OJMO

Open Journal of Mathematical Optimization

Hamed Rahimian & Sanjay Mehrotra

Frameworks and Results in Distributionally Robust Optimization

Volume 3 (2022), article no. 4 (85 pages)

https://doi.org/10.5802/ojmo.15

Article submitted on December 13, 2021, revised on June 26, 2022,
accepted on July 11, 2022.

CENTRE MERSENNE

Open Journal of Mathematical Optimization is a member of the
Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/
e-ISSN: 2777-5860

Frameworks and Results in Distributionally Robust Optimization

Hamed Rahimian
Department of Industrial Engineering, Clemson University, Clemson, SC 29634, USA
hrahimi@clemson.edu

Sanjay Mehrotra
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, USA
mehrotra@northwestern.edu

Abstract

The concepts of risk aversion, chance-constrained optimization, and robust optimization have developed significantly over the last decade. The statistical learning community has also witnessed a rapid theoretical and applied growth by relying on these concepts. A modeling framework, called *distributionally robust optimization* (DRO), has recently received significant attention in both the operations research and statistical learning communities. This paper surveys main concepts and contributions to DRO, and relationships with robust optimization, risk aversion, chance-constrained optimization, and function regularization. Various approaches to model the distributional ambiguity and their calibrations are discussed. The paper also describes the main solution techniques used to the solve the resulting optimization problems.

Digital Object Identifier 10.5802/ojmo.15

2020 Mathematics Subject Classification 90C15, 90C22, 90C25, 90C30, 90C34, 90C90, 68T37, 68T05.

Keywords Distributionally robust optimization; Robust optimization; Stochastic optimization; Risk-averse optimization; Chance-constrained optimization; Statistical learning.

Contents

1 Introduction

Many real-world decision problems arising in engineering and management have uncertain parameters. This parameter uncertainty may be due to limited observability of data, noisy measurements, implementations and prediction errors. *Stochastic optimization* (SO) and *robust optimization* (RO) frameworks have classically allowed to model this uncertainty within a decision-making framework. SO assumes that the decision maker has *complete* knowledge about the underlying uncertainty through a *known* probability distribution and minimizes a functional of the cost, see, e.g., Birge and Louveaux [57], Shapiro et al. [374]. The probability distribution of the random parameters can be inferred from prior beliefs, expert opinions, errors in predictions based on the historical data (e.g., Kim and Mehrotra [220]), or a mixture of these. In RO, on the other hand, it is assumed that the decision maker has no distributional knowledge about the underlying uncertainty, except for its support, and the model

minimizes the worst-case cost over an uncertainty set, see, e.g., Ben-Tal et al. [30], Ben-Tal and Nemirovski [22, 23], Bertsimas and Sim [43], El Ghaoui and Lebret [133], El Ghaoui et al. [134][1].

We often have partial knowledge on the statistical and/or structural properties of the underlying probability distribution of the model parameter uncertainty, such as mean, symmetry, and unimodality. Specifically, the probability distribution quantifying the model parameter uncertainty is known ambiguously[2]. A typical approach to handle this ambiguity, from a statistical point of view, is to estimate the probability distribution using statistical tools, such as the maximal likelihood estimator, minimum Hellinger distance estimator (Vidyashankar and Xu [402]), or maximum entropy principle (Grünwald and Dawid [166]). The decision-making process can then be performed with respect to the estimated distribution. Because such an estimation may be imprecise, the impact of inaccuracy in estimation (and the subsequent ambiguity in the underlying distribution) is widely studied in the literature through (1) the perturbation analysis of optimization problems, see, e.g., Bonnans and Shapiro [69], (2) stability analysis of a SO model with respect to a change in the probability distribution, see, e.g, Rachev [323], Römisch [341], or (3) input uncertainty analysis in stochastic simulation models, see, e.g., Lam [227] and references therein. The typical goals of these approaches are to quantify the sensitivity of the optimal value/solution(s) to the probability distribution and provide continuity and/or large-deviation-type results, see, e.g., Dupačová [128], Heitsch et al. [185], Pflug and Pichler [299], Rachev and Römisch [324], Schultz [353]. While these approaches quantify the input uncertainty, they do not provide a systematic decision-making framework to hedge against the ambiguity in the underlying probability distribution.

Ambiguous stochastic optimization is a systematic modeling approach that bridges the gap between data and decision-making (statistics and optimization frameworks) to protect the decision-maker from the ambiguity in the underlying probability distribution. The ambiguous stochastic optimization approach assumes that the underlying probability distribution is unknown and lies in an *ambiguity set* of probability distributions. As in robust optimization, this approach hedges against the ambiguity in the probability distribution by taking a worst-case approach. Scarf [351] is arguably the first to consider such an approach to obtain an order quantity for a newsvendor problem to maximize the worst-case expected profit, where the worst-case is taken with respect to all product demand probability distributions with a known mean and variance. Since Scarf's seminal work, and particularly in the past few years, significant research has been done on the ambiguous stochastic optimization model. This paper provides a review of the theoretical, modeling, and computational developments in this area. This paper also puts the ambiguous stochastic optimization approach in the context of risk-averse optimization, chance-constrained optimization, and robust optimization.

1.1 A General DRO Model

We now formally introduce the model formulation that we discuss in this paper. Let $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^n$ be the decision vector. On a measurable space $(\Xi, \mathcal{F})$, let us define a random vector $\boldsymbol{\xi} : \Xi \mapsto \Omega \subseteq \mathbb{R}^d$ and random functions $h_j(\boldsymbol{x}, \boldsymbol{\xi}) : \mathcal{X} \times \Xi \mapsto \mathbb{R}$, $j = 0, 1, \ldots, m$. Given this setup, a general stochastic optimization problem has the form

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \left\{ \mathcal{R}_P \left[h_0(\boldsymbol{x}, \boldsymbol{\xi}) \right] \middle| \, \mathcal{R}_P \left[h_j(\boldsymbol{x}, \boldsymbol{\xi}) \right] \leq 0, \ j = 1, \ldots, m \right\}, \tag{SO}$$

where P denotes a (known) probability measure on $(\Xi, \mathcal{F})$ and $\mathcal{R}_P : \mathcal{Z} \mapsto \mathbb{R}$ denotes a real-valued functional under P, where $\mathcal{Z}$ is a linear space of measurable functions on $(\Xi, \mathcal{F})$. The functional $\mathcal{R}_P$ accounts for quantifying the uncertainty in the outcomes of the decision under P. This setup represents a broad range of problems in statistics, optimization, and control, such as regression and classification models (Friedman et al. [144], James et al. [210]), simulation-optimization (Fu [145], Pasupathy and Ghosh [295]), stochastic optimal control (Bertsekas [37]), Markov decision processes (Puterman [321]), and stochastic programming (Birge and Louveaux [57], Shapiro et al. [374]).

[1] The concept of robustness in mathematical programming is developed independently in Mulvey et al. [270] and Ben-Tal and Nemirovski [22]. Both papers share the same name, robust optimization, and pursue the same goal to address uncertainty, yet in distinct ways. Mulvey et al. [270] propose robust optimization of large-scale systems to address the case that data takes values from a discrete scenario set, by using a regularization of the objective function to control its sensitivity and a penalty function to control constraint violation. On the other hand, Ben-Tal and Nemirovski [22] propose robust convex optimization for problems with data uncertainty described by an ellipsoid, by taking a worst-case approach. In this paper, we deal with uncertainty in the sense of Ben-Tal and Nemirovski [22], a worst-case approach.

[2] The concept of ambiguity is defined in the sense of Knight Knight [222], where the probability distribution of the unknown uncertain parameters is unknown/uncertain itself. This concept is different from the assumption in stochastic programming, where the probability distribution of the random parameters is known.

As a special case of (SO), we have the classical stochastic programming problems:

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_P\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right], \tag{1}$$

and

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \left\{h_0(\boldsymbol{x}) \,|\, \mathbb{E}_P\left[h_j(\boldsymbol{x},\boldsymbol{\xi})\right] \le 0,\ j=1,\dots,m\right\}, \tag{2}$$

where $\mathcal{R}_P[\,\cdot\,]$ is taken as the expected-value functional $\mathbb{E}_P[\,\cdot\,]$. Note that when

$$h_0(\boldsymbol{x},\boldsymbol{\xi}) = \inf_{\boldsymbol{y}(\boldsymbol{\xi})} \left\{\phi_0(\boldsymbol{x},\boldsymbol{y}(\boldsymbol{\xi}),\boldsymbol{\xi}) \,\middle|\, \phi_l(\boldsymbol{x},\boldsymbol{y}(\boldsymbol{\xi}),\boldsymbol{\xi}) \ge 0,\ l=1,\dots,\kappa,\ \boldsymbol{y}(\boldsymbol{\xi}) \in \mathbb{Z}^{q_1}\times\mathbb{R}^{q-q_1}\right\},$$

we have the class of two-stage stochastic programs with recourse. In particular,

$$h_0(\boldsymbol{x},\boldsymbol{\xi}) = \inf_{\boldsymbol{y}(\boldsymbol{\xi})} \boldsymbol{c}^\top\boldsymbol{x} + \boldsymbol{q}^\top(\boldsymbol{\xi})\boldsymbol{y}(\boldsymbol{\xi}) \quad \text{s.t.} \begin{cases} \boldsymbol{W}(\boldsymbol{\xi})\boldsymbol{y}(\boldsymbol{\xi}) \ge \boldsymbol{r}(\boldsymbol{\xi}) - \boldsymbol{T}(\boldsymbol{\xi})\boldsymbol{x}, \\ \boldsymbol{y}(\boldsymbol{\xi}) \in \mathbb{Z}^{q_1}\times\mathbb{R}^{q-q_1}, \end{cases} \tag{3}$$

corresponds to the class of two-stage stochastic programs with a linear recourse with mixed integer variables in the second stage.

Moreover, by taking $h_0(\boldsymbol{x},\cdot) := \mathbb{1}_{S_0(\boldsymbol{x})}(\,\cdot\,)$ in (1), where $\mathbb{1}_{S_0(\boldsymbol{x})}(\,\cdot\,)$ denotes an indicator function for an arbitrary set $S_0(\boldsymbol{x}) \subseteq \mathcal{B}(\mathbb{R}^d)$ (we define the indicator function and $\mathcal{B}(\mathbb{R}^d)$ precisely in Section 1.4), we obtain the class of problems with a probabilistic objective function, see, e.g., Prékopa [317], as follows

$$\inf_{\boldsymbol{x}\in\mathcal{X}} P\{\boldsymbol{\xi} \in S_0(\boldsymbol{x})\}, \tag{4}$$

The set $S_0(\boldsymbol{x})$ may be in the form of a single constraint[3], e.g.,

$$S_0(\boldsymbol{x}) := \left\{\boldsymbol{\xi}\in\Xi \,\middle|\, \boldsymbol{a}(\boldsymbol{x})^\top\boldsymbol{\xi} \le b(\boldsymbol{x})\right\} \text{ or } S_0(\boldsymbol{x}) := \left\{\boldsymbol{\xi}\in\Xi \,\middle|\, \boldsymbol{a}(\boldsymbol{\xi})^\top\boldsymbol{x} \le b(\boldsymbol{\xi})\right\},$$

or several constraints, e.g.,

$$S_0(\boldsymbol{x}) := \{\boldsymbol{\xi}\in\Xi \,|\, \boldsymbol{A}(\boldsymbol{x})\boldsymbol{\xi} \le \boldsymbol{b}(\boldsymbol{x})\} \text{ or } S_0(\boldsymbol{x}) := \{\boldsymbol{\xi}\in\Xi \,|\, \boldsymbol{A}(\boldsymbol{\xi})\boldsymbol{x} \le \boldsymbol{b}(\boldsymbol{\xi})\}.$$

Similarly, by taking $h_j(\boldsymbol{x},\cdot) := \mathbb{1}_{S_j(\boldsymbol{x})}(\,\cdot\,)$, $j=1,\dots,m$, for suitable indicator functions $\mathbb{1}_{S_j(\boldsymbol{x})}(\,\cdot\,)$, $j=1,\dots,m$, (2) is in the form of a chance-constrained program as follows (see, e.g., Charnes and Cooper [79], Charnes et al. [81], Dentcheva [111], Prékopa [315, 316]):

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \left\{h_0(\boldsymbol{x}) \,|\, P\{\boldsymbol{\xi}\in S_j(\boldsymbol{x})\} \le 0,\ j=1,\dots,m\right\}. \tag{5}$$

Note that the case where the set $S_j(\boldsymbol{x})$ is formed via several constraints is called *joint chance constraint* as compared to *individual chance constraint*, where the event $S_j(\boldsymbol{x})$ is formed via one constraint, $j=1,\dots,m$[4].

A robust optimization model is defined as

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \left\{\sup_{\boldsymbol{\xi}\in\mathcal{U}} h_0(\boldsymbol{x},\boldsymbol{\xi}) \,\middle|\, \sup_{\boldsymbol{\xi}\in\mathcal{U}} h_j(\boldsymbol{x},\boldsymbol{\xi}) \le 0,\ j=1,\dots,m\right\}, \tag{RO}$$

where $\mathcal{U}\subseteq\mathbb{R}^d$ denotes an *uncertainty set* for the parameters $\boldsymbol{\xi}$. Similar to (SO),

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\boldsymbol{\xi}\in\mathcal{U}} h_0(\boldsymbol{x},\boldsymbol{\xi}) \tag{6}$$

and

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \left\{h_0(\boldsymbol{x}) \,\middle|\, \sup_{\boldsymbol{\xi}\in\mathcal{U}} h_j(\boldsymbol{x},\boldsymbol{\xi}) \le 0,\ j=1,\dots,m\right\} \tag{7}$$

---

[3] We say a set of the form $S_0(\boldsymbol{x}) = \left\{\boldsymbol{\xi}\in\Xi \,\middle|\, \boldsymbol{a}(\boldsymbol{x})^\top\boldsymbol{\xi} \le b(\boldsymbol{x})\right\}$ is bi-affine in $\boldsymbol{x}$ and $\boldsymbol{\xi}$ if $\boldsymbol{a}(\boldsymbol{x})$ and $b(\boldsymbol{x})$ are both affine in $\boldsymbol{x}$. Similarly, we say a set of the form $S_0(\boldsymbol{x}) = \left\{\boldsymbol{\xi}\in\Xi \,\middle|\, \boldsymbol{a}(\boldsymbol{\xi})^\top\boldsymbol{x} \le b(\boldsymbol{\xi})\right\}$ is bi-affine in $\boldsymbol{x}$ and $\boldsymbol{\xi}$ if $\boldsymbol{a}(\boldsymbol{\xi})$ and $b(\boldsymbol{\xi})$ are both affine in $\boldsymbol{\xi}$.

[4] Observe that a bi-affine set of the form $S_0(\boldsymbol{x}) = \left\{\boldsymbol{\xi}\in\Xi \,\middle|\, \boldsymbol{a}(\boldsymbol{x})^\top\boldsymbol{\xi} \le b(\boldsymbol{x})\right\}$ can be equivalently written as a bi-affine set of the form $S_0(\boldsymbol{x}) = \left\{\boldsymbol{\xi}\in\Xi \,\middle|\, \boldsymbol{a}(\boldsymbol{\xi})^\top\boldsymbol{x} \le b(\boldsymbol{\xi})\right\}$, and vice versa. With a parallel reasoning, we can define $h_j(\,\cdot\,,\boldsymbol{\xi}) := \mathbb{1}_{S_j(\boldsymbol{\xi})}(\,\cdot\,)$, where $\mathbb{1}_{S_j(\boldsymbol{\xi})}(\,\cdot\,)$ denotes an indicator function for an arbitrary set $S_j(\boldsymbol{\xi}) \subseteq \mathbb{R}^n$, $j=0,1,\dots,m$. Hence, we obtain the probabilistic objective function $P\{\boldsymbol{x}\in S_0(\boldsymbol{\xi})\}$, and the probabilistic constraints $P\{\boldsymbol{x}\in S_j(\boldsymbol{\xi})\} \le 0$, $j=1,\dots,m$.

are two special cases of (RO).

Problem (SO), as well as (1) and (2), require the knowledge of the underlying measure P, whereas (RO), as well as (6) and (7), ignore all distributional knowledge of $\boldsymbol{\xi}$, except for its support. An ambiguous version of (SO) is formulated as

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \left\{ \sup_{P\in\mathcal{P}} \mathcal{R}_P \left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right] \,\middle|\, \sup_{P\in\mathcal{P}} \mathcal{R}_P \left[h_j(\boldsymbol{x},\boldsymbol{\xi})\right] \leq 0,\ j=1,\ldots,m \right\}. \tag{DRO}$$

Here, $\mathcal{P}$ denotes an *ambiguity set of probability measures*, e.g., a family of measures consistent with the prior knowledge about uncertainty. Note that if we consider the measurable space $\left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right)$, as opposed to $(\Xi,\mathcal{F})$, then $\mathcal{P}$ can be viewed as an ambiguity set of probability distributions $\mathbb{P}$ induced by $\boldsymbol{\xi}$[5].

As discussed before, (DRO) finds a decision that minimizes the worst-case of the functional $\mathcal{R}_P$ of the cost h_0 among all probability measures in the ambiguity set provided that the worst-case of the functional $\mathcal{R}_P$ of the function $\boldsymbol{h}_j$, $j=1,\ldots,m$, is non-positive. The ambiguous versions of (1) and (2) are formulated as follows:

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{P\in\mathcal{P}} \mathbb{E}_P \left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right], \tag{8}$$

and

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \left\{ h_0(\boldsymbol{x}) \,\middle|\, \sup_{P\in\mathcal{P}} \mathbb{E}_P \left[h_j(\boldsymbol{x},\boldsymbol{\xi})\right] \leq 0,\ j=1,\ldots,m \right\}. \tag{9}$$

Models (8) and (9) are discussed in the context of minimax stochastic optimization models, in which optimal solutions are evaluated under the worst-case expectation with respect to a family of probability distributions of the uncertain parameters, see, e.g., Scarf [351]; Žáčková [437] (a.k.a. Dupačová); Breton and El Hachem [71], Dupačová [127], Shapiro and Ahmed [369], Shapiro and Kleywegt [370]. Delage and Ye [105] refer to this approach as *distributionally robust optimization*, in short DRO, and since then, this terminology has become widely dominant in the research community. We adopt this terminology, and for the rest of the paper, we refer to the ambiguous stochastic optimization of the form (DRO) as DRO.

▶ Remark 1. Problem (8) yields a pessimistic minimax criterion for decision-making under uncertainty, which conservatively minimizes the worst-case expected cost. Alternatively, one may consider an optimistic minimin criterion $\inf_{\boldsymbol{x}\in\mathcal{X}} \inf_{P\in\mathcal{P}} \mathbb{E}_P \left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right]$, which aggressively minimizes the best-case expected cost. A trade-off between pessimistic and optimistic objectives, known as *Hurwicz* criterion, is formulated in Hurwicz [207] as

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \left\{ \lambda \sup_{P\in\mathcal{P}} \mathbb{E}_P \left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right] + (1-\lambda) \inf_{P\in\mathcal{P}} \mathbb{E}_P \left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right] \right\},$$

The minimin criterion, or the more general the Hurwicz criteria, typically leads to non-convex optimization problems.

In this paper, we mostly focus on static and two-stage problems of the form (8) and (9), while we also mention some results on their dynamic versions. For references, we introduce a dynamic version of (8), referred to as *multistage* DRO, below. To formally define the problem, assume that the uncertain parameters are realized over time and are represented by a stochastic process $\boldsymbol{\xi} := (\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_T)$, where $\boldsymbol{\xi}_t : \Omega \mapsto \Xi_t \subseteq \mathbb{R}^{d_t}$ and is composed of the random parameters in stage t. We assume that $\boldsymbol{\xi}_1$ is a degenerate random vector. Let $\boldsymbol{\xi}_{[t]} := (\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_t)$ denote the history of the stochastic process up to (and including) time t. A multistage DRO problem is formed as

$$\begin{aligned} \min_{x_1\in\mathcal{X}_1} h_1(\boldsymbol{x}_1,\boldsymbol{\xi}_1) + \max_{P_2\in\mathcal{P}_{2|\boldsymbol{\xi}_{[1]}}} \mathbb{E}_{P_2}\Bigg[\min_{\boldsymbol{x}_2\in\mathcal{X}_2} h_2(\boldsymbol{x}_2,\boldsymbol{\xi}_2) \\ + \max_{P_3\in\mathcal{P}_{3|\boldsymbol{\xi}_{[2]}}} \mathbb{E}_{P_3}\Bigg[\ldots + \max_{P_T\in\mathcal{P}_{T|\boldsymbol{\xi}_{[T-1]}}} \mathbb{E}_{P_T}\Big[\min_{x_T\in\mathcal{X}_T} h_T(\boldsymbol{x}_T,\boldsymbol{\xi}_T)\Big]\ldots\Bigg]\Bigg]. \end{aligned} \tag{10}$$

Above, in stage t, $t=1,\ldots,T$, the set-valued mapping $\mathcal{X}_t := \mathcal{X}_t(\boldsymbol{x}_{[t-1]},\boldsymbol{\xi}_{[t]}) \subset \mathbb{R}^{n_t}$ denotes a feasibility set and $h_t : \mathbb{R}^{n_t}\times\mathbb{R}^{d_t} \mapsto \mathbb{R}$ is a random function, given the decision $\boldsymbol{x}_t$ and the realized uncertainty $\boldsymbol{\xi}_t$. Moreover, $\mathcal{P}_{t+1|\boldsymbol{\xi}_{[t]}}$

[5] In this paper, we use $\mathcal{P}$ to denote both an ambiguity set of probability measures and an ambiguity set of distributions induced by $\boldsymbol{\xi}$. Whether $\mathcal{P}$ denotes an ambiguity set of probability measures or an ambiguity set of distributions induced by $\boldsymbol{\xi}$ should be understood from the context and the distinction we make between the notation of a probability measure P and a probability distribution $\mathbb{P}$.

is the *conditional ambiguity set* for the conditional distribution of $\boldsymbol{\xi}_{t+1}$, conditioned on $\boldsymbol{\xi}_{[t]}$, $t = 1, \ldots, T-1$. Also, $\mathbb{E}_{P_{t+1}}[\cdot]$ denotes the conditional expectation with respect to $P_{t+1} \in \mathcal{P}_{t+1|\boldsymbol{\xi}_{[t]}}$.

As mentioned before, (DRO) is a modeling approach that assumes only partial distributional information, whereas (SO) assumes complete distributional information. In fact, if $\mathcal{P}$ contains only the true distribution of the random vector $\boldsymbol{\xi}$, (DRO) reduces to (SO). On the other hand, if $\mathcal{P}$ contains all probability distributions on the support of the random vector $\boldsymbol{\xi}$, supported on $\mathcal{U}$, then, (DRO) reduces to (RO). Thus, a judicious choice of $\mathcal{P}$ can put (DRO) between (SO) and (RO). Consequently, (DRO) may not be as conservative as (RO), which ignores all distributional information, except for the support $\mathcal{U}$ of the uncertain parameters. (DRO) can be viewed as a unifying framework for (SO) and (RO) (see Qian et al. [322]).

Below, we discuss two different perspectives on the relationship between DRO and RO.

▶ Remark 2. Suppose that Ξ has M atoms, $\Xi = \{s_1, \ldots, s_M\}$. Then, for a fixed $x \in \mathcal{X}$, $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ has M possible outcomes $\{h_0(\boldsymbol{x}, \boldsymbol{\xi}(s_1)), \ldots, h_0(\boldsymbol{x}, \boldsymbol{\xi}(s_M))\}$. For short, let us write these outcomes as a vector $\boldsymbol{h}(\boldsymbol{x}) \in \mathbb{R}^M$, where $h_m(\boldsymbol{x}) := h_0(\boldsymbol{x}, \boldsymbol{\xi}(s_m))$. In (8), $\mathcal{P}$ is a subset of all probability measures on $\boldsymbol{\xi}$. So, one can think of $\mathcal{P}$ as a subset of all discrete probability distributions $\mathbb{P}$ on $\mathbb{R}^d$ induced by $\boldsymbol{\xi}$. That is, $\mathbb{P}$ can be identified with a vector $\boldsymbol{p} \in \mathbb{R}^M$. Consequently, $\mathcal{P}$ may be interpreted as a subset of $\mathbb{R}^M$. With this interpretation, (8) is written as

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \sup_{\boldsymbol{p} \in \mathcal{P}} \boldsymbol{p}^\top \boldsymbol{h}(\boldsymbol{x}). \tag{11}$$

Setting $f(\boldsymbol{x}, \boldsymbol{p}) := \boldsymbol{p}^\top \boldsymbol{h}(\boldsymbol{x})$, we can rewrite (11) as $\inf_{\boldsymbol{x} \in \mathcal{X}} \sup_{\boldsymbol{p} \in \mathcal{P}} f(\boldsymbol{x}, \boldsymbol{p})$. This problem has the form of (6), where the probability vector $\boldsymbol{p}$ takes values in an "uncertainty set" $\mathcal{P}$. For a through treatment of different nonlinear functions $f(\boldsymbol{x}, \boldsymbol{p})$ and different uncertainty sets $\mathcal{P}$, we refer to Ben-Tal et al. [33]. The other perspective on the relationship between DRO and RO arises when type of the probability distribution is known but its parameters are unknown. For example, the probability distribution may be assumed to be normal, but its mean and variance are unknown, see, e.g., Goldfarb and Iyengar [161]. With the above-mentioned perspectives, it can be said that problem-driven and statistical techniques that are applicable for specifying the uncertainty set in a RO model may now be used to specify $\mathcal{P}$ in (11), see, e.g., Ben-Tal and Nemirovski [23, 25], Bertsimas et al. [51, 46], Chen et al. [91], Long et al. [252] (see also Section 3.1.2). We also refer to Xu et al. [424] for a distributional interpretation of RO. DRO has the richness that allows Ξ to be continuous without specifying the type of the distribution a priori.

1.2 Motivation and Contributions

In this paper, we provide an overview of the main contributions to DRO within both operations research and machine learning communities. This paper is an adaptation of authors' unpublished manuscript Rahimian and Mehrotra [326]. While there are separate review papers on RO, see, e.g., Bertsimas et al. [49], Gabrel et al. [146], Gorissen et al. [163], to the best of our knowledge, there are only a few tutorials and survey papers on DRO within these communities. A tutorial on DRO, its connection to risk-averse optimization, and the use of ϕ-divergence to construct the ambiguity set is presented in Bayraksan and Love [17]. Kuhn et al. [224] study DRO models with Wasserstein ambiguity sets and discuss their applications in machine learning. Shapiro [368] provides a general tutorial on DRO and its connection to risk-averse optimization. Postek et al. [311] survey different papers that address distributionally robust risk constraints, with a variety of risk functionals and ambiguity sets. Similar to Bayraksan and Love [17], Postek et al. [311], Shapiro [368], in this paper, we show the connection between DRO and risk aversion. However, the current review is different from those in the literature from many different perspectives. We outline our contributions as follows:

- We bring together the research done on DRO within the operations research and machine learning communities. This motivation is materialized throughout the paper as we take a holistic view of DRO, from modeling, to solution techniques and to machine learning applications.
- We provide a detailed discussion on how DRO models are connected to different concepts such as game theory, risk-averse optimization, chance-constrained optimization, robust optimization, and function regularization in statistical learning.
- From the algorithmic perspective, we review techniques to solve a DRO model.
- From the modeling and theoretical perspectives, we categorize different approaches to model the distributional ambiguity and discuss results for each of these ambiguity sets, by focusing on the structure of functions $h_j(\boldsymbol{x}, \boldsymbol{\xi})$, $j = 0, 1, \ldots, m$. Moreover, we discuss the calibration of different parameters used in these ambiguity sets of distributions.

1.3 Organization of this Paper

This paper is organized as follows. Section 2 discusses the notion of optimality gap and generalization bound in optimization. Section 3 reviews the connection of DRO to different concepts: risk aversion and chance-constrained optimization with its relationship to robust optimization in Section 3.1, and regularization in statistical learning in Section 3.2. In Section 4, we review three main solution techniques to solve a DRO model by introducing tools in semi-infinite programming and duality. In Section 5, we discuss different functionals that quantify uncertainty in the outcomes of a fixed decision. This includes regret functions in Section 5.1, risk measures in Section 5.2, and utility functions in Section 5.3. In Section 6, we discuss different models to construct the ambiguity set of distributions. This includes discrepancy-based models in Section 6.1, moment-based models in Section 6.2, shape-preserving-based models in Section 6.3, and kernel-based models in Section 6.4. In Section 7, we discuss the calibration of different parameters used in the ambiguity set of distributions. In Section 8, we introduce some modeling toolboxes for a DRO model. We end the paper in Section 9 with some conclusions and directions of future research.

1.4 Notation and Basic Definitions

In this section, we introduce additional notation used throughout the paper. In order to keep the paper self-contained, we also introduce all definitions used in this paper. For a given space Ξ and a σ-field $\mathcal{F}$ of that space, we define an underlying measurable space $(\Xi, \mathcal{F})$. In particular, let us define $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, where $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-field on $\mathbb{R}^d$. Let $\mathbb{1}_A : \Xi \mapsto \{0, 1\}$ indicate the indicator function of set $A \in \mathcal{F}$ where $\mathbb{1}_A(s) = 1$ if $s \in A$, and 0 otherwise. Let $\mathfrak{M}_+(\Xi, \mathcal{F})$ and $\mathfrak{M}(\Xi, \mathcal{F})$ denote the set of all nonnegative measures and the set of all probability measures $Q : \mathcal{F} \mapsto [0, 1]$ defined on $(\Xi, \mathcal{F})$, respectively. A measure ν_2 is preferred over a measure ν_1, denoted as $\nu_2 \succeq \nu_1$ if $\nu_2(A) \geq \nu_1(A)$ for all measurable sets $A \in \mathcal{F}$. We denote by $Q\{A\}$ the probability of event $A \in \mathcal{F}$, with respect to $Q \in \mathfrak{M}(\Xi, \mathcal{F})$. A random vector on the measurable space $(\Xi, \mathcal{F})$ is defined as $\boldsymbol{\xi} : \Xi \mapsto \mathbb{R}^d$. We use the same notation to denote a realization of the random vector, and the distinction should be understood from the context. For a probability measure $Q \in \mathfrak{M}(\Xi, \mathcal{F})$, we define a probability space $(\Xi, \mathcal{F}, Q)$. We denote by $\mathbb{Q} := Q \circ \boldsymbol{\xi}^{-1}$ the probability distribution induced by a random vector $\boldsymbol{\xi}$ under Q, where $\boldsymbol{\xi}^{-1}$ denotes the inverse image of $\boldsymbol{\xi}$. That is, $\mathbb{Q} : \mathcal{B}(\mathbb{R}^d) \mapsto [0, 1]$ is a probability distribution on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Let $\mathfrak{P}(\cdot, \cdot)$ denote the set of all such probability distributions. For example, $\mathfrak{P}(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ denotes the set of all probability distributions of $\boldsymbol{\xi}$. Note that in our notation, we make a distinction between a probability measure $Q \in \mathfrak{M}(\Xi, \mathcal{F})$ and a probability distribution $\mathbb{Q} \in \mathfrak{P}(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$. Nevertheless, we have always an appropriate transformation, so we might use the terminology of probability measure and probability distribution interchangeably. Given this, for a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we may write $\int_\Xi f(\boldsymbol{\xi}(s))Q(\mathrm{d}s)$ equivalently as $\int_{\mathbb{R}^d} f(s)\mathbb{Q}(\mathrm{d}s)$ with a change of measure. As we shall see later, we may denote $f(\boldsymbol{\xi}(s))$ with $f(s)$ in this transformation. For two random variables $Z, Z' : \Xi \mapsto \mathbb{R}$, we use $Z \geq Z'$ to denote $Z(s) \geq Z'(s)$ almost everywhere (a.e.) on Ξ. A random variable Z is Q-integrable if $\|Z\|_1 := \int_\Xi |Z| \mathrm{d}Q$ is finite. Two random variables Z, Z' are distributionally equivalent, denoted by $Z \overset{\mathrm{d}}{\sim} Z'$, if they induce the same distribution, i.e., $Q\{Z \leq z\} = Q\{Z' \leq z\}$. We also denote by $\mathcal{S}(\Xi, \mathcal{F})$ the collection of all $\mathcal{F}$-measurable functions $Z : (\Xi, \mathcal{F}) \mapsto (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$, where $\overline{\mathbb{R}}$ denotes the extended real line $\mathbb{R} \cup \{-\infty, +\infty\}$.

For Ξ with M atoms $\Xi = \{s_1, \ldots, s_M\}$ and $\mathcal{F} = 2^\Xi$, let $\{q(s_1), \ldots, q(s_M)\}$ be the probabilities of the corresponding elementary events under probability measure $Q \in \mathfrak{M}(\Xi, \mathcal{F})$. As a shorthand notation, we use $\boldsymbol{q} = [q_1, \ldots, q_M]^T \in \mathbb{R}^M$, where $q_i := q(s_i)$, $i \in \{1, \ldots, M\}$. An $\mathcal{F}$-measurable function $Z : \Xi \mapsto \mathbb{R}$ has M outcomes $\{Z(s_1), \ldots, Z(s_M)\}$ with probabilities $\{q_1, \ldots, q_M\}$. For short, we identify Z as a vector in $\mathbb{R}^M$, i.e., $\boldsymbol{z} = [z_1, \ldots, z_M]^T$ with $z_i := Z(s_i)$, $i \in \{1, \ldots, M\}$.

Consider a linear space $\mathcal{V}$, paired with a dual linear space $\mathcal{V}^*$, in the sense that a (real-valued) bilinear form $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V}^* \mapsto \mathbb{R}$ is defined. That is, for any $v \in \mathcal{V}$ and $v^* \in \mathcal{V}^*$, we have that $\langle \cdot, v^* \rangle : \mathcal{V} \mapsto \mathbb{R}$ and $\langle v, \cdot \rangle : \mathcal{V}^* \mapsto \mathbb{R}$ are linear functionals on $\mathcal{V}$ and $\mathcal{V}^*$, respectively. Similarly, we define $\mathcal{W}$ and $\mathcal{W}^*$. For a linear mapping $\boldsymbol{A} : \mathcal{V} \mapsto \mathcal{W}$, we define the adjoint mapping $\boldsymbol{A}^* : \mathcal{W}^* \mapsto \mathcal{V}^*$ by means of the equation $\langle w^*, \boldsymbol{A}v \rangle = \langle \boldsymbol{A}^* w^*, v \rangle$, $\forall\, v \in \mathcal{V}$. For two linear mappings, defined by finite dimensional matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \bullet \boldsymbol{B} = Tr(\boldsymbol{A}^\top \boldsymbol{B})$ denotes the Frobenius inner product between matrices. Moreover, $\boldsymbol{A} \odot \boldsymbol{B}$ denotes the Hadamard (i.e., componentwise) product between matrices.

For a function $f : \mathcal{V} \mapsto \overline{\mathbb{R}}$, the (convex) conjugate $f^* : \mathcal{V}^* \mapsto \overline{\mathbb{R}}$ is defined as $f^*(v^*) = \sup_{v \in \mathcal{V}} \{\langle v^*, v \rangle - f(v)\}$. Similarly, the biconjugate $f^{**} : \mathcal{V} \mapsto \overline{\mathbb{R}}$ is defined as $f^{**}(v) = \sup_{v^* \in \mathcal{V}^*} \{\langle v^*, v \rangle - f^*(v^*)\}$. The characteristic function $\delta(\cdot | \mathcal{A})$ of a nonempty set $\mathcal{A} \in \mathcal{V}$ is defined as $\delta(v | \mathcal{A}) = 0$ if $v \in \mathcal{A}$, and $+\infty$ otherwise. The support

function of a nonempty set $\mathcal{A} \in \mathcal{V}$ is defined as the convex conjugate of the characteristic function $\delta(\cdot|\mathcal{A})$: $\delta^*(v^*|\mathcal{V}) = \sup_{v\in\mathcal{V}}\{\langle v^*, v\rangle - \delta(v|\mathcal{A})\} = \sup_{v\in\mathcal{V}}\langle v^*, v\rangle$.

For $Q \in \mathfrak{M}(\Xi, \mathcal{F})$, let $\mathcal{L}_\infty(\Xi, \mathcal{F}, Q)$ be the linear space of all essentially bounded $\mathcal{F}$-measurable functions Z. A function Z is essentially bounded if $\|Z\|_\infty := \operatorname{ess\,sup}_{s\in\Omega}|Z(s)|$ is finite, where

$$\operatorname*{ess\,sup}_{s\in\Xi}|Z(s)| := \inf\left\{\sup_{s\in\Xi}|Z'(s)| \;\middle|\; Z(s) = Z'(s)\text{a.e.}s \in \Xi\right\}.$$

We denote by $\|\cdot\|_p : \mathbb{R}^d \mapsto \mathbb{R}$ the ℓ_p-norm on $\mathbb{R}^d$. That is, for a vector $\boldsymbol{u} \in \mathbb{R}^d$, $\|\boldsymbol{u}\|_p = \left(\sum_{i=1}^d |u_i|^p\right)^{\frac{1}{p}}$. We use Δ^d to denote the simplex in $\mathbb{R}^d$, i.e., $\Delta^d = \left\{\boldsymbol{u} \in \mathbb{R}^d \,\middle|\, \boldsymbol{e}^\top\boldsymbol{u} = 1,\ \boldsymbol{u} \geq \boldsymbol{0}\right\}$, where $\boldsymbol{e}$ is a vector of ones in $\mathbb{R}^d$.

For a proper cone $\mathcal{K}$, the relation $x \preccurlyeq_\mathcal{K} y$ indicates that $y - x \in \mathcal{K}$. For simplicity, we drop $\mathcal{K}$ from the notation, when $\mathcal{K}$ is the positive semidefinite cone. Let $\mathcal{S}^n_+$ denote the cone of symmetric positive semidefinite matrices in the $n\times n$ matrix spaces $\mathbb{R}^{n\times n}$. For a cone $\mathcal{K} \subset \mathcal{V}$, we define its dual cone as $\mathcal{K}' := \{v^* \in \mathcal{V}^* \,|\, \langle v^*, v\rangle \geq 0,\ \forall\, v \in \mathcal{K}\}$. The negative of the dual cone is called polar cone and is denoted by $\mathcal{K}^\circ$. The $\mathcal{K}$-epigraph of a function $\boldsymbol{f} : \mathbb{R}^N \mapsto \mathbb{R}^M$ and a proper cone $\mathcal{K}$ is conic-representable if the set $\left\{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^N \times \mathbb{R}^M \,\middle|\, \boldsymbol{f}(\boldsymbol{x}) \preccurlyeq_\mathcal{K} \boldsymbol{y}\right\}$ can be expressed via conic inequalities, possibly involving a cone different from $\mathcal{K}$ and additional auxiliary variables.

For a set $\mathcal{K}$, we use $\operatorname{conv}(\mathcal{K})$ and $\operatorname{int}(\mathcal{K})$ to denote the convex hull and the interior of $\mathcal{K}$, respectively. We let $[d]$ denote the index set $\{1, \ldots, d\}$. We also let $(\,\cdot\,)_+$ denote $\max\{0, \cdot\}$.

Acknowledgement

The authors thank the Associate Editor and two anonymous referees for their insightful comments.

2 Optimality Gap and Performance Guarantees

In this section, we first discuss the optimality gap and performance guarantees in solving (1) using a data-driven approach. Then, we discuss these concepts in solving (8) using a data-drive approach.

For every approach that uses a set of (training) samples to prescribe a solution or to predict an outcome, it is important to assess the *out-of-sample* quality of the prescriber/predictor under a new set of (test) samples, independent from the training samples $\{\boldsymbol{\xi}^i\}_{i=1}^N$. Let $\widehat{\mathbb{P}}_N$ be a nominal probability distribution estimated based on $\{\boldsymbol{\xi}^i\}_{i=1}^N$. Because $\widehat{\mathbb{P}}_N$ is a function of $\{\boldsymbol{\xi}^i\}_{i=1}^N$, hence, it is random itself. Suppose that $\mathbb{P}^N$ denotes the sampling distribution of $\widehat{\mathbb{P}}_N$, or N training samples. Also, let $\mathbb{P}^{\text{true}}$ be the unknown true distribution. Let us consider a data-driven solution (or, in-sample solution) $\boldsymbol{x}_N$ that is constructed using $\{\boldsymbol{\xi}^i\}_{i=1}^N$. The in-sample risk of $\boldsymbol{x}_N$ is defined as $\mathcal{R}_{\hat{\mathbb{P}}_N}[h_0(\boldsymbol{x}_N, \boldsymbol{\xi})]$. Additionally, The out-of-sample risk of $\boldsymbol{x}_N$ is defined as $\mathcal{R}_{\mathbb{P}^{\text{true}}}[h_0(\boldsymbol{x}_N, \boldsymbol{\xi})]$, which is the risk of $\boldsymbol{x}_N$ given a new (test) sample that is independent of $\{\boldsymbol{\xi}^i\}_{i=1}^N$, drawn from the unknown true distribution $\mathbb{P}^{\text{true}}$.

