TUTORIALS IN
OPERATIONS RESEARCH
INFORMS 2015

 | ISBN 978-0-9843378-8-0
http://dx.doi.org/10.1287/educ.2015.0134

Data-Driven Stochastic Programming Using Phi-Divergences

Güzin Bayraksan
Department of Integrated Systems Engineering, the Ohio State University, Columbus, Ohio 43210, bayraksan.1@osu.edu

David K. Love
American Express Company, New York, New York 10285, love.david.k@gmail.com

Abstract Most of classical stochastic programming assumes that the distribution of uncertain parameters are known, and this distribution is an input to the model. In many applications, however, the true distribution is unknown. An ambiguity set of distributions can be used in these cases to hedge against the distributional uncertainty. Phi-divergences (Kullback–Leibler divergence, χ^2-distance, etc.) provide a measure of distance between two probability distributions. They can be used in data-driven stochastic optimization to create an ambiguity set of distributions that are centered on a nominal distribution. The nominal distribution can be determined by collected observations, expert opinions, simulations, etc. Many phi-divergences are widely used in statistics; therefore they provide a natural way to create an ambiguity set of distributions from available data and expert opinions. In this tutorial, we present two-stage models with distributional uncertainty using phi-divergences and tie them to risk-averse optimization. We examine the value of collecting additional data. We present a classification of phi-divergences to elucidate their use for models with different sources of data and decision makers with different risk preferences. We illustrate these ideas on several examples.

Keywords stochastic programming; distributionally robust optimization; phi-divergences

1. Introduction

Consideration of uncertainty is paramount in real-world applications. Significant work has been expended in the operations research and management science literature on developing new optimization models under uncertainty, designing efficient algorithms to solve these models, and investigating the theoretical properties of the resulting models and algorithms. Most of these models assume that a probability distribution of the unknown parameters is given as an input. However, in many cases, such a probability distribution is unknown, or its approximations are not fully trusted. For instance, we might have few data points when dealing with a new system design. Alternatively, we might have a lot of data but not much trust in the data. As an example of the latter case, consider solving long-term water resources management problems that incorporate climate models. There are many climate models developed by earth scientists, and these models provide temperature and precipitation values 50–100 years into the future. Even though there are many data points, the trust in these data can be low, especially as we look farther into the future.

To handle the ambiguities in the probability distributions of uncertain parameters, consider a set of possible distributions. Instead of one (assumed known) probability distribution, a set of possible distributions centered on the available data can be used. The optimization model then hedges against the distributional ambiguities within this set. This approach

dictates some (but not all) knowledge about the distribution family such as a common expected value and/or variance across the family of distributions. Other information that could be used includes higher moments, limits on the parameters of a distribution family, distributional information from existing data, and expert opinions, among others. This approach is not new by any means. One of the earliest models is the newsvendor problem studied by Scarf [33] in 1958. In stochastic programming, these models have been examined by Dupačová (as Žáčková) [40] in a minimax framework, dating back to the 1960s; see also Dupačová [10], Shapiro and Ahmed [34], and Shapiro and Kleywegt [35] for minimax stochastic programs. Such models have been referred to as *ambiguous stochastic programs*; see, for instance, Pflug and Wozabal [27] and Erdoğan and Iyengar [11]. More recently, this approach has also been called the *distributionally robust optimization*; see, for instance, Delage and Ye [9], Goh and Sim [13], Mehrotra and Papp [23], and Hanasusanto et al. [14]. In this tutorial we use the term ambiguous stochastic programs.

This tutorial introduces and examines the properties of two-stage ambiguous stochastic programs, where the distributional uncertainty is handled via phi-divergences. Such models were first systematically analyzed by Ben-Tal et al. [3]. Phi-divergences measure distances between distributions. We will shortly review them in §3. Specifically, the distributional uncertainty is modeled using a set of distributions that are sufficiently close to a *nominal distribution* with respect to phi-divergences. A nominal distribution is typically obtained by analyzing historical data, collecting expert opinions, tallying results of simulations, and so forth.

There are several advantages of using phi-divergences. First, many common phi-divergences are used in statistics, for instance, to conduct goodness-of-fit tests (Pardo [24]). Therefore, they provide natural ways to deal with data and distributions. Another attractive feature of phi-divergences is that they preserve convexity, resulting in computationally tractable models. Using phi-divergences can also be more data driven—many phi-divergences use more distributional information than just the first and second moments. Completely different distributions might have the same (say, first or first two) moments. Consequently, considering all distributions with the same moments could result in overly conservative solutions.

The rest of this tutorial is organized as follows. We begin by narrowing down the problem class in §2, followed by background information on phi-divergences in §3. The two-stage ambiguous stochastic program formulation is presented in §4. This section also discusses the relation to risk-averse optimization and other properties of the two-stage ambiguous stochastic programs. In a data-driven setting, it is important to study how the models change as more data are collected. We examine the value of additional data in §6. There are many phi-divergences, and this raises the question as to which one to choose and why. Toward answering this question, we provide a classification of phi-divergences with respect to their use in a data-driven stochastic optimization setting in §7. We then discuss which class of phi-divergences is appropriate for which data type and model, relating to the risk preferences of the modeler. We end the tutorial in §8 with future research directions. Throughout the tutorial, we present the results without formal proofs; for detailed results and applications, we refer readers to Love [20] and Love and Bayraksan [21, 22].

2. Problem Class

We begin with a general stochastic program to emphasize that similar ideas can be used for a larger class of problems. In fact, some work has already been conducted in this direction, to which we refer later in this section. We will then narrow down the focus to two-stage models and list our assumptions.

Consider a general stochastic optimization problem of the form

$$\min_{\mathbf{x}\in X}\left\{\mathbb{E}_Q[f_0(\mathbf{x},\tilde{\boldsymbol{\xi}})]:\ \mathbb{E}_Q[f_k(\mathbf{x},\tilde{\boldsymbol{\xi}})]\le 0,\ k\in\kappa\right\}, \tag{SP}$$

where f_k, $k \in \{0\} \cup \kappa$ are extended real-valued functions with inputs being the decision vector $\mathbf{x}$ and a realization of the random vector $\tilde{\boldsymbol{\xi}}$. Set $X \subset \mathbb{R}^{d_x}$ represents the deterministic constraints $\mathbf{x}$ must satisfy, and set $\Xi \subset \mathbb{R}^{d_\xi}$ denotes the support of $\tilde{\boldsymbol{\xi}}$. Here, d_x and d_ξ are the dimensions of the vectors $\mathbf{x}$ and $\tilde{\boldsymbol{\xi}}$, respectively. The expectations $\mathbb{E}_Q[\cdot]$ in (SP) are taken with respect to Q, the distribution of $\tilde{\boldsymbol{\xi}}$. We assume that all expectations are well defined and finite for all $\mathbf{x} \in X$.

The (SP) formulation requires several important implicit assumptions. First, the probability distribution of $\tilde{\boldsymbol{\xi}}$ is assumed to be known, which may not be realistic in some situations as discussed above. Second, the objective function and the constraints are also assumed to be known or observable for a given decision vector $\mathbf{x}$ and a realization of $\tilde{\boldsymbol{\xi}}$. This implicit assumption is necessary to be able to write down the optimization model. Third, any function that depends on the random parameters must be summarized via a probability functional such as an expectation or a probability. Whereas (SP) above is written using expectations, by changing the form of f, these expectations could represent regular expectations, probabilities, conditional values at risk (CVaR), risk measures, etc. The probability functional could be adjusted depending on the risk preference of the decision maker, where a regular expectation is considered to be risk neutral.

The (SP) formulation represents a large class of problems depending on κ, X, and f_k, $k \in \{0\} \cup \kappa$. Several examples are as follows:

1. *Simulation-Optimization Problem*: In mathematical programming under uncertainty, functions that form the objective and constraints can typically be written in analytical form, even though the resulting problem may not be solved exactly. By contrast, the objective function and/or constraints of a simulation optimization problem can only be observed after (typically expensive) "black-box" simulations. These simulations could, for instance, mimic the operation of a system under random inputs, leading to random outputs. These random outputs are then related to objective functions and/or constraints. The system might have design variables that could be adjusted to optimize system performance. An example problem in this category maximizes the expected revenue of a manufacturing plant made up of a series of machines (Buchholz and Thümmler [5]). The expected revenue depends on the throughput of the plant—i.e., the time-averaged number of products leaving the last machine—which is observed through simulations. The manufacturing plant could be viewed as a queueing network, with random inputs being the arrival process of the parts and processing times of the machines. Each machine has a buffer that can hold the incoming parts. When the buffer of a particular machine is full, the parts before this machine can no longer move forward until space frees up. The decision variables of this problem could be, for instance, the buffer sizes of the machines, with a total buffer budget (Patsis et al. [26]).

