\left(\frac{p_\omega}{q_\omega^N}\right) \geq \min\left\{\phi\left(\frac{q_\omega^N+\frac{\epsilon}{2}}{q_\omega^N}\right), \phi\left(\frac{q_\omega^N-\frac{\epsilon}{2}}{q_\omega^N}\right)\right\}$ for at least one ω, and applying the inequalities $\frac{a+\eta}{a} \geq 1+\eta$ and $\frac{a-\eta}{a} \leq 1-\eta$.

Finally, choose N' to satisfy $\bar{s}\mathbb{I}_{\bar{s}<\infty}\sum_{\omega\in Z} p_\omega + \min_{\omega\notin Z}\left\{q_\omega^{\text{true}} - \frac{\epsilon}{2}\right\}\cdot\min\left\{\phi\left(1+\frac{\epsilon}{2}\right), \phi\left(1-\frac{\epsilon}{2}\right)\right\} \geq \frac{\rho_0}{N'}$. $\square$

PROOF OF THEOREM 2. The theorem can be proven by showing the epiconvergence of the objective function of ϕLP-2 to that of SLP-2. To this end, we apply Proposition 2 to Theorem 3.7 of Dupacová and Wets (1988), which establishes the epiconvergence of (5) under the evident conditions that the objective function (under the worst-case distribution) is continuous with respect to ω (because ω is discrete) and lower semicontinuous and locally lower Lipschitz with respect to x (because ϕLP-2 is convex). □

References

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** 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.

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

Calafiore, G.C., M.C. Campi. 2005. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming* **102** 25–46.

Calafiore, G.C., L. El Ghaoui. 2006. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications* **130** 1–22.

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

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

Dupacová, J., R. Wets. 1988. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *The Annals of Statistics* **16** 1517–1549.

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

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** 69–95.

Jiang, R., Y. Guan. 2013. Data-driven chance constrained stochastic program. Tech. rep., University of Florida. Available at: Optimization Online `www.optimization-online.org`.

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

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** 435–442.

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.

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

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

Wang, Z., P.W. Glynn, Y. Ye. 2010. Likelihood robust optimization for data-driven newsvendor problems. Tech. rep., Department of Management Science and Engineering, Stanford University, USA.

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