▶ **Proposition 3.** *Suppose that $\widehat{\mathbb{P}}_N$ is a nominal probability distribution estimated based on $\{\boldsymbol{\xi}^i\}_{i=1}^N$, governed by the distribution $\mathbb{P}^N$. For any data-driven solution $\boldsymbol{x}_N$, suppose that the in-sample risk $\mathcal{R}_{\hat{\mathbb{P}}_N}[h_0(\boldsymbol{x}_N, \boldsymbol{\xi})]$ underestimates the out-of-sample performance $\mathcal{R}_{\mathbb{P}^{true}}[h_0(\boldsymbol{x}_N, \boldsymbol{\xi})]$ on average, i.e., $\mathbb{E}_{\mathbb{P}^N}\left[\mathcal{R}_{\hat{\mathbb{P}}_N}[h_0(\boldsymbol{x}_N, \boldsymbol{\xi})]\right] \leq \mathcal{R}_{\mathbb{P}^{true}}[h_0(\boldsymbol{x}_N, \boldsymbol{\xi})]$. Then, any minimizer $\boldsymbol{x}^*_N$ of $\inf_{\boldsymbol{x}\in\mathcal{X}}\mathcal{R}_{\hat{\mathbb{P}}_N}[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$ yields an optimistically biased risk on average.*

Proof. See Appendix A. ◀

Having a negative optimality gap on average is a known result in SO, see, e.g., Bayraksan and Morton [18, 19], Homem-de-Mello and Bayraksan [192]. A data-driven solution $\boldsymbol{x}^*_N$ for a problem of the form (1) can be obtained by solving a *sample average approximation* (SAA) of that problem, where the underlying distribution is chosen to be $\widehat{\mathbb{P}}_N$ Shapiro et al. [374]. It is well-known that SAA yields an optimistically biased optimal value on average even if $\widehat{\mathbb{P}}_N$ is an unbiased estimator of $\mathbb{P}^{\text{true}}$.

Nonetheless, as $\mathbb{P}^{\text{true}}$ is unknown, one need to establish performance guarantees. One such guarantee, referred to as *finite-sample performance guarantee* is defined for any fixed N and $\delta > 0$ as

$$\mathbb{P}^N\left\{\mathcal{R}_{\mathbb{P}^{\text{true}}}[h_0(\boldsymbol{x}^*_N, \boldsymbol{\xi})] \leq \widehat{V}_N + \delta\right\} \geq 1 - \alpha, \tag{12}$$

which guarantees that an (in-sample) *certificate* $\widehat{V}_N$ provides a $(1-\alpha)$ confidence (with respect to the probability distribution $\mathbb{P}^N$) on the out-of-sample performance of $\boldsymbol{x}^*_N$, and of course, $\inf_{\boldsymbol{x}\in\mathcal{X}}\mathcal{R}_{\mathbb{P}^{\text{true}}}[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$. The certificate $\widehat{V}_N$ can be chosen as $\mathcal{R}_{\hat{\mathbb{P}}_N}[h_0(\boldsymbol{x}^*_N, \boldsymbol{\xi})] = \inf_{\boldsymbol{x}\in\mathcal{X}}\mathcal{R}_{\hat{\mathbb{P}}_N}[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$. The other guarantee, referred to as *asymptotic*

consistency, guarantees that as N increases, the certificate $\widehat{V}_N$ and the data-driven solution $\boldsymbol{x}_N^*$ converges (in some sense) to the optimal value and an optimal solution of the true problem $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathcal{R}_{\mathbb{P}^{\text{true}}}[h_0(\boldsymbol{x},\boldsymbol{\xi})]$.

Now, we introduce the analog of such performance guarantees that are used to assess the quality of a solution in the context of a DRO model. For the ease of exposition, let us focus on a DRO problem of the form $\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{P\in\mathcal{P}} \mathcal{R}_P[h_0(\boldsymbol{x},\boldsymbol{\xi})]$. As before consider a data-driven solution $\boldsymbol{x}_N^* \in \mathcal{X}$. Such a solution may be obtained by solving a data-driven version of the DRO model, where the ambiguity set $\mathcal{P}$ is constructed using data, namely $\mathcal{P}_N$. The finite-sample and asymptotic performance guarantees are defined in similar manners as before except for that now the certificate $\widehat{V}_N$ may be chosen as the optimal value of the inner problem in DRO, where the worst-case is taken within $\mathcal{P}_N$, evaluated at $\boldsymbol{x}_N^*$, i.e., $\widehat{V}_N = \sup_{P\in\mathcal{P}_N} \mathcal{R}_P[h_0(\boldsymbol{x}_N^*,\boldsymbol{\xi})] = \inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{P\in\mathcal{P}_N} \mathcal{R}_P[h_0(\boldsymbol{x},\boldsymbol{\xi})]$.

3 Relationship with Risk Aversion, Chance-Constrained Optimization, and Regularization

The modeling approach DRO is closely related to other concepts in operations research and statistical learning, such as robust optimization, risk aversion, chance-constrained optimization, and function regularization. The connection to risk aversion, chance-constrained optimization, and robust optimization is discussed in Section 3.1, and regularization in statistical learning is discussed in Section 3.2.

3.1 Relationship with Risk Aversion

In this section, we discuss the relationship of DRO with risk aversion in operations research. In Section 3.1.1, we discuss the connection between DRO and coherent risk measures. In Section 3.1.2, we explain how DRO connects with chance-constrained optimization through their relationship with robust optimization.

3.1.1 Relationship between DRO and Coherent and Law Invariant Risk Measures

Under mild conditions (e.g., real-valued cost functions, a convex and compact ambiguity set), the worst-case expectations given in (8) or (9) are equivalent to a *coherent* risk measure (Artzner et al. [10], Rockafellar [337], Ruszczyński and Shapiro [348]). Furthermore, under mild conditions, the worst-case expectations given in (8) or (9) are equivalent to a *law invariant* risk measure (Shapiro [367]). These results imply that a DRO model may have an equivalent risk-averse optimization problem. In order to explain the relationship between (8) and (9) and risk-averse optimization more precisely, we present some definitions and fundamental results, all with respect to measuring losses and a reference probability space $(\Xi, \mathcal{F}, Q)$.

▶ **Definition 4** (Artzner et al. [10, Definition 2.4], Shapiro et al. [374, Definition 6.4]). *A (real-valued) risk measure $\rho : \mathcal{Z} \mapsto \mathbb{R}$ is called coherent if it satisfies the following axioms:*

- Translation Equivariance: *If $a \in \mathbb{R}$ and $Z \in \mathcal{Z}$, then $\rho(Z + a) = \rho(Z) + a$.*
- Positive Homogeneity: *If $t \geq 0$ and $Z \in \mathcal{Z}$, then $\rho(tZ) = t\rho(Z)$.*
- Monotonicity: *If $Z, Z' \in \mathcal{Z}$ and $Z \geq Z'$, then $\rho(Z) \geq \rho(Z')$.*
- Convexity: $\rho(tZ + (1-t)Z') \leq t\rho(Z) + (1-t)\rho(Z')$, *for all $Z, Z' \in \mathcal{Z}$ and all $t \in [0,1]$.*

A risk measure ρ is called convex if it satisfies all the above axioms besides the positive homogeneity condition.

▶ Remark 5. In Definition 4, the convexity axiom can be replaced with the *subadditivity* axiom: $\rho(Z + Z') \leq \rho(Z) + \rho(Z')$, for all $Z, Z' \in \mathcal{Z}$. This is true because the convexity and positive homogeneity axioms imply the subadditivity axiom, and conversely, the positive homogeneity and subadditivity axioms imply the convexity axiom. Artzner et al. [10, Definition 2.4] defines a coherent risk measure with the subadditivity axiom, whereas Shapiro et al. [374, Definition 6.4] defines a coherent risk measure with the convexity axiom.

▶ **Definition 6** (Shapiro [367, Definition 2.1]). *A (real-valued) risk measure $\rho : \mathcal{Z} \mapsto \mathbb{R}$ is called law invariant (with respect to the reference probability measure Q) if for all $Z, Z' \in \mathcal{Z}$, $Z \overset{d}{\sim} Z'$ implies that $\rho(Z) = \rho(Z')$.*

▶ **Definition 7** (Shapiro [367, Definition 2.2]). *A set $\mathcal{M}$ is called law invariant (with respect to the reference probability measure Q) if $\zeta \in \mathcal{M}$ and $\zeta \overset{d}{\sim} \zeta'$ implies that $\zeta' \in \mathcal{M}$.*

To relate the worst-case expectation with respect to a set of probability distributions induced by $\boldsymbol{\xi}$ to coherent risk measures, we adopt the following result from Shapiro et al. [374, Theorem 6.7], Shapiro [364, Theorem 3.1].

▶ **Theorem 8.** *Let $\mathcal{Z}$ be the linear space of all essentially bounded $\mathcal{F}$-measurable functions $Z : \Xi \mapsto \mathbb{R}$ that are P-integrable for all $P \in \mathfrak{M}(\Xi, \mathcal{F})$. Let $\mathcal{Z}^*$ be the space of all signed measures P on $(\Xi, \mathcal{F})$ such that $\int_\Xi |\mathrm{d}P| < \infty$. Suppose that $\mathcal{Z}$ is paired with $\mathcal{Z}^*$ such that the bilinear form $\mathbb{E}_P[Z]$ is well-defined. Moreover, suppose that $\mathcal{Z}$ and $\mathcal{Z}^*$ are equipped with the sup norm $\|\cdot\|_\infty$ and variation norm $\|\cdot\|_1$, respectively[6]. Let $\mathfrak{M}(\Xi, \mathcal{F})$ denotes the space of all probability measures on $(\Xi, \mathcal{F})$: $\mathfrak{M}(\Xi, \mathcal{F}) = \left\{P \in \mathcal{Z}^* \,\middle|\, \int_\Xi \mathrm{d}P = 1,\ P \succcurlyeq 0\right\}$. Let $\rho : \mathcal{Z} \mapsto \overline{\mathbb{R}}$. Then, ρ is a real-valued coherent risk measure if and only if there exists a convex compact set $\mathcal{M} \subseteq \mathfrak{M}(\Xi, \mathcal{F})$ (in the weakly* topology of $\mathcal{Z}^*$) such that*

$$\rho(Z) = \sup_{P \in \mathcal{M}} \mathbb{E}_P[Z],\ \forall\ Z \in \mathcal{Z}. \tag{13}$$

Moreover, given a real-valued coherent risk measure, the set $\mathcal{M}$ in (13) *can be written in the form*

$$\mathcal{M} = \{P \in \mathfrak{M}(\Xi, \mathcal{F}) \,|\, \mathbb{E}_P[Z] \le \rho(Z),\ \forall\ Z \in \mathcal{Z}\}.$$

Proof. See Appendix A. ◀

Before we proceed, let us characterize the set $\mathcal{M}$, as described in Theorem 8, for three well-studied coherent risk measures, namely *conditional Value-at-Risk* (CVaR), see, e.g., Rockafellar [337], Rockafellar and Uryasev [339, 340], convex combination of expectation and CVaR, see, e.g., Zhang et al. [438], and *mean-upper-absolute semideviation*, see, e.g., Shapiro et al. [374]. CVaR at level β, $0 < \beta < 1$, denoted by $\mathrm{CVaR}^Q_\beta[\,\cdot\,]$, is defined as $\mathrm{CVaR}^Q_\beta[Z] := \frac{1}{1-\beta}\int_\beta^1 \mathrm{VaR}^Q_\alpha[Z]\,\mathrm{d}\alpha$, where $\mathrm{VaR}^Q_\alpha[Z] := \inf\{u \,|\, Q\{Z \le u\} \ge \alpha\}$ is the (left-side) α-quantile or Value-at-Risk (VaR) at level α. The mean-upper-absolute semideviation is defined as $\mathbb{E}_Q[Z] + c\mathbb{E}_Q[(Z - \mathbb{E}_P[Z])_+]$, where $c \in [0, 1]$.

▶ **Example 9.** Consider a probability space $(\Xi, \mathcal{F}, Q)$ and $\mathcal{Z} = \mathcal{L}_\infty(\Xi, \mathcal{F}, Q)$. Suppose that Ξ is a finite space with M atoms. For a coherent risk measure ρ, we have $\rho(Z) = \sup_{\boldsymbol{p} \in \mathcal{M}} \sum_{m=1}^M z_m p_m,\ \forall\ Z \in \mathcal{Z}$, where $\mathcal{M}$ is closed convex subset of

$$\mathcal{D} := \left\{\boldsymbol{p} \in \mathbb{R}^M \,\middle|\, \boldsymbol{p}^\top \boldsymbol{e} = 1,\ \boldsymbol{p} \ge \boldsymbol{0}\right\},$$

and $\boldsymbol{e}$ is a vector of ones.

- When $\rho(Z) = \mathrm{CVaR}^Q_\beta[Z]$[7], we have

$$\mathcal{M} = \left\{\boldsymbol{p} \in \mathcal{D} \,\middle|\, p_m \in \left[0, \frac{q_m}{1-\beta}\right],\ m \in [M]\right\}.$$

- When $\rho(Z) = \mathbb{E}_Q[Z] + \inf_{\tau \in \mathbb{R}} \mathbb{E}_Q[(1-\gamma_1)(\tau - Z)_+ + (\gamma_2 - 1)(Z - \tau)_+]$, with $\gamma_1 \in [0, 1)$ and $\gamma_2 > 1$, we have

$$\mathcal{M} = \{\boldsymbol{p} \in \mathcal{D} \,|\, p_m \in [q_m\gamma_1, q_m\gamma_2],\ m \in [M]\}.$$

  The above risk measure is also equivalent to $\gamma_1\mathbb{E}_Q[Z] + (1-\gamma_1)\mathrm{CVaR}^Q_\beta[Z]$, where $\beta := \frac{1-\gamma_1}{\gamma_2 - \gamma_1}$.
- When $\rho(Z) = \mathbb{E}_Q[Z] + c\mathbb{E}_Q[(Z - \mathbb{E}_P[Z])_+]$, we have

$$\mathcal{M} = \left\{\boldsymbol{p}' \in \mathcal{D} \,\middle|\, \boldsymbol{p}' = \boldsymbol{q} + \boldsymbol{\zeta} \odot \boldsymbol{q} - (\boldsymbol{\zeta}^\top \boldsymbol{q}) \odot \boldsymbol{q},\ \|\boldsymbol{\zeta}\|_\infty \le c\right\},$$

  where $\boldsymbol{a} \odot \boldsymbol{b}$ denotes the componentwise product of two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. ◀

Theorem 8 relates problems (8) and (9) to risk-averse optimization problems, involving the coherent risk-measure ρ. Consider a fixed $\boldsymbol{x} \in \mathcal{X}$. With an appropriate transformation of measure $\mathbb{P} = P \circ \boldsymbol{\xi}^{-1}$, we can write the inner problem $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$ in (8) as $\sup_{P \in \mathcal{P}} \mathbb{E}_P[h_0(\boldsymbol{x}, s)]$, where in the former, $\mathcal{P}$ is a set of probability distributions induced by $\boldsymbol{\xi}$, while in the latter, $\mathcal{P}$ is a set of probability measures on $(\Xi, \mathcal{F})$. Then, by applying Theorem 8 and setting $Z = h_0(\boldsymbol{x}, \boldsymbol{\xi})$, $\sup_{P \in \mathcal{P}} \mathbb{E}_P[h_0(\boldsymbol{x}, s)]$ evaluates a (real-valued) coherent risk measure $\rho[h_0(\boldsymbol{x}, s)]$, provided that $\mathcal{P} \subset \mathfrak{M}(\Xi, \mathcal{F})$ is a convex compact set. It is easy to verify that such a function ρ is coherent:

---

[6] Recall that for a function $Z \in \mathcal{Z}$, $\|Z\|_\infty = \operatorname{ess\,sup}_{s \in \Omega} |Z(s)|$, where $\operatorname{ess\,sup}_{s \in \Xi} |Z(s)| = \inf\{\sup_{s \in \Xi} |Z'(s)| \mid Z(s) = Z'(s) \text{ a.e. } s \in \Xi\}$. Also, for a measure $P \in \mathcal{Z}^*$, $\|P\|_1 = \int_\Xi |\mathrm{d}P|$.

[7] It is known that $\rho(Z) = \mathrm{CVaR}^Q_\beta[Z]$ can be written equivalently as $\mathrm{CVaR}^Q_\beta[Z] = \min\left\{\eta + \frac{1}{1-\beta}\mathbb{E}_Q[(Z-\eta)_+] \,\middle|\, \eta \in \mathbb{R}\right\}$ Rockafellar and Uryasev [339, 340]. In particular, when Ξ is a finite set with M atoms, $\mathrm{CVaR}^Q_\beta[Z]$ can be equivalently written as a linear program: $\mathrm{CVaR}^Q_\beta[Z] = \min\left\{\eta + \frac{1}{1-\beta}\sum_{m=1}^M q_m v_m \,\middle|\, \eta \in \mathbb{R},\ v_m \ge z_m - \eta,\ v_m \ge 0,\ m = 1, \ldots, M\right\}$.

- **Translation Equivariance.** Consider $\boldsymbol{x} \in \mathcal{X}$ and $a \in \mathbb{R}$. Then, $\rho\left[h_0(\boldsymbol{x}, s) + a\right] = \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[h_0(\boldsymbol{x}, s) + a\right] = \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[h_0(\boldsymbol{x}, s)\right] + a = \rho\left[h_0(\boldsymbol{x}, s)\right] + a$.
- **Positive Homogeneity.** Consider $\boldsymbol{x} \in \mathcal{X}$ and $t \geq 0$. Then, $\rho\left[th_0(\boldsymbol{x}, s)\right] = \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[th_0(\boldsymbol{x}, s)\right] = t\sup_{P\in\mathcal{P}} \mathbb{E}_P\left[h_0(\boldsymbol{x}, s)\right] = t\rho\left[h_0(\boldsymbol{x}, s)\right]$.
- **Monotonicity.** Consider $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ such that $h_0(\boldsymbol{x}, s) \geq h_0(\boldsymbol{x}', s)$. Thus, $\mathbb{E}_P\left[h_0(\boldsymbol{x}, s)\right] \geq \mathbb{E}_P\left[h_0(\boldsymbol{x}', s)\right]$ for any $P \in \mathcal{P}$, which implies $\rho\left[h_0(\boldsymbol{x}, s)\right] = \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[h_0(\boldsymbol{x}, s)\right] \geq \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[h_0(\boldsymbol{x}', s)\right] = \rho\left[h_0(\boldsymbol{x}', s)\right]$.
- **Convexity.** Consider $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and $t \in [0, 1]$. Then, we have

$$\begin{aligned}
\rho\left[th_0(\boldsymbol{x}, s) + (1-t)h_0(\boldsymbol{x}', s)\right] &= \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[th_0(\boldsymbol{x}, s) + (1-t)h_0(\boldsymbol{x}', s)\right] \\
&\leq \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[th_0(\boldsymbol{x}, s)\right] + \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[(1-t)h_0(\boldsymbol{x}', s)\right] \\
&= t\sup_{P\in\mathcal{P}} \mathbb{E}_P\left[h_0(\boldsymbol{x}, s)\right] + (1-t)\sup_{P\in\mathcal{P}} \mathbb{E}_P\left[h_0(\boldsymbol{x}', s)\right] \\
&= t\rho\left[h_0(\boldsymbol{x}, s)\right] + (1-t)\rho\left[h_0(\boldsymbol{x}', s)\right].
\end{aligned}$$

Consequently, (8) is equivalent to minimizing a coherent risk measure. Similarly, (9) is equivalent to a risk-averse optimization problem, subject to coherent risk constraints. Thus, a convex and compact ambiguity set of distributions gives rise to a coherent risk measure. Conversely, Theorem 8 implies that given a risk preference that can be expressed in the form of a coherent risk measure as a primitive, we can construct a corresponding convex and compact ambiguity set $\mathcal{P}$ of probability distributions in a DRO framework. Thus, the ambiguity set becomes a consequence of the particular risk measure the decision maker selects.

It is worth noting that if h_0 is a convex random function in (8), i.e., $h_0(\,\cdot\,, \boldsymbol{\xi})$ is convex in $\boldsymbol{x}$ for almost every $\boldsymbol{\xi}$, then, $\rho\left[h_0(\,\cdot\,, \boldsymbol{\xi})\right]$ is convex in $\boldsymbol{x}$. Convexity of h_j, $j = 1, \ldots, m$, in (9) also implies the convexity of the region induced by the risk constraints $\rho\left[h_j(\,\cdot\,, \boldsymbol{\xi})\right] \leq \mathbf{0}$, $j = 1, \ldots, m$. In our setup, neither $h(\,\cdot\,, \boldsymbol{\xi})$ nor $h_j(\,\cdot\,, \boldsymbol{\xi})$, $j = 1, \ldots, m$, need to be convex as for example in the case where they are indicator functions.

We now state the connection between the worst-case expectation with respect to a set of probability distributions induced by $\boldsymbol{\xi}$ to law invariant risk measures.

▶ **Theorem 10** (Shapiro [367, Theorem 2.3])**.** *Consider $\mathcal{Z}$ and $\mathcal{Z}^*$ as defined in Theorem 8. Also, consider $\rho : \mathcal{Z} \mapsto \mathbb{R}$, defined as $\rho(Z) = \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[Z\right]$, $\forall\, Z \in \mathcal{Z}$. If the set $\mathcal{P}$ is law invariant, then the corresponding risk measure ρ is law invariant. Conversely, if the risk measure ρ is law invariant, and the set $\mathcal{P}$ is convex and weakly* closed, then the set $\mathcal{P}$ is law invariant.*

For the connection between a general multistage DRO model, risk-averse multistage programming with conditional coherent risk mappings, and the concept of time consistency of the problem and policies, we refer to Shapiro [364, 366, 368].

▶ Remark 11. Recall that DRO is related to RO in a finite sample space. Thus, it is no surprise to conclude that the analogous of the relationship between DRO and coherent risk measures holds for RO and coherent risk measures when the sample space is finite. More precisely, one can apply Theorem 8 to the setting that the sample space is finite to conclude the result. On one hand, given a coherent risk measure as a primitive, we can construct a corresponding convex uncertainty set $\mathcal{U}$ in a RO framework (Bertsimas and Brown [38]). A converse implication also holds; a convex uncertainty set induces a coherent risk measure (Natarajan et al. [275]).

3.1.2 Relationship with Chance-Constrained Optimization

In the previous section, we discussed how DRO is connected to risk-averse optimization. In this section, we present another perspective that connects DRO to risk-averse optimization through a proper choice of the uncertainty set of the random variables $\boldsymbol{\xi}$, as in RO.

Many approaches in RO construct the uncertainty set for the parameters $\boldsymbol{\xi}$ such that the uncertainty set implies a probabilistic guarantee with respect to the true unknown distribution. To explain how this construction is related to risk and DRO, consider the uncertain constraints $g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0$ for a fixed $\boldsymbol{x}$. Suppose that $\boldsymbol{\xi}$ belongs to a bounded uncertainty set $\mathcal{U} \subseteq \mathbb{R}^d$, i.e., $\mathcal{U}$ is the support of $\boldsymbol{\xi}$. The RO counterpart of this constraint then can be formulated as

$$g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0,\ \forall\, \boldsymbol{\xi} \in \mathcal{U}. \tag{14}$$

Two criticisms of (14) are that: (1) it treats the constraints for all parameters $\boldsymbol{\xi} \in \mathcal{U}$ with equal importance and (2) all the parameterized constraints are hard, i.e., no violation is accepted. An alternative framework to

reduce the conservatism caused by this approach is to use a chance constraint framework that allows a small probability of violation (with respect to the probability distribution of $\boldsymbol{\xi}$) instead of enforcing the constraint to be satisfied almost everywhere. Under the assumption that $\boldsymbol{\xi}$ is defined on a probability space $(\Xi, \mathcal{F}, P^{\text{true}})$, the chance constraint framework can be represented as follows:

$$P^{\text{true}}\{g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0\} \geq 1 - \epsilon, \tag{15}$$

for some $0 < \epsilon < 1$. The parameter ϵ controls the risk of violating the uncertain constraint $g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0$. In fact, as ϵ goes to zero, the set

$$\mathcal{X}_\epsilon := \left\{\boldsymbol{x} \in \mathcal{X} \,\middle|\, P^{\text{true}}\{g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0\} \geq 1 - \epsilon\right\}$$

decreases to

$$\mathcal{X}(\mathcal{U}) := \{\boldsymbol{x} \in \mathcal{X} \,|\, g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0,\ \forall\, \boldsymbol{\xi} \in \mathcal{U}\}\,.$$

Motivated by the chance constraint framework (15), many approaches in RO construct an uncertainty set $\mathcal{U}_\epsilon$ such that a feasible solution to a problem of the form (14) will also be feasible with probability at least $1-\epsilon$ with respect to P^{true}. More precisely, for any fixed $\boldsymbol{x}$, these constructions guarantee that the following implication holds:

$$\text{If } g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0,\ \forall\, \boldsymbol{\xi} \in \mathcal{U}_\epsilon,\ \text{ then, } P^{\text{true}}\{g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0\} \geq 1 - \epsilon. \tag{C1}$$

However, as we argued before, the probability measure P^{true} cannot be known with certainty. As far as it is relevant to the scope and interest of this paper, there are two streams of research in order to handle the ambiguity about the true probability distribution and obtain a safe (or, conservative) approximation[8] to (15)[9]: (1) scenario approximation scheme of (14) based on Monte Carlo sampling, see, e.g., Ben-Tal and Nemirovski [24], Calafiore and Campi [73], Campi and Calafiore [75], Campi and Garatti [76], Luedtke and Ahmed [258], Nemirovski and Shapiro [280], and (2) DRO approach to (15), see, e.g., Erdoğan and Iyengar [138], Nemirovski and Shapiro [279]. Research on scenario approximation of (14) focuses on providing probabilistic guarantee (with respect to the sample probability measure) that a solution to the sampled problem of (14) is feasible to (15) with a high probability.

The DRO approach, on the other hand, forms a version of (15) as follows:

$$P\{g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0\} \geq 1 - \epsilon,\ \forall\, P \in \mathcal{P} \equiv \inf_{P \in \mathcal{P}} P\{g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0\} \geq 1 - \epsilon. \tag{16}$$

Let $\bar{\mathcal{X}}_\epsilon$ denote the feasibility set induced by (16):

$$\bar{\mathcal{X}}_\epsilon := \left\{\boldsymbol{x} \in \mathcal{X} \,\middle|\, \inf_{P \in \mathcal{P}} P\{g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0\} \geq 1 - \epsilon\right\}.$$

If $P^{\text{true}} \in \mathcal{P}$, then, $\boldsymbol{x} \in \bar{\mathcal{X}}_\epsilon$ implies $\boldsymbol{x} \in \mathcal{X}_\epsilon$. That is, $\bar{\mathcal{X}}_\epsilon$ provides a conservative approximation to $\mathcal{X}_\epsilon$[10]. By leveraging a goodness-of-fit test, Bertsimas et al. [51] construct a $(1-\alpha)$-confidence region $\mathcal{P}(\alpha)$ for P^{true}. Such a construction leads to an uncertainty set $\mathcal{U}_\epsilon(\alpha)$ that guarantees the implication (C1) with probability at least $(1-\alpha)$ with respect to the sample probability measure.

▶ **Theorem 12** (Bertsimas et al. [51, Theorem 2])**.** *Suppose that for any fixed $\boldsymbol{x}$, $g(\boldsymbol{x}, \boldsymbol{\xi})$ is concave in $\boldsymbol{\xi}$. Consider a set of data $\{\boldsymbol{\xi}^i\}_{i=1}^N$, drawn independently and identically distributed (i.i.d.) according to P^{true}. Let $\mathcal{P}_\epsilon(\alpha)$ be a $(1-\alpha)$-confidence region for P^{true}, constructed from a goodness-of-fit test on data. Moreover, for any $\boldsymbol{y} \in \mathbb{R}^d$, let $l_\epsilon(\boldsymbol{y}; \alpha)$ be a closed, convex, finite-valued, and positively homogeneous (in $\boldsymbol{y}$) upper bound to the worst-case VaR of $\boldsymbol{y}^\top \boldsymbol{\xi}$ at level $1-\epsilon$ over $\mathcal{P}_\epsilon(\alpha)$, i.e., $\sup_{P \in \mathcal{P}_\epsilon(\alpha)} \text{VaR}^P_{1-\epsilon}\left[\boldsymbol{y}^\top \boldsymbol{\xi}\right] \leq l_\epsilon(\boldsymbol{y}; \alpha)$, $\boldsymbol{y} \in \mathbb{R}^d$. Then, the closed, convex set $\mathcal{U}_\epsilon(\alpha)$ for which $\delta^*\big(\boldsymbol{y} | \mathcal{U}_\epsilon(\alpha)\big) = l_\epsilon(\boldsymbol{y}; \alpha)$ guarantees the implication* (C1) *with probability at least $(1-\alpha)$ (with respect to the sample probability measure).*

As a byproduct of Theorem 12, $\delta^*\big(\boldsymbol{y} | \mathcal{U}_\epsilon(\alpha)\big) \leq \boldsymbol{b}$ provides a safe approximation to $\sup_{P \in \mathcal{P}_\epsilon(\alpha)} P\{\boldsymbol{y}^\top \boldsymbol{\xi} \leq \boldsymbol{b}\} \geq 1 - \epsilon$. That is, there is a correspondence between the uncertainty set $\mathcal{U}_\epsilon(\alpha)$ that satisfies the probabilistic guarantee (C1) and safe approximations to $\sup_{P \in \mathcal{P}_\epsilon(\alpha)} P\{\boldsymbol{y}^\top \boldsymbol{\xi} \leq \boldsymbol{b}\} \geq 1 - \epsilon$.

---

[8] A set of constraints is called a safe or conservative approximation of the chance constraint if the feasible region induced by the approximation is a subset of the feasible region induced by the chance constraint.

[9] There is another stream of research that approximates (15) by CVaR or its approximations, see, e.g., Chen and Sim [87], Chen et al. [88, 91] and references there in.

[10] One can in turns seek a safe approximation to (16). For example, one stream of such approximations includes using Chebyshev's inequality, see, e.g., Bertsimas and Popescu [42], Popescu [309], Bernstein's inequality, see, e.g., Nemirovski and Shapiro [279], or Hoeffding's inequality. We review such safe approximations to (16) in Section 6.

3.2 Relationship with Function Regularization

The goal of this section is to discuss the relationship of DRO/RO with the function regularization commonly used in machine learning.

Some papers have shown that DRO problems via the *optimal transport discrepancy* and *ϕ-divergences* are connected to regularization. When the optimal transport discrepancy is used, as shown in Blanchet et al. [64], Gao and Kleywegt [148], Shafieezadeh-Abadeh et al. [355], many mainstream machine learning classification and regression models, including support vector machine (SVM), regularized logistic regression, and Least Absolute Shrinkage and Selection Operator (LASSO), have a direct distributionally robust interpretation that connects regularization to the protection from the disturbance in data. To state this result, we first present a duality theorem, due to Blanchet and Murthy [62], and we relegate the technical details and assumptions to Section 6. On the other hand, when ϕ-divergences are used, the DRO problem is connected to variance regularization, see, e.g., Duchi et al. [123], Namkoong and Duchi [273].

Let us begin by defining the optimal transport discrepancy. Consider two probability measures $P_1, P_2 \in \mathfrak{M}(\Xi, \mathcal{F})$. Let $\Pi(P_1, P_2)$ denote the set of all probability measures on $(\Xi \times \Xi, \mathcal{F} \times \mathcal{F})$ whose marginals are P_1 and P_2:

$$\Pi(P_1, P_2) = \left\{ \pi \in \mathfrak{M}(\Xi \times \Xi, \mathcal{F} \times \mathcal{F}) \,\middle|\, \begin{array}{l} \pi(A \times \Xi) = P_1(A),\ \forall\, A \in \mathcal{F}, \\ \pi(\Xi \times A) = P_2(A),\ \forall\, A \in \mathcal{F} \end{array} \right\}.$$

An element of the above set is called a *coupling* or *transport plan*. Furthermore, suppose that there is a lower semicontinuous function $c : \Xi \times \Xi \mapsto \mathbb{R}_+ \cup \{\infty\}$ with $c(s_1, s_2) = 0$ if $s_1 = s_2$. Then, the optimal transport discrepancy between P_1 and P_2 is defined as[11]:

$$\mathfrak{d}_c^{\mathrm{W}}(P_1, P_2) := \inf_{\pi \in \Pi(P_1, P_2)} \int_{\Xi \times \Xi} c(s_1, s_2) \pi(\mathrm{d}s_1 \times \mathrm{d}s_2).$$

▶ **Theorem 13** (Blanchet and Murthy [62, Remark 1]). *Consider an ambiguity set of probability measures as*

$$\mathcal{P}^W(P_0; \epsilon) := \left\{ P \in \mathfrak{M}(\Xi, \mathcal{F}) \,\middle|\, \mathfrak{d}_c^W(P, P_0) \le \epsilon \right\},$$

formed via the optimal transport discrepancy $\mathfrak{d}_c^W(P, P_0)$, where c is the transportation cost function, ϵ is the size of the ambiguity set (i.e., level of robustness), and P_0 is a nominal probability measure. Then, for a fixed $\boldsymbol{x} \in \mathcal{X}$, we have

$$\sup_{P \in \mathcal{P}^W(P_0; \epsilon)} \mathbb{E}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] = \inf_{\lambda \ge 0} \left\{ \lambda \epsilon + \mathbb{E}_{P_0}\left[\sup_{s' \in \Xi} \{ h_0(\boldsymbol{x}, s') - \lambda c(s, s') \} \right] \right\}.$$

We can use Theorem 13 to explicitly state the connection between DRO and regularization. We adopt the following two theorems from Blanchet and Murthy [62], due to their generality. However, similar results are obtained in other papers, see, e.g., Gao and Kleywegt [148], Shafieezadeh-Abadeh et al. [355].

▶ **Theorem 14** (Blanchet et al. [64, Theorems 2–3]). *Consider a dataset $\{\boldsymbol{\xi}^i := (\boldsymbol{u}^i, y^i)\}_{i=1}^N$, where $\boldsymbol{u}^i \in \mathbb{R}^n$ is a vector of covariates and $y^i \in \mathbb{R}$ is the response variable. Suppose that $\widehat{\mathbb{P}}_N$ is the empirical probability distribution on $\{\boldsymbol{\xi}^i\}_{i=1}^N$, $c(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2) := \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_q^2$ if $y^1 = y^2$, and $c(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2) = \infty$, otherwise. Let $\frac{1}{p} + \frac{1}{q} = 1$. Then,*

- *For a linear regression model with a square loss function $h_0(\boldsymbol{x}, \boldsymbol{\xi}) := (y - \boldsymbol{x}^\top \boldsymbol{u})^2$, we have*

$$\inf_{\boldsymbol{x} \in \mathbb{R}^n} \sup_{\mathbb{P} \in \mathcal{P}^W(\hat{\mathbb{P}}_N; \epsilon)} \mathbb{E}_{\mathbb{P}}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] = \inf_{\boldsymbol{x} \in \mathbb{R}^n} \left\{ \epsilon^{\frac{1}{2}} \|\boldsymbol{x}\|_p + \left(\mathbb{E}_{\hat{\mathbb{P}}_N}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] \right)^{\frac{1}{2}} \right\}^2,$$

- *For a logistic regression model with cost function $h_0(\boldsymbol{x}, \boldsymbol{\xi}) := \log(1 + e^{-y \boldsymbol{x}^\top \boldsymbol{u}})$, we have*

$$\inf_{\boldsymbol{x} \in \mathbb{R}^n} \sup_{\mathbb{P} \in \mathcal{P}^W(\hat{\mathbb{P}}_N; \epsilon)} \mathbb{E}_{\mathbb{P}}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] = \inf_{\boldsymbol{x} \in \mathbb{R}^n} \left\{ \epsilon \|\boldsymbol{x}\|_p + \mathbb{E}_{\hat{\mathbb{P}}_N}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] \right\},$$

- *For a SVM with Hinge loss $h_0(\boldsymbol{x}, \boldsymbol{\xi}) := (1 - y \boldsymbol{x}^\top \boldsymbol{u})_+$, we have*

$$\inf_{\boldsymbol{x} \in \mathbb{R}^n} \sup_{\mathbb{P} \in \mathcal{P}^W(\hat{\mathbb{P}}_N; \epsilon)} \mathbb{E}_{\mathbb{P}}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] = \inf_{\boldsymbol{x} \in \mathbb{R}^n} \left\{ \epsilon \|\boldsymbol{x}\|_p + \mathbb{E}_{\hat{\mathbb{P}}_N}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] \right\}.$$

[11] One can similarly define the optimal transport discrepancy between two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ induced by $\boldsymbol{\xi}$.

As stated in Theorem 14, we can rewrite an *unconstrained* DRO model with the optimal transport discrepancy as a minimization problem, in which the objective function, in one hand, includes an expected-cost term with respect to the empirical distribution, and on the other hand, includes a regularization term. Two other interesting results can be inferred from Theorem 14 about the connection between DRO and regularization: (i) the shape of the transportation cost function c in the definition of the optimal transport discrepancy directly implies the type of regularization, and (ii) the size of the ambiguity set is related to the regularization parameter. An important implication of these results is that one can judiciously choose an appropriate regularization parameter for the problem in hand by using the equivalent DRO reformulation. We review the papers that draw this conclusion in Section 6.1.

Now, let us focus on DRO problems formulated via ϕ-divergences. For two probability measures $P_1, P_2 \in \mathfrak{M}(\Xi, \mathcal{F})$, the ϕ-divergence between P_1 and P_2 is defined as $\mathfrak{d}^\phi(P_1, P_2) := \int_\Xi \phi\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}\right) \mathrm{d}P_2$, where the ϕ-divergence function $\phi : \mathbb{R}_+ \to \mathbb{R}_+ \cup \{+\infty\}$ is convex, and it satisfies the following properties: $\phi(1) = 0$, $0\phi\left(\frac{0}{0}\right) := 0$, and $a\phi\left(\frac{a}{0}\right) := a \lim_{t\to\infty} \frac{\phi(t)}{t}$ if $a > 0$[12].

▶ **Theorem 15** (Duchi et al. [123, Theorem 2])**.** *Consider an ambiguity set of probability distributions as*

$$\mathcal{P}^\phi(\widehat{\mathbb{P}}_N; \epsilon) := \left\{ \mathbb{P} \in \mathfrak{P}(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d)) \,\middle|\, \mathfrak{d}^\phi(\mathbb{P}, \widehat{\mathbb{P}}_N) \le \epsilon \right\},$$

formed via the ϕ-divergence $\mathfrak{d}^\phi(\mathbb{P}, \widehat{\mathbb{P}}_N)$, where ϵ is the size of the ambiguity set and $\widehat{\mathbb{P}}_N$ is the empirical probability distribution on a set of independently and identically distributed (i.i.d) data $\{\boldsymbol{\xi}^i\}_{i=1}^N$, according to $\mathbb{P}^{true}$. Furthermore, suppose that $\mathcal{X}$ is compact, there exists a measurable function $M : \Omega \mapsto \mathbb{R}_+$ such that for all $\boldsymbol{\xi} \in \Omega$, $h(\cdot, \boldsymbol{\xi})$ is $M(\boldsymbol{\xi})$-Lipschitz with respect to some norm $\|\cdot\|$ on $\mathcal{X}$, $\mathbb{E}_{\mathbb{P}^{true}}\left[M(\boldsymbol{\xi})^2\right] < \infty$, and $\mathbb{E}_{\mathbb{P}^{true}}\left[|h_0(\boldsymbol{x}_0, \boldsymbol{\xi})|\right] < \infty$ for some $\boldsymbol{x}_0 \in \mathcal{X}$. Then,

$$\sup_{\mathbb{P} \in \mathcal{P}^\phi(\hat{\mathbb{P}}_N; \frac{\epsilon}{N})} \mathbb{E}_{\mathbb{P}}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] = \mathbb{E}_{\hat{\mathbb{P}}_N}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] + \left(\frac{\epsilon}{N} \mathrm{Var}_{\hat{\mathbb{P}}_N}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]\right)^{\frac{1}{2}} + \gamma_N(\boldsymbol{x}),$$

where $\gamma_N(\boldsymbol{x})$ is such that $\sup_{\boldsymbol{x}\in\mathcal{X}} \sqrt{N}|\gamma_N(\boldsymbol{x})| \to 0$ in probability.

According to Theorem 15, we can rewrite the inner problem of a model of the form (8) with ϕ-divergences as the expected cost plus a regularization term that accounts for the standard deviation of the cost, under the empirical distribution. This type of trade-off between the bias (approximation error) and standard deviation (estimation error) in a DRO model via ϕ-divergences is obtained in Gotoh et al. [164]. Additionally, similar results to Theorem 15 are obtained in Dupuis et al. [130], Lam [228, 229] in the context of robust sensitivity analysis of stochastic systems, where the square root of the divergence, ϵ, is the correct scaling of the standard deviation to capture the misspecification effect of the true probability distribution. For an ambiguity set constructed via Kullback–Leibler divergence and around any nominal distribution $\mathbb{P}_0$, Lam [228] derives an asymptotic expansion on the worst-case objective value as the divergence shrinks to zero. Such an expansion contains terms involving up to the third-order cumulant of $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ under $\mathbb{P}_0$. A similar expansion is derived in Lam [229] for an ambiguity set constructed via χ^2-distance.

4 General Solution Techniques to Solve DRO Models

In this section, we discuss four solution approaches to handle (DRO). In Section 4.1 and 4.2, we discuss the cutting-surface and dualization methods, respectively. Then, in Section 4.3, we explain numerical methods to solve (DRO). Finally, in Section 4.4, we describe approximation based on decision rules as a widely used method.