For more information on simulation optimization, please see the recent book by Fu [12] and the tutorial by Pasupathy and Ghosh [25]. The simulation optimization library contains a variety of other examples (e.g., Henderson and Pasupathy [15]).

2. *Two-Stage Stochastic Linear Program with Recourse* (*SLP*-2): In SLP-2, $X = \{\mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$, $\kappa = \varnothing$, and $f_0(\mathbf{x}, \tilde{\boldsymbol{\xi}}) = \mathbf{c}\mathbf{x} + h(\mathbf{x}, \tilde{\boldsymbol{\xi}})$. Thus, SLP-2 can be written as

$$\min_{\mathbf{x} \in X} \{\mathbf{c}\mathbf{x} + \mathbb{E}_Q[h(\mathbf{x}, \tilde{\boldsymbol{\xi}})]\}, \tag{1}$$

where $h(\mathbf{x}, \tilde{\boldsymbol{\xi}})$ is the optimal value of the linear program $\min_{\mathbf{y}}\{\tilde{\mathbf{g}}\mathbf{y}\colon \tilde{\mathbf{D}}\mathbf{y} = \tilde{\mathbf{B}}\mathbf{x} + \tilde{\mathbf{d}}, \mathbf{y} \geq 0\}$ for a given $\mathbf{x}$ and a given realization of the random vector $\tilde{\boldsymbol{\xi}}$, which is composed of the random elements $\tilde{\mathbf{g}}$, $\tilde{\mathbf{D}}$, $\tilde{\mathbf{B}}$, and $\tilde{\mathbf{d}}$. Decisions $\mathbf{x}$ need to be determined before knowing the outcome of the random events. These are called the *first-stage decisions*. For instance, an energy application allocates capacities $\mathbf{x}$ to different electric generators in the first stage before knowing the demand and electricity prices. Later, as random elements become known, a second set of decisions $\mathbf{y}$, called the *second-stage decisions* or *recourse decisions*, can be taken. The recourse decisions can be corrective actions, or they can be operational decisions

that can only be made after uncertainties become known. Going back to the energy example, when the demand and prices become known, the second-stage problem distributes electricity in a minimal-cost fashion to the demand sites. Energy can be bought at a higher price from other sources to satisfy any unmet demand. Observe that although the first-stage decisions are made before the random elements become known, they do consider uncertainty through the expected second-stage cost term $\mathbb{E}_Q[h(\mathbf{x},\tilde{\boldsymbol{\xi}})]$ in (1).

SLP-2 can be extended to multiple stages, and it can include integrality constraints and nonlinear objective terms and constraint functions, leading to other types of stochastic programs with recourse. For more examples and further information on stochastic programs with recourse, please see, e.g., the books by Birge and Louveaux [4] and Shapiro et al. [36].

3. *Chance-Constrained Stochastic Program*: Consider a stochastic program with a single chance constraint, $\kappa = \{1\}$. Suppose further that $f_0(\mathbf{x},\tilde{\boldsymbol{\xi}}) = \mathbf{cx}$ and $f_1(\mathbf{x},\tilde{\boldsymbol{\xi}}) = \alpha - \mathbb{I}(\tilde{\mathbf{A}}\mathbf{x} \geq \tilde{\mathbf{b}})$, where $\mathbb{I}(\cdot)$ denotes the indicator function that takes value 1 if its argument is true and 0 otherwise, and $\alpha \in (0,1)$ is a desired probability level. In this case, $\tilde{\boldsymbol{\xi}}$ is composed of random elements of $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{b}}$. Because expectation of an indicator function is a probability, this (SP) can be written as a chance-constrained problem:

$$\min_{\mathbf{x}\in X}\{\mathbf{cx}\colon \Pr_Q\{\tilde{\mathbf{A}}\mathbf{x} \geq \tilde{\mathbf{b}}\} \geq \alpha\}.$$

Above, the probability is taken with respect to Q. The chance-constrained stochastic program aims to satisfy the constraints subject to uncertainty by at least a desired probability α. A prototypical chance-constrained model allocates resources in a minimal-cost fashion while meeting the demand, say, 95% of the time.

Chance-constrained stochastic programs were first analyzed by Charnes et al. [8] and Charnes and Cooper [7]. For further information on this class of problems, we again refer readers to Birge and Louveaux [4] and Shapiro et al. [36].

The above formulations assume that the distribution of $\tilde{\boldsymbol{\xi}}$ is known. However, if the distribution used in the model is incorrect, (SP) can give highly suboptimal results. Such problems have led to the development of a modeling technique that replaces the probability distribution by a set of distributions. Optimization is done relative to the worst distribution in the uncertainty set.

In this tutorial, we refer to this uncertainty set as the *ambiguity set of distributions* and denote it as $\mathcal{P}$. An ambiguous version of SLP-2 is then given by

$$\min_{\mathbf{x}\in X}\max_{P\in\mathcal{P}}\{\mathbf{cx} + \mathbb{E}_P[h(\mathbf{x},\tilde{\boldsymbol{\xi}})]\}, \tag{2}$$

which minimizes the worst-case expected cost. In (2), the expectations are taken with respect to distributions P in set $\mathcal{P}$. Similarly, the chance constraint $\Pr_Q\{\tilde{\mathbf{A}}\mathbf{x} \geq \tilde{\mathbf{b}}\} \geq \alpha$ is changed to

$$\min_{P\in\mathcal{P}} \Pr_P\{\tilde{\mathbf{A}}\mathbf{x} \geq \tilde{\mathbf{b}}\} \geq \alpha \tag{3}$$

so that the constraint is satisfied under each distribution in the ambiguity set $\mathcal{P}$.

There are different ways to form the ambiguity sets. One common method uses moments of a distribution and either fixes the moments (typically the first and second moments) to certain values or considers all distributions with certain bounds on these moments; see, e.g., Wiesemann et al. [38] and Delage and Ye [9]. Probability metrics have also been used: Erdoğan and Iyengar [11] use the Prokhorov metric, and Pflug and Wozabal [27] use the Kantorovich metric. Hanasusanto et al. [14] provide a comprehensive review of different types of ambiguity sets. We refer readers to this paper and the references therein for more details on different types of ambiguity sets.

We are specifically interested in building the ambiguity set of distributions using existing (but potentially incomplete or limited) data via phi-divergences. As mentioned before,

Ben-Tal et al. [3] have studied models with phi-divergence-based ambiguity sets and examined their computational tractability. Some papers have studied specific phi-divergences such as the Kullback–Leibler divergence (Calafiore [6], Hu and Hong [16], Wang et al. [37]) and χ^2-distance (Klabjan et al. [19]). For data-driven, ambiguous, chance-constrained stochastic programs with phi-divergences, we refer readers to Jiang and Guan [17] and Yanıkoğlu and den Hertog [39]. In particular, Jiang and Guan [17] present an exact approach to solve ambiguous chance-constrained problems with phi-divergences and examine the value of data. Yanıkoğlu and den Hertog [39] present a safe approximation method.

Throughout the rest of the tutorial, we mainly focus on SLP-2 as given in (1) and its ambiguous version presented in (2). We further assume that $\tilde{\boldsymbol{\xi}}$ has a finite distribution with n realizations. We equivalently refer to realizations of $\tilde{\boldsymbol{\xi}}$ as *scenarios*. Each realization of $\tilde{\boldsymbol{\xi}}$ is indexed by ω, $\omega = 1, 2, \ldots, n$, and scenario ω has probability q_ω. To emphasize the dependence on scenario ω, we rewrite SLP-2 as

$$\min_{\mathbf{x}\in X}\left\{\mathbf{c}\mathbf{x} + \sum_{\omega=1}^{n} q_\omega h_\omega(\mathbf{x})\right\},$$

and $h_\omega(\mathbf{x}) = \min_{\mathbf{y}}\{\mathbf{g}^\omega\mathbf{y}\colon \mathbf{D}^\omega\mathbf{y} = \mathbf{B}^\omega\mathbf{x} + \mathbf{d}^\omega, \mathbf{y} \geq 0\}$. 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 each ω, $\omega = 1, 2, \ldots, n$. This ensures that expectations taken with respect to distributions formed on these scenarios are finite. We allow distributions where the probability mass on a particular scenario ω may be zero. We will come back to this later.

As a final note, our main classification of phi-divergences in §7 is based on the (geometric) properties of phi-divergences and therefore holds for a larger class of problems than considered here.

3. Background

We begin with a definition of phi-divergences in the discrete case and discuss their properties, largely based on Ben-Tal et al. [2, 3]. For a comprehensive review of phi-divergences, we refer readers to Pardo [24]. Then, we discuss how to use phi-divergences to form an ambiguity set of distributions.

3.1. Phi-Divergences

Phi-divergences measure the distance between two nonnegative vectors $\mathbf{p} = (p_1, \ldots, p_n)^T$ and $\mathbf{q} = (q_1, \ldots, q_n)^T$. We are interested in discrete probability distributions. In this case, $\mathbf{p}$ and $\mathbf{q}$ additionally satisfy $\sum_{\omega=1}^{n} p_\omega = \sum_{\omega=1}^{n} q_\omega = 1$. In our setup, $\mathbf{q}$ denotes the nominal distribution's probabilities, and we will quantify distributions close to the nominal distribution via phi-divergences.

The phi-divergence is defined by

$$I_\phi(\mathbf{p}, \mathbf{q}) = \sum_{\omega=1}^{n} q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right), \tag{4}$$

where $\phi(t)$, called the *phi-divergence function*, is a convex function on $t \geq 0$. It can be extended to $\mathbb{R}$ by setting $\phi(t) = +\infty$ for $t < 0$. The phi-divergence function takes value 0 when both $p_\omega > 0$ and $q_\omega > 0$ have the same value; i.e., $\phi(1) = 0$. When $q_\omega = 0$, the terms of (4) are interpreted as $0\phi(a/0) = a\lim_{t\to\infty}(\phi(t)/t)$, and $0\phi(0/0) = 0$. When both $\mathbf{p}$ and $\mathbf{q}$ are probability vectors, we can assume $\phi(t) \geq 0$ without loss of generality. In addition, we assume that $\phi(t)$ is a closed function—hence, it is lower semicontinuous. This assumption is

TABLE 1. Examples of phi-divergences, 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
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-distance	$\phi_{m\chi^2}$	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	$(\sqrt{t}-1)^2$	$\sum(\sqrt{p_\omega} - \sqrt{q_\omega})^2$	$\frac{s}{1-s}$, $s<1$

satisfied by many common phi-divergences (see, e.g., Table 1). It is also a natural assumption because we will use ϕ in an optimization context, and lower semicontinuity is a desirable property in this setting.

Even though phi-divergences can quantify distances between distributions, they are not, in general, metrics or semidistances. Most phi-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).

The *adjoint* of a phi-divergence, defined by $\tilde{\phi}(t) = t\phi(1/t)$, satisfies $I_{\tilde{\phi}}(\mathbf{p},\mathbf{q}) = I_\phi(\mathbf{q},\mathbf{p})$. The adjoint itself is a phi-divergence. If a phi-divergence is symmetric, it is called *self-adjoint*. Table 1 shows that the Kullback–Leibler divergence and Burg entropy are adjoints, and J-divergence is the sum of the two, which is self-adjoint; χ^2-distance and modified χ^2-distance, related to the famous χ^2 statistical test, are also adjoints. The variation distance and the Hellinger distance are self-adjoint. Most of these common phi-divergences are widely used in statistics and information theory. We present other phi-divergences that result in frequently used risk models in §5.

An important related function is the *conjugate function*, which we use for problem reformulations in §4.1 and for classifying phi-divergences in §7. The conjugate $\phi^*\colon \mathbb{R} \to \mathbb{R}\cup\{\infty\}$ is defined as

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

which is itself a convex function. Because $\phi(\cdot)$ is a proper closed convex function, we have the relationship that $t \in \partial\phi^*(s)$ if and only if $s \in \partial\phi(t)$ (Rockafellar [29, Corollary 23.5.1]) and $\phi^{**} = \phi$. Table 1 lists common examples of phi-divergences, along with their adjoints and conjugates. The value of the conjugate is listed only in its domain; i.e., $\{s\colon \phi^*(s) < \infty\}$.

3.2. Ambiguity Set of Distributions

We replace the nominal distribution of $\tilde{\boldsymbol{\xi}}$, $\{q_\omega\}_{\omega=1}^n$, with an ambiguity set of distributions by considering all distributions $\{p_\omega\}_{\omega=1}^n$ whose phi-divergence from the nominal distribution is sufficiently small. A remark is in order here. Although we refer to $\mathbf{p}$ or $\mathbf{q}$ as a "distribution"

with some abuse of terminology, these might not be probability mass functions on the whole set of scenarios $\omega = 1, 2, \ldots, n$ in the classical sense. That is, we allow $p_\omega = 0$ and $q_\omega = 0$ for some ω. All we require is that $0 \le p_\omega \le 1$ for all ω and $\sum_{i=1}^{n} p_\omega = 1$, and we require the same conditions on elements of $\mathbf{q}$.

The ambiguity set in formulations (2) and (3) using phi-divergences is given by

$$\mathcal{P} = \Bigg\{ \mathbf{p}\colon \sum_{\omega=1}^{n} q_\omega \phi\left(\frac{p_\omega}{q_\omega}\right) \le \rho, \tag{5}$$

$$\sum_{\omega=1}^{n} p_\omega = 1, \tag{6}$$

$$p_\omega \ge 0, \forall \omega \Bigg\}. \tag{7}$$

We refer to constraint (5) as the phi-divergence constraint, and constraints (6) and (7) simply ensure a probability measure.

It is important to have the ambiguity set as small as possible but not too small. For example, setting $\rho = 0$ would typically admit only one distribution in this set—the nominal distribution. (Recall that phi-divergence between $\mathbf{q}$ and $\mathbf{q}$ is zero.) In this case, the ambiguous problem (2) is equivalent to the regular SLP-2. On the other hand, as $\rho \to \infty$, the ambiguity set $\mathcal{P}$ admits all possible distributions. The worst case among all possible distributions is to put all probabilities only on the highest objective function values and to set the probabilities of other scenarios to zero. This clearly shows the role of ρ as a *risk-level parameter*. Setting $\rho = 0$, one obtains the *risk-neutral* expected value minimization of SLP-2 using the nominal distribution. At the other extreme, setting $\rho = \infty$ gives the *most risk-averse approach* (*minimizing the worst outcome*), making overly conservative decisions. The idea is to select the ambiguity set to reflect the perceived risk from the data. This is especially important in a data-driven setting. That is, in a data-rich environment, ρ could be smaller. By contrast, if there are little data or not much trust in the data, then ρ should be larger. Ideally, one would like to have a probabilistic guarantee that, given the data, the ambiguity set contains the true distribution with a desired level of confidence. Although such a probabilistic guarantee could be too much to ask for at all levels of data collection, an asymptotic statistical result exists. We discuss this next.

In a data-driven setting, the empirical distribution typically serves as the nominal distribution. That is, the nominal probability of scenario ω is set to $q_\omega^N = N_\omega / N$, where N_ω is the number of observations of scenario ω and $N = \sum_{\omega=1}^{n} N_\omega$ is the total number of observations. When ϕ is twice continuously differentiable around 1 with $\phi''(1) > 0$, Theorem 3.1 of Pardo [24] shows that the statistic $(2N/\phi''(1)) I_\phi(\mathbf{q}^N, \mathbf{q}^{\text{true}})$ converges in distribution to a χ^2-distribution with $n-1$ degrees of freedom, where $\mathbf{q}^N = (q_1^N, q_2^N, \ldots, q_n^N)^T$ denotes the empirical distribution and $\mathbf{q}^{\text{true}} = (q_1^{\text{true}}, q_2^{\text{true}}, \ldots, q_n^{\text{true}})^T$ denotes the underlying true distribution. Most phi-divergences in Table 1 satisfy this differentiability condition. Ben-Tal et al. [3] then use this result to suggest the asymptotic value

$$\rho = \frac{\phi''(1)}{2N} \chi^2_{n-1,\,1-\alpha}, \tag{8}$$

where $\chi^2_{n-1,\,1-\alpha}$ is the $1-\alpha$ quantile of a χ^2 distribution with $n-1$ degrees of freedom. This value of ρ provides an approximate $1-\alpha$ confidence region on the true distribution. For corrections for small sample sizes and more details, we refer readers to Pardo [24] and Ben-Tal et al. [3].