In order to explain the first two approaches, let us first reformulate (DRO) as follows:

$$\inf_{\boldsymbol{x}\in\mathcal{X},\theta} \theta \quad \text{s.t.} \begin{cases} \theta \ge \mathcal{R}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right], \ \forall\, P \in \mathcal{P} \\ \mathcal{R}_P\left[h_j(\boldsymbol{x}, \boldsymbol{\xi})\right] \le 0, \ \forall\, P \in \mathcal{P}, \ j \in [m]. \end{cases} \tag{17}$$

Reformulation (17) is a semi-infinite program (SIP), and at a first glance, obtaining an optimal solution to this problem looks difficult[13]. It is well-known that even convex SIPs cannot be solved directly with numerical

[12] One can similarly define the ϕ-divergence between two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ induced by $\boldsymbol{\xi}$.

[13] The study of SIPs is pioneered by Haar [170], and followed up in Charnes et al. [82, 83, 84], which focus on linear SIP s. The first- and second-order optimality conditions of a general SIP are also obtained in Hettich and Jongen [186, 187], Hettich and Still [189], Nürnberger [289, 290], Still [386]. For reviews of the theory and methods for SIP s, we refer the readers to Hettich and Kortanek [188], López and Still [253], Reemtsen and Görner [333].

methods, and in particular are not amenable to the use of methods such as interior point method. Therefore, a key step of the solution techniques to handle the semi-infinite qualifier (i.e., $\forall\, P \in \mathcal{P}$) is to reformulate (17) as an optimization problem that is amenable to the use of available optimization techniques and off-the-shelf solvers. Of course, the complexity and tractability of such SIPs and their reformulations depend on the geometry and properties of both the ambiguity set $\mathcal{P}$ and the functions $h_j(\boldsymbol{x}, \boldsymbol{\xi})$, $j \in \{0\} \cup [m]$. As we shall see in details in Section 6, proper assumptions on $\mathcal{P}$ and these functions are important in most studies on DRO in order to obtain a solvable reformulation or approximation of (17).

In the context of DRO, there are two main approaches to handle the semi-infinite quantifier $\forall\, P$ and to numerically solve (17): cutting-surface method and dual method. Both approaches have their roots in the SIP literature, and they both aim at getting rid of the quantifier $\forall\, P$, but in different ways. It is worth noting that numerical methods to solve a SIP, such as penalty methods, see, e.g., Lin et al. [246], Yang et al. [430], smooth approximation and projection methods, see, e.g., Xu et al. [426], and primal methods, see, e.g., Wang and Yuan [405], have not been popular to solve (17).

4.1 Cutting-Surface Method

The first approach replaces the quantifier $\forall\, P$ by *for some finite atomic subset of* $\mathcal{P}$. The idea is to successively solve a relaxed problem of (17) over a finitely generated inner approximations of the ambiguity set $\mathcal{P}$. To be precise, this approach approximates the semi-infinite constraints for all $P \in \mathcal{P}$ by finitely many ones over a finite set of probability distributions. In each iteration of this approach, a new probability distribution is added to this finite set until optimality criteria are met. We refer to this as a *cutting-surface* method (also known as *exchange method*, following the terminology in the SIP literature, see, e.g., Hettich and Kortanek [188], Mehrotra and Papp [262]. We refer to Bansal et al. [15], Pflug and Wozabal [302], Rahimian et al. [327] as examples of this approach in the context of DRO.

The key requirements in order to use the cutting-surface method are the abilities to (i) solve a relaxation of (17) with a finite number of probability distributions to optimality and (ii) generate an ϵ-optimal solution[14] to a distribution separation subproblem Luo and Mehrotra [259].

▶ **Theorem 16** (Luo and Mehrotra [259, Theorem 3.2])**.** *Suppose that $\mathcal{X} \times \mathcal{P}$ is a compact set, and $\mathcal{R}_P[h_j(\boldsymbol{x}, \boldsymbol{\xi})]$, $j \in \{0\} \cup [m]$, are continuous on $\mathcal{X} \times \mathcal{P}$. Moreover, suppose that we have an oracle that generates an optimal solution $(\boldsymbol{x}_k, \theta_k)$ to a relaxation of problem* (17) *for any finite set $\mathcal{P}_k \subseteq \mathcal{P}$, and an oracle that generates an ϵ-optimal solution of the distribution generation subproblem*

$$\sup_{P \in \mathcal{P}} \max \left\{ \mathcal{R}_P[h_0(\boldsymbol{x}, \boldsymbol{\xi})] - \theta_k, \max_{j \in [m]} \mathcal{R}_P[h_j(\boldsymbol{x}, \boldsymbol{\xi})] \right\}$$

for any $\boldsymbol{x} \in \mathcal{X}$ and $\epsilon > 0$. Suppose that iteratively the relaxed master problem is solved to optimally and yields the solution $(\boldsymbol{x}_k, \theta_k)$, and the distribution separation subproblem is solved to $\frac{\epsilon}{2}$-optimality and yields the solution P_k. Then, the stopping criteria $\mathcal{R}_{P_k}[h_0(\boldsymbol{x}_k, \boldsymbol{\xi})] \leq \theta_k + \frac{\epsilon}{2}$ and $\mathcal{R}_{P_k}[h_j(\boldsymbol{x}_k, \boldsymbol{\xi})] \leq \frac{\epsilon}{2}$, $j \in [m]$, guarantee that an ϵ-feasible solution[15] to problem (17)*, yielding an objective function value lower bounding the optimal value of* (17)*, can be obtained in a finite number of iterations.*

It is worth noting that the distribution separation subproblem in the cutting-surface method may be a nonconvex optimization problem. One may efficiently solve (DRO) through the cutting-surface method if the ambiguity set $\mathcal{P}$ can be convexified without causing a change to the optimal value. In this case, the distribution separation subproblem may be solved through interior point methods. The following lemma states that if $\mathcal{R}_P[\,\cdot\,]$ is convex in P on $\mathfrak{M}(\Xi, \mathcal{F})$, then, it can be assumed without loss of generality that $\mathcal{P}$ is convex.

▶ **Lemma 17.** *Consider* (DRO)*. For a fixed $x \in \mathcal{X}$, suppose that $\mathcal{R}_P[\,\cdot\,]$ is convex in P on $\mathfrak{M}(\Xi, \mathcal{F})$. Then, $\boldsymbol{x}^* \in \mathcal{X}$ is an optimal solution to* (DRO) *if and only if it is an optimal solution to the following problem:*

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \sup_{P \in \mathrm{conv}(\mathcal{P})} \left\{ \mathcal{R}_P[h_0(\boldsymbol{x}, \boldsymbol{\xi})] \,\middle|\, \sup_{P \in \mathrm{conv}(\mathcal{P})} \mathcal{R}_P[h_j(\boldsymbol{x}, \boldsymbol{\xi})] \leq 0,\ j \in [m] \right\}. \tag{18}$$

Proof. See Appendix A. ◀

[14] For an optimization problem of the form $z^* = \min\{\alpha(\boldsymbol{x}) \,|\, \beta(\boldsymbol{x}) \leq \mathbf{0}\}$, a point $\boldsymbol{x}_0$ is an ϵ-optimal solution if $\beta(\boldsymbol{x}_0) \leq \mathbf{0}$ and $\alpha(\boldsymbol{x}_0) \leq z^* + \epsilon$.

[15] For an optimization problem of the form $z^* = \min\{\alpha(\boldsymbol{x}) \,|\, \beta(\boldsymbol{x}) \leq \mathbf{0}\}$, a point $\boldsymbol{x}_0$ is an ϵ-feasible solution if $\beta(\boldsymbol{x}_0) \leq \boldsymbol{\epsilon}$.

4.2 Dual Method

The second approach to solve (DRO) handles the quantifier $\forall\, P$ through the dualization of $\sup_{P\in\mathcal{P}} \mathcal{R}_P\left[h_j(\boldsymbol{x},\boldsymbol{\xi})\right]$, $j \in \{0\}\cup[m]$. Under suitable regularity conditions, there is no duality gap between the primal problem and its dual, i.e., strong duality holds. Hence, the supremum can be replaced by an infimum which should hold for at least one corresponding solution in the dual space. We refer to this approach as a *dual method*. Most of the existing papers in the DRO literature are focused on the dual method, see, e.g., Ben-Tal et al. [32], Bertsimas et al. [48], Delage and Ye [105], Wiesemann et al. [409]. A situation where one benefits from the application of the dual method to solve (DRO) arises when the ambiguity set of probability distribution depends on decision $\boldsymbol{x}$ as formulated below, see, e.g., Luo and Mehrotra [260], Noyan et al. [288]:

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{P\in\mathcal{P}(\boldsymbol{x})} \left\{ \mathcal{R}_P\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right] \,\middle|\, \sup_{P\in\mathcal{P}(\boldsymbol{x})} \mathcal{R}_P\left[h_j(\boldsymbol{x},\boldsymbol{\xi})\right] \le 0,\ j\in[m] \right\}, \tag{19}$$

where, $\mathcal{P}(\boldsymbol{x})$ denotes a *decision-dependent* ambiguity set of the probability distributions.

The papers that rely on the dual method exploit linear duality, Lagrangian duality, convex analysis (e.g., support function, conjugate duality, Fenchel duality), and conic duality. A fundamental question is then under what conditions the strong duality holds. One such condition is the existence of a probability measure that lies in the interior of the ambiguity set, i.e., the ambiguity set satisfies a Slater-type condition. We refer the readers to the optimization textbooks for results on linear and Lagrangian duality, see, e.g., Bazaraa et al. [20], Bertsekas [36], Rockafellar [335], Ruszczyński [347]. For detailed discussions of the duality theory in infinite-dimensional convex problems, we refer to Rockafellar [335], and we refer to Isii [209] and Shapiro [363] for duality theory in conic LPs. Below, we briefly present the results from conic duality that are widely used in the dualization of DRO models.

▶ **Theorem 18** (Shapiro [363, Proposition 2.1])**.** *For a linear mapping $A : \mathcal{V} \mapsto \mathcal{W}$, recall the definition of the adjoint mapping $A^* : \mathcal{W}^* \mapsto \mathcal{V}^*$, where $\langle w^*, Av\rangle = \langle A^*w^*, v\rangle$, $\forall\, v\in\mathcal{V}$. Consider a conic linear optimization problem of the form*

$$\min_{v\in\mathcal{C}} \langle c, v\rangle \quad s.t.\ Av \succcurlyeq_{\mathcal{K}} b, \tag{20}$$

where, $\mathcal{C}$ and $\mathcal{K}$ are convex cones and subsets of linear spaces $\mathcal{V}$ and $\mathcal{W}$, respectively, such that for any $w^*\in\mathcal{W}^*$, there exists a unique $v^*\in\mathcal{V}^*$ with $\langle w^*, Av\rangle = \langle v^*, v\rangle$, with $v^* = A^*w^*$, for all $v\in\mathcal{V}$. Then, the dual problem to (20) *is written as*

$$\max_{w^*\in\mathcal{K}'} \langle w^*, b\rangle \quad s.t.\ A^*w^* \preccurlyeq_{\mathcal{C}'} c. \tag{21}$$

Moreover, there is no duality gap between (20) *and* (21) *and both problems have optimal solutions if and only if there exists a feasible pair (v, w^*) such that $\langle w^*, Av - b\rangle = 0$ and $\langle c - A^*w^*, v\rangle = 0$.*

Note that the dual method can turn the DRO model into a convex minimization problem in special cases (e.g., linear objective function in P for the inner maximization problem subject to linear constraints on $\boldsymbol{x}$). In these cases, variants of the stochastic descent algorithm or stochastic approximation (see, e.g., Newton et al. [282]) may be used to solve the resulting reformulation.

A closely related subject to the dual method that motivates the use of convex duality is a game-theoretic interpretation of DRO. For the ease of exposition, let us consider a problem of the form (8). The decision maker, the first player in this setup, makes a decision $\boldsymbol{x}\in\mathcal{X}$ whose consequences (i.e., cost h_0) depends on the outcome of the random vector $\boldsymbol{\xi}$. The decision maker assumes that $\boldsymbol{\xi}$ follows some distribution $\mathbb{P}\in\mathcal{P}$. However, he/she does not know which distribution the nature, the second player in this setup, will choose to represent the uncertainty in $\boldsymbol{\xi}$. Thus, in one hand, the decision maker is looking for a decision that minimizes the maximum expected cost with respect to $\mathcal{P}$; while on the other hand, the nature is seeking a distribution that maximizes the minimum expected cost with respect to $\mathcal{X}$. Under suitable conditions, it can be shown that these two problems are the dual of each other and the solution to one problem provides the solution to the other problem. Such a solution $(\boldsymbol{x}^*, \mathbb{P}^*)$ is called an *equilibrium* or *saddle* point. In other words, at this point, the decision maker would not change its decision $\boldsymbol{x}^*$, knowing that the nature chose $\mathbb{P}^*$. Similarly, the nature would not change its distribution $\mathbb{P}^*$, knowing that the decision maker chose $\boldsymbol{x}^*$. We state this result in the following theorem, which generalizes John von Neumann's minimax theorem.

▶ **Theorem 19** (Sion [378, Theorem 3.4]). *Suppose that*
1. $\mathcal{X}$ *and* $\mathcal{P}$ *are convex and compact spaces,*
2. $\boldsymbol{x} \mapsto \mathcal{R}_{\mathbb{P}}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]$ *is upper semicontinuous and quasiconcave on* $\mathcal{P}$ *for all* $\boldsymbol{x} \in \mathcal{X}$*, and*
3. $\mathbb{P} \mapsto \mathcal{R}_{\mathbb{P}}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]$ *is lower semicontinuous and quasiconvex on* $\mathcal{X}$ *for all* $\mathbb{P} \in \mathcal{P}$.

Then,

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}} \mathcal{R}_{\mathbb{P}}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] = \sup_{\mathbb{P}\in\mathcal{P}} \inf_{\boldsymbol{x}\in\mathcal{X}} \qquad rrisk\mathbb{P}h_0(\boldsymbol{x}, \boldsymbol{\xi}).$$

According to Theorem 19, under appropriate conditions, exchanging the order of infimum and supremum will not change the optimal value to $\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]$. We refer to Grünwald and Dawid [166] for a variety of alternative regularity conditions for this to hold. The exchange of the order between inf and sup can be interpreted as follows Grünwald and Dawid [166]: a probability distribution $\mathbb{P}^*$ that maximizes the *generalized entropy* $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathcal{R}_{\mathbb{P}}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]$ over $\mathcal{P}$ has an associated decision $\boldsymbol{x}^*$, achieving $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathcal{R}_{\mathbb{P}^*}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]$, and it achieves $\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}} \mathcal{R}_{\mathbb{P}}\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]$. We note that the assumption that $\mathcal{X}$ is convex may not hold in many models (for example, when the decision variables are integer). Nevertheless, this theorem may be useful when constructing reformulations and developing algorithms, see, e.g., Gao et al. [151].

4.3 Numerical Methods

Some researchers have used numerical methods to solve (DRO). To explain the ideas, let us consider problem (8). Assuming that Ξ is a finite sample space with M atoms, i.e., $\Xi = \{s_1, \ldots, s_M\}$, and similar to the setup in Remark 2, we can write (8) as

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\boldsymbol{p}\in\mathcal{P}} \sum_{k\in[M]} p_k h_0\big(\boldsymbol{x}, \boldsymbol{\xi}(s_k)\big). \tag{22}$$

We further assume that $\mathcal{X}$ is a convex set, h_0 is continuous (not necessarily smooth), and $h_0(\,\cdot\,, \boldsymbol{\xi})$ is convex for each $\boldsymbol{\xi} \in \Xi$. The resulting DRO model (22) is then a convex-concave saddle-point problem, i.e., a two-player game between the $\boldsymbol{x}$ player and the P player.

Numerical algorithms to solve a convex-concave saddle-point problem alternate between a variant of mirror descent and ascent algorithms (Ben-Tal and Nemirovski [25], Nemirovski et al. [281]). When the minimax formulation is constrained, variants of the stochastic approximation may project the candidate solution onto the corresponding feasible region to derive the next iterate. Another possibility is to search for a feasible direction along which the next iterate is guaranteed to belong to the feasible region. Hence, when solving the inner maximization problem for a DRO problem, it is necessary to not only guarantee the ascent step provides valid probability distribution but also the iterate satisfies other constraints in the ambiguity set.

The literature on numerical methods for DRO is small. Liu et al. [249] propose to turn (22) into a bilinear saddle-point problem as follows:

$$\inf_{(\boldsymbol{x},\boldsymbol{\theta})\in\mathcal{G}} \sup_{\boldsymbol{p}\in\mathcal{P}} \sum_{k\in[M]} p_k \theta_k, \tag{23}$$

where

$$\mathcal{G} := \left\{(\boldsymbol{x}, \boldsymbol{\theta}) \in \mathbb{R}^{n+M} \,\middle|\, \boldsymbol{x} \in \mathcal{X},\ \theta_k \geq h_0\big(\boldsymbol{x}, \boldsymbol{\xi}(s_k)\big),\ k \in [M]\right\}.$$

They propose to solve (23) with a primal-dual hybrid algorithm, proposed in Chambolle and Pock [78]. This algorithm obtains an $\mathcal{O}(1/\epsilon)$ iteration complexity bound, and it involves projecting into $\mathcal{G}$ and $\mathcal{P}$ to obtain the iterates. They showcase this method for cases where the ambiguity set is formed via the moment constraints as in (55) or the Wasserstein metric. Motivated by the decomposition scheme of the progressive hedging algorithm, Chen et al. [93] propose to reformulate (23) as the bilinear saddle-point problem

$$\inf_{\boldsymbol{x}_0\in\mathcal{X},(\boldsymbol{x}_k,\theta_k)\in\mathcal{G}_k} \sup_{\boldsymbol{p}\in\mathcal{P}} \sup_{\boldsymbol{\lambda}_k} \sum_{k\in[M]} p_k \theta_k + \sum_{k\in[M]} \boldsymbol{\lambda}_k^\top (\boldsymbol{x}_0 - \boldsymbol{x}_k), \tag{24}$$

where

$$\mathcal{G}_k := \left\{(\boldsymbol{x}_k, \theta_k) \in \mathbb{R}^{n+1} \,\middle|\, \boldsymbol{x}_k \in \mathcal{X},\ \theta_k \geq h_0\big(\boldsymbol{x}, \boldsymbol{\xi}(s_k)\big)\right\}.$$

Hence, the primal-dual hybrid algorithm Chambolle and Pock [78] can still be applied to (24), where now the projections on $\mathcal{G}_k$ can be done in parallel, although still cumbersome. This algorithm obtains an $\mathcal{O}(M/\epsilon)$ iteration complexity bound. Chen et al. [93] showcase their proposed algorithm for DRO problems with the moment constraints as in (55), Wasserstein metric, and ϕ-divergences. Zhang et al. [440] propose to use biconjugation of $h_0(\boldsymbol{x}, \cdot)$ to reformulate (22) as

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\boldsymbol{p}\in\mathcal{P}} \sum_{k\in[M]} \sup_{\boldsymbol{\pi_k}\in\Pi_k} p_k \left(\boldsymbol{\pi}_k^\top \boldsymbol{x} - h_0^*\big(\boldsymbol{\pi}_k, \boldsymbol{\xi}(s_k)\big)\right), \tag{25}$$

where Π_k denotes the domain of the conjugate function $h_0^*\big(\,\cdot\,, \boldsymbol{\xi}(s_k)\big)$. Note that the objective function in (25) is no longer jointly concave in $\boldsymbol{p}$ and $\boldsymbol{\pi} := [\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_M]$. Nevertheless, given that $\boldsymbol{p} \geq 0$, the inner maximization problem in (25) can be solved sequentially in $\boldsymbol{\pi}$ and $\boldsymbol{p}$, in order. Taking advantage of this property and by exploiting the geometry of $\mathcal{P}$, Zhang et al. [440] propose two algorithms to solve (25). The first algorithm, referred to as sequential dual (SD), relies on a decomposition of the primal-dual gap into individual optimality gaps of $\boldsymbol{\pi}$, $\boldsymbol{p}$, and $\boldsymbol{x}$ blocks in the form of simple convex functions. Hence, standard first-order methods can be applied by iterative proximal updates. The second algorithm, referred to as sequential smoothing level (SSL), builds an adaptively smoothed approximation of the objective function of (25) based on the $\boldsymbol{\pi}$ and $\boldsymbol{p}$ blocks. Then, it applies a bundle-level type method to the approximation. In both SD and SSL, proximal updates for the $\boldsymbol{x}$ and $\boldsymbol{\pi}$ are done via the Euclidean distance, while both the Euclidean and entropy distances are explored for the $\boldsymbol{p}$ updates. Zhang et al. [440] show that both the SD and SSL algorithms attain the iteration complexity bounds $\mathcal{O}(\sqrt{M}/\epsilon)$ and $\mathcal{O}(\sqrt{\log M}/\epsilon)$ when the Euclidean and entropy distances are used for the $\boldsymbol{p}$ updates, respectively. Zhang et al. [440] adapt their proposed algorithms for the case that the ambiguity set is formed via the Wasserstein metric, where they apply a proximal update of the transportation plan instead of the proximal update of $\boldsymbol{p}$ to avoid the expensive projection onto $\mathcal{P}$.

Namkoong and Duchi [272] propose a variant of mirror descent SA algorithm, proposed in Nemirovski et al. [281], for ambiguity sets formed via the Cressie–Read divergences over discrete probability distributions (Cressie–Read divergences belongs to the class of ϕ-divergences. See Section 6.1.2 for more information). The proposed algorithm is shown to require almost the same computational cost as the stochastic gradient descent algorithm, and it involves proper projections to obtain the iterates of $\boldsymbol{x}$ and P. Alternatively, as mentioned before, Namkoong and Duchi [272] could have solved the minimization problem resulted from the dualization of the inner maximization over the divergence-based set using a variant of the stochastic gradient descent algorithm. However, a motivation to directly handle the minimax formulation is that the stochastic gradient descent becomes unstable as one of the dual variables goes to zero (see Theorem 24). This observation is also made in Namkoong and Duchi [271], where they also observe that focusing on the entire data points make the method computationally cumbersome. They propose a primal subgradient descent algorithm to solve the DRO model, where they obtain the solution to the inner problem via a bisection method. However, at each step the inner maximization problem is solved over a geometrically increasing subset of the data points, obtained from sampling without replacement. The step on updating $\boldsymbol{x}$ then involves a subgradient of the approximated inner objective function value.

4.4 Approximation Schemes with Decision Rules

Decision-making problems under uncertainty, particularly those in a dynamic setting, typically suffer from the curse of dimensionality and are computationally intractable (Shapiro and Nemirovski [371]). This is partly because (recourse) decisions depend on the realizations of the random vector $\boldsymbol{\xi}$ and one need to optimize over the space of all functions (adapted to $\boldsymbol{\xi}$). To overcome these challenges, a common approximation scheme is to restrict (recourse) decisions to a space that leads to a computationally more tractable problem.

Decision rules, proposed in Charnes and Cooper [79, 80], Charnes et al. [81] in the context of chance-constrained programming, and popularized in Ben-Tal et al. [28] in the context of adjustable robust optimization, is a common approximation scheme in the DRO literature, see, e.g., Bertsimas et al. [53], Goh and Sim [159]. Thus, rather than optimizing over the space of all functions, one can optimize over a finite collection of decision variables to achieve a conservative approximation. One may apply decision rules to the dual problem as well Kuhn et al. [223]. Decision rules in RO are inspired by linear feedbacks in controlled dynamical systems Ben-Tal et al. [28], and we refer the readers to Georghiou et al. [152], Yanıkoğlu et al. [432] for an overview of decision rules in optimization under uncertainty.

Approximation schemes based on decision rules are typically achieved via *linear* and *nonlinear* rules. As its name suggests, linear decision rules (LDR) are those that linearly depend on $\boldsymbol{\xi}$, see, e.g., Bertsimas et al.

[55], Peng and Delage [296]. Bertsimas et al. [53] propose a linear approximation by incorporating auxiliary variables associated with the lifted ambiguity set in the decision rule (Wiesemann et al. [409]).

Nonlinear decision rules, on the other hand, do not linearly depend on $\boldsymbol{\xi}$. Hence, by allowing a richer class of functions, nonlinear decision rules may lead to a less conservative approximation than LDR, albeit at the price of a typically more computationally demanding problem. Approximations based on piece-wise linear decision rules are proposed in Goh and Sim [159]. These approximations, termed as bideflected LDR, generalize the deflected and segregated LDR of Chen et al. [92], Chen and Zhang [90] and truncated LDR of See and Sim [354]. Finite adaptability is another special case of nonlinear decision rules and allows for constant or linear decisions over a finite partition of the support set, see, e.g., Bertsimas et al. [55], Peng and Delage [296]. This method is widely used when recourse decisions are discrete, see, e.g., Bertsimas and Caramanis [39], Hanasusanto et al. [176], Subramanyam et al. [387].

A criticism of decision rules is that, in general, they are not optimal. Hence, quantifying the suboptimality of decision rules becomes essential. Despite this, decisions rules appear to perform reasonably well for some applications, see, e.g., Bertsimas et al. [53], Peng and Delage [296]. We also note that while decision rules provides a conservative approximation, they do not eliminate the quantifier $\forall\, P$, explained at the beginning of Section 4. Hence, one would still need to rely on the techniques described in Section 4.1–4.3 to address the resulting approximate problem.

5 Cost Function of the Inner Problem

Recall formulation (DRO) and the functional $\mathcal{R}_P : \mathcal{Z} \mapsto \mathbb{R}$. This functional accounts for quantifying the uncertainty in the outcomes of a fixed decision $\boldsymbol{x} \in \mathcal{X}$ and for a given fixed probability measure $P \in \mathfrak{M}(\Xi, \mathcal{F})$. As pointed out before in Section 1.1 for (1) and (2), one choice for this functional is the expectation operator. Other functionals, such as *regret function*, *risk measure*, and *utility function* have also been used in the DRO literature. These functionals are closely related concepts and we refer to Ben-Tal and Teboulle [27] and Rockafellar and Royset [338] for a comprehensive treatment and how one can induce one from the other. In this section, we review some notable works, where regret function, risk measure, and utility function are used to capture the uncertainty in the outcomes of the decision.

5.1 Regret Function

As an alternative to the worst-case expected criteria, the modeling approach DRO allows for a decision criterion that optimizes the disappointment or *regret* of finding out that another decision would have achieved a better cost under the realized uncertainty. This decision criteria for decision-making under uncertainty is introduced in Savage [350] and we refer to Blackwell and Girshick [58] for general information.

Given a decision $\boldsymbol{x} \in \mathcal{X}$ and a probability measure $P \in \mathfrak{M}(\Xi, \mathcal{F})$, a regret functional $\mathcal{V}_P$ may quantify the expected displeasure or disappointment of the current decision with respect to a possible mix of future outcomes as follows:

$$\mathcal{V}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] := \mathbb{E}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi}) - \min_{\boldsymbol{x}' \in \mathcal{X}} h_0(\boldsymbol{x}', \boldsymbol{\xi})\right]. \tag{26}$$

In other words, $\mathcal{V}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]$, defined in (26), calculates the expected additional loss that could have been avoided by acting optimally. Definition (26) of regret function is used in Natarajan et al. [278] and Hu et al. [196] in the context of combinatorial optimization and multicriteria decision-making, respectively. Another, and perhaps more popular, way for formulating a regret function may be as

$$\mathcal{V}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] := \mathbb{E}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] - \min_{\boldsymbol{x}' \in \mathcal{X}} \mathbb{E}_P\left[h_0(\boldsymbol{x}', \boldsymbol{\xi})\right], \tag{27}$$

which quantifies the (absolute) difference between the cost of decision $\boldsymbol{x}$ and the optimal decision (e.g., of a clairvoyant) under distribution P. The regret function (27) can be interpreted as the expected value of additional information or the value that the decision maker is willing to pay to acquire information about the underlying distribution. The worst-case regret resulting from the distributional ambiguity can be stated as $\max_{P \in \mathcal{P}} \mathcal{V}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]$, and one may consider the minimax regret criterion

$$\min_{\boldsymbol{x} \in \mathcal{X}} \max_{P \in \mathcal{P}} \left\{\mathbb{E}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] - \min_{\boldsymbol{x}' \in \mathcal{X}} \mathbb{E}_P\left[h_0(\boldsymbol{x}', \boldsymbol{\xi})\right]\right\}.$$

It can be shown that the minimax regret criterion leads to an optimal value not greater than the difference between those of optimistic and pessimistic criteria (recall Remark 1). Hence, the minimax regret criterion can lead to decisions that are neither too conservative nor aggressive. This type of regret function is used in Chen and Xie [94], Perakis and Roels [297], Yue et al. [436] in the context of the newsvendor problem. Yue et al. [436] extend the work of Scarf [351] with a minimax regret criteria. Perakis and Roels [297] obtain closed form solutions to distributionally robust single-item newsvendor problems that minimize the worst-case regret, where only (1) support, (2) mean, (3) mean and median, and (4) mean and variance information is available. This information can be captured with the ambiguity set $\mathcal{P}^{\text{MM}}$, to be defined in 6.2.3. Perakis and Roels [297] also study the ambiguity sets that preserve the shape of the distribution, including information on (1) mean and symmetry, (2) support and unimodality with a given mode, (3) median and unimodality with a given mode, and (4) mean, symmtery, and unimodality with a given mode. Chen and Xie [94] study a newsvendor problem with minimax regret criteria, where the distributional ambiguity is modeled via Wassrstein distance, to be reviewed in Section 6.1.1. For general information on regret-based DRO models, we refer to Lim et al. [245]. Absolute and relative regret functions are also studied in RO, see, e.g., Bertsimas and Dunning [40], Poursoltani and Delage [314].

5.2 Risk Measure

As introduced in Section 3.1.1, a functional that quantifies the uncertainty in the outcomes of a decision is a risk measure (Acerbi [1], Artzner et al. [10], Kusuoka [225], Shapiro [365]). A risk measure ρ_P usually satisfies some *averseness* property, i.e., $\rho_P[\cdot] > \mathbb{E}_P[\cdot]$, and imposes a preference order on random variables, i.e., if $Z, Z' \in \mathcal{Z}$ and $Z \geq Z'$, then $\rho_P[Z] \geq \rho_P[Z']$. Explicit incorporation of a risk measure into a DRO model has also received attention in the literature. We refer to Pflug et al. [303], Pichler [305], Pichler and Xu [307], Wozabal [412] for spectral and distortion risk measures, Calafiore [72] for variance, Calafiore [72] for mean absolute-deviation, Hanasusanto et al. [178], Wiesemann et al. [410] for optimized certainty equivalent, Hanasusanto et al. [175] for CVaR, and Postek et al. [311] for a variety of risk measures. Delage and Li [103] study a risk minimization problem, where there is ambiguity on the underlying risk measure. However, the decision maker can state her risk preferences using a set of properties such as monotonicity, convexity, translation invariance, positive homogeneity, law invariance, and partial ordering.

5.3 Utility Function

An alternative to using risk measures to compare random variables is to evaluate their expected utility Gilboa and Schmeidler [154]. As before, let us consider a probability space $(\Xi, \mathcal{F}, P)$. A random variable $Z \in \mathcal{Z}$ is preferred over a random variable $Z' \in \mathcal{Z}$ if $\mathbb{E}_P[u(Z)] \geq \mathbb{E}_P[u(Z')]$ for a given univariate utility function u[16]. A bounded utility function u can be normalized to take values between 0 and 1, and hence, it can be interpreted as a cdf of a random variable ζ independent of Z, i.e., $u(t) = P\{\zeta \leq t\}$ for $t \in \mathbb{R}$. Under this interpretation, Z is preferred over Z' if $P\{Z \geq \zeta\} \geq P\{Z' \geq \zeta\}$ because

$$\mathbb{E}_P[u(Z)] = \mathbb{E}_P[P\{\zeta \leq Z|Z\}] = \mathbb{E}_P\left[\mathbb{E}_P\left[\mathbb{1}_{\{\zeta \leq Z\}}|Z\right]\right] = \mathbb{E}_P\left[\mathbb{1}_{\{\zeta \leq Z\}}\right] = P\{\zeta \leq Z\}.$$

However, as in decision theory, it is difficult to have a complete knowledge of a decision maker's preference (i.e., utility function), it is also difficult to have a complete knowledge of the cdf of ζ. The notion of (second-order) *stochastic dominance* handles this issue by comparing the expected utility of random variables, for a given family $\mathcal{U}$ of utility functions, or equivalently, compare the probability of exceeding the target random variable ζ for a given family of cdfs Dentcheva and Ruszczynski [112], Dentcheva and Ruszczyński [113]. Consequently, to address the problem of ambiguity in decision maker's utility or equivalently, cdf of the random variable ζ, one can study

$$\min_{\boldsymbol{x} \in \mathcal{X}} \max_{\zeta \in \mathcal{U}} P\{h_0(\boldsymbol{x}, \boldsymbol{\xi}) \geq \zeta\}, \tag{28}$$

and

$$\min_{\boldsymbol{x} \in \mathcal{X}} \left\{ h_0(\boldsymbol{x}) \,\middle|\, \max_{\zeta \in \mathcal{U}} P\{h_j(\boldsymbol{x}, \boldsymbol{\xi}) \geq \zeta\} \leq 0,\ j \in [m] \right\}, \tag{29}$$

[16] For definitions in a multivariate case, we refer to Hu et al. [197, 198].

where $\mathcal{U}$ denotes a given family of normalized and nondecreasing utility functions, or equivalently, a given family of cdfs. Note that problems (28) and (29) have the form of problems (8) and (9), respectively. Robust preference optimization with respect to utility function is studied in several papers. Hu and Mehrotra [195] study problem of the form (28), where $\mathcal{U}$ is further restricted to include concave utility functions or equivalently, cdf, and satisfy functional bounds on the utility and marginal utility functions (cdf and pdf of ζ) as in (54). They provide a linear programming formulation of a particular case where the bounds on the utility function are piecewise-linear increasing concave functions, and the bounds on all other functions are step functions. For the general continuous case, they study an approximation problem by discretisizing the continuous functions, and analyze the convergence properties of the approximated problem. They apply their results to a portfolio optimization problem. Unlike Hu and Mehrotra [195], in Hu et al. [200], no shape restrictions on the utility function is assumed and only functional bounds on the utility function are enforced. Hu et al. [200] show that an SAA approach to the Lagrangian dual of the resulting problem can be used while solving a mixed-integer linear program. Bertsimas et al. [48] study a DRO model of the form (8), where a convex nondecreasing disutility function is used to quantify the uncertainty in decision. A utility function is closely related to risk measures (Hu and Mehrotra [195]). For instance, for a given probability measure, the expected utility might have the form of a combination of expectation and expected excess beyond a target, or an optimized certainty equivalent risk measure. As shown in Ben-Tal and Teboulle [27], under appropriate choices of utility functions, an optimized certainty equivalent risk measure can be reduced to the mean-variance and the mean-CVaR formulations. Wiesemann et al. [410] study a DRO model via a cross-moment or nested moment ambiguity set, to be reviewed in Section 6.2.5, where the decision maker is risk-averse via a nondecreasing convex piecewise-affine disutility function. In particular, they investigate shortfall risk and optimized certainty equivalent risk measures. Armbruster and Delage [7] assume that the nondecreasing utility function is ambiguous but satisfies certain properties. In particular, they consider the case that the utility function is concave to capture risk aversion or it is S-shaped (convex on loses and concave on gains) to capture risk aversion and risk receptiveness. Armbruster and Delage [7] also consider the case that the derivative of the utility function exists and is convex to capture risk prudence, i.e., the decision maker is more risk tolerant when the cost is lower. For an ambiguous risk minimization problem and given certain properties such as coherence and convexity on the risk measure, Delage et al. [107] derive an equivalent shortfall risk minimization problem where the utility function lies in an ambiguity set.

As we mentioned earlier, the notion of stochastic dominance handles the ambiguity in decision maker's preference by comparing the expected utility of random variables for a given family $\mathcal{U}$ of utility functions. However, the underlying probability distribution itself might be ambiguous. This naturally leads to a distributionally robust stochastic dominance constraint, first introduced in Dentcheva and Ruszczyński [115]. An axiomatic definition of a distributionally robust stochastic dominance constraint is presented in Peng and Delage [296]. Peng and Delage [296] and Mei et al. [264] study stochastic programs with a distributionally robust stochastic dominance constraint when the distributional ambiguity is modeled via Wasserstein distance, to be reviewed in Section 6.1.1. Mei et al. [264] study lower and upper approximations to this problem and establish their convergence analysis. Finite-sample and asymptotic guarantees as well as a tractable conservative approximation and solution algorithms are also studied in Peng and Delage [296].

Unlike the above discussion, many decision-making problems involve comparing random vectors. One can generalize the notion of utility-based comparison to random vectors by using multivariate utility functions Armbruster and Luedtke [8]. Another approach to compare random vectors is based on the idea of the weighted scalarization of random vectors. For the case that the weights are deterministic and take value in an arbitrary set, we refer to Dentcheva and Ruszczyński [114] for unrestricted sets, Homem-de-Mello and Mehrotra [193], Hu et al. [196], Hu and Mehrotra [194] for polyhedral sets, and Hu et al. [197] for convex sets. For instance, Hu et al. [196] study a weighted sum approach to a multiobjective budget allocation problem under uncertain performance indicators of projects. They assume that the weights take value in the convex hull of the weights suggested by experts and study a minimax approach to the expected weighted sum problem, where the expectation is taken with respect to the uncertainty in the performance indicators and the worst-case is taken with respect to the weights. Note that the problem studied in Hu et al. [196] is in the framework of RO as the weights are deterministic.

The idea of using stochastic weights, governed by a probability measure that determines the relative importance of each vector of weights, is also introduced in Hu and Mehrotra [194] and Hu et al. [198]. For instance, Hu and Mehrotra [194] study a DRO approach to stochastically weighted multiobjective deterministic and stochastic optimization problems, where the weights are perturbed along different rays from a reference weight vector.

They study the reformulations of the deterministic problem for the cases where the weights take values in (1) a polyhedral set, including those induced by a simplex, ℓ_1-norm, and ℓ_∞-norm, and (2) a conic-representable set, including those induced by a single cone (e.g., ℓ_p-norm, ellipsoids), intersection of multiple cones, and union of multiple cones. They further study the stochastic optimization problem. For the case that the weights and random parameters are independent, and the ambiguity in the probability distribution of weights is modeled via ellipsoid and matrix inequality ambiguity set, introduced in Delage and Ye [105], they obtain a reformulation of the problem. For the case that the weights and random parameters are dependent, they also obtain reformulations of the resulting problem by utilizing the result from the deterministic case.

6 Ambiguity Sets of Probability Distributions

The ambiguity set of distributions in a DRO model provides a flexible framework to model uncertainty by allowing the modelers to incorporate partial information about the uncertainty, obtained from historical data or domain-specific knowledge. This information includes, but it is not limited to, support of the uncertainty, discrepancy from a reference distribution, descriptive statistics, and structural properties, such as symmetry and unimodality. Early DRO models considered ambiguity sets based on the support and moment information, for which techniques in global optimization for polynomial optimization problems and problem of moments are applied to obtain reformulations, see, e.g., Bertsimas et al. [47], Bertsimas and Popescu [42], Gilboa and Schmeidler [154], Lasserre [236], Popescu [309, 310]. Since then, many researchers have incorporated information such as descriptive statistics as well as the structural properties of the underlying unknown true distribution into the ambiguity set.

There are usually two principles to model the ambiguity set $\mathcal{P}$:

1. $\mathcal{P}$ should be chosen as small as possible,
2. $\mathcal{P}$ should be chosen so that the prescribed solution provides good out-of-sample statistical performances.

It is worth noting that high quality solutions with good out-of-sample performances may be obtained by enforcing the ambiguity set $\mathcal{P}$ to contain the unknown true distribution with certainty (or at least, with a high confidence). While this holds for most research papers, it is also known that $\mathcal{P}$ does not necessarily need to contain the unknown true distribution with high confidence to provide high quality solutions. In fact, even if $\mathcal{P}$ contains the true distribution with exactly zero confidence, the prescribed solution may still provide good statistical performances, see, e.g., Blanchet et al. [64], Duchi et al. [123], Lam [230], Lam and Zhou [232, 233]. The principle in attaining statistical guarantees, without enforcing $\mathcal{P}$ to be a confidence region, is to use the empirical likelihood or the profile likelihood; see Section 7 for more details.

Abiding by the above two principles not only reduces the conservatism of the problem but it also robustifies the problem against the unknown true distribution. These two principles, in turns, give rise to three questions:

1. what distributional information should the ambiguity set contain?
2. what are the nominal values of the included distributional information?
3. what should be the size of the ambiguity set?

We discuss the last two questions in Section 7, and focus on the information that is incorporated into the ambiguity set in this section.

With a few exceptions, the common practice in constructing the ambiguity set is that first, the type of information that should be incorporated into the ambiguity set is determined by decision makers/modelers. In this step, data does not directly affect the choice of information. Then, the nominal values of the included distributional information are chosen based on available data or belief. Finally, the parameters that control the size of the ambiguity set are chosen in a data-driven fashion. We emphasize that albeit being a common practice, the type of information in the ambiguity set, their nominal values, and the size of the ambiguity set might be chosen neither separately nor in a data-driven fashion. To make the transition between Section 6 and 7 somewhat smoother, we devote Section 6.4 to review those papers that address these three questions somewhat concurrently.