4. Formulations and Basic Properties

4.1. Primal and Dual Formulations

The primal problem is formulated as

$$\min_{\mathbf{x}\in X}\max_{\mathbf{p}\in\mathcal{P}}\left\{\mathbf{cx}+\sum_{\omega=1}^{n}p_\omega h_\omega(\mathbf{x})\right\}, \tag{9}$$

where $\mathcal{P}$ is given by (5)–(7). For a given $\mathbf{x}\in X$, the inner maximization is a convex optimization problem. Suppose ρ in the phi-divergence constraint (5) is positive. When $\rho>0$, the nominal distribution $\mathbf{q}$ strictly satisfies the phi-divergence constraint (5), $I_\phi(\mathbf{q},\mathbf{q})=0<\rho$. So the Slater condition holds, and we have strong duality.

The dual formulation of ambiguous SLP-2 uses the dual variables λ and μ for constraints (5) and (6), respectively. For simplicity of exposition, throughout the rest of the tutorial, we use s_ω to denote

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

Taking the Lagrangian dual of the inner problem and combining with the outer minimization results in the dual formulation:

$$\min_{\mathbf{x},\lambda,\mu}\ \left\{\mathbf{cx}+\mu+\rho\lambda+\lambda\sum_{\omega=1}^{n}q_\omega\phi^*(s_\omega)\right\} \tag{10}$$

$$\text{s.t. } \mathbf{x}\in X,$$

$$s_\omega\le\lim_{t\to\infty}\phi(t)/t,\quad\forall\omega, \tag{11}$$

$$\lambda\ge 0.$$

In the objective function of (10), the last term has the following interpretations when $\lambda=0$: $0\phi^*(a/0)=0$ if $a\le 0$ and $0\phi^*(a/0)=+\infty$ if $a>0$. Some phi-divergences, such as the J-divergence, do not have closed-form representations of the conjugate ϕ^*. However, if they can be expressed as the sum of other phi-divergences with closed-form conjugates, their dual can be formed. For instance, the J-divergence can be written as the sum of the Burg entropy and Kullback–Leibler divergence and admits a dual form; see Ben-Tal et al. [3] for details on its formulation. Note in particular that the dual formulation is accurate even for $q_\omega=0$ for some ω (Love [20], Love and Bayraksan [22]). The right-hand side of constraint (11) contains a limit. This constraint results from an implicit feasibility consideration. When this limit is finite, i.e., $\lim_{t\to\infty}\phi(t)/t=\bar{s}<\infty$, then for any $s>\bar{s}$, $\phi^*(s)=\infty$. This happens, for instance, for the Hellinger distance. In this case, constraint (11) is added to the formulation. On the other hand, this limit is ∞ for some phi-divergences, such as the Kullback–Leibler divergence. In this case, constraint (11) is redundant and can be removed.

It is possible to obtain the optimal worst-case probabilities from the dual optimal solution. When the dual is solved, an optimal solution $(\mathbf{x}^*,\lambda^*,\mu^*)$ is obtained, yielding s_ω^* for all $\omega=1,2,\ldots,n$. By using the properties of the convex conjugate and the Karush-Kuhn-Tucker conditions for the inner problem, we can obtain the worst-case probabilities $\mathbf{p}^*$ from the following set of equations:

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

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 $p_\omega^*=q_\omega\phi^{*\prime}(s_\omega^*)$. For further details on special cases when $\lambda^*=0$ or $q_\omega=0$ for some ω, we refer readers to Love and Bayraksan [22] and Love [20].

4.2. Basic Properties

The ambiguous SLP-2 formulated in the previous section has three important basic properties:

i. equivalence to minimizing a coherent risk measure,
ii. convexity, and
iii. decomposability.

The first property formalizes its relation to risk-averse optimization. We discuss these properties in detail below.

4.2.1. Relation to Risk-Averse Optimization. The ambiguous SLP-2 using phi-divergences is equivalent to minimizing a coherent risk measure. Rockafellar [30] defines a coherent risk measure in the basic sense as a functional $\mathcal{R}$: $L^2 \to (-\infty, \infty]$ on random variables (e.g., $Y, Y' \in L^2$) such that

1. $\mathcal{R}(C) = C$ for all constants C;
2. $\mathcal{R}((1-\lambda)Y + \lambda Y') \le (1-\lambda)\mathcal{R}(Y) + \lambda\mathcal{R}(Y')$ for all $\lambda \in [0,1]$, i.e., $\mathcal{R}$ is convex;
3. $\mathcal{R}(Y) \le \mathcal{R}(Y')$ when $Y \le Y'$, i.e., $\mathcal{R}$ is monotonic;
4. $\mathcal{R}(Y) \le 0$ when $\|Y^k - Y\|_2 \to 0$ with $\mathcal{R}(Y^k) \le 0$, i.e., $\mathcal{R}$ is closed; and
5. $\mathcal{R}(\lambda R) = \lambda\mathcal{R}(Y)$ for $\lambda > 0$, i.e., $\mathcal{R}$ is positively homogeneous.

There are other definitions of coherent risk measures. Artzner et al. [1], who have initiated the work on coherent risk measures, defined coherent risk measures $\mathcal{R}$ to be subadditive, be positively homogeneous, be monotonic, and satisfy the condition $\mathcal{R}(Y + C) = \mathcal{R}(Y) + C$ for any constant C, which is referred to as translation equivariance. Later definitions replace subadditivity in the Artzner et al. [1] definition with convexity; see, e.g., the one presented by Shapiro et al. [36] with extension to L^p for $p \in [1, \infty)$. Conditions 1, 2, and 4 above imply translation equivariance, and convex and positively homogeneous is equivalent to subadditive and positively homogeneous. Closedness (condition 4) is equivalent to lower semicontinuity, and it is automatically satisfied for risk measures $\mathcal{R}$ under other definitions of coherency if they are finite valued.

It is well known that coherent risk measures can be interpreted as worst-case expectations from a set of probability measures. The dual representation (also known as the envelope representation) of coherent risk measures states that they can be written as

$$\mathcal{R}(Y) = \max_{A \in \mathcal{A}} \mathbb{E}_A[Y], \tag{13}$$

where $\mathcal{A}$ is a closed convex set of probability measures, and the expectation is taken with respect to an element A in this set. This result was first established for finite-dimensional case by Artzner et al. [1] and was later refined by several researchers; see, e.g., Ruszczyński and Shapiro [32], Rockafellar and Uryasev [31], and references therein.

It is easy to show that ambiguous SLP-2 is equivalent to minimizing a coherent risk measure. Setting $Y = h(\mathbf{x})$, $\mathcal{A} = \mathcal{P}$, and $A = \mathbf{p}$ in (13), we can view the inner maximization problem in (9) as a coherent risk measure. Here, $h(\mathbf{x})$ is a random variable that takes on values $h_\omega(\mathbf{x})$ for each ω for a given $\mathbf{x} \in X$. As a result, the ambiguous SLP-2 can be equivalently written as

$$\min_{\mathbf{x} \in X} \{\mathbf{c}\mathbf{x} + \mathcal{R}_\phi(h(\mathbf{x}))\}, \tag{14}$$

where $\mathcal{R}_\phi$ is a coherent risk measure induced by the specific phi-divergence. We will shortly provide some examples.

Several remarks are in order. First, note that viewing ambiguous SLP-2 as minimizing a coherent risk measure is true even when $q_\omega = 0$ for some ω. The case of $q_\omega = 0$ will become important when we present our classification in §7. Second, the ambiguous SLP-2 hedges against the risk that the nominal distribution may not be the true distribution. In practice, what is available are data, not the distribution. The distribution is typically inferred by statistical analysis. However, there is a chance that the assumed (nominal) distribution is

incorrect, especially when there are little data or the data source is not fully trusted. The ambiguous SLP-2 tries to balance the risk of making an error in this estimation and its ramifications for making decisions under uncertainty. As a final remark, in addition to the specific risk measure induced by a particular ϕ, the parameter ρ also plays a role in the coherent risk minimization view presented in (14). As discussed in §3.2, a small value of ρ indicates less risk aversion, and a large value of ρ makes the problem more risk averse. Several examples in §5 will illustrate this further.

4.2.2. Convexity. Convexity of ambiguous SLP-2 follows immediately from minimizing a coherent risk measure over a polyhedron X. Coherent risk measures are convex and monotone nondecreasing by definition, and $h_\omega(\mathbf{x})$ is convex in $\mathbf{x}$ for all ω. Therefore, their composition $\mathcal{R}(h(\mathbf{x}))$ is convex. Convexity of these problems have also been noted by Ben-Tal et al. [3].