When dealing with the question of what distributional information should the ambiguity set contain, most researchers, on one hand, have focused on the ambiguity sets that facilitate a tractable (exact or conservative approximate) formulation, such as linear program (LP), second-order cone program (SOCP), or to a lesser degree, semidefinite program (SDP), so that efficient computational techniques can be developed. On the other hand, many researchers have focused on the expressiveness of the ambiguity set by incorporating information such as descriptive statistics as well as the structural properties of the underlying unknown true distribution.

In what follows in this section, we review different approaches to model the distributional ambiguity. We acknowledge that the ambiguity sets in the literature are typically categorized in two groups: *discrepancy-based* and *moment-based* ambiguity sets. In short, discrepancy-based ambiguity sets contain distributions that are close to a nominal distribution in the sense of some *discrepancy* measure, while moment-based ambiguity sets contain distributions whose moments satisfy certain properties. Within these two groups, some specific ambiguity sets have been given names, see, e.g., Hanasusanto et al. [177]. For example,

- *Markov* ambiguity set contains all distributions with known mean and support,
- *Chebyshev* ambiguity set contains all distributions with bounds on the first- and second-order moments,
- *Gauss* ambiguity set contains all unimodal distributions from within the Chebyshev ambiguity set,
- *Median-absolute deviation* ambiguity set contains all symmetric distributions with known median and mean absolute deviation,
- *Huber* ambiguity set contains all distributions with known upper bound on the expected Huber loss function,
- *Hoeffding*ambiguity set contains all componentwise independent distributions with a box support,
- *Bernstein* ambiguity set contains all distributions from within the *Hoeffding* ambiguity set subject to marginal moment bounds,
- *Choquet* ambiguity set contains all distributions that can be expressed as an infinite convex combination of extremal distributions of the set,
- *Mixture* ambiguity set contains all distributions that can be expressed as a mixture of a parametric family of distributions.

While we use the above terminology in this paper, we categorize DRO papers into four groups:

- Discrepancy-based ambiguity sets (Section 6.1),
- Moment-based ambiguity sets (Section 6.2),
- Shape-preserving ambiguity sets (Section 6.3),
- Kernel-based ambiguity sets (Section 6.4).

We briefly mentioned what is meant by discrepancy-based and moment-based ambiguity sets. In short, shape-preserving ambiguity sets contain distributions with similar structural properties (e.g., unimodality, symmetry). Kernel-based ambiguity sets contain distributions that are formed via a kernel function in a functional space. The above groups are not necessarily disjoint from a modeling perspective and there are some overlaps between them. However, we assign papers to these categories as close as possible to what the authors might explicitly or implicitly have stated in their work. We review these four groups of ambiguity sets in Sections 6.1–6.4.

Besides these four groups, they are papers that provide a unified modeling approach. A unified scenario-wise format for ambiguity sets to contain both the moment-based and discrepancy-based distributional information about the ambiguous distribution is proposed in Chen et al. [98]. It is shown that the ambiguity sets formed via generalized moments, mixture distribution, Wasserstein metric, ϕ-divergence, k-means clustering, among other, all can be represented under this unified ambiguity set. The key feature of this scenario-wise ambiguity set is the introduction of a discrete random variable, which represents a finite number of scenarios that would affect the distributional ambiguity of the underlying nominal random variable. This ambiguity set can be characterized by a finite number of (conditional) expectation constraints based on generalized moments (Wiesemann et al. [410]). For practical purposes, they restrict the ambiguity set to be second-order conic representable. Based on the scenario-wise ambiguity set, they introduce an adaptive robust optimization format that unifies the classical SP and (distributionally) RO models with recourse. They also introduce a scenario-wise affine recourse approximation to provide tractable solutions to the adaptive robust optimization model. Besides Chen et al. [98], there are some proposals for unified models in the context of discrepancy-based, moment-based, and shape-preserving models. As mentioned before, a broad class of moment-based ambiguity sets with conic-representable expectation constraints and a collection of nested conic-representable confidence sets is proposed in Wiesemann et al. [410], and a broad class of shape-preserving ambiguity sets is proposed in Hanasusanto et al. [177].

6.1 Discrepancy-Based Ambiguity Sets

In many situations, such as financial risk measurements (Glasserman and Xu [156]) or in the control of stochastic uncertain systems (Petersen et al. [298]), we have a *nominal* or *baseline* estimate of the underlying probability distribution. A natural way to hedge against the distributional ambiguity is then to consider a neighborhood of the nominal probability distribution by allowing some perturbations around it. So, the ambiguity set can be formed with all probability distributions whose *discrepancy* or *dissimilarity* to the nominal probability

distribution is sufficiently small. More precisely, such an ambiguity set has the following generic form:

$$\mathcal{P}^{\mathfrak{d}}(P_0;\epsilon) = \{P \in \mathfrak{M}(\Xi,\mathcal{F}) \,|\, \mathfrak{d}(P,P_0) \le \epsilon\}, \tag{30}$$

where P_0 denotes the nominal probability measure, and $\mathfrak{d} : \mathfrak{M}(\Xi,\mathcal{F}) \times \mathfrak{M}(\Xi,\mathcal{F}) \mapsto \mathbb{R}_+ \cup \{\infty\}$ is a functional that measures the discrepancy between two probability measure $P, P_0 \in \mathfrak{M}(\Xi,\mathcal{F})$, dictating the shape of the ambiguity set. Moreover, parameter $\epsilon \in [0,\infty)$ controls the size of the ambiguity set, and it can be interpreted as the decision maker's belief in P_0. Parameter ϵ is also referred to as the *level of robustness*.

A generic ambiguity set of the form (30) has been widely studied in the DRO literature. We relegate the discussion about P_0 and ϵ to Section 7. In this section, we review different discrepancy functionals $\mathfrak{d}(\,\cdot\,,\cdot\,)$ that are used in the literature. These include (i) *optimal transport discrepancy*, (ii) *ϕ-divergences*, (iii) *total variation metric*, (iv) *goodness-of-fit test*, (v) *Prohorov metric*, (vi) *ℓ_p-norm*, (vii) *ζ-structure metric*, (viii) *Levy metric*, and (ix) *contamination neighborhood*.

We emphasize that although all studied functionals $\mathfrak{d}$ can quantify the discrepancy between two probability measures, they may or may not be a metric. For example, Prohorov and total variation are probability metrics, see, e.g., Gibbs and Su [153], while *Kullback–Leibler* and χ^2-distance from the family of ϕ-divergences are not a probability metric. Thus, we refer to the models of the form (30) collectively as *discrepancy-based* ambiguity sets, as opposed to distance-based or metric-based terminologies, which are prevailed in the literature.

6.1.1 Optimal Transport Discrepancy

We begin this section by providing more details on the optimal transport discrepancy. Consider two probability measures $P_1, P_2 \in \mathfrak{M}(\Xi,\mathcal{F})$. Let $\Pi(P_1,P_2)$ denote the set of all probability measures on $(\Xi\times\Xi, \mathcal{F}\times\mathcal{F})$ whose marginals are P_1 and P_2:

$$\Pi(P_1,P_2) = \left\{ \pi \in \mathfrak{M}(\Xi\times\Xi, \mathcal{F}\times\mathcal{F}) \,\middle|\, \begin{array}{l} \pi(A\times\Xi) = P_1(A),\ \forall\, A\in\mathcal{F}, \\ \pi(\Xi\times A) = P_2(A),\ \forall\, A\in\mathcal{F} \end{array} \right\}.$$

Furthermore, suppose that there is a lower semicontinuous function $c : \Xi\times\Xi \mapsto \mathbb{R}_+\cup\{\infty\}$ with $c(s_1,s_2)=0$ if $s_1 = s_2$. Then, the optimal transport discrepancy between P_1 and P_2 is defined as:

$$\mathfrak{d}_c^{\mathrm{W}}(P_1,P_2) := \inf_{\pi\in\Pi(P_1,P_2)} \int_{\Xi\times\Xi} c(s_1,s_2)\pi(\mathrm{d}s_1\times\mathrm{d}s_2). \tag{31}$$

If, in addition, function c is symmetric (i.e., $c(s_1,s_2)=c(s_2,s_1)$) and $c^{\frac{1}{p}}(\,\cdot\,)$ satisfies a triangle inequality for some $1\le p<\infty$ (i.e., $c^{\frac{1}{p}}(s_1,s_2) \le c^{\frac{1}{p}}(s_1,s_3) + c^{\frac{1}{p}}(s_3,s_2)$), then, $\mathfrak{d}_{c^{\frac{1}{p}}}^{\mathrm{W}}(P_1,P_2)$ metricizes the weak convergence in $\mathfrak{M}(\Xi,\mathcal{F})$, see, e.g., Villani [403, Theorem 6.9]. If Ξ is equipped with a metric d and $c(\,\cdot\,) = d^p(\,\cdot\,)$, then $\mathfrak{d}_c^{\mathrm{W}}(P_1,P_2)$ is called *Wasserstein metric of order p or p-Wasserstein metric*, for short[17].

The optimal transport discrepancy (31) can be interpreted as the minimum cost of transporting one pile of dirt from a source, represented by P_1, to a sinkhole, represented by P_2. The cost of transporting a unit mass from $s_1\in\Xi$ to $s_2\in\Xi$ is captured by $c(s_1,s_2)$. The optimal transportation plan $\pi(A_1\times A_2)$ is then the amount of mass that is moved from the source A_1 to the sink A_2, where $A_1, A_2 \in \mathcal{F}$. When P_1 and P_2 are two discrete distributions, then finding the optimal transport discrepancy can be translated as solving a LP, that may be solved in polynomial time. Otherwise, if at least one of the probability measures P_1 and P_2 is continuous, then finding the optimal transport discrepancy might become computationally very challenging.

The optimal transport discrepancy (31) can be used to form an ambiguity set of probability measures as follows:

$$\mathcal{P}^{\mathrm{W}}(P_0;\epsilon) := \left\{P\in\mathfrak{M}(\Xi,\mathcal{F}) \,\middle|\, \mathfrak{d}_c^{\mathrm{W}}(P,P_0)\le\epsilon\right\}. \tag{32}$$

In general, there are two types of ambiguity sets of the form (32): (1) *discrete* ambiguity set, where there is only ambiguity in the probability distribution of $\boldsymbol{\xi}$, while the realizations are fixed, and (2) *continuous* ambiguity set, where there is ambiguity in both the probability distribution of $\boldsymbol{\xi}$ and its realizations. The discrete ambiguity set translates to the case that P_0 and any $P\in\mathcal{P}^{\mathrm{W}}(P_0;\epsilon)$ are discrete. On the other hand, the continuous ambiguity

[17] Wasserstein metric of order 1 is sometimes referred to as *Kantorovich* metric. Wasserstein metric of order ∞ is defined as $\inf_{\pi\in\Pi(P_1,P_2)} \pi\text{-}\operatorname{ess\,sup} c(s_1,s_2)$, where $\pi\text{-}\operatorname{ess\,sup}_{\Xi\times\Xi}[\,\cdot\,]$ is the essential supremum with respect to measure π: $\pi\text{-}\operatorname{ess\,sup}_{\Xi\times\Xi} c(s_1,s_2) = \inf\{a\in\mathbb{R} : \pi(s_1\times s_2\in\Xi\times\Xi : c(s_1,s_2)>a)=0\}$.

set translates to the case that either P_0 or any $P \in \mathcal{P}^{\mathrm{W}}(P_0;\epsilon)$, or both, are continuous. As we shall shortly see below, most of the literature focuses on the discrete ambiguity set and the continuous ambiguity set for the case that P_0 is discrete.

Pioneered by the work of Pflug and Wozabal [302], most of the literature on DRO has focused on the Wasserstein metric. Over the past few years, there has been a significant growth in the popularity of the optimal transport discrepancy to model the distributional ambiguity in DRO, in both operations research and machine learning communities, see, e.g., Blanchet et al. [67], Chen et al. [96], Gao and Kleywegt [148], Lee and Mehrotra [238], Lee and Raginsky [239], Luo and Mehrotra [259], Mehrotra and Zhang [263], Mohajerin Esfahani and Kuhn [266], Shafieezadeh-Abadeh et al. [355, 356], Singh and Póczos [376], Sinha et al. [377]. Wasserstein metric of order p, for $p = 1, 2$, are more popular due to their theoretical and empirical aspects. A DRO model via 1-Wasserstein metric is usually used when the function $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ is bounded or has linear growth, leading to LP reformulation when ℓ_1-norm or ℓ_∞-norm is utilized. A DRO model via 2-Wasserstein metric may be used for a larger class of functions such as quadratic forms. Before we review these papers, we present a duality result on $\sup_{P\in\mathcal{P}^{\mathrm{W}}(P_0;\epsilon)} \mathbb{E}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]$, proved in a general form in Blanchet and Murthy [62].

Because the infimum in the defintion of (31) is attained for a lower semicontinuous function c Rachev and Rüschendorf [325], Villani [403], we can rewrite $\sup_{P\in\mathcal{P}^{\mathrm{W}}(P_0;\epsilon)} \mathbb{E}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]$ as follows:

$$\sup_{\pi\in\Phi_{P_0,\epsilon}} \int_{\Xi} h_0(\boldsymbol{x}, s)\pi(\Xi \times \mathrm{d}s), \tag{33}$$

where

$$\Phi_{P_0,\epsilon} := \left\{ \pi \in \mathfrak{M}\left(\Xi\times\Xi, \mathcal{F}\times\mathcal{F}\right) \,\middle|\, \begin{array}{l} \pi \in \cup_{P\in\mathfrak{M}(\Xi,\mathcal{F})}\Pi(P_0, P), \\ \displaystyle\int_{\Xi\times\Xi} c(s_1, s_2)\pi(\mathrm{d}s_1 \times \mathrm{d}s_2) \le \epsilon \end{array} \right\}.$$

Recall that $\mathcal{S}\left(\Xi, \mathcal{F}\right)$ is the collection of all $\mathcal{F}$-measurable functions $\phi : (\Xi, \mathcal{F}) \mapsto (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$. With the primal problem (33), we have a dual problem

$$\inf_{(\lambda,\phi)\in\Lambda_{c,h_0(\boldsymbol{x},\cdot)}} \left\{ \lambda\epsilon + \int_{\Xi} \phi(s) P_0(\mathrm{d}s) \right\}, \tag{34}$$

where

$$\Lambda_{c,h_0(\boldsymbol{x},\cdot)} := \left\{ (\lambda, \phi) \,|\, \lambda \ge 0, \phi \in \mathcal{S}\left(\Xi, \mathcal{F}\right), \phi(s_1) + \lambda c(s_1, s_2) \ge h_0(\boldsymbol{x}, s_2), \forall\, s_1, s_2 \in \Xi \right\}.$$

▶ **Theorem 20** (Blanchet and Murthy [62, Theorem 1])**.** *For a fixed $\boldsymbol{x} \in \mathcal{X}$, suppose that $h_0(\boldsymbol{x}, \cdot)$ is upper semicontinuous and P_0-integrable, i.e., $\int_\Xi |h_0(\boldsymbol{x}, \boldsymbol{\xi}(s))| P_0(\mathrm{d}s) < \infty$. Let $\Phi_{P_0,\epsilon}$ and $\Lambda_{c,h_0(\boldsymbol{x},\cdot)}$ be defined as in* (33) *and* (34)*, respectively. Then,*

$$\sup_{\pi\in\Phi_{P_0,\epsilon}} \int_{\Xi} h_0(\boldsymbol{x}, s)\pi(\Xi \times \mathrm{d}s) = \inf_{(\lambda,\phi)\in\Lambda_{c,\boldsymbol{g}(\boldsymbol{x},\cdot)}} \left\{ \lambda\epsilon + \int_{\Xi} \phi(s) P_0(\mathrm{d}s) \right\}.$$

Moreover, there exists a dual optimal solution of the form (λ, ϕ_λ), for some $\lambda \ge 0$, where $\phi_\lambda(s_1) := \sup_{s_2\in\Xi}\{h_0(\boldsymbol{x}, s_2) - \lambda c(s_1, s_2)\}$. In addition, any feasible $\pi^* \in \Phi_{P_0,\epsilon}$ and $(\lambda^*, \phi_{\lambda^*}) \in \Lambda_{c,h_0(\boldsymbol{x},\cdot)}$ are primal and dual optimizers, satisfying

$$\int_{\Xi} h_0(\boldsymbol{x}, s)\pi^*(\Xi \times \mathrm{d}s) = \lambda^*\epsilon + \int_{\Xi} \phi_{\lambda^*}(s) P_0(\mathrm{d}s),$$

if and only if

$$h_0(\boldsymbol{x}, s_2) - \lambda^* c(s_1, s_2) = \sup_{s_3\in\Xi}\{h_0(\boldsymbol{x}, s_3) - \lambda^* c(s_1, s_3)\}, \pi^*\text{-almost surely}, \tag{35a}$$

$$\lambda^*\Big(\int_{\Xi\times\Xi} c(s_1, s_2)\pi(\mathrm{d}s_1 \times \mathrm{d}s_2) - \epsilon \Big) = 0. \tag{35b}$$

▶ **Corollary 21.** *Suppose that $h_0(\boldsymbol{x}, \cdot)$ is upper semicontinuous and P_0-integrable. Then,*

$$\sup_{P\in\mathcal{P}^{W}(P_0;\epsilon)} \mathbb{E}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right] = \inf_{\lambda\ge 0} \left\{ \lambda\epsilon + \mathbb{E}_{P_0}\left[\sup_{s'\in\Xi}\{h_0(\boldsymbol{x}, s') - \lambda c(s, s')\} \right] \right\}. \tag{36}$$

The importance of Theorem 20 and Corollary 21 is that (1) the transportation cost $c(\cdot,\cdot)$ is only known to be lower semicontinuous, (2) function $h_0(\boldsymbol{x},\cdot)$ is assumed to be upper semicontinuous and integrable, and (3) Ξ is a general Polish space. In fact, there are only mild conditions on $h_0(\boldsymbol{x},\cdot)$ and function c, and P_0 can be any probability measure. Moreover, $\sup_{P\in\mathcal{P}^{\mathrm{W}}(P_0;\epsilon)}\mathbb{E}_P[h_0(\boldsymbol{x},\boldsymbol{\xi})]$ can be obtained by solving a univariate reformulation of the dual problem (34), where it involves an expectation with respect to P_0 and a linear term in the level of robustness ϵ. We shall shortly comment on similar results in the literature but under stronger assumptions. As shown in Section 3.2, by using Theorem 20 or its weaker forms, researchers have shown that many mainstream machine learning algorithms, such as regularized logistic regression and LASSO, have a DRO representation, see, e.g., Blanchet and Kang [59, 60], Blanchet et al. [64], Gao et al. [150], Shafieezadeh-Abadeh et al. [355, 357].

While a strong duality result for DRO formed via the optimal transport discrepancy is provided in Blanchet and Murthy [62] under mild assumptions by utilizing Fenchel duality, Gao and Kleywegt [148], Kuhn et al. [224], Luo and Mehrotra [259], Mohajerin Esfahani and Kuhn [266], Zhao and Guan [443] are also among notable papers in this area. Below, we first highlight the main differences of these papers with Blanchet and Murthy [62] and comment on their main contributions for the specific functions $h_0(\boldsymbol{x},\boldsymbol{\xi})$ under study. A summary of main duality contributions and assumptions for optimal transport discrepancy-based DRO models is presented in Table 1.

Table 1 Duality and reformulation contributions for p-Wasserstein-based ambiguity sets. "usc" stands for upper semicontinuous and "2SLP" stands for a two-stage stochastic LP, i.e., (3) with $q_1 = 0$. Note that a 2SLP with stochasticity on the right-hand side or objective is a special case of function $h_0(\boldsymbol{x},\boldsymbol{\xi})$ that is convex or concave in $\boldsymbol{\xi}$, respectively.
Note: (a) Blanchet et al. [64]; (b): Gao and Kleywegt [148]; (c): Kuhn et al. [224]; (d): Luo and Mehrotra [259]; (e): Mohajerin Esfahani and Kuhn [266]; (f): Zhao and Guan [443].

		Function $h_0(\boldsymbol{x},\boldsymbol{\xi})$				
Ω	p	2SLP	concave and usc in $\boldsymbol{\xi}$	usc in $\boldsymbol{\xi}$	convex in $\boldsymbol{\xi}$	convex in $\boldsymbol{x}$ and $\boldsymbol{\xi}$
Convex	1	(f)	(d), (e)	(a), (b)	-	(d)
	≥ 1	-	(c)	(a), (b)	-	-
$\mathbb{R}^d$	1	(f)	(d), (e)	(a), (b)	(d), (e)	(d)
	≥ 1	-	(c)	(a), (b)	(c)	-
	2	-	(c)	(a), (b)	(c)	-
Polish	≥ 1	-	-	(a), (b)	-	-

Zhao and Guan [443] reformulate the studied problem as a semi-infinite linear two-stage robust optimization problem. In addition, they derive a closed-form expression of the worst-case distribution whose parameters can be obtained by solving a traditional two-stage robust optimization model. Using conic linear duality, Luo and Mehrotra [259] reformulate the studied problem as a SIP. In order to solve the resulting SIP, they propose a finitely convergent exchange method when the cost function $h_0(\cdot,\boldsymbol{\xi})$ is a general nonlinear function in $\boldsymbol{x}$, and a central cutting-surface method with a linear rate of convergence when the cost function $h_0(\cdot,\boldsymbol{\xi})$ is convex in $\boldsymbol{x}$ and $\mathcal{X}$ is convex. By utilizing Lagrangian duality, Gao and Kleywegt [148] prove a strong duality result for the studied DRO problems. They also show data-driven DRO problems can be approximated by robust optimization problems. The key to this is approximating the worst-case distributions (or obtaining a worst-case distribution, if it exists) via the first-order optimality conditions of the dual reformulation. Mohajerin Esfahani and Kuhn [266] study data-driven DRO problems formed via 1-Wasserstein metric. They reformulate the problem as a finite-dimensional convex program for different cost functions. This contribution is of importance as most of the previous research on DRO formed via Wasserstein ambiguity sets reformulates the problem as a finite-dimensional nonconvex program and relies on global optimization techniques, such as difference of convex programming, to solve the problem, see, e.g., Wozabal [411, Theorem 6]. Kuhn et al. [224] extend these results to the case that $p \geq 1$. We present their duality result for the case that function $h_0(\boldsymbol{x},\boldsymbol{\xi})$ is concave in $\boldsymbol{\xi}$ and Ω is a convex set.

▶ **Theorem 22** (Kuhn et al. [224, Theorem 8]). *Suppose that the uncertainty set Ω is convex and closed, and $h_0(\boldsymbol{x},\boldsymbol{\xi}) := \max_{j\in[J]} l_j(\boldsymbol{\xi})$, where $-l_j$ is a proper, convex, and lower semicontinuous function for all $j \in [J]$. Moreover, suppose that $\widehat{\mathbb{P}}_N$ is the empirical distribution with N data points $\{\boldsymbol{\xi}^i\}_{i=1}^N$. Suppose that the transportation cost $c(\cdot,\cdot)$ in the definition of $\mathcal{P}^W(\widehat{\mathbb{P}}_N;\epsilon)$ is $\|\cdot\|^p$. Then, we can rewrite $\sup_{\mathbb{P}\in\mathcal{P}^W(\widehat{\mathbb{P}}_N;\epsilon)}\mathbb{E}_{\mathbb{P}}[h_0(\boldsymbol{x},\boldsymbol{\xi})]$ as the optimal*

value of a finite convex minimization problem

$$\inf_{\gamma,t_i,\boldsymbol{u}_{ij},\boldsymbol{v}_{ij}} \gamma\epsilon + \frac{1}{N}\sum_{i=1}^{N} t_i$$
$$s.t. \begin{cases} \gamma \geq 0,\ t_i \in \mathbb{R},\ \boldsymbol{u}_{ij}, \boldsymbol{v}_{ij} \in \mathbb{R}^d, & i \in [N],\ j \in [J], \\ [-l_j]^*(\boldsymbol{u}_{ij} - \boldsymbol{v}_{ij}) + \delta^*(\boldsymbol{v}_{ij}|\Omega) - \boldsymbol{u}_{ij}^\top \boldsymbol{\xi}^i + \phi(q)\gamma \left\| \frac{\boldsymbol{u}_{ij}}{\gamma} \right\|_*^q \leq t_i, & i \in [N],\ j \in [J], \end{cases} \tag{37}$$

where $\frac{1}{p} + \frac{1}{q} = 1$, $\phi(q) = \frac{(q-1)^{q-1}}{q^q}$ *for* $q > 1$, *and* $\phi(1) = 1$. *For* $\gamma = 0$, *the expression* $0\|\frac{\boldsymbol{u}_{ij}}{0}\|_*^q$ *is interpreted as* $\lim_{\gamma\to 0} \gamma \left\| \frac{\boldsymbol{u}_{ij}}{\gamma} \right\|_*^q$.

As can be seen from Theorem 22, conjugate function of $-l_j$, support function of the set Ω, as well as dual-norm used in the definition of $\mathcal{P}^{\mathrm{W}}(\widehat{\mathbb{P}}_N; \epsilon)$ play important roles in the equivalent finite convex minimization problem. Special cases of Theorem 22 can be obtained for the case that (1) Ω is a polytope and $h_0(\boldsymbol{x}, \boldsymbol{\xi}) := \max_{j\in[J]} l_j(\boldsymbol{\xi})$, where $l_j(\boldsymbol{\xi})$ is affine in $\boldsymbol{\xi}$, (2) Ω is a polytope and $h_0(\boldsymbol{x}, \boldsymbol{\xi}) := \min_{j\in[J]} l_j(\boldsymbol{\xi})$, where $l_j(\boldsymbol{\xi})$ is affine in $\boldsymbol{\xi}$, (3) $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ is the second stage of a two-stage stochastic program with objective uncertainty, (4) $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ is the second stage of a two-stage stochastic program with right-hand side uncertainty, and (5) $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ is the indicator function of a set, e.g., in uncertainty quantification and chance constraints.

In addition to Blanchet and Murthy [62], Gao and Kleywegt [148], Mohajerin Esfahani and Kuhn [266], Zhao and Guan [443], there are other research on DRO problems formed via the optimal transport discrepancy, but under more restricted assumptions, that move the frontier of research in this area. Kuhn et al. [224] develop duality results for the case that P_0 is an elliptical probability distribution and present SDP reformulation under some conditions. Hanasusanto and Kuhn [174] study (8), where $h_0(\boldsymbol{x}, \xi)$ is defined as (3) with $q_1 = 0$ and with stochasticity on the right-hand side $\boldsymbol{T}(\boldsymbol{\xi})$ and $\boldsymbol{r}(\boldsymbol{\xi})$, and objective coefficients $\boldsymbol{q}(\boldsymbol{\xi})$. They model the distributional ambiguity via 2-Wasserstein metric utilizing ℓ_2-norm. By relying on the strong duality result from Mohajerin Esfahani and Kuhn [266] and Gao and Kleywegt [148], Hanasusanto and Kuhn [174] show that when the ambiguity set is formed around a discrete distribution, the resulting model is equivalent to a copositive program of polynomial size (if the problem has complete recourse) or it can be approximated by a sequence of copositive programs of polynomial size (if for any fixed $\boldsymbol{x}$ and $\boldsymbol{\xi}$, the dual of the second-stage problem is feasible). Moreover, by using nested hierarchies of semidefinite approximations of the (intractable) copositive cones from the inside, they obtain sequences of tractable conservative approximations to the problem. They also show if the stochasticity is only on the right-hand side, the ambiguity set is formed via the 1-Wasserstein metric around a discrete distribution utilizing ℓ_1-norm, and $\Xi = \mathbb{R}^d$, then the DRO model is equivalent to a LP.

Wozabal [411] study a DRO approach to single-stage stochastic programs, where the distributional ambiguity in the constraints and objective function is modeled via 1-Wasserstein metric utilizing ℓ_1-norm around the empirical distribution. Because such a model has a higher complexity than that of those with fixed atoms (see, e.g., Mehrotra and Zhang [263], Pflug and Pichler [300]), Wozabal [411] propose to reformulate the problem into an equivalent finite-dimensional, nonconvex saddle-point optimization problem, under appropriate conditions. The key ideas in Wozabal [411] to obtain such a reformulation are that (i) at any level of precision and in the sense of Kantorovich distance, every distribution in the ambiguity set can be represented via a discrete probability distribution supported on at most $(N + 3)$ atoms, and (ii) considering only the extremal distributions in the ambiguity set suffices to obtain the equivalent reformulation. This considerable reduction of complexity, without sacrificing optimality, motivates designing numerical methods to solve the problem. Hence, they propose to solve such a finite-dimensional reformulated problem via the exchange method, proposed in Pflug and Wozabal [302].

Pflug and Pichler [300] and Mehrotra and Zhang [263] allow varying the probabilities on atoms identical to those of the nominal distribution. Hence, the ambiguity set can be represented as a subset of a finite-dimensional space. Pflug and Pichler [300] study a DRO approach to single- and two-stage stochastic programs formed via the p-Wasserstein metric utilizing an arbitrary norm. To solve the resulting problem, they apply the exchange method, proposed in Pflug and Wozabal [302]. Mehrotra and Zhang [263] study a distributionally robust ordinary least squares problem, where the ambiguity set of probability distribution is formed via 1-Wasserstein metric utilizing ℓ_1-norm. They show that the resulting problem can be solved by using an equivalent SOCP reformulation.

Motivated by the drawback of moment-based DRO problems, Gao and Kleywegt [149] study DRO models formed via various ambiguity sets of probability distributions that incorporate the dependence structure between the uncertain parameters. In the case that there exists a linear dependence structure, they consider probability

distributions around a nominal distribution, in the sense of p-Wasserstein metric utilizing an arbitrary norm, satisfying a second-order moment constraint. They also study cases with different rank dependencies between the uncertain parameters, and obtain tractable reformulations of these models. Along the same lines as Gao and Kleywegt [149], Pflug and Pohl [301] study a DRO approach to portfolio optimization via the 1-Wasserstein metric utilizing an arbitrary norm. They address the case where the dependence structure between the assets is uncertain while the marginal distributions of the assets are known. Noyan et al. [288] study DRO model with decision-dependent ambiguity set, where the ambiguity set is formed via the p-Wasserstein metric utilizing ℓ_p-norm. Rujeerapaiboon et al. [346] study continuous and discrete scenario reduction (Arpón et al. [9], Dupačová et al. [129], Heitsch and Römisch [182, 183, 184]), where p-Wasserstein metric utilizing ℓ_p-norm is used as a measure of discrepancy between distributions.

While most of the literature with optimal transport discrepancy is focused on p-Wasserstein with $p \in [1, \infty)$, a few papers study DRO problems with ∞-Wassersterin ambiguity, see, e.g., Chen and Xie [94], Gao et al. [150], Gao and Kleywegt [149]. Bertsimas et al. [55] shows that a distribution $\mathbb{P}$ in a ∞-Wassersterin ball, centered around the empirical probability distribution $\widehat{\mathbb{P}}_N$ of N data points, is characterized as a mixture probability distribution $\mathbb{P} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_i$. In this representation, $\mathbb{P}_i$, $i \in [N]$, is a probability distribution with a support in an ϵ-ball, centered around $\boldsymbol{\xi}^i$. Using this observation, Bertsimas et al. [55] study a data-driven approach, referred to as sample robust optimization, to multistage stochastic linear optimization that consists of constructing multiple uncertainty sets around data points. It is shown that this approach can be interpreted as a DRO problem with ∞-Wasserstein when uncertainty sets are formed via the norm used in the description of the Wasserstein distance. We refer to Bertsimas et al. [55] for asymptotic optimality and feasibility guarantees for DRO problems with ∞-Wasserstein ambiguity. Bertsimas et al. [56] study a two-stage version of the problem studied in Bertsimas et al. [55]. Xie [413] derive tractable reformulations for two-stage LPs.

6.1.1.1 Finite-Sample and Asymptotic Guarantees

In this section, we discuss how the choice of the size of the ambiguity set results in finite-sample and asymptotic performance guarantees. A sufficient condition to achieve such performances is to ensure that the true unknown distribution is contained in the constructed Wasserstein ball with a high confidence.

When the ambiguity set contains all discrete distributions around the empirical distribution in the sense of the Wasserstein metric, Pflug and Wozabal [302] and Pflug et al. [303] propose to choose the level of robustness based on a probabilistic statement on the Wasserstein metric between the empirical and true distributions, due to Dudley [124], as $\epsilon = \frac{CN^{-\frac{1}{d}}}{\alpha}$. This choice of ϵ guarantees that $\mathbb{P}^N\{\mathfrak{d}_c^{\mathrm{W}}(\mathbb{P}, \widehat{\mathbb{P}}_N) \geq \epsilon\} \leq \alpha$, and consequently, a finite-sample guarantee with confidence $1-\alpha$ can be achieved. To achieve a finite-sample guarantee, Mohajerin Esfahani and Kuhn [266], Zhao and Guan [442] for $p = 1$ and Kuhn et al. [224] for $p \geq 1$ propose to use a modern convergence result for light-tail distributions, established in Fournier and Guillin [143]. This modern convergence result implies that to guarantee that the true unknown distribution is contained in the Wassesrtein ball with confidence $1-\alpha$, and consequently, achieving a finite-sample guarantee, the radius ϵ should decay as $\mathcal{O}(N^{-\frac{p}{d}})$, where d is the dimension of $\boldsymbol{\xi}$. That is, to reduce the radius by a factor of 2, the sample size N must increase by $2^{\frac{d}{p}}$. Nevertheless, such data-independent theoretical results usually lead to an overly conservative ambiguity set with size ϵ, i.e., $\mathbb{P}^N\{\mathfrak{d}_c^{\mathrm{W}}(\mathbb{P}, \widehat{\mathbb{P}}_N) \geq \epsilon\} \leq \beta$, where $\beta \ll \alpha$. Moreover, even if the unknown true probability distribution does not belong to the ambiguity set, the optimal value to the resulting DRO model may still provide an upper bound on the true optimal value. Thus, it is more practical to calibrate the size of the ambiguity set randomly and perhaps through cross-validation and bootstrapping. To alleviate these issues and choosing the size of the ambiguity set judiciously, Blanchet et al. [64] propose a mechanism based on *robust Wasserstein profile* (RWP) function that take advantage of the optimization framework and data. Their proposed size for the ambiguity set is such that it does not necessarily contain the unknown true distribution and decays faster than $\mathcal{O}(N^{-\frac{p}{d}})$, while achieving the desired finite-sample guarantee. Asymptotic performance guarantees under appropriate condition such as upper semicontinuity of $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ in $\boldsymbol{\xi}$ and its linear growth rate, as well as lower semicontinuity of $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ in $\boldsymbol{\xi}$ and closedness of $\mathcal{X}$ is shown in Kuhn et al. [224], Mohajerin Esfahani and Kuhn [266].

6.1.1.2 Worst-Case Distribution

In this section, we discuss a worst-case distribution that may attain the optimal value $\sup_{\mathbb{P}\in\mathcal{P}^{\mathrm{W}}(\hat{\mathbb{P}}_N;\epsilon)} \mathbb{E}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]$. As we mentioned, for $p = 1$, Wozabal [411] shows that any extremal distribution of the Wasserstein ball around the empirical measure is a discrete measure supported on at most $(N + 3)$ atoms for a continuous bounded

function $h_0(\boldsymbol{x}, \boldsymbol{\xi})$. This results is further improved to $(N+1)$ atoms in Gao and Kleywegt [148] for an upper semicontinuous function $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ and N atoms in Kuhn et al. [224], Mohajerin Esfahani and Kuhn [266] for an upper semicontinuous and concave function $h_0(\boldsymbol{x}, \boldsymbol{\xi})$. A characterization of the worst-case probability distribution for the optimal transport-based ambiguity sets around an N-atom empirical measure is listed in Table 2. In order to obtain a worst-case distribution that attains the optimal value of $\sup_{\mathbb{P}\in\mathcal{P}^{\mathrm{W}}(\hat{\mathbb{P}}_N;\epsilon)} \mathbb{E}_{\mathbb{P}}[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$, one may dualize the dual equivalent reformulation.

▶ **Theorem 23** (Kuhn et al. [224, Theorem 9], Mohajerin Esfahani and Kuhn [266, Theorem 4.4])**.** *Suppose that assumptions in Theorem 22 hold. Then,* $\sup_{\mathbb{P}\in\mathcal{P}^{W}(\hat{\mathbb{P}}_N;\epsilon)} \mathbb{E}_{\mathbb{P}}[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$ *is equivalent to the optimal value of the finite convex program*

$$\sup_{\alpha_{ij}, \boldsymbol{q}_{ij}} \frac{1}{N} \sum_{i\in[N]} \sum_{j\in[J]} \alpha_{ij} l_j \left(\boldsymbol{\xi}^i + \frac{\boldsymbol{q}_{ij}}{\alpha_{ij}}\right) \quad s.t. \begin{cases} \alpha_{ij} \in \mathbb{R}_+,\ \boldsymbol{q}_{ij} \in \mathbb{R}^d, & \forall\, i \in [N],\ j \in [J], \\ \boldsymbol{\xi}^i + \frac{\boldsymbol{q}_{ij}}{\alpha_{ij}} \in \Omega, & \forall\, i \in [N],\ j \in [J], \\ \sum_{j\in[J]} \alpha_{ij} = 1, & \forall\, i \in [N], \\ \frac{1}{N} \sum_{i\in[N]} \sum_{j\in[J]} \alpha_{ij} \left\| \frac{\boldsymbol{q}_{ij}}{\alpha_{ij}} \right\|^p \le \epsilon, \end{cases} \tag{38}$$

where $0 l_j \left(\boldsymbol{\xi}^i + \frac{\boldsymbol{q}_{ij}}{0}\right)$ *is defined as the value that makes the function* $\alpha_{ij} l_j \left(\boldsymbol{\xi}^i + \frac{\boldsymbol{q}_{ij}}{\alpha_{ij}}\right)$ *upper semicontinuous at* $(\boldsymbol{q}_{ij}, \alpha_{ij}) = (\boldsymbol{q}_{ij}, 0)$*. Similarly, the constraint* $\boldsymbol{\xi}^i + \frac{\boldsymbol{q}_{ij}}{0} \in \Omega$ *means that* $\boldsymbol{q}_{ij}$ *belongs to the recession cone of* Ω*, and* $0 \left\| \frac{\boldsymbol{q}_{ij}}{0} \right\|^p$ *is interpreted as* $\lim_{\alpha_{ij}\downarrow 0} \alpha_{ij} \left\| \frac{\boldsymbol{q}_{ij}}{\alpha_{ij}} \right\|^p$*. Moreover, let* $\{(\alpha_{ij}^r, \boldsymbol{q}_{ij}^r)\}_{r\in\boldsymbol{N}}$ *be a sequence of feasible decisions whose objective values converge to* $\sup_{\mathbb{P}\in\mathcal{P}^{W}(\hat{\mathbb{P}}_N;\epsilon)} \mathbb{E}_{\mathbb{P}}[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$*. Then, the discrete probability distributions*

$$\mathbb{Q}^r := \frac{1}{N} \sum_{i\in[N]} \sum_{j\in[J]} \alpha_{ij}^r \delta_{\boldsymbol{\xi}_{ij}^r},$$

with $\boldsymbol{\xi}_{ij}^r := \boldsymbol{\xi}^i + \frac{\boldsymbol{q}_{ij}^r}{\alpha_{ij}^r}$ *and* $\delta_{\boldsymbol{\xi}_{ij}^r}$ *denoting the Dirac point mass on* $\boldsymbol{\xi}_{ij}^r$*, belong to* $\mathcal{P}^W(\widehat{\mathbb{P}}_N;\epsilon)$ *and attains* $\sup_{\mathbb{P}\in\mathcal{P}^{W}(\hat{\mathbb{P}}_N;\epsilon)} \mathbb{E}_{\mathbb{P}}[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$ *asymptotically, i.e.,*

$$\sup_{\mathbb{P}\in\mathcal{P}^{W}(\hat{\mathbb{P}}_N;\epsilon)} \mathbb{E}_{\mathbb{P}}[h_0(\boldsymbol{x}, \boldsymbol{\xi})] = \lim_{r\to\infty} \mathbb{E}_{\mathbb{Q}^r}[h_0(\boldsymbol{x}, \boldsymbol{\xi})] = \lim_{r\to\infty} \frac{1}{N} \sum_{i\in[N]} \sum_{j\in[J]} \alpha_{ij}^r l(\boldsymbol{\xi}_{ij}^r).$$

As Theorem 23 reveals, a DRO model with a Wasserstein ambiguity set may result in future protection against unobserved realizations in Ω Gao and Kleywegt [148]. For the case that $p > 1$, it can be verified that the last constraint in (38) enforces that $\boldsymbol{q}_{ij} = \boldsymbol{0}$ when $\alpha_{ij} = 0$. Hence, it can be shown that the optimal value of (38) is attained at $(\alpha_{ij}^*, \boldsymbol{q}_{ij}^*)$, with the worst-case distribution $\mathbb{Q}^* := \frac{1}{N} \sum_{(i,j)\in\mathcal{S}_+} \alpha_{ij}^* \delta_{\boldsymbol{\xi}^i + \frac{\boldsymbol{q}_{ij}^*}{\alpha_{ij}^*}}$, where $\mathcal{S}_+ := \{(i,j) \in [N] \times [J] \,|\, \alpha_{ij} > 0\}$.

Thus, if $(i,j) \in \mathcal{S}_\infty$, it is possible to transport that atom to infinity along with a recession direction $\boldsymbol{q}_{ij}^*$ of Ω with a decaying probability, i.e., $\frac{1}{r}$.

Table 2 Construction of a worst-case probability distribution for optimal transport-based ambiguity sets around an N-atom empirical measure.
Note: (a): Gao and Kleywegt [148]; (b): Mohajerin Esfahani and Kuhn [266]; (c): Kuhn et al. [224]; (d) Wozabal [411].

		Function $h_0(\boldsymbol{x}, \boldsymbol{\xi})$		
Ω	p	continuous and bounded in $\boldsymbol{\xi}$	concave and usc in $\boldsymbol{\xi}$	usc in $\boldsymbol{\xi}$
Convex	1	(d)- $(N+3)$-atoms	(b)- N atoms	(a)- $(N+1)$ atoms
	≥ 1	-	(c)- N atoms	(a)- $(N+1)$ atoms
Polish	≥ 1	-	(a)- $(N+1)$ atoms	(a)- $(N+1)$ atoms

6.1.1.3 Choice of the Transportation Cost

When forming a Wasserstein ambiguity set, the transportation cost function $c(\cdot, \cdot)$ should be chosen in addition to the nominal probability measure P_0 and the size of the ambiguity set ϵ. Blanchet et al. [65] propose a

comprehensive approach for designing the ambiguity set in a data-driven way, using the role of the transportation cost $c(\cdot,\cdot)$ in the definition of the p-Wasserstein metric. They apply various metric-learning procedures to estimate $c(\cdot,\cdot)$ from the training data, where they associate a relatively high transportation cost to two locations if transporting mass between these locations substantially impacts performance. This mechanism induces enhanced out-of-sample performance by focusing on regions of relevance, while improving the generalization error. Moreover, this approach connects the metric-learning procedure to estimate the parameters of adaptive regularized estimators. Blanchet et al. [63] propose a data-driven robust optimization approach to optimally inform the transportation cost in the definition of the p-Wasserstein metric. This additional layer of robustification within a suitable parametric family of transportation costs does not exist in the metric-learning approach, proposed in Blanchet et al. [65], and it allows to enhance the generalization properties of regularized estimators while reducing the variability in the out-of-sample performance error.

6.1.1.4 Discrete Problems

We now review DRO models over Wasserstein ambiguity sets where there are discrete decisions. Bansal et al. [15] study a two-stage integer program, i.e., (8) with $h_0(\boldsymbol{x},\boldsymbol{\xi})$ defined as (3), with pure binary first-stage and mixed-binary second-stage variables on a finite set of scenarios as follows:

$$\min_{\boldsymbol{x}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \max_{P\in\mathcal{P}} \mathbb{E}_P\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right] \,\middle|\, \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b},\ \boldsymbol{x}\in\{0,1\}^n \right\},$$

where

$$h_0(\boldsymbol{x},\boldsymbol{\xi}) = \min_{\boldsymbol{y}} \left\{ \boldsymbol{q}^\top(\boldsymbol{\xi})\boldsymbol{y}(\boldsymbol{\xi}) \,\middle|\, \boldsymbol{W}(\boldsymbol{\xi})\boldsymbol{y}(\boldsymbol{\xi}) \geq \boldsymbol{r}(\boldsymbol{\xi}) - \boldsymbol{T}(\boldsymbol{\xi})\boldsymbol{x},\ \boldsymbol{y}(\boldsymbol{\xi}) \in \{0,1\}^{q_1}\times\mathbb{R}^{q-q_1} \right\}.$$

For the case that the ambiguity set of distributions is formed via 1-Wasserstein metric utilizing an arbitrary norm, they propose a finitely convergent decomposition-based L-shaped algorithm and a cutting surface algorithm to solve the resulting model. The results in Bansal et al. [15] are extended in Bansal and Zhang [14] to the case that the second-stage problem have p-order conic constraints as well as integer variables, and in Bansal and Mehrotra [13] to the case with disjunctive constraints in both stages. Xu and Burer [420] study a mixed-binary LP, where the coefficients of the objective functions are affinely dependent on the random vector $\boldsymbol{\xi}$. They seek a bound on the worst-case expected optimal value to this problem, where the worst-case is taken with respect to an ambiguity set of discrete distributions formed via 2-Wasserstein metric utilizing ℓ_2-norm around the empirical distribution of data. Under mild assumptions, they reformulate the problem into a copositive program, which leads to a tractable semidefinite-based approximation. Wang et al. [406] study a distributionally robust chance-constrained bin-packing problem with a finite number of scenarios, where the safe region of the chance constraint is bi-affine in $\boldsymbol{x}$ and $\boldsymbol{\xi}$, with a random technology matrix. They present a binary bilinear reformulation of the problem, where the feasible region is modeled as the intersection of multiple binary bilinear knapsack constraints, a cardinality constraint, and a general (probability) knapcksack constraint. They propose lifted cover valid inequalities for the binary bilinear knapsack substructure induced by a given bin and scenario, and they further obtain lifted cover inequalities that are valid for the substructure induced by each bin. They obtain valid probability cuts and incorporate them with the lifted cover inequalities in a branch-and-cut framework to solve the model. They show that the proposed algorithm is finitely convergent if a distribution separation subproblem can be solved in a finite number of iterations.

Recently, there has been interest in randomized policies, as opposed to deterministic policies, in SO and DRO problems (Delage et al. [106]). For a mixed-integer DRO problem, Delage and Saif [104] study the value of using a randomized policy. They show that the value of randomization for such DRO models with a convex cost function $h_0(\cdot,\boldsymbol{\xi})$ and a convex risk measure is bounded by the difference between the optimal values of the nominal DRO problem and that of its convex relaxation. They show that when the risk measure is an expectation and the cost function is affine in the decision vector, this bound is tight.

6.1.1.5 Risk and Chance Constraints

For a portfolio selection problem complemented via a broad class of convex risk measures appearing in the constraints, Wozabal [411] obtain an equivalent finite-dimensional, nonconvex, semidefinite saddle-point optimization problem. Pichler and Xu [307] study a DRO model with a distortion risk measure and form the ambiguity set of distributions via p-Wasserstein metric utilizing an arbitrary norm. They quantitatively investigate the effect of

the variation of the ambiguity set on the optimal value and the optimal solution in the resulting optimization problem, as the number of data points increases. They illustrate their results in the context of (8) with $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ defined as (3).

A class of distributionally robust fractional optimization problems with a finite sample space, representing a reward-risk ratio, is studied in Ji and Lejeune [211] as follows:

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{P\in\mathcal{P}} \frac{\mathcal{R}_P^1\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]}{\mathcal{R}_P^2\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\right]}. \tag{39}$$

Above, $\mathcal{R}_P^1 : \mathcal{Z} \mapsto \mathbb{R}$ is a reward measure and $\mathcal{R}_P^2 : \mathcal{Z} \mapsto \mathbb{R}_+$ is a nonnegative risk measure. They focus on the cases that (1) the reward measure is linear and the risk measure is concave in the probability vector (e.g., Sharpe ratio) and (2) both reward and risk measures are linear in the probability vector (e.g., Omega ratio). They model the ambiguity about discrete distributions using the 1-Wasserstein metric utilizing ℓ_1-norm, and provide a nonconvex reformulation for the resulting model by relying on the support function of the ambiguity set and the convex conjugate of the ratio function Postek et al. [311].

Ji and Lejeune [212] study a distributionally robust individual chance constraint, where the ambiguity set of distributions is formed via 1-Wasserstein metric utilizing ℓ_1-norm, and $h_j(\boldsymbol{x}, \boldsymbol{\xi})$, $j \in [m]$, in (9) is defined as

$$h_j(\boldsymbol{x}, \boldsymbol{\xi}) := \mathbb{1}_{[\boldsymbol{a}(\boldsymbol{\xi})^\top \boldsymbol{x} \le \boldsymbol{b}(\boldsymbol{\xi})]}(\boldsymbol{\xi}).$$

For the case that the underlying distribution is supported on the same atoms as those of the empirical distribution, they provide mixed-integer LP reformulations for the linear random right-hand side case, i.e., $h_j(\boldsymbol{x}, \boldsymbol{\xi}) := \mathbb{1}_{[\boldsymbol{a}^\top \boldsymbol{x} \le \boldsymbol{\xi}]}(\boldsymbol{\xi})$, and the linear random technology matrix case, i.e., $h_j(\boldsymbol{x}, \boldsymbol{\xi}) := \mathbb{1}_{[\boldsymbol{\xi}^\top \boldsymbol{x} \le \boldsymbol{b}]}(\boldsymbol{\xi})$. For the case that the underlying distribution is infinitely supported, they propose an exact mixed-integer SOCP reformulation for models with random right-hand side, while a relaxation is proposed for constraints with a random technology matrix. They show that this mixed-integer SOCP relaxation is exact when the decision variables are binary or bounded general integer.

Chen et al. [96] study data-driven distributionally robust chance constrained programs, where the ambiguity set of distributions is formed via p-Wasserstein metric utilizing an arbitrary norm. For individual linear chance constraints with affine dependency on the uncertainty, and for joint chance constraints with right-hand side affine uncertainty, they provide an exact deterministic reformulation as a mixed-integer conic program. When ℓ_1-norm or ℓ_∞-norm are used as the transportation cost in the definition of Wasserstein metric, the chance-constrained program can be reformulated as a mixed-integer LP. They leverage the structural insights into the worst-case distributions, and show that both the CVaR and the Bonferroni approximation may give solutions that are inferior to the optimal solution of their proposed reformulation. For other studies, we refer to Chen et al. [96], Ho-Nguyen et al. [190, 191], Jiang and Guan [213], Xie [414], Yang [428].

6.1.1.6 Statistical Learning

DRO problems formed via the optimal transport discrepancy has been widely studied in the context of statistical learning. Below, we review the latest developments of DRO in the context of statistical learning. Problems in this section are generally modeled as (8), where $h(\boldsymbol{x}, \boldsymbol{\xi})$ is interpreted as the loss function. Moreover, a set of data $\{\boldsymbol{\xi}^i := (\boldsymbol{u}^i, y^i)\}_{i=1}^N$ is available, where $\boldsymbol{u}^i \in \mathbb{R}^n$ is a vector of covariates and $y^i \in \mathbb{R}$ is the response variable. Given this setup, for example, for a linear regression model with a squared loss we have $h_0(\boldsymbol{x}, \boldsymbol{\xi}) := (y - \boldsymbol{x}^\top \boldsymbol{u})^2$, for a logistic regression model we have $h_0(\boldsymbol{x}, \boldsymbol{\xi}) := \log(1 + e^{-y\boldsymbol{x}^\top \boldsymbol{u}})$, and for a SVM with Hinge loss we have $h_0(\boldsymbol{x}, \boldsymbol{\xi}) := (1 - y\boldsymbol{x}^\top \boldsymbol{u})_+$. Under the assumption that the nominal measure P_0 is the empirical distribution on $\{\boldsymbol{\xi}^i\}_{i=1}^N$, (36) can be written as the following semi-infinite program:

$$\min_{\lambda, \boldsymbol{\theta}} \lambda\epsilon + \frac{1}{N}\sum_{i=1}^N \theta_i \quad \text{s.t.} \theta_i \ge h_0(\boldsymbol{x}, s) - \lambda c(\boldsymbol{\xi}^i, s), \quad i \in [N], s \in \Xi.$$

A data-driven distributionally robust maximum likelihood estimation model to infer the inverse of the covariance matrix of a normal random vector is proposed in Nguyen et al. [284]. They form the ambiguity set of distributions with all normal distributions close enough to a nominal distribution characterized by the sample mean and sample covariance matrix, in the sense of the 2-Wasserstein metric utilizing ℓ_1-norm. By leveraging an analytical formula for the Wasserstein distance between two normal distributions, they obtain an equivalent SDP reformulation of the problem. When there is no prior sparsity information on the inverse covariance matrix, they

propose a closed-form expression for the estimator that can be interpreted as a nonlinear shrinkage estimator. Otherwise, they propose a sequential quadratic approximation algorithm to obtain the estimator by solving the equivalent SDP.

Lee and Mehrotra [238] study a distributionally robust framework for finding support vector machines via the 1-Wasserstein metric. They provide a SIP formulation of the resulting model and propose a cutting-plane algorithm to solve the problem. Lee and Raginsky [239] study a distributionally robust statistical learning problem formed via the p-Wasserstein metric utilizing ℓ_p-norm, motivated by a domain (i.e., measure) adaption problem. This problem arises when training data are generated according to an unknown source domain $\mathbb{P}$, but the learned hypothesis is evaluated on another unknown but related target domain $\mathbb{Q}$. In this problem, it is assumed that a set of labeled data (covariates and responses) is drawn from $\mathbb{P}$ and a set of unlabeled covariates is drawn from $\mathbb{Q}$. It is further assumed that the domain drift is due to an unknown deterministic transformation on the covariates space that preserves the distribution of the response conditioned on the covariates. Under these assumptions and some further regularity conditions, they prove generalization bound and generalization error guarantees for the problem.

Gao et al. [151] develop a novel distributionally robust framework for hypothesis testing where the ambiguity set of distribution is constructed by 1-Wasserstein metric utilizing an arbitrary norm, around the empirical distribution. The goal is to obtain the optimal decision rule as well the least favorable distribution by minimizing the maximum of the worst-case type-I and type-II errors. They develop a convex safe approximation of the resulting problem and show that such an approximation renders a nearly optimal decision rule among the family of all possible tests. By exploiting the structure of the least favorable distribution, they also develop a finite-dimensional convex programming reformulation of the safe approximation.

We now turn our attention to the connection between DRO and regularization in statistical learning. Pflug et al. [303], Pichler [305], Wozabal [412] draw the connection between robustification and regularization, where as in Theorem 14, the shape of the transportation cost in the definition of the optimal transport discrepancy directly implies the type of regularization, and (ii) the size of the ambiguity set dictates the regularization parameter. Pichler [305] studies worst-case values of lower semicontinuous and law-invariant risk measures, including spectral and distortion risk measures, over an ambiguity set of distributions formed via the p-Wasserstein metric utilizing an arbitrary norm around the empirical distribution. They show that when the function $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ is linear in $\boldsymbol{\xi}$, the worst-case value is the sum of the risk of $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ under the nominal distribution and a regularization term. Pflug et al. [303] and Wozabal [412] show the worst-case value of a convex law-invariant risk measure over an ambiguity set of distributions, formed via the p-Wasserstein metric utilizing ℓ_p-norm around the empirical distribution, reduces to the sum of the nominal risk and a regularization term whenever the function $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ is affine in $\boldsymbol{\xi}$. They provide closed-form expressions for risk measures such as expectation, sum of expectation and standard deviation, CVaR, distortion risk measure, Wang transform, proportional hazards transform, the Gini measure, and sum of expectation and mean absolute deviation from the median. Important parts of the derivation of results in Pflug et al. [303], Pichler [305], Wozabal [412] are Kusuoka's representation of risk measures (Kusuoka [225], Shapiro [365]) and Fenchel–Moreau theorem (Rockafellar [336], Ruszczyński and Shapiro [348]).

In the context of statistical learning, the connection between DRO and regularization was first made in Shafieezadeh-Abadeh et al. [355], to the best of our knowledge. In fact, they study a distributionally robust logistic regression, where an ambiguity set of probability distributions, supported on an open set, is formed around the empirical distribution of data and via the 1-Wasserstein metric utilizing an arbitrary norm. They show that the resulting problem admits an equivalent reformulation as a tractable convex program. As stated in Theorem 14, this problem can be interpreted as a standard regularized logistic regression, where the size of the ambiguity set dictates the regularization parameter. They further propose a distributionally robust approach based on Wasserstein metric to compute upper and lower confidence bounds on the misclassification probability of the resulting classifier, based on the optimal values of two LPs.

Shafieezadeh-Abadeh et al. [357] extend the work of Shafieezadeh-Abadeh et al. [355] and study distributionally robust supervised learning (regression and classification) models. They introduce a new generalization technique using ideas from DRO, whose ambiguity set contains all infinite-dimensional distributions in the Wasserstein neighborhood of the empirical distribution. They show that the classical robust and the distributionally robust learning models are equivalent if the data satisfies a dispersion condition (for regression) or a separability condition (for classification). By imposing bound on the decision (i.e., hypothesis) space, they improve the upper confidence bound on the out-of-sample performance proposed in Mohajerin Esfahani and Kuhn [266] and prove a generalization bound that does not rely on the complexity of the hypothesis space. This is unlike

the traditional generalization bounds that are derived by controlling the complexity of the hypothesis space, in terms of Vapnik–Chervonenkis (VC)-dimension, covering numbers, or Rademacher complexities (Bartlett and Mendelson [16], Shalev-Shwartz and Ben-David [358]), which are usually difficult to calculate and interpret in practice. They extend their results to the case that the unknown hypothesis is searched from the space of nonlinear functionals. Given a symmetric and positive definite kernel function, such a setting gives rise to a lifted DRO problem that searches for a linear hypothesis over a *reproducing kernel Hilbert space* (RKHS). We refer to Section 6.4 for more details.

Gao et al. [150] study DRO problems formed via the p-Wasserstein metric utilizing an arbitrary norm, around the empirical distribution. They identify a broad class of cost functions, for which such a DRO is asymptotically equivalent to a regularization problem with a gradient-norm penalty under the nominal distribution. For linear function class, this equivalence is exact and results in a new interpretation for discrete choice models, including multinomial logit, nested logit, and generalized extreme value choice models. They also obtain lower and upper bounds on the worst-case expected cost in terms of regularization.

Mohajerin Esfahani et al. [267] study a data-driven inverse optimization problem to learn the objective function of the decision maker, given the historical data on uncertain parameters and decisions. In an environment with imperfect information, they propose a DRO model formed via the p-Wasserstein metric utilizing an arbitrary norm to minimize the worst-case risk of the predicted error. Such a model can be interpreted as a regularization of the corresponding empirical risk minimization problem. They present exact (or safe approximation) tractable convex programming reformulation for different combinations of risk measures and error functions.

Blanchet and Kang [59] study group-square-root LASSO (group LASSO focuses on variable selection in settings where some predictive variables, if selected, must be chosen as a group). They model this problem as a DRO problem formed via the p-Wasserstein metric utilizing an arbitrary norm. A method for (semi-) supervised learning based on data-driven DRO via p-Wasserstein metric utilizing an arbitrary norm, is proposed in Blanchet and Kang [60]. This method enhances the generalization error by using the unlabeled data to restrict the support of the worst-case distribution in the resulting DRO. They select the level of robustness using cross-validation, and they discuss the nonparametric behavior of an optimal selection of the level of robustness.

Chen and Paschalidis [86] study a DRO approach to linear regression using an ℓ_1-norm cost function, where the ambiguity set of distributions is formed via p-Wasserstein metric utilizing an arbitrary norm. They show that this DRO formulation can be relaxed to a convex optimization problem. By selecting proper norm spaces for the Wasserstein metric, they are able to recover several commonly used regularized regression models. They establish performance guarantees on both the out-of-sample behavior (prediction bias) and the discrepancy between the estimated and true regression planes (estimation bias), which elucidate the role of the regularizer. We refer to Staib and Jegelka [383] for distributionally robust deep learning for adversarial training, Derman and Mannor [116] for distributionally robust reinforcement learning, and Staib and Jegelka [384] for distributionally robust kernel methods.

6.1.1.7 Multistage Setting

The single- and two-stage stochastic programs in Pflug and Pichler [300] are extended in Analui and Pflug [5] to the multistage case, i.e., (10), where the reference data and information structure is represented as a tree. In these papers it is assumed that the tree structure and scenario values are fixed, while the probabilities are changing only in an ambiguous neighborhood of the reference model by utilizing the multistage *nested distance*, formed via the Wasserstein metric. Built upon the above results, Glanzer et al. [155] show that a scenario tree can be constructed out of data such that it converges (in terms of the nested distance) to the true model in probability at an exponential rate. Duque and Morton [131] study a stochastic dual dynamic programming (SDDP) approach to solve a multistage DRO model formed via the Wasserstein metric.

6.1.2 ϕ-Divergences

Another popular way to model the distributional ambiguity is to use *ϕ-divergences*, a class of measures used in information theory. A ϕ-divergence measures the discrepancy between two probability measures $P_1, P_2 \in \mathfrak{M}(\Xi, \mathcal{F})$ as $\mathfrak{d}^{\phi}(P_1, P_2) := \int_{\Xi} \phi\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}\right) \mathrm{d}P_2$, where the ϕ-divergence function $\phi : \mathbb{R}_+ \to \mathbb{R}_+ \cup \{+\infty\}$ is convex, and satisfy the following properties: $\phi(1) = 0$[18], $0\phi\left(\frac{0}{0}\right) := 0$, and $a\phi\left(\frac{a}{0}\right) := a \lim_{t\to\infty} \frac{\phi(t)}{t}$ if $a > 0$. Note that a ϕ-divergence

[18] The assumption $\phi(1) = 0$ is without loss of generality because the function $\psi(t) = \phi(t) + c(t-1)$ yields identical discrepancy measure to ϕ Pardo [293].

Table 3 Examples of ϕ-divergence functions, their conjugates $\phi^*(a)$, and their DRO counterparts (see Ben-Tal et al. [32])

Divergence	$\phi(t)$	$\phi(t),\ t \geq 0$	$\mathfrak{d}^\phi(P_1, P_2)$	$\phi^*(a)$	DRO Counterpart
Kullback–Leibler	$\phi_{\rm kl}(t)$	$t\log t - t + 1$	$\int_\Xi \log\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}\right)\mathrm{d}P_1$	$e^a - 1$	Convex program
Burg entropy	$\phi_{\rm b}(t)$	$-\log t + t - 1$	$\int_\Xi \log\left(\frac{\mathrm{d}P_2}{\mathrm{d}P_1}\right)\mathrm{d}P_2$	$-\log(1-a),\ a<1$	Convex program
J-divergence	$\phi_{\rm j}(t)$	$(t-1)\log t$	$\int_\Xi \log\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}\right)(\mathrm{d}P_1 - \mathrm{d}P_2)$	No closed form	Convex program
χ^2-distance	$\phi_{\rm c}(t)$	$\frac{1}{t}(t-1)^2$	$\int_\Xi \frac{(\mathrm{d}P_1 - \mathrm{d}P_2)^2}{\mathrm{d}P_1}$	$2 - 2\sqrt{1-a},\ a<1$	SOCP
Modified χ^2-distance	$\phi_{\rm mc}(t)$	$(t-1)^2$	$\int_\Xi \frac{(\mathrm{d}P_1 - \mathrm{d}P_2)^2}{\mathrm{d}P_2}$	$\begin{cases} -1 & a < -2 \\ a + \frac{a^2}{4} & a \geq -2 \end{cases}$	SOCP
Hellinger distance	$\phi_{\rm h}(t)$	$(\sqrt{t}-1)^2$	$\int_\Xi (\sqrt{\mathrm{d}P_1} - \sqrt{\mathrm{d}P_2})^2$	$\frac{a}{1-a},\ a<1$	SOCP
χ-divergence of order $\theta > 1$	$\phi_{\rm ca^\theta}(t)$	$\lvert t-1\rvert^\theta$	$\int_\Xi \lvert 1 - \frac{\mathrm{d}P_1}{\mathrm{d}P_2}\rvert^\theta \mathrm{d}P_2$	$a + (\theta-1)\left(\frac{\lvert a\rvert}{\theta}\right)^{\frac{\theta}{\theta-1}}$	SOCP
Variation distance	$\phi_{\rm v}(t)$	$\lvert t-1\rvert$	$\int_\Xi \lvert \mathrm{d}P_1 - \mathrm{d}P_2\rvert$	$\begin{cases} -1 & a \leq -1 \\ a & -1 \leq a \leq 1 \end{cases}$	LP
Cressie–Read	$\phi_{\rm cr^\theta}(t)$	$\frac{1-\theta+\theta t - t^\theta}{\theta(1-\theta)}$	$\frac{1}{\theta(1-\theta)}\left(1 - \int_\Xi \mathrm{d}P_1^\theta \mathrm{d}P_2^{1-\theta}\right)$	$\frac{1}{\theta}\left(1 - a(1-\theta)\right)^{\frac{\theta}{1-\theta}} - \frac{1}{\theta},\ a < \frac{1}{1-\theta}$	SOCP

does not necessarily induce a metric on the underlying space. For detailed information on ϕ-divergences, we refer to Pardo [293], Read and Cressie [332], Vajda [394].

A ϕ-divergence can be used to model the distributional ambiguity as follows:

$$\mathcal{P}^\phi(P_0;\epsilon) := \left\{P \in \mathfrak{M}\left(\Xi, \mathcal{F}\right) \,\middle|\, \mathfrak{d}^\phi(P, P_0) \leq \epsilon\right\}, \tag{40}$$

where as before P_0 is a nominal probability measure and ϵ controls the size of the ambiguity set. Table 3 presents a list of commonly used ϕ-divergence functions in DRO and their conjugate functions ϕ^*.

Before we review the papers that model the distributional ambiguity via the ϕ-divergences, we present a duality result on $\sup_{P\in\mathcal{P}^\phi(P_0;\epsilon)} \mathbb{E}_P\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right]$ on a general probability space.

▶ **Theorem 24** (Shapiro [367], Ahmadi-Javid [3, Theorem 3], Ahmadi-Javid [4, Theorem 5.1])**.** *Suppose that $\epsilon > 0$ in (40). Then, for a fixed $\boldsymbol{x} \in \mathcal{X}$, we have*

$$\sup_{P\in\mathcal{P}^\phi(P_0;\epsilon)} \mathbb{E}_P\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right] = \inf_{(\lambda,\mu)\in\Lambda_{\phi,h_0(\boldsymbol{x},\cdot)}} \left\{\mu + \lambda\epsilon + \int_\Xi (\lambda\phi)^*(h_0(\boldsymbol{x},s) - \mu)P_0(\mathrm{d}s)\right\},$$

where $\Lambda_{\phi,h_0(\boldsymbol{x},\cdot)} := \left\{(\lambda,\mu) \,\middle|\, \lambda \geq 0, h_0(\boldsymbol{x},s) - \mu - \lambda\lim_{t\to\infty}\frac{\phi(t)}{t} \leq 0, \forall\ s \in \Xi\right\}$, with the interpretation that $(\lambda\phi)^*(a) = \lambda\phi^*(\frac{a}{\lambda})$ for $\lambda \geq 0$. Here, $(0\phi)^*(a) = 0\phi^*(\frac{a}{0})$, which equals to 0 if $a \leq 0$ and $+\infty$ if $a > 0$.

Note that in the context of DRO, ϕ-divergences are mostly used to model distributional ambiguity for discrete distributions. Hence, the integrations in this section may be interpreted as summations. The above result is first obtained in Ahmadi-Javid [3, 4] for an essentially bounded measurable function $h_0(\boldsymbol{x},\cdot)$ on a general probability space. An extension to a function $h_0(\boldsymbol{x},\cdot)$ with a finite p-th order moment is derived in Shapiro [367]. Theorem 24 can be obtained by taking the Lagrangian dual of $\sup_{P\in\mathcal{P}^\phi(P_0;\epsilon)} \mathbb{E}_P\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right]$, as used in Shapiro [367], or using Donsker–Varadhan variation formula, as done in Ahmadi-Javid [3, 4]. We refer the readers to Bayraksan and Love [17], Ben-Tal et al. [32], Love and Bayraksan [255] for a detailed derivation in the discrete setting.

Based on the duality result in Theorem 24, Ahmadi-Javid [3, 4] introduce a class of law invariant coherent risk measure, referred to as *ϕ-entropic* risk measure. CVaR belong to this class. Moreover, the risk measure corresponding to Kullback–Leibler diveregnce is called *entropic value-at-risk*, that provides an upper bound on CVaR.

The robust counterpart of linear and nonlinear optimization problems with an uncertainty set of parameters defined via general ϕ-divergences is derived in Ben-Tal et al. [32]. As it is presented in Table 3, when the uncertain parameter is a finite-dimensional probability vector, the robust counterpart is tractable for most choices of ϕ-divergence function considered in the literature. The use of ϕ-divergence to model the distributional ambiguity in DRO is systematically introduced in Bayraksan and Love [17] and Love and Bayraksan [256]. To elucidate the use of ϕ-divergences for models with different sources of data and decision makers with different risk preferences, they present a classification of ϕ-divergences based on the notions of *suppressing* and *popping* a scenario. The situation that a scenario with a positive nominal probability ends up having a zero worst-case probability is called suppressing. On the contrary, the situation that a scenario with a zero nominal probability ends up having a positive worst-case probability is called popping. These notions give rise to four categories of ϕ-divergences.

For example, they show that the variation distance can both suppress and pop scenarios, while Kullback–Leibler divergence can only suppress scenarios. Furthermore, they propose a decomposition algorithm to solve the dual of the resulting DRO model formed via a general ϕ-divergence.

Motivated by the difficulty in choosing the ambiguity set and the fact that all probability distributions in the set are treated equally (while those outside the set are completely ignored), Ben-Tal et al. [31] propose to minimize the expected cost under the nominal distribution while the maximum expected cost over an infinite nested family of ambiguity sets, parameterized by ϵ, is bounded from above. More specifically, they allow a varying level of feasibility for each family of probability distributions, where the maximum allowed expected cost for distributions in a set with parameter ϵ is proportional to ϵ. They refer to this approach as *soft robust optimization* and relate the feasibility region induced by this approach to convex risk measures. They illustrate that the ambiguity sets formed via ϕ-divergences are related to an optimized certainty equivalent risk measure formed via ϕ-functions Ben-Tal and Teboulle [27]. Furthermore, they show that the complexity of the soft robust approach is equivalent to that of solving a small number of standard corresponding DRO (i.e., DRO with one ambiguity set) problems. In fact, by showing that standard DRO is concave in ϵ, they solve the soft robust model by a bisection method. They also investigate how much larger a feasible region implied by the soft robust approach can cover compared to the standard DRO, without compromising the objective value.

6.1.2.1 Worst-Case Distribution

For a fixed $\boldsymbol{x} \in \mathcal{X}$, $\sup_{P \in \mathcal{P}^\phi(P_0;\epsilon)} \mathbb{E}_P [h_0(\boldsymbol{x}, \boldsymbol{\xi})]$ is reformulated into a convex optimization problem in (λ, μ) by using Theorem 24. One can obtain an optimal worst-case probability distribution at $\boldsymbol{x}$ using an optimal solution (λ^*, μ^*) to the dual reformulation and by exploiting the KKT conditions.

▶ **Theorem 25** (Love and Bayraksan [256, Property 4])**.** *Suppose that Ξ is a finite sample space with M atoms: $\Xi = \{s_1, \ldots, s_M\}$, and let $\mathbb{P}_0 = \sum_{k=1}^M q_k \delta_k$, where δ_k is the Dirac point mass on s_k, $k \in [M]$. For a fixed $x \in \mathcal{X}$, let (λ^*, μ^*) be an optimal dual solution to the dual reformulation in Theorem 24. An optimal worst-case probability distribution $\mathbb{P}^* = \sum_{k=1}^M p_k^* \delta_k$ to $\sup_{\mathbb{P} \in \mathcal{P}^\phi(\mathbb{P}_0;\epsilon)} \mathbb{E}_{\mathbb{P}} [h_0(\boldsymbol{x}, \boldsymbol{\xi})]$ can be calculated with the equations*

$$\frac{p_k^*}{q_k} \in \partial\phi^* \left(\frac{h_0(\boldsymbol{x}, s_k) - \mu^*}{\lambda^*} \right), \quad \sum_{k \in [M]} q_k \phi \left(\frac{p_k^*}{q_k} \right) = \epsilon, \quad \sum_{k \in [M]} q_k = 1,$$

when $\lambda^* > 0$ and $q_k > 0$. With $\lambda^* > 0$ and $q_k = 0$, we have $p_k^* \in q_k \partial\phi^* \left(\frac{h_0(\boldsymbol{x}, s_k) - \mu^*}{\lambda^*} \right)$ (i.e., $p_k^* = 0$) when $\frac{h_0(\boldsymbol{x}, s_k) - \mu^*}{\lambda^*} < \lim_{t \to \infty} \frac{\phi(t)}{t}$, and otherwise the last two equations above must be used. With $\lambda^* = 0$, we have $p_k^* = 0$ when $h_0(\boldsymbol{x}, s_k) - \mu^* < 0$, otherwise the equations $\sum_{k \in [M]} q_k \phi \left(\frac{p_k^*}{q_k} \right) \leq \epsilon$, $\sum_{k \in [M]} q_k = 1$ must be used.

6.1.2.2 Risk and Chance Constraints

A data-driven DRO approach to chance-constrained problems modeled via ϕ-divergences is studied in Yanıkoğlu and den Hertog [431]. They propose safe approximations to these ambiguous chance constraints. Their approach is capable of handling joint chance constraints, dependent uncertain parameters, and a general nonlinear function $h_j(\boldsymbol{x}, \boldsymbol{\xi})$, $j \in [m]$. Hu et al. [202] and Jiang and Guan [213] show that distributionally robust chance-constrained programs formed via ϕ-divergences can be transformed into a chance-constrained problem under the nominal distribution but with an adjusted risk level. For a general ϕ-divergence, a bisection line search algorithm to obtain the perturbed risk level is proposed in Hu et al. [202], Jiang and Guan [213]. In addition, closed-form expressions for the adjusted risk level are obtained for the case of the variation distance (see, Hu et al. [202] and Jiang and Guan [213]), and Kullback–Leibler divergence and χ^2-distance (see, Jiang and Guan [213]). For the ambiguous probabilistic programs formed via ϕ-divergences, similar results to the chance-constrained programs are shown in Hu et al. [202]. Hu et al. [202] show that the ambiguous probability minimization problem can be transformed into a corresponding problem under the nominal distribution. In particular, they show that these problems have the same complexity as the corresponding pure probabilistic programs.

6.1.2.3 Statistical Learning

Hu et al. [200] study distributionally robust supervised learning, where the ambiguity set of distributions is formed via ϕ-divergences. They prove that such a DRO model for a classification problem gives a classifier that is optimal for the training set distribution rather than being robust against all distributions in the ambiguity

set. They argue such a pessimism comes from two sources: the particular losses used in classification and the over-conservation of the ambiguity set formed via ϕ-divergences. Motivated by this observation, they propose an ambiguity set that incorporates prior expert structural information on the distribution. More precisely, they introduce a latent variable from a prior distribution. While such a distribution can change in the ambiguity set, they leave the ambiguous joint distribution of data conditioned on the latent variable intact. Duchi et al. [123] show that the inner problem of a data-driven DRO formed around the empirical distribution, with $\epsilon = \frac{\chi^2_{1,1-\alpha}}{N}$ has an almost-sure asymptotic expansion. Such an expansion is equivalent to the expected cost under the empirical distribution plus a regularization term that accounts for the standard deviation of the objective function. They also show that the set of the optimal solutions of the DRO model converges to that of the stochastic program under the true underlying distribution, provided that $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ is lower-semicontinuous. As mentioned in Section 3.2, similar results to Duchi et al. [123] are obtained in Dupuis et al. [130], Lam [228, 229] in the context of robust sensitivity analysis of stochastic systems.

6.1.2.4 Specific ϕ-Divergences

In this section, we review papers that consider specific ϕ-divergences.

Kullback–Leibler Divergence Calafiore [72] investigates the optimal robust portfolio and worst-case distribution for a data-driven distributionally robust portfolio optimization problem with a mean-risk objective. Motivated by the application, they consider the variance and absolute deviation as measures of risk. Hu and Hong [201] study a variety of distributionally robust optimization problems, where the ambiguity is in either the objective function or constraints. They show that the ambiguous chance-constrained problem can be reformulated as a chance-constrained problem under the nominal distribution but with an adjusted risk level. They further show that when the chance safe region is bi-affine in $\boldsymbol{x}$ and $\boldsymbol{\xi}$, and the nominal distribution belongs to the exponential families of distributions, then both the nominal and worst-case distribution belong to the same distribution family.

Blanchet et al. [67] study a DRO approach to extreme value analysis in order to estimate the tail distributions and consequently, extreme quantiles. They form the ambiguity set of distributions by the class of Rényi divergences Pardo [293], that includes Kullback–Leibler as a special case[19]. Kullback–Leibler is also used for the DRO approach to hypothesis testing in Gül [167], Gül and Zoubir [168], Levy [240]. Guo et al. [169] study the impacts of the variation of the ambiguity set of probability distributions on the optimal value and optimal solution of the stochastic programs with distributionally robust chance constraints. To establish the results, they present conditions under which a sequence of approximated ambiguity sets converges to the true ambiguity set, for some discrepancy measures, including Kolmogorov and the total variation distance. They apply their convergence results to the ambiguity sets formed via Kullback–Leibler divergence.

Burg Entropy Wang et al. [407] model the distributional ambiguity via the Burg entropy to consider all probability distributions that make the observed data achieve a certain level of likelihood. They present statistical analyses of their model using Bayesian statistics and empirical likelihood theory.

Wiesemann et al. [409] study Markov decision processes where the transition Kernel is known. They use Burg entropy to construct a confidence region that contains the unknown probability distribution with a high probability, based on an observation history. It is shown in Lam [230] that a DRO model formed via the Burg entropy around the empirical distribution of data gives rise to a confidence bound on the expected cost that recovers the exact asymptotic statistical guarantees provided by the Central Limit Theorem.

χ^2-Distance Hanasusanto and Kuhn [173] propose a robust data-driven dynamic programming approach which replaces the expectations in the dynamic programming recursions with worst-case expectations over an ambiguity set of distributions. Their motivation to propose such a scheme is to mitigate the poor out-of-sample performance of the data-driven dynamic programming approach under sparse training data. The proposed method combines convex parametric function approximation methods (to model the dependence on the endogenous state) with nonparametric kernel regression method (to model the dependence on the exogenous state). They show the conditions under which the resulting DRO model, formed via χ^2-distance, reduces to a tractable conic program.

---

[19] The class of Rényi divergences is defined as $\mathfrak{d}_r^{\mathrm{R}}(P_1, P_2) := \frac{1}{1-r} \int_{\Xi} \left(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}\right)^{r-1} \mathrm{d}P_1$. This class is not a ϕ-divergence, but $\mathfrak{d}_r^{\mathrm{R}}(P_1, P_2)$ can be rewritten as $h(\mathfrak{d}^{\phi}(P_1, P_2))$, where $h(t) = \frac{1}{r-1}\log[(r-1)t+1]$ and $\phi(t) = \frac{t^r - r(t-1) - 1}{r-1}$ Pardo [293].

Klabjan et al. [221] study optimal inventory control for a single-item multiperiod periodic review stochastic lot-sizing problem under uncertain demand, where the distributional ambiguity is modeled via χ^2-distance. They show that the resulting model generalizes the Bayesian model, and it can be interpreted as minimizing demand-history-dependent risk measures.

Modified χ^2-Distance A SDDP algorithm to solve a distributionally robust multistage optimization model formed via the modified χ^2-distance is proposed in Philpott et al. [304].

Variation Distance Variation distance, or ℓ_1-norm, as defined in Table 3, can be used to safely approximate several ambiguity sets formed via ϕ-divergences, including χ-divergence of order 2, J-divergence, Kullback–Leibler divergence, and Hellinger distance. The following lemma states the above result more formally.

▶ **Lemma 26.** *The following relationship holds between ϕ-divergences, as defined in Table 3:*

$$\frac{1}{4}\left(\mathfrak{d}^{\phi_v}(P,P_0)\right)^2 \le \mathfrak{d}^{\phi_h}(P,P_0) \le \mathfrak{d}^{\phi_{kl}}(P,P_0) \le \mathfrak{d}^{\phi_j}(P,P_0) \le \mathfrak{d}^{\phi_{ca^2}}(P,P_0), \tag{41}$$

which implies

$$\mathcal{P}^{\phi_{ca^2}}(P_0;\epsilon) \subseteq \mathcal{P}^{\phi_j}(P_0;\epsilon) \subseteq \mathcal{P}^{\phi_{kl}}(P_0;\epsilon) \subseteq \mathcal{P}^{\phi_h}(P_0;\epsilon) \subseteq \mathcal{P}^{\phi_v}(P_0;2\epsilon^{\frac{1}{2}}). \tag{42}$$

Proof. See Appendix A. ◀

6.1.3 Total Variation Distance

For two probability measures P_1, $P_2 \in \mathfrak{M}(\Xi,\mathcal{F})$, the total variation distance is defined as

$$d_{\mathrm{TV}}(P_1,P_2) := \sup_{A\in\mathcal{F}} |P_1(A) - P_2(A)|.$$

When P_1 and P_2 are absolutely continuous with respect to a measure $\nu \in \mathfrak{M}(\Xi,\mathcal{F})$, with Radon–Nikodym derivatives f_1 and f_2, respectively, then, $\mathfrak{d}^{\mathrm{TV}}(P_1,P_2) = \frac{1}{2}\int_\Xi |f_1(s) - f_2(s)|\nu(\mathrm{d}s)$. Note that the total variation distance can be obtained from other classes of probability metrics: (1) it is a ϕ-divergence with $\phi(t) = \frac{1}{2}|t-1|$, (2) it is half of the ℓ_1-norm, and (3) it is obtained from the optimal transport discrepancy (31) with

$$c(s_1,s_2) = \begin{cases} 0, & \text{if} s_1 = s_2, \\ 1, & \text{if} s_1 \neq s_2. \end{cases} \tag{43}$$

The total variation distance can be used to model the distributional ambiguity as follows:

$$\mathcal{P}^{\mathrm{TV}}(P_0;\epsilon) := \left\{P \in \mathfrak{M}(\Xi,\mathcal{F}) \,\middle|\, \mathfrak{d}^{\mathrm{TV}}(P,P_0) \le \epsilon\right\}. \tag{44}$$

The total variation distance between P_1 and P_2 is also related to the *one-sided* variation distances $\frac{1}{2}\int_\Xi (f_1(s) - f_2(s))_+\nu(\mathrm{d}s)$ and $\frac{1}{2}\int_\Xi (f_2(s) - f_1(s))_+\nu(\mathrm{d}s)$ Rahimian et al. [327], which correspond ϕ-divergences with $\phi(t) = \frac{1}{2}(t-1)_+$ and $\phi(t) = \frac{1}{2}(1-t)_+$, respectively. However, unlike the total variation distance, the one-sided variation distances are not a probability metric.