Because ambiguous SLP-2s are convex optimization problems, they can be solved efficiently. Ben-Tal et al. [3] examine the computational complexity of solving these problems for different phi-divergences. However, it is also possible to solve them via decomposition-based methods. Decomposition can help shorten solution times, especially when the number of distinct scenarios n is large. We discuss this next.

4.2.3. Decomposability. Decomposability can be directly seen from the dual formulation (10). Rewriting it slightly, we obtain

$$\min_{\mathbf{x}\in X,\,\lambda\geq 0,\,\mu} \Big\{ \mathbf{cx} + \mu + \rho\lambda + \mathbb{E}_{\mathbf{q}}[h^\dagger(\mathbf{x},\lambda,\mu)]\colon s_\omega \leq \lim_{t\to\infty} \phi(t)/t \Big\}. \tag{15}$$

The above formulation preserves the two-stage structure of the SLP-2. The first-stage variables can now be viewed as $\mathbf{x}$, λ, and μ. The expectation is taken with respect to the nominal distribution. That is, $\mathbb{E}_{\mathbf{q}}[h^\dagger(\mathbf{x},\lambda,\mu)] = \sum_{\omega=1}^{n} q_\omega h_\omega^\dagger(\mathbf{x},\lambda,\mu)$. Finally, $h_\omega^\dagger(\mathbf{x},\lambda,\mu) = \lambda\phi^*((h_\omega(\mathbf{x})-\mu)/\lambda)$, or equivalently, $\lambda\phi^*(s_\omega)$, and $h_\omega(\mathbf{x})$ are defined as before.

A decomposition method replaces the convex function $\mathbb{E}_{\mathbf{q}}[h^\dagger(\mathbf{x},\lambda,\mu)]$ by its lower approximation through affine cutting planes using the (sub)gradients of $h^\dagger(\mathbf{x},\lambda,\mu)$. Luckily, it is easy to generate (sub)gradients of $h_\omega^\dagger(\mathbf{x},\lambda,\mu)$ by translating the (sub)gradients $h_\omega(\mathbf{x})$ through the chain rule. This means that only linear subproblems need to be solved. We illustrate this idea on the cut coefficients of μ and $\mathbf{x}$. Suppose at the current iteration, $\hat{\mathbf{x}}, \hat{\lambda}$, and $\hat{\mu}$ solve the master problem. Let us assume $\hat{\lambda} > 0$ and ϕ^* is differentiable for simplicity. The master solution $\hat{\mathbf{x}}$ is then passed to the subproblems. The solution of the (linear) subproblems yields $h_\omega(\hat{\mathbf{x}})$ and dual solutions $\hat{\boldsymbol{\pi}}_\omega$, which allow us to compute $\hat{s}_\omega = (h_\omega(\hat{\mathbf{x}}) - \hat{\mu})/\hat{\lambda}$. The cut coefficient of μ from scenario ω is found by $\partial h_\omega^\dagger/\partial\mu = (\partial h_\omega^\dagger/\partial s_\omega)\cdot(\partial s_\omega/\partial\mu)$. This means the cut coefficient of μ is $-\phi^{*\prime}(\hat{s}_\omega)$. Similarly, the cut coefficient of $\mathbf{x}$ from scenario ω is found by the chain rule as $\phi^{*\prime}(\hat{s}_\omega)\cdot(\hat{\pi}_\omega\mathbf{B}^\omega)$, where $\mathbf{B}^\omega$ is the so-called technology matrix that appears in the second-stage constraints discussed at the end of §2. The cut coefficient of λ and the cut intercept are found in the usual way to obtain an affine cutting plane.

Recall that the constraint with the limit in (15) is not present for phi-divergences for which this limit is ∞. When this constraint is present, i.e., $\lim_{t\to\infty}\phi(t)/t = \bar{s} < \infty$, the master problem will have a nonlinear constraint. To simplify computation, this constraint can be removed from the master, and affine feasibility cuts can be generated when infeasibility is detected. Again, the feasibility cuts can be generated when the values of $h_\omega(\mathbf{x})$ are obtained (that is, when the linear subproblems are solved). To illustrate this, consider the same setup as for optimality cuts discussed above. After solving the subproblems, if $\hat{s}_\omega = (h_\omega(\hat{\mathbf{x}}) - \hat{\mu})/\hat{\lambda} > \bar{s}$ for some scenario ω, the current master solution is infeasible. We need $h_\omega(x) - \mu - \bar{s}\lambda \leq 0$ for feasibility. Again, we can use the lower approximation of this function to generate a feasibility cut. The cut coefficients are $(\hat{\pi}_\omega\mathbf{B}^\omega - 1 - \bar{s})$ for $(\mathbf{x}\ \mu\ \lambda)$. Once this cut is added to the master problem, the current solution will become infeasible. We can find a μ that satisfies the feasibility constraint easily and continue with the algorithm. The discussed

decomposition algorithm will solve a linear master problem and linear subproblems for each $\omega = 1, 2, \ldots, n$ at each iteration. For details and methods to deal with complications that arise when, e.g., $\lambda = 0$, see Love and Bayraksan [22] and Love [20].

5. Examples

We now go through some examples to show how the application of phi-divergences results in different models, some of which are well-known risk optimization models.

Example 1 (Likelihood Robust Optimization; Wang et al. [37]). Wang et al. [37] have proposed likelihood robust optimization as an attractive data-driven approach because likelihood functions are used extensively to conduct statistical analysis in the presence of data. Given N total number of observations, the likelihood robust optimization forms an ambiguity set of observations by considering all distributions whose empirical likelihood function is above a certain threshold. This constraint can be formulated as $\prod_{\omega=1}^{n} p_\omega^{N_\omega} \geq e^{\gamma}$. Let $0 \leq \gamma' \leq 1$ be the *relative likelihood parameter* that expresses γ as a proportion of the maximum likelihood; i.e., $\gamma = \log(\gamma' \prod_\omega (N_\omega/N)^{N_\omega})$. Using the log-likelihood function, this constraint can be expressed equivalently as

$$\sum_{\omega=1}^{n} \frac{N_\omega}{N} \log\left(\frac{N_\omega/N}{p_\omega}\right) \leq -\frac{1}{N} \log \gamma' =: \rho. \tag{16}$$

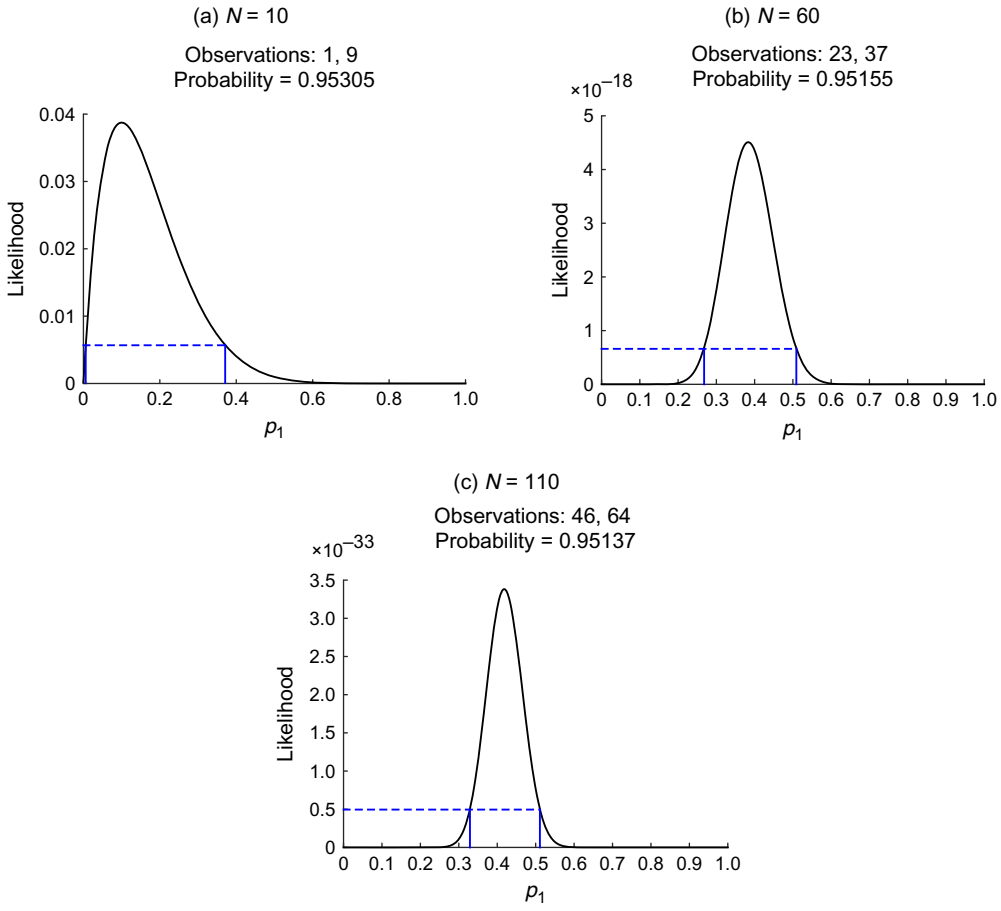
The above constraint is equivalent to the phi-divergence constraint (5) using the Burg entropy and taking $\mathbf{q}$ to be the empirical measure ($q_\omega = N_\omega/N$). We remark that Calafiore [6], Hu and Hong [16], and Wang et al. [37] all use a different naming convention than the one given here, referring to the Burg entropy as the "Kullback–Leibler divergence"—reversing the order of the arguments $\mathbf{p}$ and $\mathbf{q}$ relative to the notation presented here. (Recall here $\mathbf{q}$ denotes the nominal distribution.)