Before we review the papers that model the distributional ambiguity via the total variation distance, we present a duality result on $\sup_{P\in\mathcal{P}^{\mathrm{TV}}(P_0;\epsilon)} \mathbb{E}_P[h_0(\boldsymbol{x},\boldsymbol{\xi})]$.

▶ **Theorem 27** (Jiang and Guan [214, Theorems 1–2], Rahimian et al. [327, Proposition 3], Shapiro [367])**.** *For a fixed $\boldsymbol{x} \in \mathcal{X}$, we have*

$$\sup_{P\in\mathcal{P}^{TV}(P_0;\epsilon)} \mathbb{E}_P[h_0(\boldsymbol{x},\boldsymbol{\xi})] = \begin{cases} \mathbb{E}_{P_0}[h_0(\boldsymbol{x},\boldsymbol{\xi})], & \epsilon = 0, \\ \epsilon\nu\text{-ess}\sup_{s\in\Xi} h_0(\boldsymbol{x},\boldsymbol{\xi}(s)) + (1-\epsilon)\mathrm{CVaR}_\epsilon^{P_0}[h_0(\boldsymbol{x},\boldsymbol{\xi})], & 0<\epsilon<1, \\ \nu\text{-ess}\sup_{s\in\Xi} h_0(\boldsymbol{x},\boldsymbol{\xi}(s)), & \epsilon \ge 1, \end{cases}$$

where $\nu\text{-ess}\sup_{s\in\Xi} h_0(\boldsymbol{x},\boldsymbol{\xi}(s)) = \inf\left\{a\in\mathbb{R} : \nu\{s\in\Xi : h_0(\boldsymbol{x},\boldsymbol{\xi}(s)) > a) = 0\}\right\}$.

▶ Remark 28 (Rahimian et al. [327, Proposition 3], Shapiro [367]). Let $\mathcal{P}^{\mathrm{OTV}}(P_0;\epsilon)$ denote the ambiguity set formed via either of the one-sided variation distances. Then, for a fixed $\boldsymbol{x} \in \mathcal{X}$, $\sup_{P\in\mathcal{P}^{\mathrm{TVO}}(P_0;\frac{\epsilon}{2})}$ can be obtained by the right-hand side of the result in Theorem 27.

It is possible to obtain a worst-case probability distribution at $x \in \mathcal{X}$, using the dual formulation in Theorem 27.

▶ **Theorem 29** (Rahimian et al. [327, Proposition 4])**.** *Suppose that Ξ is a finite sample space with M atoms: $\Xi = \{s_1, \ldots, s_M\}$, and let $\mathbb{P}_0 = \sum_{k=1}^{M} q_k \delta_k$, where δ_k is the Dirac point mass on s_k, $k \in [M]$. For a fixed $x \in \mathcal{X}$, let (λ^*, μ^*) be optimal dual variables as follows:*

$$\lambda^* = \sup_{k\in[M]} h(\boldsymbol{x},\boldsymbol{\xi}(s_k)) - \mathrm{VaR}_\epsilon\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right],\ \mu^* = \frac{1}{2}\left(\sup_{k\in[M]} h_0(\boldsymbol{x},\boldsymbol{\xi}(s_k)) + \mathrm{VaR}_\epsilon\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right]\right).$$

An optimal worst-case probability distribution $\mathbb{P}^* = \sum_{k=1}^{M} p_k^* \delta_k$ to $\sup_{\mathbb{P}\in\mathcal{P}^{TV}(\mathbb{P}_0;\epsilon)} \mathbb{E}_{\mathbb{P}}\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right]$ is calculated with

$$\begin{cases} p_k^* = 0, & h_0(\boldsymbol{x},\boldsymbol{\xi}(s_k)) < \mathrm{VaR}_\epsilon\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right], \\ p_k^* \le q_k, & h_0(\boldsymbol{x},\boldsymbol{\xi}(s_k)) = \mathrm{VaR}_\epsilon\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right], \\ p_k^* = q_k, & \mathrm{VaR}_\epsilon\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right] < h_0(\boldsymbol{x},\boldsymbol{\xi}(s_k)) < \sup_{k\in[M]} h(\boldsymbol{x},\boldsymbol{\xi}(s_k)), \\ p_k^* \ge q_k, & h_0(\boldsymbol{x},\boldsymbol{\xi}(s_k)) = \sup_{k\in[M]} h(\boldsymbol{x},\boldsymbol{\xi}(s_k)), \end{cases}$$

coupled with constraints

$$\sum_{k\in[k:h_0(\boldsymbol{x},\boldsymbol{\xi}(s_k))=\mathrm{VaR}_\epsilon[h_0(\boldsymbol{x},\boldsymbol{\xi})]]} p_k^* = \sum_{k\in[k:h_0(\boldsymbol{x},\boldsymbol{\xi}(s_k))\le\mathrm{VaR}_\epsilon[h_0(\boldsymbol{x},\boldsymbol{\xi})]]} q_k - \epsilon,$$

$$\sum_{k\in[k:h_0(\boldsymbol{x},\boldsymbol{\xi}(s_k))=\sup_{k\in[M]} h(\boldsymbol{x},\boldsymbol{\xi}(s_k))]} p_k^* = \epsilon + \sum_{k\in[k:h_0(\boldsymbol{x},\boldsymbol{\xi}(s_k))=\sup_{k\in[M]} h(\boldsymbol{x},\boldsymbol{\xi}(s_k))]} q_k,$$

in addition to $\sum_{k\in[M]} p_k^* = 1$ and $p_k^* \ge 0$, $k \in [M]$, when $\lambda^* > 0$. When $\lambda^* = 0$, the conditions can be written as$p_k^* \ge 0$ if $h_0(\boldsymbol{x},\boldsymbol{\xi}(s_k)) = \sup_{k\in[M]} h(\boldsymbol{x},\boldsymbol{\xi}(s_k))$, and $p_k^* = 0$ otherwise, in addition to $\sum_{k\in[M]} p_k^* = 1$ and $\frac{1}{2}\sum_{k\in[M]} |p_k^* - q_k| \le \epsilon$.

Jiang and Guan [214] study a two-stage stochastic program with $h_0(\boldsymbol{x},\boldsymbol{\xi})$ defined as (3), formed via the total variation distance. They discuss how to find the nominal probability distribution and analyze the convergence of the problem to the corresponding stochastic program under the true unknown probability distribution. Rahimian et al. [327] study distributionally robust convex optimization problems with a finite sample space. They study how the uncertain parameters affect the optimization. In order to do so, they define the notion of "effective" and "ineffective" scenarios. According to their definitions, a subset of scenarios is effective if their removal from the support of the worst-case distribution, by forcing their probabilities to zero in the ambiguity set, changes the optimal value of the DRO problem. They propose easy-to-check conditions to identify effective and ineffective scenarios for the case that the distributional ambiguity is modeled via the total variation distance. Rahimian et al. [329] extend the work of Rahimian et al. [327] to the multistage setting, where they define the notions of effectiveness of scenario paths and the conditional effectiveness of realizations along a scenario path for a general class of multistage DRO problems. They propose easy-to-check conditions to identify the effectiveness of scenario paths in the multistage setting when the distributional ambiguity is modeled via the total variation distance. Rahimian et al. [328] extends Rahimian et al. [327] to distributionally robust newsvendor problems with a continuous sample space. They derive a closed-form expression for the optimal solution and identify the maximal effective subsets of demands.

6.1.4 Goodness-of-Fit Test

Various statistical hypothesis tests have been used to construct an ambiguity set of distributions. Bertsimas et al. [51] and Bertsimas et al. [52] propose a systematic view on how to choose statistical goodness-of-fit test to construct an ambiguity set of distributions that guarantee the implication (C1) (recall Theorem 12). For a null hypothesis $H_0 : P = P_0$ that makes a claim about an unknown probability distribution P^{true}, a set of data $\mathcal{D}_N$ with N data points, a significance level α, a test statistics $T(\mathcal{D}_N, P_0)$, and a critical value $\Gamma(\alpha, \mathcal{D}_N, P_0)$,

Bertsimas et al. [51] propose an ambiguity set of probability distributions, constructed as the $(1-\alpha)$ confidence region, as follows

$$\mathcal{P}^{\mathrm{GoF}}(P_0;\epsilon) := \{P \in \mathfrak{M}(\Xi,\mathcal{F}) \,|\, T(\mathcal{D}_N, P_0) \leq \Gamma(\alpha, \mathcal{D}_N, P_0)\}. \tag{45}$$

For instance, for two univariate probability measures $P_1, P_2 \in \mathfrak{M}(\Xi,\mathcal{F})$, Kolmogorov–Smirnov distance is defined as

$$\mathfrak{d}^{\mathrm{KS}}(P_1,P_2) := \sup_t |P_1\{(-\infty,t]\} - P_2\{(-\infty,t]\}|.$$

The Kolmogorov–Smirnov distance can be generalized to multivariate random vectors and be used to model the distributional ambiguity as follows:

$$\mathcal{P}^{\mathrm{p}}(P_0;\epsilon) := \left\{P \in \mathfrak{M}(\Xi,\mathcal{F}) \,\middle|\, \mathfrak{d}^{\mathrm{KS}}(P,P_0) \leq \epsilon\right\}. \tag{46}$$

A reformulation of the resulting DRO model (8) formed via (46) is given in Luo and Mehrotra [260] for the case that the random vector have a discrete and continuous supports. Postek et al. [311] review and derive computationally tractable reformulations of distributionally robust risk constraints over discrete probability distributions for various risk measures and ambiguity sets formed using statistical goodness-of-fit tests or probability metrics, including ϕ-divergences, Kolmogorov–Smirnov, Wasserstein, Anderson–Darling, Cramer-von Mises, Watson, and Kuiper. For each pair of risk measure and ambiguity set, they obtain a tractable reformulation by relying on the conjugate duality for the risk measure and the support function of the ambiguity set. Bertsimas et al. [51] and Bertsimas et al. [52] consider the situation that (i) $\mathbb{P}^{\mathrm{true}} = P^{\mathrm{true}} \circ \boldsymbol{\xi}^{-1}$ may have continuous support, and the components of $\boldsymbol{\xi}$ are independent, (ii) $\mathbb{P}^{\mathrm{true}}$ may have continuous support, and data are drawn from its marginal distributions asynchronously, and (iii) $\mathbb{P}^{\mathrm{true}}$ may have continuous support, and data are drawn from its joint distribution. They also study a wide range of statistical hypothesis tests, including χ^2, G, Kolmogorov–Smirnov, Kuiper, Cramer-von Mises, Watson, and Anderson–Darling goodness-of-fit tests, and they characterize the geometric shape of the corresponding ambiguity sets. For instance, they show that G, Kolmogorov–Smirnov, and Kuiper tests result in polyhedral sets, while χ^2, Cramer-von Mises, and Watson result in SOC sets.

6.1.5 Prohorov Metric

For two probability measures $P_1, P_2 \in \mathfrak{M}(\Xi,\mathcal{F})$, the Prohorov metric is defined as

$$\mathfrak{d}^{\mathrm{p}}(P_1,P_2) := \inf\{\gamma > 0 \,|\, P_1\{A\} \leq P_2\{A^\gamma\} + \gamma \mathrm{and} P_2\{A\} \leq P_1\{A^\gamma\} + \gamma \ \ \forall\, A \in \mathcal{F}\},$$

where $A^\gamma := \{s \in \Xi \,|\, \inf_{s' \in A} d(s,s') \leq \gamma\}$ Gibbs and Su [153]. The Prohorov metric takes values in $[0,1]$ and can be used to model the distributional ambiguity as follows:

$$\mathcal{P}^{\mathrm{p}}(P_0;\epsilon) := \{P \in \mathfrak{M}(\Xi,\mathcal{F}) \,|\, \mathfrak{d}^{\mathrm{p}}(P,P_0) \leq \epsilon\}. \tag{47}$$

A specialization of the Prohorov metric to the univariate distributions is called *Levy* metric, which is defined in Gibbs and Su [153] as

$$\mathfrak{d}^{\mathrm{L}}(P_1,P_2) := \inf\{\gamma > 0 \,|\, P_2\{(-\infty,t-\gamma]\} - \gamma \leq P_1\{(-\infty,t]\} \leq P_2\{(-\infty,t+\gamma]\} + \gamma,\ \forall\, t \in \mathbb{R}\}.$$

The Levy metric can be used to model the distributional ambiguity as follows:

$$\mathcal{P}^{\mathrm{L}}(P_0;\epsilon) := \left\{P \in \mathfrak{M}(\Xi,\mathcal{F}) \,\middle|\, \mathfrak{d}^{\mathrm{L}}(P,P_0) \leq \epsilon\right\}. \tag{48}$$

Erdoğan and Iyengar [138] study an optimization problem subject to a set of parameterized convex constraints. Similar to the argument in Section 3.1.2, they study a DRO approach to this problem, where the distributional ambiguity is modeled by the Prohorov metric. They also consider a scenario approximation scheme of the problem. By extending the work of Calafiore and Campi [73], Campi and Calafiore [75], they provide an upper bound on the number of samples required to guarantee that the sampled problem is a good approximation for the associated ambiguous chance-constrained problem with a high probability.

6.1.6 ℓ_p-Norm

Calafiore and El Ghaoui [74] study distributionally robust individual linear chance constrained problem, and provide convex conditions that guarantee the satisfaction of the chance constraint within the family of radially-symmetric nonincreasing densities whose supports are defined by means of the ℓ_1- and ℓ_∞-norm. Interestingly, they show that the worst-case distribution is attained by a uniform distribution on the respective support. To be more precise, consider the sets $\mathcal{H}(\boldsymbol{A},\boldsymbol{\xi}_0) := \{\boldsymbol{\xi} = \boldsymbol{\xi}_0 + \boldsymbol{A}\omega \,|\, \|\omega\|_\infty \le 1\}$ and $\mathcal{E}(\boldsymbol{B},\boldsymbol{\xi}_0) := \{\boldsymbol{\xi} = \boldsymbol{\xi}_0 + \boldsymbol{B}\omega \,|\, \|\omega\|_1 \le 1\}$, where $\boldsymbol{A}$ is a diagonal positive-definite matrix and $\boldsymbol{B}$ is a positive-definite matrix. A random vector $\boldsymbol{\xi}$ has a probability distribution $\mathbb{P}$ within the class of radially-symmetric nonincreasing densities supported on $\mathcal{H}(\boldsymbol{A},\boldsymbol{\xi}_0)$ (respectively, $\mathcal{E}(\boldsymbol{B},\boldsymbol{\xi}_0)$) if $\boldsymbol{\xi} - \mathbb{E}_{\mathbb{P}}[\boldsymbol{\xi}] = \boldsymbol{A}\omega$ (respectively, $\boldsymbol{\xi} - \mathbb{E}_{\mathbb{P}}[\boldsymbol{\xi}] = \boldsymbol{B}\omega$), where ω is a random vector having the probability density f_ω such that $f_\omega(\omega) = t(\|\omega\|_\infty)$ for $\|\omega\|_\infty \le 1$ and 0 otherwise (respectively, $f_\omega(\omega) = t(\|\omega\|_1)$ for $\|\omega\|_1 \le 1$ and 0 otherwise) and $t(\,\cdot\,)$ is a nonincreasing function. Calafiore and El Ghaoui [74] show that the worst-case chance within the class of radially-symmetric nonincreasing densities supported on $\mathcal{H}(\boldsymbol{A},\boldsymbol{\xi}_0)$ (respectively, $\mathcal{E}(\boldsymbol{B},\boldsymbol{\xi}_0)$) is attained at a uniform distribution supported on $\mathcal{H}(\boldsymbol{A},\boldsymbol{\xi}_0)$ (respectively, $\mathcal{E}(\boldsymbol{B},\boldsymbol{\xi}_0)$). The class of radially-symmetric distributions contains for example Gaussian, truncated Gaussian, uniform distribution on ellipsoidal support, and nonunimodal densities Calafiore and El Ghaoui [74].

Mevissen et al. [265] study distributionally robust polynomial optimization, where the distribution of the uncertain parameter is estimated using polynomial basis functions via the ℓ_p-norm. They show that the optimal value of the problem is the limit of a sequence of tractable SDP relaxations of polynomial optimization problems. They also provide a finite-sample consistency guarantee for the data-driven uncertainty sets, and an asymptotic guarantee on the solutions of the SDP relaxations.

Jiang and Guan [214] study two-stage stochastic program with $h_0(\boldsymbol{x},\boldsymbol{\xi})$ defined as (3), formed via ℓ_∞-norm. Huang et al. [203] extend the work of Jiang and Guan [214] to the multistage setting. They formulate the problem into a problem that contains a convex combination of expectation and CVaR in the objective function of each stage to remove the nested multistage minimax structure in the objective function. They analyze the convergence of the resulting DRO problem to the corresponding multistage stochastic program under the true unknown probability distribution.

6.1.7 ζ-Structure Metrics

Consider $P_1, P_2 \in \mathfrak{M}(\Xi,\mathcal{F})$ and let $\mathcal{Z}$ be a family of real-valued measurable functions $z : \left(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d)\right) \mapsto (\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. The ζ-structure metric (also referred to as *integral probability metric* Müller [269], Sriperumbudur et al. [382]) is defined as $\mathfrak{d}^{\mathcal{Z}}(P_1,P_2) := \sup_{z\in\mathcal{Z}} \left|\mathbb{E}_{P_1}[z(\boldsymbol{\xi})] - \mathbb{E}_{P_2}[z(\boldsymbol{\xi})]\right|$. A wide range of metrics in probability theory can be written as special cases of the above family of metrics (Pichler and Xu [307], Zhao and Guan [442]). Let us introduce them below.

- *Total variation metric* $\mathfrak{d}^{\mathrm{TV}}(P_1,P_2)$:

  $$\mathcal{Z} = \{z \,|\, \|z\|_\infty \le 1\},$$

  where $\|z\|_\infty = \sup_{\boldsymbol{\xi}\in\Omega} |z(\boldsymbol{\xi})|$.
- *Bounded Lipschitz metric* $\mathfrak{d}^{\mathrm{BL}}(P_1,P_2)$:

  $$\mathcal{Z} = \{z \,|\, \|z\|_\infty \le 1,\ z\text{is Lipschitz continuous},\ L_1(z) \le 1\},$$

  where $L_1(z) =: \sup\{|z(\boldsymbol{u}) - z(\boldsymbol{v})|/\|\boldsymbol{u}-\boldsymbol{v}\| \,|\, \boldsymbol{u}\neq\boldsymbol{v}\}$, is the Lipschitz modulus.
- *Kantorovich metric* $\mathfrak{d}^{\mathrm{K}}(P_1,P_2)$:

  $$\mathcal{Z} = \{z \,|\, z \text{ is Lipschitz continuous},\ L_1(z) \le 1\}.$$

- *Fortet–Mourier metric* $\mathfrak{d}^{\mathrm{FM}}(P_1,P_2)$:

  $$\mathcal{Z} = \{z \,|\, z \text{ is Lipschitz continuous},\ L_q(z) \le 1\},$$

  where

  $$L_q(z) =: \inf\left\{L \,\middle|\, |z(\boldsymbol{u}) - z(\boldsymbol{v})| \le L\cdot\|\boldsymbol{u}-\boldsymbol{v}\|\cdot\max(1,\|\boldsymbol{u}\|^{q-1},\|\boldsymbol{v}\|^{q-1}), \forall\, \boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^d\right\}.$$

  Note that when $q = 1$, Fortet–Mourier metric is the same as the Kantorovich metric.

- *Uniform (Kolmogorov) metric* $\mathfrak{d}^{\mathrm{U}}(P_1, P_2)$:

$$\mathcal{Z} = \left\{ z \,\middle|\, z = \mathbb{1}_{(-\infty,t]},\ t \in \mathbb{R}^n \right\}.$$

The following lemma, which is immediate from Zhao and Guan [442, Lemmas 1–4], establishes the relationship between ζ-structure metrics.

▶ **Lemma 30.** *Suppose that the support Ω of $\boldsymbol{\xi}$ is bounded with diameter θ, i.e., $\theta := \sup\{d(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) : \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \Omega\}$, where d is metric. Then, the following relationship holds between ζ-structure metrics:*

$$\mathfrak{d}^{BL}(P, P_0) \le \mathfrak{d}^{K}(P, P_0)$$
$$\mathfrak{d}^{K}(P, P_0) \le \theta \mathfrak{d}^{TV}(P, P_0)$$
$$\mathfrak{d}^{U}(P, P_0) \le \mathfrak{d}^{TV}(P, P_0)$$
$$\mathfrak{d}^{K}(P, P_0) \le \mathfrak{d}^{FM}(P, P_0)$$
$$\mathfrak{d}^{FM}(P, P_0) \le \max\{1, \theta^{q-1}\}\mathfrak{d}^{K}(P, P_0).$$

The class of ζ-structure metrics may be used to model the distributional ambiguity as follows:

$$\mathcal{P}^{\mathcal{Z}}(P_0; \epsilon) := \left\{ P \in \mathfrak{M}(\Xi, \mathcal{F}) \,\middle|\, \mathfrak{d}^{\mathcal{Z}}(P, P_0) \le \epsilon \right\}. \tag{50}$$

Zhao and Guan [442] study distributionally robust two-stage stochastic programs with recourse. They discuss how to construct the ambiguity set from historical data while utilizing a family of ζ-structure metrics. They propose solution approaches to solve the resulting problem, where the true unknown distribution is discrete or continuous. They further analyze the convergence of the DRO problem to the corresponding stochastic program under the true unknown probability distribution. Pichler and Xu [307] investigate how the variation of the ambiguity set would affect the optimal value and the optimal solution in the DRO problem formed via ζ-structure metric. They illustrate their results in the context of a two-stage stochastic program with recourse.

6.1.8 Contamination Neighborhood

Contamination around P_0, a nominal measure, is defined as

$$\mathcal{P}^{\mathrm{c}}(P_0; \epsilon) = \{P \in \mathfrak{M}(\Xi, \mathcal{F}) \,|\, P = (1-\epsilon)P_0 + \epsilon Q,\ Q \in \mathfrak{Q}\}, \tag{51}$$

where $\mathfrak{Q} \subseteq \mathfrak{M}(\Xi, \mathcal{F})$ and $\epsilon \in [0, 1]$.

This ambiguity set is extensively used in the context of robust statistics, see, e.g., Huber [205], Huber and Ronchetti [206], and it has also been used in the economics literature, see, e.g., Nishimura and Ozaki [286, 287]. Bose and Daripa [70] study ambiguity aversion in a mechanism design problem using a maximin expected utility model of Gilboa and Schmeidler [154]. The contamination neighborhood is also used in the context of statistical learning, see, e.g., Duchi et al. [122], and hypothesis testing, see, e.g., Huber [204].

6.2 Moment-Based Ambiguity Sets

A common approach to model the ambiguity set is moment based, in which the ambiguity set contains all probability distributions whose moments satisfy certain properties. We categorize this type of models into several subgroups, although there are some overlaps. Specifically, in this section, we review different moment-based ambiguity sets that are used in the literature. These include (i) *Chebyshev*, (ii) *ellipsoid and matrix inequality*, (iii) *generalized moment and measure inequalities*, (iv) *moment matrix inequalities*, (v) *cross-moment or nested moment*, (vi) *marginals (Fréchet)*, and (vii) *mixture distribution*.

6.2.1 Chebyshev

In this section, we review papers that model the distributional ambiguity by considering the first- and second-order moments information.

Scarf [351] models the distributional ambiguity in a single-product newsvendor problem, where only the mean and variance of the random demand is known. He obtains a closed-form expression for the optimal order quantity and shows that the worst-case probability distribution is supported on only two points. Motivated by

Table 4 Contributions for Chebyshev-based ambiguity sets

	Contributions	Assumptions
Scarf [351]	Closed-form expression for the optimal order quantity for a univariate newsvendor model, worst-case distribution is supported on only two points	Known mean and variance
El Ghaoui et al. [135]	SOCP/SDP reformulation for worst-case VaR in a one-period portfolio optimization	Mean and covariance matrix in convex polytopic sets
Lotfi and Zenios [254]	SOCP reformulation for worst-case VaR/CVaR in a one-period portfolio optimization	Mean and covariance matrix in an ellipsoid set
Goldfarb and Iyengar [161]	SOCP reformulation for various portfolio optimization problems under a linear factor model	Covariance matrix in an ellipsoid set

the Scarf's seminal work, other researchers have investigated the Chebyshev ambiguity set in the context of the newsvendor model. Gallego and Moon [147] study multiple extensions of the problem studied in Scarf [351]. These include the situations where there is a recourse opportunity, a fixed ordering cost, a random production output, and a scare resource for multiple competing products. Grünwald and Dawid [166] confine the ambiguity set to distributions with fixed first order moments $\boldsymbol{\tau}$. By varying $\boldsymbol{\tau}$, they obtain a collection of maximum generalized entropy distribution and relate it to the exponential family of distributions.

Unlike the ambiguity sets studied in Scarf [351] and Gallego and Moon [147], the mean and covariance matrix may be unknown themselves and belong to some uncertainty sets, for example:

$$\mathcal{P}^C := \left\{ P \in \mathfrak{M}(\Xi, \mathcal{F}) \;\middle|\; \begin{array}{l} \left\| \mathbb{E}_P[\boldsymbol{\xi}] - \boldsymbol{\mu}_0 \right\|_2 \leq \varrho_1, \\ \mathbb{E}_P\left[(\boldsymbol{\xi} - \boldsymbol{\mu}_0)(\boldsymbol{\xi} - \boldsymbol{\mu}_0)^\top \right] \preccurlyeq \varrho_2 \boldsymbol{\Sigma}_0 \end{array} \right\}. \tag{52}$$

Some researchers have studied this setup and obtain SDP or SOCP equivalent or approximate reformulations. We review some of these papers in this section, and provide a summary of main contributions in Table 4.

6.2.1.1 Worst-Case Distribution

Several papers has studied obtaining a worst-case distribution for DRO models with Chebyshev ambiguity sets. In his seminal paper, Scarf [351] consider a single-product newsvendor problem to decide on the optimal order quantity x under an uncertain nonnegative demand ξ, with a know mean μ and variance σ^2. The key idea in the analysis is to show that at a fixed x, the cost function $h_0(x, \xi) := r \min\{x, \xi\} - cx$, with c as the unit purchasing cost and r as the unit revenue, is supported from below by a quadratic function in ξ, with equality happens at only two points a and b. They show that there is a minimizing two-point distribution, supported on a and b, that attains the known mean μ and variance σ^2. This idea is further used for other variants of a single-item newsvendor problem, with different functions $h_0(x, \xi)$, in Gallego and Moon [147], Han et al. [172], Xie and Ahmed [415], Yu et al. [433]. In fact, it is known that moment problems with n moment constraints can be solved by optimizing only over a convex combination of $n + 1$ Dirac measures (i.e., discrete distributions with at most $n + 1$ support points) Smith [380]. In the spirit of this result, Popescu [310] show that for any univariate function $h_0(\xi) := h_0(\cdot, \xi)$, $\min_{\xi \sim (\mu, \sigma^2)} \mathbb{E}[h_0(\xi)]$ can be calculated as

$$\min p_a h_0(a) + p_b h_0(b) + p_c h_0(c) \quad \text{s.t. } a \leq \mu - \frac{\sigma^2}{c - \mu} \leq b \leq \mu + \frac{\sigma^2}{\mu - a} \leq c,$$

where

$$p_a = \begin{cases} \frac{\sigma^2 + (\mu - b)(\mu - c)}{(a - b)(a - c)} & \text{if } a < b, \\ 1 - \frac{\sigma^2 + (\mu - a)^2}{(c - a)^2} & \text{if } a = b, \end{cases}$$

$$p_c = \begin{cases} \frac{\sigma^2 + (\mu - a)(\mu - b)}{(c - a)(c - b)} & \text{if } b < c, \\ 1 - \frac{\sigma^2 + (\mu - c)^2}{(c - a)^2} & \text{if } b = c, \end{cases}$$

$p_b = 1 - p_a - p_c$.

Popescu [310] further extends the results in Scarf [351] to the class of univariate functions $h_0(\xi)$ that satisfy the two-point support property (i.e., supported from below by a quadratic function, with equality happens at two points a and b such that for any (μ, σ^2), a feasible distribution with supports a and b exists). If $h_0(\xi)$ satisfies the two-point support property, $\min_{\xi\sim(\mu,\sigma^2)} \mathbb{E}\left[h_0(\xi)\right]$ can be calculated as min $ph_0\Big(\mu + \sqrt{\frac{1-p}{p}}\sigma\Big) + (1-p)h_0\Big(\mu - \sqrt{\frac{1-p}{p}}\sigma\Big)$. Special cases of functions with a two-point support property are the cost function considered in Gallego and Moon [147], Scarf [351], functions with a decreasing and concave-convex derivative, and concave functions with a concave-convex derivative.

6.2.1.2 Risk and Chance Constraints

El Ghaoui et al. [135] study a distributionally robust one-period portfolio optimization, where the worst-case VaR of loss over an ambiguity set of distributions with a known mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} \succcurlyeq \mathbf{0}$ is minimized. They show that this problem can be reformulated as a SOCP. Moreover, El Ghaoui et al. [135] show that minimizing worst-case VaR with respect to such an ambiguity set can be interpreted as a RO model where the worst-case portfolio loss with respect to an ellipsoid uncertainty set is minimized. To be precise, let $h_0(\boldsymbol{x}, \boldsymbol{\xi}) := -\boldsymbol{x}^\top \boldsymbol{\xi}$, where $\boldsymbol{x}$ is the vector of investment and $\boldsymbol{\xi}$ the random vector of returns. Let $\text{WVaR}_\beta(\boldsymbol{x}) := \inf\{\gamma : \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\{h_0(\boldsymbol{x}, \boldsymbol{\xi}) \geq \gamma\} \leq \beta\}$ be the worst-case VaR of loss at level β, over a set of probability distributions $\mathbb{P}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} \succcurlyeq \mathbf{0}$. El Ghaoui et al. [135] show that $\text{WVaR}_\beta(\boldsymbol{x})$ is equivalent to $\kappa(\beta)\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2 - \boldsymbol{\mu}^\top \boldsymbol{x}$, where $\kappa(\beta) := \sqrt{\frac{1-\beta}{\beta}}$ (see also Bertsimas and Popescu [42, Proposition 6.3] and Calafiore and El Ghaoui [74, Theorem 3.1]). Recall that if $\boldsymbol{\xi}$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} \succcurlyeq \mathbf{0}$, then $\text{VaR}_\beta(\boldsymbol{x})$ is equivalent to $\kappa(\beta)\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2 - \boldsymbol{\mu}^\top \boldsymbol{x}$, where $\kappa(\beta) := -\phi^{-1}(\beta)$ with $\phi^{-1}(\beta)$ as the $(1-\beta)$-quantile of the standard multivariate normal distribution. If $\boldsymbol{\xi}$ does not follow a multivariate normal distribution, then, we can calculate an upper bound on $\text{WVaR}_\beta(\boldsymbol{x})$ as $\kappa(\beta)\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2 - \boldsymbol{\mu}^\top \boldsymbol{x}$, where $\kappa(\beta) := \frac{1}{\sqrt{\beta}}$. Hence, the importance of the results in El Ghaoui et al. [135] is that an exact reformulation for $\text{WVaR}_\beta(\boldsymbol{x})$ can be obtained by just replacing $\kappa(\beta)$ in the multivariate normal case by $\sqrt{\frac{1-\beta}{\beta}}$.

El Ghaoui et al. [135] extend their study to the case that the first two moments are only known to belong to a convex (bounded) uncertainty set, and they show the conditions under which the resulting model can be cast as a SDP. In particular, for independent polytopic uncertainty sets for the mean and covariance (so that the mean and covariance belong to the Cartesian product of these two sets), the problem can be reformulated as a SOCP. Also, for sets with componentwise bound on the mean and covariance, they cast the problem as a SDP (see also Halldórsson and Tütüncü [171] for a similar result). Moreover, they show that in the presence of additional information on the distribution, besides the first two moments, including constraints on the support and Kullback–Leibler divergence, an upper bound on the worst-case VaR can be obtained by solving a SDP. Motivated by the work in El Ghaoui et al. [135], Li [242] showcases the results in the context of a risk-averse portfolio optimization problem. Unlike El Ghaoui et al. [135] that considers polytopic and interval uncertainty sets for the mean and covariance, Lotfi and Zenios [254] assume that the unknown mean and covariance belong to an ellipsoidal uncertainty set. They study the worst-case VaR and worst-case CVaR optimization problems, subject to an expected ambiguous return constraint. They derive an interesting (yet surprising) result that both problems yield the same SOCP reformulation. It is worth noting that the problem studied in Lotfi and Zenios [254] is generally more difficult than the ones analyzed in El Ghaoui et al. [135] as they consider an ambiguous constraints in addition to an ambiguous objective function. El Ghaoui et al. [135] study a similar linear factor model as the one in Goldfarb and Iyengar [161], but without normality assumption and they assume that the uncertainty in the mean is not independent of the uncertainty in the covariance matrix of the returns. When the factor matrix $\boldsymbol{A}$ belongs to ellipsoidal uncertainty set, they show that an upper bound on the worst-case VaR can be computed by solving a SDP.

Zymler et al. [449] extend the work of El Ghaoui et al. [135] with known first and second order moments to a portfolio of derivatives, and develop two worst-case VaR models to capture the nonlinear dependencies between the derivative returns and the underlying asset returns. They introduce worst-case polyhedral VaR with convex piecewise-linear relationship between the derivative return and the asset returns. They also show that minimizing worst-case polyhedral VaR is equivalent to a convex SOCP. A worst-case quadratic VaR with (possibly nonconvex) quadratic relationships between the derivative return and the asset returns is also introduced, and

they show that minimizing worst-case quadratic VaR is equivalent to a convex SDP. These worst-case VaR measures are equivalent to the worst-case CVaR of the underlying polyhedral or quadratic loss function, and they are coherent. As in El Ghaoui et al. [135], Zymler et al. [449] show that optimization of these new worst-case VaR has a RO interpretation over an uncertainty set, asymmetrically oriented around the mean values of the asset returns. Using the result from Zymler et al. [448], Rujeerapaiboon et al. [344] show that the worst-case VaR of the quadratic approximation of a portfolio growth rate can be expressed as the optimal value of a SDP.

Chen et al. [88] summarize and develop different approximations to the individual chance constraint used in the robust optimization as the consequence of applying different bounds on CVaR. These bounds, in turn, can be written as an optimization problem over an uncertainty set. For instance, they show that when the uncertainties are characterized only by their means and covariance, the corresponding uncertainty set is an ellipsoid. Calafiore and El Ghaoui [74] provide explicit results for enforcement of the individual chance constraint over an ambiguity set of distributions. When only the information on the mean and covariance are considered, the worst-case chance constraint is equivalent to a convex second-order conic (SOC) constraint. With additional information on the symmetry, the worst-case chance constraint can be safely approximated via a convex SOC constraint. Additionally, when the means are known and individual elements are known to belong with probability one to independent bounded intervals, the worst-case chance constraint can be safely approximated via a convex SOC constraint.

Zymler et al. [448] study a safe approximation to distributionally robust individual and joint chance constraints based on the worst-case CVaR. Under the assumptions that the ambiguity set is formed via distributions with fixed mean and covariance, and the chance safe regions are bi-affine in $\boldsymbol{x}$ and $\boldsymbol{\xi}$, they obtain an exact SDP reformulation of the worst-case CVaR. They show that the CVaR approximation is in fact exact for individual chance constraints whose constraint functions are either convex or (possibly nonconconvex) quadratic in $\boldsymbol{\xi}$ by relying on nonlinear Farkas lemma and $\mathcal{S}$-lemma, see, e.g., Pólik and Terlaky [308].

Chen et al. [88] extend their idea to the joint chance constraint by using bounds for order statistics. They show that the resulting approximation for the joint chance constraint outperforms the Bonferroni approximation, and the constraints of the approximation are second-order conic-representable. Zymler et al. [448] show that the CVaR approximation is exact for joint chance constraints whose constraint functions depend linearly on $\boldsymbol{\xi}$.

Motivated by the fact that chance constraints do not take into account the magnitude of the violation, Xu et al. [424] study a probabilistic envelope constraint. This approach can be interpreted as a continuum of chance constraints with nondecreasing target values and probabilities. They show that when the first two order moments are known, an ambiguous probabilistic envelope constraint is equivalent to a deterministic SIP, which is called as the *globalized robust optimization*, originally referred as the *comprehensive robust optimization*, problem (Ben-Tal et al. [31, 29]). In other words, ambiguous probabilistic envelope constraint alleviates the "all-or-nothing" view of the standard RO that ignores realizations outside of the uncertainty set. We refer to Yang and Xu [429] for an extension of the work in Xu et al. [424] to the nonlinear inequalities.

Bertsimas et al. [48] study a risk-averse two-stage stochastic LP, i.e., $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ is defined as (3) with $q_1 = 0$. They assume that the mean and the covariance matrix are known, and a convex nondecreasing piecewise linear disutility function is used to model risk. When the second-stage objective function's coefficients $\boldsymbol{q}(\boldsymbol{\xi})$ are random, they obtain a tight polynomial-sized SDP formulation. They also provide an explicit construction for a sequence of (worst-case) distributions that asymptotically attain the optimal value. They prove that this problem is NP-hard when the right-hand side is random, and further show that under the special case that the extreme points of the dual of the second-stage problem are explicitly known, the problem admits a SDP reformulation.

Li and Kwon [243] study a distributionally robust approach for a single-period portfolio selection problem. They consider a set of reference means and variances, and they form the ambiguity set by all distributions whose means and variances are in a pre-specified distance from the reference means and variances set (in the regular sense of a point from a set via a norm). For the case that moments take values outside the reference region, since evaluation based on its worst-case performance can be overly conservative, they consider a penalty term that further accounts for measure discrepancy between the moments in and outside the reference region. Moreover, for the case that the reference region is a conic set, they obtain an equivalent SDP reformulation.

6.2.1.3 Discrete Problems

Under the assumption that the mean and covariance are unknown but belongs to a nonempty, closed, convex set, Natarajan et al. [277] investigate the worst-case expected value of a random function $Z(\boldsymbol{\xi})$, defined as the maximum of a linear function of nonnegative random variables, i.e., $Z(\boldsymbol{\xi}) = \max\left\{\boldsymbol{\xi}^\top \boldsymbol{x} \,\middle|\, \boldsymbol{x} \in \mathcal{X}\right\}$ with $\mathcal{X}$ is

specified as a bounded feasible region to a mixed-binary LP. They show that this problem can be reformulated as a completely positive program, i.e., an optimization problem over the convex cone of completely positive matrices. A relaxation to this problem can be obtained by using the cone of doubly nonnegative matrices, i.e., both positive semidefinite and nonnegative. When the mean and covariance matrix are known, Natarajan and Teo [274] investigate a similar problem to the one in Natarajan et al. [277] where set $\mathcal{X}$ is specified with either a finite number of points or a bounded feasible region to a mixed-integer LP. They reformulate this problem as a SDP, whose complexity is related to characterizing the convex hull of the quadratic forms of the points in the feasible region (Natarajan and Teo [274, Theorem 2]).

Xie and Ahmed [415] study a DRO approach to a two-stage stochastic program with a simple integer round-up recourse function, defined as follows:

$$\min_{\boldsymbol{x}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \max_{P \in \mathcal{P}} \mathbb{E}_P \left[h_0(\boldsymbol{x}, \boldsymbol{\xi}) \right] \,\middle|\, \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b},\ \boldsymbol{x} \in \mathbb{R}^n \right\},$$

where

$$h_0(\boldsymbol{x}, \boldsymbol{\xi}) = \min_{\boldsymbol{u}, \boldsymbol{v}} \left\{ \boldsymbol{q}^\top \boldsymbol{u} + \boldsymbol{r}^\top \boldsymbol{v} \,\middle|\, \boldsymbol{u} \geq \boldsymbol{\xi} - \boldsymbol{T}\boldsymbol{x},\ \boldsymbol{v} \geq \boldsymbol{T}\boldsymbol{x} - \boldsymbol{\xi},\ \boldsymbol{u}, \boldsymbol{v} \in \mathbb{Z}^q_+ \right\}.$$

The ambiguity set is formed by the product of one-dimensional ambiguity sets for each component of the random parameter $\boldsymbol{\xi}$, formed with marginal distributions with known support and mean. They obtain a closed-form expression for the inner problem corresponding to each component, and they reformulate the problem as a mixed-integer SOCP.