To make things more concrete, consider only two possible scenarios, scenario 1 ($\omega = 1$) and scenario 2 ($\omega = 2$). The empirical likelihood function is given by $p_1^{N_1}(1-p_1)^{N-N_1}$. Given N observations, of which N_1 is of scenario 1, SLP-2 would use the empirical distribution and set $q_1 = N_1/N$ and $q_2 = 1 - q_1 = (N - N_1)/N$. Figure 1 shows three examples as we collect more data. These data were generated randomly with $p_1 = 0.4$. In these graphs, the value of ρ (shown as blue dashed lines) in the phi-divergence constraint (5) is chosen to contain the true distribution (asymptotically) 95% of the time. Figure 1(a) shows the likelihood function with $N = 10$ total observations with only $N_1 = 1$ observation for scenario 1. The empirical estimate (or the maximum likelihood estimate) of p_1 is 0.1 in this case, but the likelihood robust ambiguity set would instead consider all values between 0 and slightly less than 0.4. Figures 1(b) and 1(c) show what happens as the number of observations increases from $N = 10$ to $N = 60$ first and then to $N = 110$. Note the change in the y-axis scale. With $N = 110$ observations, the point estimate of p_1 is updated to $46/110 = 0.4181$—quite close to the true value—and the ambiguity set contains $p_1 \in [0.32, 0.51]$.

The value of ρ in (16) demonstrates that as more observations are collected, our estimates get better, and the ambiguity region shrinks. As a result, the problems get less conservative by considering the worst-case expectation among a smaller set of distributions. □

Next, we show two examples that result in commonly used risk models in the operations research and management science literature. These two risk-averse problems—one that minimizes the CVaR and the other that minimizes a convex combination of expectation and CVaR—can be obtained via phi-divergences that are not listed in Table 1 (for more details on these phi-divergences, please see Love [20] and Love and Bayraksan [22]).

FIGURE 1. Likelihood robust ambiguity region with various total number of observations N.



Example 2 (CVaR). The coherent risk measure CVaR is well studied in financial applications. Minimizing

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

over $\mathbf{x} \in X$ is equivalent to the phi-divergence-constrained ambiguous SLP-2 with

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

for $0 < \beta < 1$. □

Example 3 (Convex Combination of CVaR and Expectation). Many models that consider risk in practice also incorporate a risk-neutral portion. This way, the average behavior as well as low-probability but high-risk situations are considered when making decisions. An attractive model in this case is to take a convex combination of CVaR and expectation. This model can be obtained by using the phi-divergence defined as

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

for $\alpha, \beta \in (0,1)$. The ambiguous SLP-2 using the above phi-divergence is equivalent to minimizing, over $\mathbf{x} \in X$,

$$\mathbf{cx} + (1-\alpha)\,\mathbb{E}[h(\mathbf{x})] + \alpha \mathrm{CVaR}_{\beta/(\alpha(1-\beta)+\beta)}[h_\omega(\mathbf{x})]. \quad \square$$

Whereas the above examples provide well-known and often-used risk measures in risk-averse optimization, in terms of phi-divergences, they are not very interesting. When viewing phi-divergences as a measure of distance, the phi-divergence functions in Examples 2 and 3 assign a distance of either 0 or ∞. For these phi-divergences, any value of $0 < \rho < \infty$ is equivalent to $\rho = 0$, hiding the effect of ρ that is typically present in ambiguous SLP-2 models. However, these examples clearly demonstrate the relationship to risk-averse optimization and that the class of problems includes well-known risk models. It is possible to create other "extreme" phi-divergences that take only 0 and ∞ to create other commonly used risk-averse optimization models. For further examples of this type, we refer to Love [20].

In the case of variation distance shown in Table 1, however, we obtain a model that both uses a regular phi-divergence and results in a convex combination of common coherent risk measures.

Example 4 (Convex Combination of CVaR and the Worst Case; Jiang and Guan [18], Rahimian et al. [28]). Using the variation distance listed in Table 1, the ambiguous SLP-2 is equivalent to minimizing a convex combination of CVaR and worst case. That is, the ambiguous SLP-2 with the variation distance is

$$\min_{\mathbf{x}\in X}\left\{\mathbf{cx} + \frac{\rho}{2}\sup_\omega h_\omega(\mathbf{x}) + \left(1-\frac{\rho}{2}\right)\mathrm{CVaR}_{\rho/2}[h_\omega(\mathbf{x})]\right\}.$$

The largest possible value of $I_{\phi_v}(\mathbf{p},\mathbf{q})$ is 2; thus the largest realistic value of ρ for this problem is 2. As $\rho \searrow 0$, the worst-case term disappears, and $\mathrm{CVaR}_{\rho/2}[\cdot]$ becomes the expected value using the nominal distribution. On the other hand, as $\rho \nearrow 2$, the CVaR term disappears, and the problem minimizes the worst-case outcome. $\square$

6. Value of Additional Data

Recall that in a data-driven setting, we have collected N_ω observations of scenario ω, and we have a total of $N = \sum_{\omega=1}^{n} N_\omega$ observations. We want to study the (potential) value of an *extra* observation. Put another way, we would like to answer the following question: If scenario k is observed (now scenario k has $N_k + 1$ observations, and we have a total of $N+1$ observations), will the optimal value of the ambiguous SLP-2 decrease, ruling out the worst-case distribution? This question is especially important in this setting. Because we are being risk averse, we might end up being overly conservative. To answer this question, we can solve the problem with the new updated nominal distribution that gives a higher nominal probability to scenario k and lower nominal probability to other scenarios. However, we would like to come up with a simple way of answering this question without solving a new problem. The condition presented in the proposition below uses the current solution and identifies scenarios for which an additional observation will rule out the current solution.

Proposition 1. *Let $(\mathbf{x}_N^*, \mu_N^*, \lambda_N^*)$ solve the N-sample problem with $\lambda_N^* > 0$, $q_\omega = N_\omega/N$, and ρ given in (8). Suppose ϕ^* is differentiable. An additional observation of scenario k will decrease the worst-case expected cost of ambiguous SLP-2 given in (9) if the following condition is satisfied:*

$$\sum_{\omega=1}^{n} q_\omega \phi^{*\prime}\left(\frac{N}{N+1}s_\omega^*\right)\left(\frac{N}{N+1}s_\omega^*\right) > \phi^*\left(\frac{N}{N+1}s_k^*\right), \tag{17}$$

where $s_\omega^* = (h_\omega(\mathbf{x}_N^*) - \mu_N^*)/\lambda_N^*$.

If an additional sample is taken from the unknown distribution and the resulting observed scenario k satisfies (17), then the $(N+1)$-sample problem will have a lower cost than the N-sample problem that was already solved. The condition in Proposition 1 provides insight into different scenarios for a decision maker. Let $L = \{k\colon \sum_{\omega=1}^{n} q_\omega \cdot \phi^{*\prime}((N/(N+1))s^*_\omega)((N/(N+1))s^*_\omega) > \phi^*((N/(N+1))s^*_k)\}$. The scenarios in L guarantee a drop in the overall cost if sampled once more. Therefore they can be considered "good" scenarios that decrease the risk aversion.