6.2.1.4 Statistical Learning

Lanckriet et al. [234] present a DRO approach to a binary classification problem to minimize the worst-case probability of missclassification where the mean and covariance matrix of each class are known. They show that for a linear hypothesis, the problem can be formulated as a SOCP. They also investigate the case where the mean and covariance are unknown and belong to convex uncertainty sets. They show that when the mean is unknown and belongs to an ellipsoid, the problem is a SOCP. On the other hand, when the mean is known and covariance belongs to a matrix norm ball, the problem is a SOCP and adopts a regularization term. For a nonlinear hypothesis, they seek a kernal function to map into a higher-dimensional covariates-response space such that a linear hypothesis in that space corresponds to a nonlinear hypothesis in the original covariate-response space. Using this idea, the model is reformulated as an SOCP.

Fathony et al. [141] study a distributonally robust approach to graphical models for leveraging the graphical structure among the variables. The proposed model in Fathony et al. [141] seeks a predictor to make a probabilistic prediction $\widehat{P}(\widehat{y}|\boldsymbol{u})$ over all possible label assignments so that it minimizes the worst-case conditional expectation of the prediction loss $l(\widehat{y}, \bar{y})$ with respect to $\bar{P}(\bar{y}|\boldsymbol{u})$ as follows:

$$\min_{\hat{P}(\hat{y}|\boldsymbol{u})} \max_{\bar{P}(\bar{y}|\boldsymbol{u})} \mathbb{E}_{\boldsymbol{U} \sim \breve{P}\hat{Y}|\boldsymbol{U} \sim \hat{P}\bar{Y}|\boldsymbol{U} \sim \bar{P}} \left[l(\widehat{Y}, \bar{Y}) \right] \quad \text{s.t. } \mathbb{E}_{\boldsymbol{U} \sim \breve{P}\bar{Y}|\boldsymbol{U} \sim \bar{P}} \left[\Phi(\boldsymbol{U}, Y) \right] = \breve{\Phi},$$

where $\Phi(\boldsymbol{U}, Y)$ is a given feature function and $\breve{\Phi} = \mathbb{E}_{(\boldsymbol{U},Y) \sim \breve{P}} \left[\Phi(\boldsymbol{U}, Y) \right]$. The worst-case in the above formulation is taken with respect to all conditional distributions of the predictor, conditioned on the covariates. This conditional distribution $\bar{P}(\bar{y}|\boldsymbol{u})$ is such that the first-order moment of the feature function $\Phi(\boldsymbol{U}, Y)$ matches the first-order moment under the empirical joint distribution of the covariates and labels, $\breve{P}$. Fathony et al. [141] show that the DRO approach enjoys the consistency guarantees of probabilistic graphical models, see, e.g., Lafferty et al. [226], and has the advantage of incorporating customized loss metrics during the training as in large margin models, see, e.g., Tsochantaridis et al. [390].

6.2.1.5 Multistage Setting

Xin and Goldberg [418] study a multistage distributionally robust newvendor problem of the form (10), where the support and the first two order moments of the demand distribution are known at each stage. They provide a formal definition of the time consistency of the optimal policies and study this phenomena in the context of the newsvendor problem. They further relate time consistency to rectangularity of measures, see, e.g., Shapiro [366], and provide sufficient conditions for time consistency. Unlike Xin and Goldberg [418] that suppose the demand process is stage-wise independent, Xin and Goldberg [419] assume that the demand process is a martingale.

They form the ambiguity set by all distributions with a known support and mean at each stage. They obtain the optimal policy and a two-point worst-case probability distribution, one of which is zero, in closed forms. They also show that for any initial inventory level, the optimal policy and random demand (distributed according to the worst-case distribution) is such that for all stages, either demand is greater than or equal to the inventory or demand is zero, meaning that all future demands are also zero. Yang [427] and Van Parys et al. [397] study a stochastic optimal control model to minimize the worst-case probability that a system remains in a safe region for all stages. Yang [427] forms the ambiguity set at each stage by all distributions for which the componentwise mean of random parameters is within an interval, while the covariance is in a positive semidefinite cone. Van Parys et al. [397] form the ambiguity set by all distributions with a known mean and covariance.

6.2.2 Ellipsoid and Matrix Inequality

Unlike the ambiguity sets studied in Scarf [351] and Gallego and Moon [147], Delage and Ye [105] allow the mean and covariance matrix to be unknown themselves and unify the studies mentioned in Section 6.2.1. This ambiguity set is defined as follows (Delage and Ye [105]):

$$\mathcal{P}^{DY} := \left\{ P \in \mathfrak{M}(\Xi, \mathcal{F}) \;\middle|\; \begin{array}{l} P\{\boldsymbol{\xi} \in \Omega\} = 1, \\ \left(\mathbb{E}_P[\boldsymbol{\xi}] - \boldsymbol{\mu}_0\right)^\top \boldsymbol{\Sigma}_0^{-1} \left(\mathbb{E}_P[\boldsymbol{\xi}] - \boldsymbol{\mu}_0\right) \leq \varrho_1, \\ \mathbb{E}_P\left[(\boldsymbol{\xi} - \boldsymbol{\mu}_0)(\boldsymbol{\xi} - \boldsymbol{\mu}_0)^\top\right] \preccurlyeq \varrho_2 \boldsymbol{\Sigma}_0 \end{array} \right\}. \tag{53}$$

The first constraint denotes the smallest closed convex set $\Omega \subseteq \mathbb{R}^d$ that contains $\boldsymbol{\xi}$ with probability one (w.p. 1), i.e., Ω is the support of $\mathbb{P} = P \circ \boldsymbol{\xi}^{-1}$ w.p.1. The second constraint ensures that the mean of $\boldsymbol{\xi}$ lies in an ellipsoid of size ϱ_1 and centered around the nominal mean estimate $\boldsymbol{\mu}_0$. Note that we can equivalently write this constraint as

$$\mathbb{E}_P\left[\begin{pmatrix} -\boldsymbol{\Sigma}_0 & \boldsymbol{\mu}_0 - \boldsymbol{\xi} \\ (\boldsymbol{\mu}_0 - \boldsymbol{\xi})^\top & -\varrho_1 \end{pmatrix}\right] \preccurlyeq \mathbf{0}.$$

The third constraint defines the second central-moment matrix of $\boldsymbol{\xi}$ by a matrix inequality. The parameters ϱ_1 and ϱ_2 control the level of confidence in $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$, respectively. Note that the ambiguity sets with a known mean and covariance matrix can be seen as a special case of (53), with $\varrho_1 = 0$ and $\varrho_2 = 1$. Delage and Ye [105] propose data-driven methods to form confidence regions for the mean and the covariance matrix of the random vector $\boldsymbol{\xi}$ using the concentration inequalities of McDiarmid [261], and provide probabilistic guarantees that the solution found using the resulting DRO model yields an upper bound on the out-of-sample performance with respect to the true distribution of the random vector $\boldsymbol{\xi}$. A conic generalization of the ambiguity set $\mathcal{P}^{\text{DY}}$, beyond the first and second moment information is also studied in Delage [102]. Below, we present a duality result for $\sup_{P \in \mathcal{P}^{\text{DY}}} \mathbb{E}_P[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$ given a fixed $\boldsymbol{x} \in \mathcal{X}$, due to Delage and Ye [105].

▶ **Theorem 31** (Delage and Ye [105, Lemma 1]). *For a fixed $\boldsymbol{x} \in \mathcal{X}$, suppose that Slater's constraint qualification conditions are satisfied, i.e., there exists a strictly feasible P to $\mathcal{P}^{DY}$, and $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ is P-integrable for all $P \in \mathcal{P}^{DY}$. Then, $\sup_{P \in \mathcal{P}^{DY}} \mathbb{E}_P[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$ is equal to the optimal value of the following semi-infinite convex conic optimization problem:*

$$\inf_{\boldsymbol{Y}, \boldsymbol{y}, r, t} r + t \quad s.t. \quad \begin{cases} r \geq h_0(\boldsymbol{x}, \boldsymbol{\xi}) - \boldsymbol{\xi}^\top \boldsymbol{Y} \boldsymbol{\xi} - \boldsymbol{\xi}^\top \boldsymbol{y}, \ \forall\, \boldsymbol{\xi} \in \Omega, \\ t \geq (\varrho_2 \boldsymbol{\Sigma}_0 + \boldsymbol{\mu}_0 \boldsymbol{\mu}_0^\top) \bullet \boldsymbol{Y} + \boldsymbol{\mu}_0^\top \boldsymbol{y} + \sqrt{\varrho_1} \|\boldsymbol{\Sigma}_0^{\frac{1}{2}} (\boldsymbol{y} + 2\boldsymbol{Y}\boldsymbol{\mu}_0)\|, \\ \boldsymbol{Y} \succcurlyeq 0, \end{cases}$$

where $\boldsymbol{Y} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{y} \in \mathbb{R}^d$.

The reformulated problem in Theorem 31 is polynomial-time solvable under the following assumptions Delage and Ye [105]:

- The sets $\mathcal{X}$ and Ω are convex and compact, and are both equipped with oracles that confirm the feasibility of a point $\boldsymbol{x}$ and $\boldsymbol{\xi}$, or provide a hyperplane that separates the infeasible point from its corresponding feasible set in time polynomial in the dimension of the set.
- Function $h_0(\boldsymbol{x}, \boldsymbol{\xi}) := \max_{j \in [J]} l_j(\boldsymbol{x}, \boldsymbol{\xi})$ is piecewise-linear and is such that for each j, $l_j(\boldsymbol{x}, \boldsymbol{\xi})$ is convex in $\boldsymbol{x}$ and concave in $\boldsymbol{\xi}$. In addition, for any given pair $(\boldsymbol{x}, \boldsymbol{\xi})$, one can evaluate $l_j(\boldsymbol{x}, \boldsymbol{\xi})$, find a supergradient of $l_j(\boldsymbol{x}, \boldsymbol{\xi})$ in $\boldsymbol{\xi}$, and find a subgradient of $l_j(\boldsymbol{x}, \boldsymbol{\xi})$ in $\boldsymbol{x}$, in time polynomial in the dimension of $\mathcal{X}$ and Ω.

As a special case when Ω is an ellipsoid, the resulting reformulation in Theorem 31 reduces to a SDP of finite size. Motivated by the computational challenges of solving a semidefinite reformulation of (8) formed via (53), Cheng et al. [99] propose an approximation method to reduce the dimensionality of the resulting DRO. This approximation method relies on the principal component analysis for the optimal lower dimensional representation of the variability in random samples. They show that this approximation yields a relaxation of the original problem and give theoretical bounds on the gap between the original problem and its approximation.

Rujeerapaiboon et al. [345] derive Chebyshev-type bounds on the worst-case right and left tail of a product of nonnegative symmetric random variables. They assume that the mean is known, but the covariance matrix might be known or bounded above by a matrix inequality. They show that if both the mean and covariance matrix are known, these bounds can be obtained by solving a SDP. For the case that the covariance matrix is bounded above, they show that (i) the bound on the left tail is equal to the bound on the left tail under the known covariance setting, and (ii) the bound on the right tail is equal to the bound on the right tail under the known mean and covariance setting, for a sufficiently large tail. They extend their results to construct Chebyshev bounds for sums, minima, and maxima of nonnegative random variables.

6.2.2.1 Risk and Chance Constraints

Risk-based DRO models formed via the ambiguity set (53) are studied in the literature. A distributionally robust approach to an individual chance constraint with binary decisions is studied in Zhang et al. [439]. They consider the following individual chance constraints with $h_j(\boldsymbol{x}, \boldsymbol{\xi})$, $j \in [m]$, in (9) is defined as $h_j(\boldsymbol{x}_j, \boldsymbol{\xi}_j) := \mathbb{1}_{[\boldsymbol{\xi}_j^\top \boldsymbol{x}_j \leq \boldsymbol{b}_j]}(\boldsymbol{\xi}_j)$, where $\boldsymbol{x}_j \in \{0,1\}^{n_j}$, $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]$, and $\boldsymbol{\xi} = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m]$. They form the ambiguity set of distributions by all joint distributions whose marginal means and covariances satisfy the constraints in (53). They reformulate the chance constraints as binary second-order conic (SOC) constraints. Li [242] obtains a closed-form expression to the worst-case of the class of law invariant coherent risk measures, where the worst case is taken with respect to all distributions with the same mean and covariance matrix.

6.2.3 Generalized Moment and Measure Inequalities

In this section we review an ambiguity set that allows to model the support of the random vector, and impose bounds on the probability measure as well as functions of the random vector as follow:

$$\mathcal{P}^{MM} := \left\{ P \in \mathfrak{M}_+(\Xi, \mathcal{F}) \,\middle|\, \nu_1 \preceq P \preceq \nu_2,\ \int_\Xi \boldsymbol{f} \mathrm{d}P \in [\boldsymbol{l}, \boldsymbol{u}] \right\}, \tag{54}$$

where $\nu_1, \nu_2 \in \mathfrak{M}_+(\Xi, \mathcal{F})$ are two given measures that impose lower and upper bounds on a measure $P \in \mathfrak{M}_+(\Xi, \mathcal{F})$, and $\boldsymbol{f} := [f_1, \ldots, f_m]$ is a vector of measurable functions on $(\Xi, \mathcal{F})$, with $m \geq 1$. The first constraint in (54) enforces a preference relationship between measures. To ensure that P is a probability measure, i.e., $P \in \mathfrak{M}(\Xi, \mathcal{F})$, we set $l_1 = u_1 = 1$ and $f_1 = 1$ in the above definition of $\mathcal{P}^{MM}$. Shapiro and Ahmed [369] propose this framework, and its special cases appear in Bansal and Mehrotra [13], Bertsimas and Popescu [42], Mehrotra and Papp [262], Perakis and Roels [297], Popescu [309], among others. Note that if the first constraint in (54) is disregarded (i.e., we only have $P \succeq 0$), then we can form the constraints of a classical problem of moments, see, e.g., Landau [235]. Using this unified set, one can impose bounds on the standard moments, by setting the ith entry of $\boldsymbol{f}$ to have the form: $f_i(\boldsymbol{\xi}) := (\xi_1)^{k_{i1}} \cdot (\xi_2)^{k_{i2}} \cdots (\xi_d)^{k_{id}}$, where k_{ij} is a nonnegative integer indicating the power of ξ_j for the ith moment function. Other possible choices for the functions $\boldsymbol{f}$ include the mean absolute deviation, the (co-)variances, semi-variance, higher order moments, and Huber loss function. Moreover, proper choices of $\boldsymbol{f}$ will give the flexibility to impose structural properties on the probability distribution, see, e.g., Popescu [309] and Perakis and Roels [297], to model the unimodality and symmetry of distributions within this framework (see also Section 6.3).

Below, we present a duality result $\sup_{P \in \mathcal{P}^{\text{MM}}} \mathbb{E}_P [h_0(\boldsymbol{x}, \boldsymbol{\xi})]$, given a fixed $\boldsymbol{x} \in \mathcal{X}$.

▶ **Theorem 32** (Shapiro and Ahmed [369, Proposition 2.1]). *For a fixed $\boldsymbol{x} \in \mathcal{X}$, suppose that $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ is ν_2-integrable, i.e., $\int_\Xi |h_0(\boldsymbol{x}, \boldsymbol{\xi})| \mathrm{d}\nu_2 < \infty$, as defined in (54). Moreover, suppose that $\boldsymbol{f}$ is ν_2-integrable, and there exists $\nu_1 \preceq P \preceq \nu_2$ such that $\int_\Xi \boldsymbol{f} \mathrm{d}P \in (\boldsymbol{l}, \boldsymbol{u})$. If $\sup_{P \in \mathcal{P}^{MM}} \mathbb{E}_P [h_0(\boldsymbol{x}, \boldsymbol{\xi})]$ is finite, then, it can be written as the optimal value of the following problem:*

$$\inf_{\boldsymbol{r}, \boldsymbol{t}} \boldsymbol{r}^\top \boldsymbol{u} - \boldsymbol{t}^\top \boldsymbol{l} + \Psi(\boldsymbol{r}, \boldsymbol{t}) \quad \textit{s.t. } \boldsymbol{r}, \boldsymbol{t} \geq \boldsymbol{0},$$

where

$$\Psi(\boldsymbol{r},\boldsymbol{t}) = \int_{\Xi} \Big(h_0(\boldsymbol{x},s) + (\boldsymbol{t}-\boldsymbol{r})^\top \boldsymbol{f}(s)\Big)_+ \nu_2(\mathrm{d}s) - \int_{\Xi} \Big(-h_0(\boldsymbol{x},s) - (\boldsymbol{t}-\boldsymbol{r})^\top \boldsymbol{f}(s)\Big)_+ \nu_1(\mathrm{d}s).$$

Shapiro and Ahmed [369] focus on a special case of (54), where the first constraint is written as $(1-\epsilon)P^* \preceq P \preceq (1+\epsilon)P^*$, for some reference measure P^*, and they identify the coherent risk measure corresponding to the studied DRO. They further study the class of problems with convex objective function $h_0(\,\cdot\,,\boldsymbol{\xi})$ and two-stage stochastic programs. Bertsimas and Popescu [42], Mehrotra and Papp [262], Popescu [309] study the classical problem of moments, i.e., ambiguity set is formed via only the second constraints in (54). When $\boldsymbol{f}$ are moment functions, Mehrotra and Papp [262] show that under mild conditions (continuous function h and compact support Ω), the optimal value of a sequence of problems of the form (8), where the ambiguity set is constructed via an increasing number of moments of the underlying probability distributions, with moments matched to those under a reference distribution, converges to the optimal value of a problem of the form (1) under the reference distribution. Moreover, using the SIP reformulation of (8), Mehrotra and Papp [262] propose a cutting surface method to solve a convex (8). This method can be applied to problems where bounds of moments are of arbitrary order, and possibly, bounds on nonpolynomial moments are available.

Chen et al. [89] consider a two-stage stochastic linear complementarity problem, where the underlying random data are continuously distributed. They study a distributionally robust approach to this problem, where the ambiguity set of distributions is formed via (54) without the first constraint, and propose a discretization scheme to solve the problem. They investigate the asymptotic behavior of the approximated solution in the number of discrete partitions of the sample space Ξ. As an application, they study robust game in a duoploy market where two players need to make strategic decisions on capacity for future production with anticipation of Nash–Cournot type competition after demand uncertainty is observed. There are studies that consider only lower order moments, up to order 2. Ardestani-Jaafari and Delage [6] study distributionally robust multi-item newsvendor problem, where the ambiguity set of distribution contains all distributions with a known budgeted support, mean, and partial first order moments. To provide a reformulation of the problem, they propose a conservative approximation scheme for maximizing the sum of piecewise linear functions over a polyhedral uncertainty set based on the relaxation of an associated mixed-integer LP. They show that for the above studied newsvendor problem such an approximation is exact and it is a LP.

Royset and Wets [343] study a DRO model with a decision-dependent ambiguity set, where the ambiguity set has the form of (54), without the second set of constraints, and the first constraint is formed via the decision-dependent cumulative distribution functions. They establish the convergence properties of the solutions to this problem by exploiting and refining results in variational analysis.

6.2.3.1 Discrete Problems

Bansal et al. [15] study a two-stage integer program, i.e., (8) with $h_0(\boldsymbol{x},\boldsymbol{\xi})$ defined as (3), with pure binary first-stage and mixed-binary second-stage variables on a finite set of scenarios. They propose a decomposition-based L-shaped algorithm and a cutting surface algorithm to solve the resulting model. They investigate the conditions and ambiguity set of distribution under which the proposed algorithm is finitely convergent. They show that ambiguity set of distributions formed via (54) without the first constraint, satisfy these conditions. Hanasusanto et al. [178] study a finite adaptability scheme to approximate the following two-stage distributionally robust program, with binary recourse decisions and optimized certainty equivalent as a risk measure:

$$\min_{\boldsymbol{x}} \max_{P\in\mathcal{P}} \Big\{ \boldsymbol{\xi}^\top \boldsymbol{C}\boldsymbol{x} + \mathcal{R}_P\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right] \,\Big|\, \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b},\ \boldsymbol{x} \in \{0,1\}^{q_1} \times \mathbb{R}^{n-q_1} \Big\},$$

where

$$h_0(\boldsymbol{x},\boldsymbol{\xi}) = \min_{\boldsymbol{y}} \left\{ \boldsymbol{q}^\top \boldsymbol{Q}\boldsymbol{y}(\boldsymbol{\xi}) \,\middle|\, \boldsymbol{W}\boldsymbol{y}(\boldsymbol{\xi}) \geq \boldsymbol{R}\boldsymbol{\xi} - \boldsymbol{T}\boldsymbol{x},\ \boldsymbol{y}(\boldsymbol{\xi}) \in \{0,1\}^{q_2} \right\},$$

and $\mathcal{R}_P\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right]$ is an optimized certainty equivalent risk measure corresponding to the utility function u: $\mathcal{R}_P\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right] = \inf_{\eta\in\mathbb{R}} \eta + \mathbb{E}_P\left[u\big(h_0(\boldsymbol{x},\boldsymbol{\xi}) - \eta\big)\right]$ (Ben-Tal and Teboulle [26, 27]). As an alternative to the affine recourse approximation, they pre-determine a set of finite recourse decisions here-and-now, and implement the best among them after the realization is observed. They form the ambiguity set of distributions as in (54) but without the first constraint, where the support is assumed to be a polytope and functions f_i are also convex piecewise linear in $\boldsymbol{\xi}$. They derive an equivalent mixed-integer LP for the resulting model. They also obtain

upper and lower bounds on the probability with which any of these recourse decisions is chosen under any ambiguous distribution as LPs. Postek et al. [313] study (8), with $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ defined as (3) with $q_1 > 0$. They model the distributional ambiguity by all distributions whose mean and mean-absolute deviation are known. While they show that the problem reduces to a two-stage stochastic program when there is no discrete variables, they develop a general approximation framework for the DRO problem with integer variables.

6.2.3.2 Risk and Chance Constraints

Bertsimas and Popescu [42] study the worst-case bound on the probability of a multivariate random vector falling outside a semialgebraic confidence region (i.e., a set described via polynomial inequalities) over an ambiguity set of the form (54), where functions $\boldsymbol{f}$ are represented by all polynomials of up to kth-order. For the univariate case, they obtain the result as a SDP. In particular, they obtain closed-form bounds, when $k \leq 3$. For the multivariate case, they show that such a bound can be obtained via a family of SDP relaxations, yielding a sequence of increasingly stronger, asymptotically exact upper bounds, each of which is calculated via a SDP. A special case of Bertsimas and Popescu [42] appears in Vandenberghe et al. [400], where the confidence region is described via linear and quadratic inequalities, and the first two order moments are assumed to be known within the ambiguity set.

Building from Chen et al. [89], Liu et al. [248] study a distributionally robust reward-risk ratio model, based on a variation of the Sharpe ratio. The ambiguity set contains all distributions whose componentwise means and covariances are restricted to intervals. They turn this problem into a model with a distributionally robust inequality constraint, and further reformulate this model as a nonconvex SIP. They approximate the semi-infinite constraint with an entropic risk measure approximation[20] and provide an iterative method to solve the resulting model. They provide statistical analysis to assess the likelihood of the true probability distribution lying in the ambiguity set, and provide a convergence analysis of the optimal value and solutions of the data-driven distributionally robust reward-risk ratio problems.

Natarajan et al. [278] study a distributionally robust approach to minimize the worst-case CVaR of regret in combinatorial optimization problems with uncertainty in the objective function coefficients, defined as follows:

$$\min_{\boldsymbol{x} \in \mathcal{X}} \mathrm{WCVaR}_{\alpha}^{P} \left[h_0(\boldsymbol{x}, \boldsymbol{\xi}) \right],$$

where $h_0(\boldsymbol{x}, \boldsymbol{\xi}) = -\boldsymbol{\xi}^{\top} \boldsymbol{x} + \max_{\boldsymbol{y} \in \{0,1\}^{q_1}} \boldsymbol{\xi}^{\top} \boldsymbol{y}$ and $\mathrm{WCVaR}_{\alpha}^{P} \left[h_0(\boldsymbol{x}, \boldsymbol{\xi}) \right] = \sup_{P \in \mathcal{P}} \mathrm{CVaR}_{\alpha}^{P} \left[h_0(\boldsymbol{x}, \boldsymbol{\xi}) \right]$. It is assumed that the ambiguity set is formed with the knowledge of marginal distributions, where the ambiguity for each marginal distribution is formed via (54). They reformulate the resulting problem as a polynomial sized mixed-integer LP when (i) the support is known, (ii) the support and mean are known, and (iii) the support, mean, and mean absolute deviation are known; and as a mixed-integer SOCP when the support, mean, and standard deviation are known.

Hanasusanto et al. [179] study a distributionally robust joint chance constrained stochastic program where each chance constraint is linear in $\boldsymbol{\xi}$, and the technology matrix and right hand-side are affine in $\boldsymbol{x}$. They form the ambiguity set of distributions as in (54) without the first constraint. They show that the pessimistic model (i.e., the chance constraint holds for every distribution in the set) is conic-representable if the technology matrix is constant in $\boldsymbol{x}$, the support set is a cone, and f_i is positively homogeneous. They also show the optimistic model (i.e., the chance constraint holds for at least one distribution in the set) is also conic-representable if the technology matrix is constant in $\boldsymbol{x}$. For other research in chance-constrained optimization problem, we refer to Xie and Ahmed [416], Xie et al. [417].

6.2.4 Moment Matrix Inequalities

In this section, we review an ambiguity set that generalizes both the ambiguity set $\mathcal{P}^{\mathrm{DY}}$ (53) and the ambiguity set $\mathcal{P}^{\mathrm{MM}}$ (54) as follows:

$$\mathcal{P}^{MMI} := \left\{ P \in \mathfrak{M}_{+}(\Xi, \mathcal{F}) \,\middle|\, \boldsymbol{L} \preccurlyeq \int_{\Xi} \boldsymbol{F} \mathrm{d}P \preccurlyeq \boldsymbol{U} \right\}, \tag{55}$$

where $\boldsymbol{F} := [\boldsymbol{F}_1, \ldots, \boldsymbol{F}_m]$, with $\boldsymbol{F}_i$ be a symmetric matrix in $\mathbb{R}^{n_i \times n_i}$ or scalar with measurable components on $(\Xi, \mathcal{F})$. Similarly, let $\boldsymbol{L} := [\boldsymbol{L}_1, \ldots, \boldsymbol{L}_m]$ and $\boldsymbol{U} := [\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m]$ be the vectors of symmetric matrices or

[20] For a measurable function $Z \in \mathcal{Z}_{\infty}(Q)$, the entropic risk meaure is defined as $\frac{1}{\gamma} \ln \mathbb{E}_Q \left[\exp\left(-\gamma Z\right) \right]$, where $\gamma > 0$ Liu et al. [248].

scalars. As in (54), to ensure that P is a probability measure, i.e., $P \in \mathfrak{M}(\Xi, \mathcal{F})$, we set $\boldsymbol{L}_1 = \boldsymbol{U}_1 = [1]_{1\times 1}$ and $\boldsymbol{F}_1 = [1]_{1\times 1}$ in the above definition of $\mathcal{P}^{MMI}$. We generalize this ambiguity set from the ambiguity set proposed in Xu et al. [425], where the moment constraint are either in the form of equality or upper bound. Note that as a special case of $\mathcal{P}^{MMI}$, we can set $\boldsymbol{F}_i$, $\boldsymbol{L}_i$, and $\boldsymbol{U}_i$ to be scalars, $i = 2, \ldots, m$, to recover the second constraint in the ambiguity set $\mathcal{P}^{MM}$, defined in (54). Moreover, by setting $\boldsymbol{F}_2$ to be a matrix as $\begin{pmatrix} -\boldsymbol{\Sigma}_0 & \boldsymbol{\mu}_0 - \boldsymbol{\xi} \\ (\boldsymbol{\mu}_0 - \boldsymbol{\xi})^\top & -\varrho_1 \end{pmatrix}$, $\boldsymbol{F}_3$ to be a matrix as $(\boldsymbol{\xi} - \boldsymbol{\mu}_0)(\xi - \boldsymbol{\mu}_0)^\top$, $\boldsymbol{L}_2 = -\boldsymbol{\infty}$, $\boldsymbol{U}_2 = \boldsymbol{L}_3 = \boldsymbol{0}$, and $\boldsymbol{U}_3 = \varrho_2 \boldsymbol{\Sigma}_0$, we can recover (53).

Below, we present a duality result on $\sup_{P\in\mathcal{P}^{\text{MMI}}} \mathbb{E}_P[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$, given a fixed $\boldsymbol{x} \in \mathcal{X}$.

▶ **Theorem 33.** *For a fixed $\boldsymbol{x} \in \mathcal{X}$, suppose that $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ and $\boldsymbol{F}$ are integrable for all $P \in \mathcal{P}^{MMI}$. In addition, suppose that the following Slater-type condition holds:*

$$(-\boldsymbol{U}, \boldsymbol{L}) \in \operatorname{int}\left(\left\{\Big(-\int_\Xi \boldsymbol{F}\mathrm{d}P, \int_\Xi \boldsymbol{F}\mathrm{d}P\Big) - \mathcal{K} \;\middle|\; P \in \mathfrak{M}_+(\Xi, \mathcal{F})\right\}\right),$$

where $\mathcal{K} := \mathcal{S}_+^{n_1} \times \ldots \mathcal{S}_+^{n_m} \times \mathcal{S}_+^{n_1} \times \ldots \mathcal{S}_+^{n_m}$. If $\sup_{P\in\mathcal{P}^{MM}} \mathbb{E}_P[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$ is finite, then, it can be written as the optimal value of the following problem:

$$\begin{aligned} &\inf_{\boldsymbol{W},\boldsymbol{Y}} \sum_{i=1}^m \boldsymbol{W}_i \bullet \boldsymbol{U}_i - \sum_{i=1}^m \boldsymbol{Y}_i \bullet \boldsymbol{L}_i \\ &s.t. \begin{cases} \sum_{i=1}^m \boldsymbol{W}_i \bullet \int_\Xi \boldsymbol{F}_i(s) P(\mathrm{d}s) - \sum_{i=1}^m \boldsymbol{Y}_i \bullet \int_\Xi \boldsymbol{F}_i(s) P(\mathrm{d}s) \geq \int_\Xi h_0(\boldsymbol{x}, \boldsymbol{\xi}(s)) P(\mathrm{d}s), \ \forall\, P \in \mathfrak{M}_+(\Xi, \mathcal{F}), \\ \boldsymbol{W}, \boldsymbol{Y} \succcurlyeq \boldsymbol{0}. \end{cases} \end{aligned}$$

Proof. See Appendix A. ◀

Suppose that every finite subset of Ξ is $\mathcal{F}$-measurable, i.e., for every $s \in \Xi$, the corresponding Dirac measure δ_s (of mass one at point s) belongs to $\mathfrak{M}_+(\Xi, \mathcal{F})$. Then, the first constraint in Theorem 33 can be written as follows:

$$\sum_{i=1}^m \boldsymbol{W}_i^* \bullet \boldsymbol{F}_i(s) - \sum_{i=1}^m \boldsymbol{Y}_i^* \bullet \boldsymbol{F}_i(s) \geq h_0(\boldsymbol{x}, \boldsymbol{\xi}(s)), \quad \forall\, s \in \Xi.$$

Motivated by the difficulty in verifying the Slater-type conditions to guarantee strong duality for $\sup_{P\in\mathcal{P}^{\text{MMI}}} \mathbb{E}_P[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$ and its dual, Xu et al. [425] investigate the duality conditions from the perspective of lower semicontinuity of the optimal value function of the inner maximization problem, with a perturbed ambiguty set. While these conditions are restrictive in general, they show that they are satisfied in the case of compact Ξ or bounded $\boldsymbol{F}_i$. Xu et al. [425] present two discretization schemes to solve the resulting DRO model: (1) a cutting-plane-based exchange method that discretizes the ambiguity set $\mathcal{P}^{\text{MMI}}$ and (2) a cutting-plane-based dual method that discretizes the semi-infinite constraint of the dual problem. For both methods, they show the convergence of the optimal values and optimal solutions as sample size increases. This type of ambiguity sets is also considered in Chen et al. [93], Liu et al. [249].

6.2.5 Cross-Moment or Nested Moment

In an attempt to unify modeling and solving DRO models, Wiesemann et al. [410] propose a framework for modeling the ambiguity set of probability distributions as follows:

$$\mathcal{P}^{\text{WKS}} := \left\{ \mathbb{P} \in \mathfrak{P}\left(\mathbb{R}^d \times \mathbb{R}^r, \mathfrak{B}(\mathbb{R}^d) \times \mathfrak{B}(\mathbb{R}^r)\right) \;\middle|\; \begin{array}{l} \mathbb{E}_{\mathbb{P}}[\boldsymbol{A}\boldsymbol{\xi} + \boldsymbol{B}\boldsymbol{u}] = \boldsymbol{b}, \\ P\{(\boldsymbol{\xi}, \boldsymbol{u}) \in \mathcal{C}_i\} \in [\underline{p}_i, \overline{p}_i], \ i \in \mathcal{I} \end{array} \right\}, \tag{56}$$

where $\mathbb{P}$ represents a joint probability distribution of $\boldsymbol{\xi}$ and some auxiliary random vector $\boldsymbol{u} \in \mathbb{R}^r$. Moreover, $\boldsymbol{A} \in \mathbb{R}^{s\times d}$, $\boldsymbol{B} \in \mathbb{R}^{s\times r}$, $\boldsymbol{b} \in \mathbb{R}^s$, and $\mathcal{I} = \{1, \ldots, I\}$, while the confidence sets $\mathcal{C}_i$ are defined as

$$\mathcal{C}_i := \left\{(\boldsymbol{\xi}, \boldsymbol{u}) \in \mathbb{R}^d \times \mathbb{R}^r \;\middle|\; \boldsymbol{C}_i\boldsymbol{\xi} + \boldsymbol{D}_i\boldsymbol{u} \preccurlyeq_{\mathcal{K}_i} \boldsymbol{c}_i\right\}, \tag{57}$$

with $\boldsymbol{C}_i \in \mathbb{R}^{L_i\times d}$, $\boldsymbol{D}_i \in \mathbb{R}^{L_i\times r}$, $\boldsymbol{c} \in \mathbb{R}^{L_i}$, and $\mathcal{K}_i$ being a proper cone. By setting $\underline{p}_I = \overline{p}_I = 1$, they ensure that $\mathcal{C}_I$ contains the support of the joint random vector $(\boldsymbol{\xi}, \boldsymbol{u})$. This set contains all distributions with prescribed conic-representable confidence sets and with mean values residing on an affine manifold. An important aspect

of (56) is that the inclusion of an auxiliary random vector $\boldsymbol{u}$ gives the flexibility to model a rich variety of structural information about the marginal distribution of $\boldsymbol{\xi}$ in a unified manner (see Section 6.3 for more details). Using this framework, Wiesemann et al. [410] show that many ambiguity sets studied in the literature can be represented by a projection of the ambiguity set (56) on the space of $\boldsymbol{\xi}$. In other words, these ambiguity sets are special cases of the ambiguity set $\mathcal{P}^{\text{WKS}}$. This development is based on the following lifting result.

▶ **Theorem 34.** *(Wiesemann et al. [410, Theorem 5]) Let $\boldsymbol{f} \in \mathbb{R}^N$ and $\boldsymbol{l} : \mathbb{R}^d \mapsto \mathbb{R}^N$ be a function with a conic-representable $\mathcal{K}$-epigraph, and consider the following ambiguity set:*

$$\mathcal{P}' := \left\{ \mathbb{P} \in \mathfrak{P}(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d)) \;\middle|\; \begin{array}{l} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{l}(\boldsymbol{\xi})\right] \preccurlyeq_{\mathcal{K}} \boldsymbol{f}, \\ \mathbb{P}\{\boldsymbol{\xi} \in \mathcal{C}_i\} \in [\underline{p}_i, \overline{p}_i],\ i \in \mathcal{I} \end{array} \right\},$$

as well as the lifted ambiguity set

$$\mathcal{P} := \left\{ \mathbb{P} \in \mathfrak{P}\left(\mathbb{R}^d \times \mathbb{R}^N, \mathfrak{B}(\mathbb{R}^d) \times \mathfrak{B}(\mathbb{R}^N)\right) \;\middle|\; \begin{array}{l} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{u}\right] = \boldsymbol{f}, \\ P\{\boldsymbol{l}(\boldsymbol{\xi}) \preccurlyeq_{\mathcal{K}} \boldsymbol{u}\} = 1, \\ P\{\boldsymbol{\xi} \in \mathcal{C}_i\} \in [\underline{p}_i, \overline{p}_i],\ i \in \mathcal{I} \end{array} \right\},$$

which involves the auxiliary random vector $\boldsymbol{u} \in \mathbb{R}^N$. We have that (i) $\mathcal{P}'$ is the union of all marginal distributions of $\boldsymbol{\xi}$ under all $\mathbb{P} \in \mathcal{P}$ and (ii) $\mathcal{P}$ can be formulated as an instance of the ambiguity set $\mathcal{P}^{WKS}$ in (56).

Using Theorem 34, Wiesemann et al. [410] show how an ambiguity set of the form $\mathcal{P}^{\text{WKS}}$, defined in (56), with conic-representable expectation constraints and a collection of conic-representable confidence sets, can represent ambiguity sets formed via (1) ϕ-divergences, (2) mean, (3) mean and upper bound on the covariance matrix (i.e., a special case of the ambiguity set (53)), (4) coefficient of variation (i.e., the inverse of signal-to-noise ratio from information theory), (5) absolute mean spread, and (6) higher-order moment information. Moreover, they illustrate that (56) can capture information from robust statistics, such as (7) marginal median, (8) marginal median-absolute deviation, and (9) known upper bound on the expected Huber loss function. It is worth noting that (56) does not cover ambiguity sets that impose infinitely many moment restrictions that would be required to describe symmetry, independence, or unimodality characteristics of the distributions Chen et al. [97].

Wiesemann et al. [410] determine conditions under which distributionally robust expectation constraints, formed via the proposed ambiguity set (56), can be solved in polynomial time as follows: (i) the cost function $h_j(\boldsymbol{x}, \boldsymbol{\xi})$, $j \in [m]$, is convex and piecewise affine in $\boldsymbol{x}$ and $\boldsymbol{\xi}$ (i.e., $h_j(\boldsymbol{x}, \boldsymbol{\xi}) := \max_{k \in [K]} l_{jk}(\boldsymbol{x}, \boldsymbol{\xi})$ with $l_{jk}(\boldsymbol{x}, \boldsymbol{\xi}) := s_{jk}(\boldsymbol{\xi})\boldsymbol{x} + t_{jk}(\boldsymbol{\xi})$ such that $s_{jk}(\boldsymbol{\xi})$ and $t_{jk}(\boldsymbol{\xi})$ are affine in $\boldsymbol{\xi}$) and (ii) the confidence sets $\mathcal{C}_i$'s satisfy a strict nesting condition. Below, we present a duality result under above assumptions and additional regularity conditions.

▶ **Theorem 35** (Wiesemann et al. [410, Theorem 1])**.** *Consider a fixed $\boldsymbol{x} \in \mathcal{X}$. Then, under suitable regularity conditions, $\sup_{\mathbb{P} \in \mathcal{P}^{WKS}} \mathbb{E}_{\mathbb{P}}\left[h_j(\boldsymbol{x}, \boldsymbol{\xi})\right] \leq 0$, $j \in [m]$, is satisfied if and only if there exists $\boldsymbol{\beta} \in \mathbb{R}^K$, $\boldsymbol{\kappa}, \boldsymbol{\lambda} \in \mathbb{R}^I_+$, and $\boldsymbol{\alpha}_{ik} \in \mathcal{K}'_i$, $i \in \mathcal{I}$ and $k \in [K]$, that satisfy the following systems:*

$$\begin{aligned}
&\boldsymbol{b}^\top \boldsymbol{\beta} + \sum_{i \in \mathcal{I}} (\overline{p}_i \kappa_i - \underline{p}_i \lambda_i) \leq 0, \\
&\boldsymbol{c}_i^\top \boldsymbol{\alpha}_{ik} + \boldsymbol{s}_k^\top \boldsymbol{x} + \boldsymbol{t}_k \leq \sum_{i' \in \{i\} \cup \mathcal{A}(i)} (\kappa_{i'} - \lambda_{i'}), \quad && \forall\, i \in \mathcal{I}, k \in [K], \\
&\boldsymbol{C}_i^\top \boldsymbol{\alpha}_{ik} + \boldsymbol{A}^\top \boldsymbol{\beta} = \boldsymbol{S}_k^\top \boldsymbol{x} + \boldsymbol{t}_k, && \forall\, i \in \mathcal{I}, k \in [K], \\
&\boldsymbol{D}_i^\top \boldsymbol{\alpha}_{ik} + \boldsymbol{B}^\top \boldsymbol{\beta} = 0, && \forall\, i \in \mathcal{I}, k \in [K],
\end{aligned}$$

where $\mathcal{A}(i)$ denote the set of all $i' \in \mathcal{I}$ such that $\mathcal{C}_{i'}$ is strictly contained in the interior of $\mathcal{C}_i$.

The tractability of the resulting system in Theorem 35 depends on how the confidence sets $\mathcal{C}_i$ are described, and hence, they give rise to linear, conic-quadratic, or semidefinite programs for the corresponding confidence sets $\mathcal{C}_i$. Wiesemann et al. [410] also provide tight tractable conservative approximations for problems that violate the nesting condition by proposing an outer approximation of (56).