In Proposition 1, differentiability is assumed for simplicity. A subgradient of ϕ^* can be used in condition (17) instead. Subgradients are typically obtained as part of a solution algorithm; therefore they do not require any additional work. In some cases, it is possible to simplify this condition and write it in terms of the optimal worst-case probabilities $\mathbf{p}^*$. We present a simplified condition for the likelihood robust optimization below. A final remark on Proposition 1 is that scenarios that do not satisfy this condition might also cause a decrease in the worst-case expected value. This condition only provides partial information. In our computations, we found that it identifies most of the "good" scenarios (Love and Bayraksan [22]). Let us now illustrate this condition for the likelihood robust optimization.

Example 5 (Worst-Case Expected Cost Decrease Condition for the Likelihood Robust Optimization of Example 1). Let p^*_k be the worst-case probability found by solving the likelihood robust optimization problem with N total observations and N_k observations of scenario k. If

$$q_k = \frac{N_k}{N} > \frac{N+1}{N} p^*_k,$$

then an additional observation of scenario k will rule out the current worst-case distribution and result in a lower overall cost.

The above condition provides a relationship between the maximum-likelihood and worst-case probabilities, and it is an easy condition to check. If the ambiguous problem assigns a probability p^*_k to scenario k that is slightly below the maximum likelihood estimate of q_k (i.e., $p^*_k < (N/(N+1))q_k$), then an additional observation from this scenario will decrease the overall cost. Observe that because of the inner maximization in (9), the problem tends to assign higher probabilities to those scenarios with higher costs. If a scenario is assigned a lower probability than the nominal probability, then increasing our belief that this scenario is more likely makes the problem less risk averse. In other words, it decreases the overall cost. □

7. Classification of Phi-Divergences

Given the large number of phi-divergences, we want to examine their behavior when used in an optimization setting. Our aim is to illuminate the type of decision maker who might prefer a particular phi-divergence under certain data characteristics. To this end, we provide a classification of phi-divergences based on their geometric properties and how these properties affect the ambiguity set $\mathcal{P}$. We begin our discussion by defining two new terms: suppression of a scenario and popping of a scenario. These two types of behavior give rise to four different types of phi-divergences. Then, we examine suppression and popping in more detail. We return to the examples presented in §5 to illustrate these behaviors. Finally, we provide some guidelines on which type of phi-divergence could be used under what data type and decision-making preferences.

7.1. Suppression and Popping

Recall the definition of the ambiguity set—in particular, the phi-divergence constraint

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

Suppose $0 < \rho < \infty$. In this constraint, the phi-divergence function $\phi(\cdot)$ has arguments given by ratios of probabilities, p_ω/q_ω, and the limits $t \to 0$ and $t \to \infty$ correspond to the cases when $p_\omega = 0$, $q_\omega > 0$ and $q_\omega = 0$, $p_\omega > 0$, respectively. Why are these limits important? For instance, consider the case where $q_\omega > 0$ but $p_\omega = 0$, i.e., the case where $t \to 0$. This means that the nominal distribution has assigned a positive probability to scenario ω ($q_\omega > 0$), but the ambiguous counterpart problem has eliminated this scenario by assigning it a zero probability ($p_\omega = 0$). This scenario, however, could be of particular importance to the decision maker and needs to be considered with positive probability even in a worst-case distribution dictated by the ambiguity set $\mathcal{P}$. The other limiting case, $t \to \infty$, corresponds to $q_\omega = 0$ and $p_\omega > 0$. This could mean that scenario ω has never been observed before, so it has a zero probability in the nominal distribution ($q_\omega = 0$), but the ambiguous counterpart problem has assigned it a positive probability ($p_\omega > 0$). This option provides an interesting modeling choice, where one could include unobserved scenarios in the nominal distribution and let the model decide an appropriate risk-averse probability to these scenarios.

Let us discuss all the limiting cases in detail.

- *Case* 1 ($q_\omega > 0$, *but* $p_\omega = 0$). We call this the *suppression* behavior because a scenario with a positive probability in the nominal distribution ($q_\omega > 0$) can take zero probability in the ambiguous problem ($p_\omega = 0$). In other words, this scenario has been suppressed. 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 $q_\omega > 0$. We say that such a phi-divergence *cannot suppress a scenario*.

  —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 phi-divergence *can suppress scenario* ω.

- *Case* 2 ($q_\omega = 0$, *but* $p_\omega > 0$). We call this the *popping* behavior because a scenario with a zero probability in the nominal distribution ($q_\omega = 0$) can have a positive probability (or, "pop") in the ambiguous problem ($p_\omega > 0$). Recall that, by definition, $0\phi(a/0) = a \lim_{t \to \infty}(\phi(t)/t)$, and so in this case, we need to examine $\lim_{t \nearrow \infty}(\phi(t)/t)$:

  —If $\lim_{t \nearrow \infty}(\phi(t)/t) = \infty$, the ambiguity region can never contain distributions with $p_\omega > 0$ but $q_\omega = 0$. We say that such a phi-divergence *cannot pop a scenario*.

  —On the other hand, if $\lim_{t \nearrow \infty}(\phi(t)/t) < \infty$, the ambiguity region will admit sufficiently small p_ω. We say that these phi-divergences *can pop scenario* ω.

- *Case* 3 ($p_\omega = 0$, AND $q_\omega = 0$). Such a situation has no contribution since, by definition, $0\phi(0/0) = 0$.

The two limiting cases describing suppressing and popping behavior in phi-divergences create four distinct categories. Examples of phi-divergences in each category are given in Table 2.

Examining the suppression in more detail gives rise to two distinct suppressing behaviors. Recall the optimal primal-dual variable relation (12), which specifies that $p^*_\omega/q_\omega \in \partial\phi^*(s^*_\omega)$, where $s^*_\omega = (h_\omega(\mathbf{x}^*) - \mu^*)/\lambda^*$. Note that suppression ($p^*_\omega = 0$, $q_\omega > 0$) can occur only when $0 \in \partial\phi^*(s^*_\omega)$. Also, because of our assumptions on ϕ, we have $p^*_\omega/q_\omega \in \partial\phi^*(s^*_\omega)$ if and only if $s^*_\omega \in \partial\phi(p^*_\omega/q_\omega)$. For convenience, assume ϕ and ϕ^* are differentiable. We can examine suppression in more detail by looking at $\phi'(t)$ as $t \searrow 0$, as this dictates when $0 = \phi^{*\prime}(s^*_\omega)$. This analysis yields two subcategories within the phi-divergences that can suppress scenarios—the first subcategory tends to suppress scenarios one at a time, and the second subcategory suppresses all but the most costly scenario(s) simultaneously.

- SUBCATEGORY 1 ($\lim_{t \searrow 0} \phi'(t) > -\infty$). There are nonpositive constants $c, \underline{s}$ such that $\phi^*(s) = c$ for all $s < \underline{s}$. Here, $\lim_{t \searrow 0} \phi'(t) = \underline{s}$, and for all $s < \underline{s}$, $\phi^{*\prime}(s) = 0$. As an example of a phi-divergence in this subcategory, see Table 1 and consider the modified χ^2-distance. For the modified χ^2-distance, $c = -1$ and $\underline{s} = -2 = \lim_{t \searrow 0} \phi'(t)$. For this subcategory of suppressing phi-divergences, $\phi^{*\prime}(s_\omega) = 0$ when $s_\omega < \underline{s}$, suppressing all such scenarios. In other words, all scenarios that satisfy the relation $s_\omega = (h_\omega(\mathbf{x}) - \mu)/\lambda < \underline{s}$ are suppressed. As ρ increases, scenarios tend to be suppressed one at a time as they each reach $\underline{s}$.

TABLE 2. Examples of phi-divergences fitting into each category.

	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

Note. The number in parentheses under the "Can suppress scenarios" column denotes the subcategory.

• *Subcategory* 2 ($\lim_{t\searrow 0}\phi'(t) = -\infty$). In this case, $\phi^*(s) \searrow c$ as $s \to -\infty$ asymptotically, but it never reaches the bound. As an example of a phi-divergence in this subcategory, see Table 1 and consider the Kullback-Leibler divergence. For the Kullback–Leibler divergence, $\lim_{s\to-\infty}\phi^*(s) = c = -1$. Because such a constant c is reached only as $s \to -\infty$, 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 because of the $\lambda = 0$ in the denominator of s_ω.

Table 2 lists phi-divergences that belong to these subcategories under the "Can suppress scenarios" column, with the number in parentheses indicating the subcategory.

Examining the popping behavior in more detail, we find that a scenario ω with $q_\omega = 0$ can only be popped if it has the highest cost (Love and Bayraksan [22]). Otherwise, such a scenario will remain to have zero probability in the ambiguous SLP-2 solution.

7.2. Revisiting the Examples

We now return to the examples presented in §5 and discuss their suppression and popping behaviors.

Example 6 (Revisiting Likelihood Robust Optimization of Example 1). This phi-divergence cannot suppress scenarios but can pop scenarios. □

Example 7 (Revisiting CVaR of Example 2). 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, $\mathbb{E}[[h(\mathbf{x}) - \eta]^+]$. Scenarios cannot be popped because the expectation is taken with respect to the nominal distribution. □

Example 8 (Revisiting the Convex Combination of CVaR and Expectation of Example 3). This phi-divergence will neither pop (because both the expectation and CVaR terms are taken with respect to the nominal distribution) nor suppress (because the expectation term includes every scenario). Of course, this behavior can also be detected by looking at the respective limits described above. □

Example 9 (Revisiting the Convex Combination of CVaR and Worst Case of Example 4). The variation distance has $\lim_{t\searrow 0}\phi_v(t) = 1 < \infty$, indicating that it can suppress scenarios. This is evident in the CVaR term because the positive part in the expected value, $\mathbb{E}[[h(\mathbf{x}) - \eta]^+]$, will suppress some scenarios. Furthermore, $\lim_{t\searrow 0}\phi_v'(t) = -1 > -\infty$, indicating that suppression will occur one at a time. This behavior can be observed from the dependence of the CVaR term on ρ. As ρ increases, the CVaR term considers further tail probabilities and ignores more scenarios. The variation distance also has $\lim_{t\nearrow\infty}(\phi_v(t)/t) = 1 < \infty$; therefore, it can pop scenarios. The popping behavior shows up in the $\sup_\omega h_\omega(\mathbf{x})$ term. This term will pop the most expensive scenario(s). □

7.3. Modeling Considerations When Choosing a Phi-Divergence

We offer the following suggestions for choosing an appropriate phi-divergence type for the data available.

- *Phi-Divergences That Cannot Suppress Scenarios:* If the problem scenarios come from high-quality data or collected observations that should be considered with positive probability in the final model, the decision maker may wish to avoid phi-divergences that can suppress scenarios.
- *Phi-Divergences That Can Suppress Scenarios:* If the data are poorly sampled or come 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 the robustness level on the optimal solutions.
- *Phi-Divergences That Cannot Pop Scenarios:* If the problem scenarios come strictly from observation, with little theoretical understanding of the problem, choosing a phi-divergence that cannot pop scenarios will eliminate any unobserved scenarios from showing up in the solution.
- *Phi-Divergences That Can Pop Scenarios:* If the problem scenarios, however, 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.

Finally, a decision maker who has the highest flexibility may wish to use phi-divergences that can both pop and suppress scenarios so that the model finds the appropriate probabilities in a risk-averse manner. In Table 2, the variation distance and the Hellinger distance belong to this category. In this case, suppression one at a time again is preferred to examine the effect of the level of robustness on individual scenarios.

8. Conclusion and Future Directions

In this tutorial, we introduced two-stage ambiguous stochastic linear programs with recourse, where the distributional ambiguity is handled via phi-divergences. We formulated the problem and discussed its properties. An attractive feature of these models is that they can be solved efficiently via decomposition-based methods. Using phi-divergences in this setting has several advantages. One important aspect of phi-divergences is that they preserve convexity. Therefore, many of the tools of convex analysis and convex optimization can be used to analyze them and solve the resulting models. Furthermore, many phi-divergences are already being used in statistics, making them attractive when dealing with data directly.

Given that there are many phi-divergences to use, we examined their differences when used in an optimization context. This led to a classification of phi-divergences, with four main categories and two subcategories within the phi-divergences that can suppress scenarios. We illustrated this classification on several examples and provided guidelines on what class of phi-divergence to use under what data type.

There are many avenues of further research. For instance, (i) extensions to continuous distributions, (ii) extensions to multistage programs, and (iii) examination of rates of convergence are among the areas that merit further research. Although some applications have started appearing in the literature—some for specific phi-divergences (e.g., Calafiore [6], Klabjan et al. [19]) and some recently for general phi-divergences (Love and Bayraksan [21])—further applications of this class of problems would be beneficial. Finally, refining the classification presented here and examining the properties of different phi-divergence classes for other types of problems would be valuable.

Acknowledgments

This work is supported in part by the National Science Foundation [Grant CMMI-1345626].

References

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

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

[3] A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2):341–357, 2013.

[4] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, New York, 2011.

[5] P. Buchholz and A. Thümmler. Enhancing evolutionary algorithms with statistical selection procedures for simulation optimization. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds. *Proceedings of the 2005 Winter Simulation Conference*, IEEE, Piscataway, NJ, 842–852, 2005.

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

[7] A. Charnes and W. W. Cooper. Chance-constrained programming. *Management Science* 6(1):73–79, 1959.

[8] A. Charnes, W. W. Cooper, and G. H. Symonds. Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil. *Management Science* 4(3):235–263, 1958.

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

[10] J. Dupačová. The minimax approach to stochastic programming and an illustrative application. *Stochastics* 20(1):73–88, 1987.

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

[12] M. C. Fu, ed. *Handbook of Simulation Optimization*, International Series in Operations Research and Management Science, Vol. 216. Springer, New York, 2015.

[13] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research* 58(4, Part 1):902–917, 2010.

[14] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann. A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming Series B* 151(1):35–62, 2015.

[15] S. Henderson and R. Pasupathy. Simulation optimization library home page. http://simopt.org, accessed June 1, 2015.

[16] Z. Hu and L. J. Hong. Kullback-Leibler divergence constrained distributionally robust optimization. Technical report, Hong Kong University of Science and Technology, Clear Water Bay, 2013.

[17] R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. Technical report, University of Florida, Gainesville, 2013.

[18] R. Jiang and Y. Guan. Risk-averse two-stage stochastic program with distributional ambiguity. Technical report, University of Florida, Gainesville, 2015.

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

[20] D. K. Love. Data-driven methods for optimization under uncertainty with application to water allocation. Ph.D. dissertation, University of Arizona, Tucson, 2013.

[21] D. Love and G. Bayraksan. A data-driven method for robust water allocation under uncertainty. Technical report, the Ohio State University, Columbus, 2015.

[22] D. Love and G. Bayraksan. Phi-divergence constrained ambiguous stochastic programs for data-driven optimization. Technical report, the Ohio State University, Columbus, 2015.

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

[24] L. Pardo. *Statistical Inference Based on Divergence Measures*. Chapman and Hall/CRC, Boca Raton, FL, 2005.

[25] R. Pasupathy and S. Ghosh. Simulation optimization: A concise overview and implementation guide. H. Topaloglu, ed. *Theory Driven by Influential Applications*, Tutorials in Operations Research. INFORMS, Hanover, MD, 122–150, 2013.

[26] N. Patsis, C. Chen, and M. Larson. SIMD parallel discrete-event dynamic system simulation. *IEEE Transactions on Control Systems Technology* 5(1):30–41, 1997.

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

[28] H. Rahimian, G. Bayraksan, and T. Homem de Mello. Ambiguous stochastic programs with variation distance. Working paper, Ohio State University, Columbus, 2015.

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

[30] R. T. Rockafellar. Coherent approaches to risk in optimization under uncertainty. T. Klastorin, ed. *OR Tools and Applications: Glimpses of Future Technologies*, Tutorials in Operations Research. INFORMS, Hanover, MD, 38–61, 2007.

[31] R. T. Rockafellar and S. Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science* 18(1–2):33–53, 2013.

[32] A. Ruszczyński and A. Shapiro. Optimization of convex risk functions. *Mathematics of Operations Research* 31(3):433–452, 2006.

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

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

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

[36] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*, MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2009.

[37] Z. Wang, P. Glynn, and Y. Ye. Likelihood robust optimization for data-driven problems. Technical report, Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN, 2013.

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

[39] I. Yanıkoğlu and D. den Hertog. Safe approximations of ambiguous chance constraints using historical data. *INFORMS Journal on Computing* 25(4):666–681, 2013.

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