There are several papers that use the ambiguity set (56) and consider its generalization or special cases. Chen et al. [97] introduce an ambiguity set of probability distributions that is characterized by conic-representable expectation constraints and a conic-represetable support set, similar to the one studied in Wiesemann et al. [410]. However, unlike Wiesemann et al. [410], an infinite number of expectation constraints can be incorporated into

the ambiguity set to describe stochastic dominance, entropic dominance, and dispersion, among other. A main result in this work is that for any ambiguity set, there exists an infinitely constrained ambiguity set, such that worst-case expected $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ over both sets are equal, provided that the objective function $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ is tractable and conic-representable in $\boldsymbol{\xi}$ for any $\boldsymbol{x} \in \mathcal{X}$. Reformulation of the resulting DRO model formed via this infinitely constrained ambiguity set yields a conic optimization problem. To solve the model, Chen et al. [97] propose a procedure that consists of solving a sequence of relaxed DRO problems (each of which considers a finitely constrained ambiguity set, and results in a conic optimization reformulation) and converges to the optimal value of the original DRO model. By incorporating covariance and fourth-order moment information into the ambiguity set, they show that the relaxed DRO is a SOCP. This is different from Delage and Ye [105] which shows that a DRO problem formed via a fixed mean and an upper bound on covariance is reformulated as a SDP.

Postek et al. [312] derive exact reformulation of the worst-case expected constraints when function $h_j(\boldsymbol{x}, \cdot)$, $j \in [m]$, is convex in $\boldsymbol{\xi}$, and the ambiguity set of distributions consists of all distributions of componentwise independent $\boldsymbol{\xi}$ with known support, mean, and mean-absolute deviation information. They also obtain exact reformulation of the resulting model when $h_j(\boldsymbol{x}, \cdot)$, $j \in [m]$, is concave in $\boldsymbol{\xi}$ and there is additional information on the probability that a component is greater than or equal to its mean. These results heavily depend on the tight lower and upper bounds on the expectation of a function of a random variable, derived in Ben-Tal and Hochman [21]–extending the well-known Jensen and Edmundson–Madansky bounds to the case that additional information on the dispersion is available. Postek et al. [312] show once random variables are linearly aggregated and function $h_j(\boldsymbol{x}, \cdot)$, $j \in [m]$, is convex, upper bounds can be constructed without the independence restriction. More importantly, under the assumption of independent random variables, Postek et al. [312] use the above results for the worst-case expected constraints to derive Bernstein-type safe approximations to a chance constraint. Long and Qi [251] study a distributionally robust binary stochastic program to minimize the entropic VaR (i.e., Bernstein approximation for the chance constraint). They propose an approximation algorithm to solve the problem via solving a sequence of problems.

To reduce the conservatism of RO, Roos and den Hertog [342] propose an approach that bounds worst-case expected total violation of constraints from above and condense all constraints into a single constraint. They form the ambiguity set with all distributions of $\boldsymbol{\xi}$ with known support, mean, and mean-absolute deviation information. When the right-hand side is uncertain, they use the results in Postek et al. [312] to show that the proposed formulation is tractable. When the left-hand side is uncertain, they use the aggregation approach introduced in Postek et al. [312] to derive tractable reformulations. We also refer to Sun et al. [388] for a two-stage quadratic stochastic optimization problem and DeMiguel and Nogales [109] for a portfolio optimization problem.

Bertsimas et al. [53] develop a modular and tractable framework for solving an adaptive distributionally robust two-stage linear optimization problem with recourse, i.e., $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ is defined as (3) with $q_1 = 0$. They assume that and the function $\boldsymbol{r}(\boldsymbol{\xi})$ and $\boldsymbol{T}(\boldsymbol{\xi})$ are affinely dependent on $\boldsymbol{\xi}$. Both the ambiguity set of probability distributions $\mathcal{P}$ and the support set are assumed to be second-order conic-representable. Such an ambiguity set is a special case of the conic-representable ambiguity set (56). They show that the studied DRO model can be formulated as a classical RO problem with a second-order conic-representable uncertainty set. To obtain a tractable formulation, they replace the recourse decision functions $\boldsymbol{y}(\boldsymbol{\xi})$ with generalized linear decision rules that have affine dependency on the uncertain parameters $\boldsymbol{\xi}$ and some auxiliary random variables[21]. By adopting the approach of Wiesemann et al. [410] to lift the ambiguity set to an extended one by introducing additional auxiliary random variables, they improve the quality of solutions and show that one can transform the adaptive DRO problem to a classical RO problem with a second-order conic-representable uncertainty set. Bertsimas et al. [53] discuss extension to the conic-representable ambiguity set (56) and multistage problems.

Following the approach in Bertsimas et al. [53], Zhen et al. [445] reformulate an adaptive distributionally robust two-stage linear optimization problem with recourse into an adaptive robust two-stage optimization problem with recourse. Then, using Fourier–Motzkin elimination, they reformulate this problem into an equivalent problem with a reduced number of adjustable variables at the expense of an increased number of constraints. Although from a theoretical perspective, every adaptive robust two-stage optimization problem with recourse admits an equivalent static reformulation, they propose to eliminate some of the adjustable variables, and for the remaining adjustable variables, they impose linear decision rules to obtain an approximated solution. They show that for problems with simplex uncertainty sets, linear decision rules are optimal, and for problems with box

[21] Restricting the recourse decision function $\boldsymbol{y}(\boldsymbol{\xi})$ to the class of functions that are affinely-dependent on $\boldsymbol{\xi}$, referred to as *linear decision rules*, is an approach to derive computationally tractable problems to approximate stochastic programming and robust optimization models Ben-Tal et al. [28], Chen et al. [91, 92]. Whether or not the linear decision rules are optimal depends on the problem Shapiro and Nemirovski [372].

uncertainty sets, there exists convex two-piecewise affine functions that are optimal for the adjustable variables. They illustrate that their approach improves the solutions obtained in Bertsimas et al. [53].

6.2.5.1 Statistical Learning

Gong et al. [162] study a distributionally robust multiple linear regression model with the least absolute value cost function. They form the ambiguity set of distributions using expectation constraints over a conic-representable support set as in (56). They reformulate the resulting model as a conic optimization problem, based on the results in Wiesemann et al. [410].

6.2.5.2 Multistage Setting

A Markov decision process with unknown distribution for the transition probabilities and rewards for each state is studied in Xu and Mannor [421, 422]. It is assumed that the parameters are statewise independent and each state belongs to only one stage. Moreover, the parameters of each state are constrained to a sequence of nested sets, such that the parameters belong to the largest set with probability one, and there is a lower bound on the probability that they should belong to other sets, in a increasing manner. Yu and Xu [434] extends the work in Xu and Mannor [421, 422] by forming the ambiguity set of distributions as in (56). Shapiro et al. [373] study a multistage stochastic program, where the data process can be naturally separated into two components: one can be modeled as a random process, with a known probability distribution, and the other can be treated as a random process, with a known support and no distributional information. They propose a variant of the SDDP method to solve this problem.

6.2.6 Marginals (Fréchet)

Most of the moment-based ambiguity sets discussed so far, study the ambiguity of the joint probability distribution of the random vector $\boldsymbol{\xi}$. Papers reviewed in this section assume that only the marginals of a multivariate distribution are constrained, while the dependence structure is unconstrained. The proposed ambiguity set is of the following generic form

$$\mathcal{P}^F := \{P \in \mathfrak{M}(\Xi, \mathcal{F}) \,|\, P \in \Gamma(\mu_1, \ldots, \mu_d)\}, \tag{58}$$

where $\Gamma(\mu_1, \ldots, \mu_d)$ denotes the set of distributions with marginals $\mu_1, \ldots, \mu_d$. We refer to the class of joint distributions with fixed marginal distributions as the *Fréchet* class of distributions Doan et al. [121]. Obtaining worst-case bounds on the distribution and on the tails for functions of dependent random vectors with fixed multivariate marginals are studied in the literature, see, e.g., Embrechts and Puccetti [136, 137], Puccetti and Rüschendorf [318, 319], Puccetti et al. [320], Wang and Wang [404]. The ambiguity sets studied in this section relate to quantitative risk management or extreme-event analysis, in which obtaining or modeling the dependence structure from data can be much more challenging than the marginals.

6.2.6.1 Discrete Problems

Chen et al. [85] study a problem of the form (8), where the cost function $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ denotes the optimal value of a linear or discrete optimization problem with random linear objective coefficients. They assume the ambiguity set of distribution is formed by all distributions with known marginals. Using techniques from optimal transport theory, they identify a set of sufficient conditions for the polynomial time solvability of this class of problems. This generalizes the tractability results under marginal information from 0-1 polytopes, studied in Bertsimas et al. [45], to a class of integral polytopes.

6.2.6.2 Risk and Chance Constraints

Dhara et al. [118] provide bounds on the worst-case CVaR over an ambiguity set of discrete distributions, where the ambiguity set contains all joint distributions whose univariate marginals are fixed and their bivariate marginals are within a minimum Kullback–Leibler distance from the nominal bivariate marginals. They develop a convex reformulation for the resulting DRO. Doan et al. [121] study a DRO model of the form (8) with a convex piecewise linear objective function in $\boldsymbol{\xi}$ and affine in $\boldsymbol{x}$. They form the ambiguity set of joint distributions via a Fréchet class of discrete distributions with multivariate marginals, where the components of the random vector are partitioned such that they have overlaps. They show that the resulting DRO model for a portfolio

optimization problem is efficiently solvable with a LP. In particular, they develop a tight LP reformulation to find a bound on the worst-case CVaR over such an ambiguity set, provided that the structure of the marginals satisfy a regularity condition.

Zhang et al. [441] study a distributionally robust approach to a stochastic bin-packing problem subject to chance constraints on the total item sizes in the bins. They form the ambiguity set by all discrete distributions with known marginal means and variances for each item size. By showing that there exists a worst-case distribution that is at most a three-point distribution, they obtain a closed-form expression for the chance constraint and they reformulate the problem as a mixed-binary program.

6.2.6.3 Statistical Learning

Farnia and Tse [140] study a DRO approach in the context of supervised learning problems to infer a function (i.e., decision rule) that predicts a response variable given a set of covariates. Motivated by the game-theoretic interpretation of Grünwald and Dawid [166] and the principle of maximum entropy, they seek a decision rule that predicts the response based on a distribution that maximizes a generalized entropy function over a set of probability distributions. However, because the covariate information is available, they apply the principle of maximum entropy to the conditional distribution of the response given the covariates, see also Globerson and Tishby [158] for the case of Shannon entropy. Farnia and Tse [140] form the ambiguity set of distributions by matching the marginal of covariates to the empirical marginal of covariates while keeping the cross-moments between the response variables and covariates close enough (with respect to some norm) to that of the joint empirical distribution. They show that the DRO approach adopts a regularization interpretation for the maximum likelihood problem under the empirical distribution. As a result, Farnia and Tse [140] recover the regularized maximum likelihood problem for generalized linear models for the following loss functions: linear regression under quadratic loss function, logistic regression under logarithmic loss function, and SVM under the 0-1 loss function.

Eban et al. [132] study a DRO approach to a classification problem to minimize the worst-case hinge loss of missclassification, where the ambiguity set of the joint probability distributions of the discrete covariates and response should contain all distributions that agree with nominal pair-wise marginals. They show that the proposed classifier provides a 2-approximation upper bound on the worst-case expected loss using a zero-one hinge loss. Razaviyayn et al. [331] study a DRO approach to the binary classification problem, with an ambiguity set similar to that of Eban et al. [132], to minimize the worst-case missclassification probability. By changing the order of inf and sup, and smoothing the objective function, they obtain a probability distribution, based on which they propose a randomized classifier. They show that this randomized classifier enjoys a 2-approximation upper bound on the worst-case missclassification probability of the optimal solution to the studied DRO.

6.2.7 Mixture Distribution

In this section, we study DRO models, where the ambiguity set is formed via *mixture distribution*. A mixture distribution is defined as a convex combination of pdfs, known as the *mixture components*. The weights associated with the mixture components are called *mixture probabilities* Kapsos et al. [219]. For example, a mixture model can be defined as the set of all mixtures of normal distributions with mean μ and standard deviation σ with parameter $\boldsymbol{a} = (\mu, \sigma)$ in some compact set $\mathcal{A} \subset \mathbb{R}^2$. In a more generic framework, the distribution P can be any mixture of probability distributions $Q_{\boldsymbol{a}} \in \mathfrak{M}(\Xi, \mathcal{F})$, for some family of distributions $\{Q_{\boldsymbol{a}}\}_{\boldsymbol{a}\in\mathcal{A}} \in \mathfrak{M}(\Xi, \mathcal{F})$, that depends on the parameter vector $\boldsymbol{a} \in \mathcal{A}$ as follows:

$$P(B) = \int_{\mathcal{A}} Q_{\boldsymbol{a}}(B) M(\mathrm{d}\boldsymbol{a}), \quad B \in \mathcal{F}, \tag{59}$$

where M is any probability distribution on $\mathcal{A}$ (Lasserre and Weisser [237]). Hence, modeling the ambiguity in the mixture probabilities may give rise to a DRO model over the *resultant or barycenter* P of M (Popescu [309]).

6.2.7.1 Risk and Chance Constraints

Lasserre and Weisser [237] study a distributionally robust (individual and joint) chance-constrained program with a polynomial objective function, over a mixture ambiguity set and a semi-algebraic deterministic set. They approximate the ambiguous chance constraint with a polynomial whose vector coefficients is an optimal solution of a SDP. They show that the induced feasibility set by a nested sequence of such polynomial optimization

approximation problems converges to that of the ambiguous chance constraints as the degree of approximate polynomials increases.

Kapsos et al. [219] introduce a probability Omega ratio for portfolio optimization (i.e., a probability weighted ratio of gains versus losses for some threshold return target). They study a distributionally robust counterpart of this ratio, where each distribution of the ratio can be represented through a mixture of some known prespecified distributions with unknown mixture probabilities. In particular, they study a mixture model for a nominal discrete distribution, where the mixture probabilities are modeled via the box uncertainty and ellipsoidal uncertainty models. In the former case, they reformulate the problem as a LP, and in the latter case, they reformulate the problem as a SOCP.

Hanasusanto et al. [175] study a distributionally robust newsvendor model with a mean-risk objective, as a convex combination of the worst-case CVaR and the worst-case expectation. The worst case is taken over all demand distributions within a *multimodal* ambiguity set, i.e., a mixture of a finite number of modes, where the conditional information on the ellipsoid support, mean, and covariance of each mode is known. The ambiguity in each mode is modeled via (53). They cast the resulting model as an exact SDP, and obtain a conservative semidefinite approximation by using quadratic decision rules to approximate the recourse decisions. Hanasusanto et al. [175] further robustify their model against ambiguity in estimating the mean-covariance information, caused from ambiguity about the mixture weights. They assume that the mixture weights are close to a nominal probability vector in the sense of χ^2-distance. For this case, they also obtain exact SDP reformulation as well as a conservative SDP approximation.

6.3 Shape-Preserving Models

A few papers propose to model the distributional ambiguity in a way that all distributions in the ambiguity set share similar structural properties. We refer to such models as *shape-preserving* models to form the ambiguity set of probability distributions.

Popescu [309] propose to incorporate structural distributional information, such as symmetry, unimodality, and convexity, into a moment-based ambiguity set. The proposed ambiguity set is of the following generic form:

$$\mathcal{P}^{SP} := \left\{ P \in \mathfrak{M}_+(\Xi, \mathcal{F}) \,\middle|\, \int_\Xi \boldsymbol{f} \mathrm{d}P = \boldsymbol{a} \right\} \cap \{P \text{ satisfies structural properties}\}. \tag{60}$$

By relying on information from classical statistics as well as robust statistics, Hanasusanto et al. [177] propose a unifying canonical ambiguity set that contains many ambiguity sets studied in the literature as special cases, including Gauss and median-absolute deviation ambiguity sets. Such a canonical framework is characterized through intersecting the cross-moment ambiguity set, proposed in Wiesemann et al. [410], and a structural ambiguity set on the marginal distributions, representing information such as symmetry and α-unimodality.

Hu et al. [199] study a data-driven newsvendor problem to decide on the optimal order quantity and price. They assume that demand depends on the pricing, however, there is ambiguity about the price-demand function. To hedge against the misspecification of the demand function, they introduce a novel approach to this problem, called *functionally robust* approach, where the demand-price function is only known to be decreasing convex or concave. The proposed modeling approach in Hu et al. [199] also provides a systematic view on the risk-reward trade-off of coordinating pricing and order quantity decisions based on the size of the ambiguity set. To solve the resulting minimax model, Hu et al. [199] reduce the problem into a univariate problem that seeks the optimal pricing and develop a two-sided cutting surface algorithm that generates function cuts to shrink the set of admissible functions.

6.3.1 Risk and Chance Constraints

Popescu [309] obtains upper and lower bounds on a generalized moment of a random vector (e.g., tail probabilities), given the moments and structural constraints in a convex subset of the proposed ambiguity set (60). Popescu [309] uses conic duality to evaluate such lower and upper bounds via SDPs. The key to the development in Popescu [309] is to focus on ambiguity sets that posses a *Choquet representation*, where every distribution in the ambiguity set can be written as a mixture (i.e., an infinite convex combination) of measures in a generating set and in the virtue of (59). For univariate distributions, it is assumed that the generating set is defined by a Markov kernel. It is shown that if the optimal value of the problem is attained, there exists a worst-case probability measure that is a convex combination of $m + 1$ (recall m is the dimension of $\boldsymbol{f}$) (extremal) probability measures

from the generating set. Popescu [309] uses the above result to obtain generalized Chebyshev's inequalities bounds for distributions of a univariate random variable that are (1) symmetric, (2) unimodal with a given mode, (3) unimodal with bounds on the mode, (4) unimodal and symmetric, or (5) convex/concave monotone densities with bounds on the slope of densities. Popescu [309] further derives generalized Chebyshev's inequality for symmetric and unimodal distributions of multivariate random variables.

Nemirovski and Shapiro [279] study a convex approximation, referred to as *Bernstein* approximation, to an ambiguous joint chance-constrained problem of the form

$$\min_{\boldsymbol{x}\in\mathcal{X}} h_0(\boldsymbol{x}) \quad \text{s.t.} \ \inf_{P\in\mathcal{P}} P\left\{\boldsymbol{\xi} : g_{i0}(\boldsymbol{x}) + \sum_{j=1}^{d} \xi_j g_{ij}(\boldsymbol{x}) \leq 0,\ i\in[m]\right\} \geq 1-\epsilon. \tag{61}$$

▶ **Theorem 36** (Nemirovski and Shapiro [279, Theorem 6.2])**.** *Suppose that the ambiguous joint chance-constrained problem* (61) *is such that (i) the components of the random vector $\boldsymbol{\xi}$ are independent of each other, with finite-valued moment generating functions, (ii) function $h_0(\boldsymbol{x})$ and all functions $g_{ij}(\boldsymbol{x})$, $i\in[m]$, $j\in[d]$, are convex and well defined on $\mathcal{X}$, and (iii) the ambiguity set of probability distributions $\mathcal{P}$ forms a convex set. Let ϵ_i, $i\in[m]$, be positive real values such that $\sum_{i=1}^m \epsilon_i \leq \epsilon$. Then, the problem*

$$\min_{\boldsymbol{x}\in\mathcal{X}} h_0(\boldsymbol{x}) \quad \textit{s.t.} \ \inf_{t>0} \left[g_{i0}(\boldsymbol{x}) + t\widehat{\Psi}(t^{-1}\boldsymbol{z}^i[\boldsymbol{x}]) - t\log\epsilon_i\right] \leq 0,\ i\in[m],$$

where $z^i(\boldsymbol{x}) = \big(g_{i1}(\boldsymbol{x}),\ldots,g_{id}(\boldsymbol{x})\big)$ and $\widehat{\Psi}(\boldsymbol{z}) := \max_{Q_1\times\ldots\times Q_d\in\mathcal{P}} \sum_{j=1}^d \log\big(\int_\Xi \exp\{z_j s\}\mathrm{d}Q_j(s)\big)$, is a conservative approximation of problem (61)*, i.e., every feasible solution to the approximation is feasible for the chance-constrained problem* (61)*. This approximation is a convex program and is efficiently solvable, provided that all g_{ij} and $\widehat{\Psi}$ are efficiently computable, and $\mathcal{X}$ is computationally tractable.*

Nemirovski and Shapiro [279] obtain closed-form expressions for $\max_{Q_j\in\mathcal{P}_j} \log\big(\int_\Xi \exp\{z_j s\}\mathrm{d}Q_j(s)\big)$ for some families of univariate distributions, including those with structural properties.

A related notion to unimodality is α-unimodality, which is defined as follows:

▶ **Definition 37** (Dharmadhikari and Joag-Dev [119])**.** *For $\alpha>0$, a distribution $\mathbb{P}\in\mathfrak{P}(\mathbb{R}^d,\mathfrak{B}(\mathbb{R}^d))$ is called α-unimodal with mode a if $\frac{\mathbb{P}\{t(A-a)\}}{t^\alpha}$ is nonincreasing in $t>0$ for all $A\in\mathcal{B}(\mathbb{R}^d)$.*

Van Parys et al. [396] further extend the work of Popescu [309] to obtain worst-case probability bounds over α-unimodal multivariate distributions with the same mode and within the class of distributions in $\mathcal{P}^{\mathrm{DY}}$, defined in (53), and on a polytopic support. They show that when the support of the random vector is an open polyhedron, this generalized Gauss bound can be obtained via a SDP. Similar to Popescu [309], Van Parys et al. [396] derive semidefinite representations for worst-case probability bounds using Choquet representation of the ambiguity set. They demonstrate that classical generalized Chebyshev and Gauss bounds[22] can be obtained as special cases of their result. They also show how to obtain a SDP reformulation to obtain the worst-case bound over α-multimodal multivariate distributions, defined via a mixture distribution.

As in Popescu [309], the key to the development in Hanasusanto et al. [177] is to focus on structural ambiguity sets that posses a Choquet representation. They study distributionally robust uncertainty quantification (i.e., a probabilistic objective function) and chance-constrained programs over the proposed ambiguity sets, where the safe region is characterized by a bi-affine expression in $\boldsymbol{\xi}$ and $\boldsymbol{x}$. They study the ambiguity sets over which the resulting problems are reformulated as conic programming formulations. A summary of these results can be found in Hanasusanto et al. [177, Table 2]. A by-product of their study is to recover some results from probability theory. For instance, by studying the worst-case probability of an event over the Chebyshev ambiguity set with a known mean and upper bound on the covariance matrix, they recover the generalized Chebyshev inequality, discovered in Popescu [309], Vandenberghe et al. [400]. Similarly, they recover the generalized Gauss inequality, discovered in Van Parys et al. [396], by considering the Gauss ambiguity set. Furthermore, they propose computable conservative approximations for the chance-constrained problem. Recognizing that the uncertainty quantification problem is tractable over a broad range of ambiguity sets, their key idea for the proposed approximation scheme is to decompose the chance-constrained problem into an uncertainty quantification problem that evaluates the worst-case probability of the chance constraint for a fixed decision $\boldsymbol{x}$, followed by a decision improvement procedure.

[22] The random variable differs from its mean by more than k standard deviations.

Li et al. [241] study distributionally robust chance- and CVaR-constrained stochastic programs, where the ambiguity set contains all α-unimodal distributions with the same first two order moments, and the safe region is bi-affine in both $\boldsymbol{\xi}$ and $\boldsymbol{x}$. They show that these two ambiguous risk constraints can be cast as an infinite set of SOC constraints. They propose a separation approach to find the violated SOC constraints in an algorithmic fashion. They also derive conservative and relaxation approximations of the two SOC constraints by a finite number of constraints. These approximations for the CVaR-constrained problem are based on the results in Van Parys et al. [398].

To overcome the difficulty in evaluating extremal performance due to the lack of data, Lam and Mottet [231] study the computation of worst-case bounds under the geometric premise of the tail convexity. They show that the worst-case convex tail behavior is in a sense either extremely light-tailed or extremely heavy-tailed.

6.4 Kernel-Based Models

Kernel smoothing methods have shown robustness properties in regression (Christmann et al. [101]) and classification (Christmann and Steinwart [100], Steinwart [385], Xu et al. [423]), demonstrating relationships between robustness and regularization. On the other hand, the space associated with kernel functions are used for comparing probability distributions Smola et al. [381]. These have motivated researchers to study a DRO model in the functional space produced by a kernel function Zhu et al. [447]. This framework unifies many discrepancy-based and moment-based DRO models. We present some basic results in this section.

Consider a metric space $(\mathcal{S}, d)$ and a positive definite symmetric kernel function $K : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}$[23]. It is known that for any positive definite symmetric kernel K, there is a mapping $\Phi : \mathcal{S} \mapsto \mathcal{H}$ such that $K(s,t) = \langle \Phi(s), \Phi(t) \rangle_{\mathcal{H}}$ defines an inner product on $\mathcal{H}$, see, e.g., Schölkopf and Smola [352] and Mohri et al. [268, Theorem 5.2]. The mapping Φ is called a *feature map*, and $\mathcal{H}$ is called a *feature space* of K, containing real-valued functions on $\mathcal{S}$. The canonical feature map $\Phi(s) = K(s, \cdot)$ gives rise to the canonical feature space $\mathcal{H}$, or the so-called *reproducing kernel Hilbert space* (RKHS), with a reproducing property: $f(s) = \langle f, \Phi(s) \rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$ and $s \in \mathcal{S}$, and norm $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$ for any $f \in \mathcal{H}$. Many continuous kernel functions are universal in the sense that their corresponding RKHSs are dense in the space of continuous bounded functions on the compact space $\mathcal{S}$ Steinwart [385].

Smola et al. [381] study embedding probability measures into a RKHS by defining the kernel mean mapping $\mu_P := \int_{\mathcal{S}} K(s, \cdot) \mathrm{d}P$. Provided that $\int_{\mathcal{S}} K(s,s) \mathrm{d}P < \infty$, then μ_P is an element of $\mathcal{H}$, for any probability measure P on $\mathcal{S}$. By the reproducing property of $\mathcal{H}$, we have $\mathbb{E}_P[f(s)] = \langle f, \mu_P \rangle_{\mathcal{H}}$. Embedding probability measures into $\mathcal{H}$ allows one to measure the discrepancy between the probability measures using the norm defined on $\mathcal{H}$. Moreover, if kernel K is universal, then the mapping $P \mapsto \mu_P$ is injective. Thus, given two probability measure P_1 and P_2, $\|\mu_{P_1} - \mu_{P_2}\|_{\mathcal{H}}$ defines a probability metric. This metric can be seen as an instance of ζ-structure or integral probability metrics; hence, can be written as $\|\mu_{P_1} - \mu_{P_2}\|_{\mathcal{H}} = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \mathbb{E}_{P_1}[f(s)] - \mathbb{E}_{P_2}[f(s)] \right|$. Motivated by these properties, Zhu et al. [447] propose a unifying kernel-based DRO model, where the distributional ambiguity is modeled with the ambiguity set

$$\mathcal{P}^{\mathrm{K}}(\mathcal{K}, \mathcal{C}) := \left\{ P \in \mathfrak{M}(\Xi, \mathcal{F}) \,\middle|\, \int \Phi \, \mathrm{d}P = \mu, \; P \in \mathcal{K}, \; \mu \in \mathcal{C} \right\}, \tag{62}$$

where $\Phi : \Xi \mapsto \mathcal{H}$ is the feature map of the RHKS associated with the kernel function $K : \Xi \times \Xi \mapsto \mathbb{R}$. Both sides of $\int \Phi \mathrm{d}P = \mu$ are functions in $\mathcal{H}$. Hence, μ can be considered as a generalized (infinite-dimensional) moment vector within the set $\mathcal{C} \subseteq \mathcal{H}$. Moreover, $\mathcal{K}$ is a subset of $\mathfrak{M}(\Xi, \mathcal{F})$. As argued in Zhu et al. [447], small RKHSs lead to a large set $\mathcal{P}^{\mathrm{K}}(\mathcal{K}, \mathcal{C})$, and thus; a conservative model. In an extreme, if $\mathcal{H} = \{0\}$, the smallest possible RKHS, then $\mathcal{H}$ does not contain any function to distinguish between distributions. Consequently, the resulting DRO model is reduced to a RO model. On the other hand, large RKHSs might lead to meaningless models with trivial solutions. Moreover, suitably chosen kernel functions and sets $\mathcal{C}$ in (62) lead to some of the models studied in previous sections. Given a set of observations $\{\boldsymbol{\xi}^i\}_{i=1}^N$, consider an empirical distribution $\widehat{\mathbb{P}}_N$. Then, $\mathcal{C} = \{\frac{1}{N} \sum_{i=1}^N \Phi(\boldsymbol{\xi}^i)\}$ reduces the resulting DRO model to a SAA model. Setting $K(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = (1 + \boldsymbol{\xi}_1^\top \boldsymbol{\xi}_2)^2$ and $\mathcal{C} = \{\mu_{\widehat{\mathbb{P}}_N}\}$ leads to an ambiguity set of distributions where the first two moments are the same as those of $\widehat{\mathbb{P}}_N$ (see Section 6.2.1). As another example, $\mathcal{C} = \{\mu : \|\mu - \mu_{\widehat{\mathbb{P}}_N}\|_{\mathcal{H}} \leq \epsilon\}$ leads to an ambiguity set $\mathcal{P}^{\mathcal{Z}}(\widehat{\mathbb{P}}_N; \epsilon)$, studied in Section 6.1.7, where $\mathcal{Z} = \{z : \|z\|_{\mathcal{H}} \leq 1\}$.

[23] A kernel is said to be positive definite symmetric if the induced kernel matrix is symmetric positive semidefinite.

Table 5 Examples of convex sets $\mathcal{C}$ and their support functions $\delta^*(f|\mathcal{C})$

Set $\mathcal{C}$	$\delta^*(f\|\mathcal{C})$
Norm-ball $\mathcal{C} = \{\mu : \|\mu - \mu_{\hat{\mathbb{P}}_N}\|_{\mathcal{H}} \le \epsilon\}$	$\frac{1}{N}\sum_{i=1}^N f(\boldsymbol{\xi}^i) + \epsilon\|f\|_{\mathcal{H}}$
Convex hull $\mathcal{C} = \text{conv}(\mathcal{C}_1, \ldots, \mathcal{C}_N)$ Example: Polytope $\mathcal{C} = \text{conv}(\Phi(\boldsymbol{\xi}^1), \ldots, \Phi(\boldsymbol{\xi}^N))$	$\max_{1\le i\le N} \delta^*(f\|\mathcal{C}_i)$ $\max_{1\le i\le N} f(\boldsymbol{\xi}^i)$
Affine combination: $\mathcal{C} = \sum_{i=1}^N \alpha_i\mathcal{C}_i,\ \sum_{i=1}^N \alpha_i = 1$ Example: Minkowski sum $\mathcal{C} = \mathcal{C}_1 + \mathcal{C}_2$ $\mathcal{C}_1 = \{\mu : \|f\|_{\mathcal{H}} \le \epsilon\}$ $\mathcal{C}_2 = \text{conv}(\Phi(\boldsymbol{\xi}^1), \ldots, \Phi(\boldsymbol{\xi}^N))$ Example: Contamination neighborhood $\mathcal{C} = \left\{P \,\middle|\, P = (1-\epsilon)\widehat{\mathbb{P}}_N + \epsilon\mathbb{Q},\ \mathbb{Q} \in \mathfrak{Q}\right\}$	$\sum_{i=1}^N \alpha_i\delta^*(f\|\mathcal{C}_i)$ $\max_{1\le i\le N} f(\boldsymbol{\xi}^i) + \epsilon\|f\|_{\mathcal{H}}$ $\frac{(1-\epsilon)}{N}\sum_{i=1}^N f(\boldsymbol{\xi}^i) + \epsilon\delta^*(f\|\mathfrak{Q})$
Intersection $\mathcal{C} = \cap_{i=1}^N \mathcal{C}_i$	$\sum_{i=1}^N \delta^*(f_i\|\mathcal{C}_i),\ f = \sum_{i=1}^N f_i$
Singleton $\mathcal{C} = \{\frac{1}{N}\sum_{i=1}^N \Phi(\boldsymbol{\xi}^i)\}$	$\frac{1}{N}\sum_{i=1}^N f(\boldsymbol{\xi}^i)$
$\mathcal{C} = \mathcal{H} = \{0\}$	0

▶ **Theorem 38** (Zhu et al. [447, Theorem 3.1]). *Consider an ambiguity set of probability measures as formed via $\mathcal{P}^K(\mathcal{K}, \mathcal{C})$, defined in (62). Suppose that $\mathcal{C}$ is a closed convex set and the relative interior of $\mathcal{P}^K(\mathcal{K}, \mathcal{C})$ is nonempty. In addition, suppose that for a fixed $\boldsymbol{x} \in \mathcal{X}$, $h_0(\boldsymbol{x}, \cdot)$ is upper semicontinuous. Then, $\sup_{P\in\mathcal{P}^K(\mathcal{K},\mathcal{C})} \mathbb{E}_P[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$ is equivalent to*

$$\min_{f_0\in\mathbb{R}, f\in\mathcal{H}} \quad f_0 + \delta^*(f|\mathcal{C}) \quad \text{s.t. } h_0(\boldsymbol{x}, \boldsymbol{\xi}) \le f_0 + f(\boldsymbol{\xi}),\ \boldsymbol{\xi} \in \Xi,$$

where $\delta^*(f|\mathcal{C}) = \sup_{\mu\in\mathcal{C}}\langle f, \mu\rangle_{\mathcal{H}}$ is the support function of $\mathcal{C}$ at $f \in \mathcal{H}$.

A list of sets $\mathcal{C}$ and their support functions are given in Zhu et al. [447]. We present a few of them in Table 5. We also refer to Staib and Jegelka [384] for a related work.

The use of kernel functions in decision-making problems has led to an emerging area of research that integrates machine learning techniques into the optimization framework, see, e.g., Ban and Rudin [11], Bertsimas and Kallus [41]. For reference, consider a given set of data $\{(\boldsymbol{u}^i, \boldsymbol{\xi}^i)\}_{i=1}^N$, where $\boldsymbol{u}^i \in \mathbb{R}^m$ is a vector of covariates associated with the uncertain parameter of interest $\boldsymbol{\xi}^i \in \mathbb{R}^d$.

Bertsimas and Kallus [41] propose a decision framework that incorporates the covariates $\boldsymbol{u}$ in addition to $\boldsymbol{\xi}$ into the optimization problem in the form of a conditional stochastic optimization problem, where the decision-maker is seeking a *predictive prescription* $\boldsymbol{x}(\boldsymbol{u})$ that minimizes the conditional expectation of $h_0(\boldsymbol{x}, \boldsymbol{\xi})$ in anticipation of the future, given the observation $\boldsymbol{u}$. However, the conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{u}$ is not known and should be learned from data. Bertsimas and Kallus [41] propose to find a data-driven predictive prescription that minimizes $\sum_{i=1}^N w^i(\boldsymbol{u})h_0(\boldsymbol{x}, \boldsymbol{\xi}^i)$ over $\mathcal{X}$. Functions $w^i(\boldsymbol{u})$ are weights learned locally from the data, in a sense that predictions are made based on the past observations that are in some way similar to the one at hand, $\boldsymbol{u}$. Bertsimas and Kallus [41] obtain these weight functions by methods that are motivated by k-nearest-neighbors regression, Nadaraya–Watson kernel regression, local linear regression (in particular, LOESS), classification and regression trees (in particular, CART), and random forests. For instance, the estimate of $\mathbb{E}_P\left[h_0(\boldsymbol{x}, \boldsymbol{\xi})\middle|\boldsymbol{u}\right]$ using the Nadaraya–Watson kernel regression is obtained as

$$\sum_{i=1}^N \frac{K_b(\boldsymbol{u} - \boldsymbol{u}^i)}{\sum_{i=1}^N K_b(\boldsymbol{u} - \boldsymbol{u}^i)} h_0(\boldsymbol{x}, \boldsymbol{\xi}^i),$$

where $K_b(\cdot) := \frac{K(\frac{\cdot}{b})}{b}$ is a kernel function with bandwidth b. Common kernel smoothing functions are

- Naive: $K(a) = \mathbb{1}_{[\|a\|\le 1]}$,
- Epanechnikov: $K(a) = (1 - \|a\|^2)\mathbb{1}_{[\|a\|\le 1]}$,
- Tri-cubic: $K(a) = (1 - \|a\|^3)^3\mathbb{1}_{[\|a\|\le 1]}$,
- Gaussian or radial basis function: $K(a) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{\|a\|^2}{2})$.

The general framework of the proposed data-driven model in Bertsimas and Kallus [41] resembles SAA, i.e., $1/N$ weights in SAA are replaced by weights $w^i(\boldsymbol{u})$, $i \in [N]$. Bertsimas and Kallus [41] show that under mild conditions, the resulting predictive prescription problem is asymptotically optimal and consistent. As illustrated

by Bertsimas and Kallus [41] the direct usage of SAA on $\{\boldsymbol{\xi}^i\}_{i=1}^N$ and ignoring $\{\boldsymbol{u}^i\}_{i=1}^N$ can result in suboptimal decisions which are neither asymptotically optimal nor consistent.

A similar modeling framework as the conditional stochastic optimization problem studied in Bertsimas and Kallus [41] is investigated in other papers to incorporate machine learning into decision-making, see, e.g., Ban et al. [12], Kannan et al. [217, 218]. Deng and Sen [110] use regression models such as k-nearest-neighbors regression to learn the conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{u}$. They study the statistical optimality of the resulting solution and its generalization error, and they provide hypothesis-based tests for model validation and selection. In Ban and Rudin [11], Hannah et al. [180], Pang Ho and Hanasusanto [292], the weights are obtained by the Nadaraya–Watson kernel regression method. For a newsvendor problem, Ban and Rudin [11] show that the SAA decision does not converge to the true optimal decision. This motivates them to derive generalization bounds for the out-of-sample performance of the cost and the finite-sample bias from the true optimal decision.

Similar to Bertsimas and Kallus [41], Bertsimas and Van Parys [44] consider the problem of finding an optimal solution to a data-driven stochastic optimization problem, where the uncertain parameter is affected by a large number of covariates. They study a distributionally robust approach to this problem formed via Kullback–Leibler divergence. By borrowing ideas from the statistical bootstrap, they propose two prescriptive methods based on the Nadaraya–Watson and nearest-neighbors learning formulation, first introduced by Bertsimas and Kallus [41], which safeguards against overfitting and lead to an improved out-of-sample performance. Both resulting prescriptive models reduce to tractable convex optimization problems. Shang and You [360] adopt the ambiguity set proposed in Wiesemann et al. [410], and propose to use principal component analysis (PCA) to calibrate the moment functions. In fact, a moment function in their model is a piecewise linear function, which is defined as a first-order deviation of the uncertain parameter along a certain projection direction, truncated at certain points. They propose to use PCA to come up with the projection directions, and choose the truncation points symmetrically around the sample mean along the direction. We refer the readers to Appendix B for discussion on kernel-based models in RO.

6.5 Choosing an Ambiguity Set of Probability Distributions

In this section, we discuss what ambiguity sets are good in what situations.

Recognizing the fact that the ambiguity set should be chosen judiciously for the application in hand, Gao and Kleywegt [148] argue that by using the Wasserstein metric the resulting distributions hedged against are more reasonable than those resulting from other popular choices of sets, such as ϕ-divergence-based sets, see Section 6.1.2.

Recall the notions of popping and suppressing scenarios (Bayraksan and Love [17], Love and Bayraksan [256]), discussed in Section 6.1.2 for ϕ-divergences. On the basis of these notions, Bayraksan and Love [17], Love and Bayraksan [256] propose some modeling considerations when choosing a ϕ-divergence. If every scenario with a positive nominal probability comes from high-quality data, and the decision maker may wish to hedge against those scenarios, then one may choose ϕ-divergences that cannot suppress scenarios, e.g., χ^2-distance and Burg entropy. However, if the data is poorly sampled or comes from opinion rather than observations or simulation, then ϕ-divergences that can suppress scenarios may be preferable, e.g., variation distance, Hellinger distance, modified χ^2-distance, and Kullback–Leibler divergence. If the scenarios strictly come from observations, with little theoretical understanding of the problem, then one may choose ϕ-divergences that cannot pop scenarios, e.g., modified χ^2-distance and Kullback–Leibler divergence. However, if the scenarios come from a mixture of observed/simulated data and expert opinions, then ϕ-divergences that can pop scenarios may be desirable, e.g., variation and Hellinger distance.

7 Calibration of the Ambiguity Set of Probability Distributions

In Section 6, we reviewed different approaches to model the distributional ambiguity. These models rely on some parameters that need to be calibrated for the problem in-hand to reach a solution with a reasonably good out-of-sample performance. In Section 7.1, we briefly discuss the choice of nominal parameters. In Section 7.2, we explain the choice of robustness parameters for a number of models described in Section 6.

7.1 Choice of the Nominal Parameters

All discrepancy-based ambiguity sets, studied in Section 6.1, and some of the moment-based ambiguity sets, studied in Section 6.2, rely on some nominal input parameters, for instance, the nominal distribution P_0 in the ambiguity set $\mathcal{P}^{\mathrm{W}}(P_0, \epsilon)$, defined in (32), and parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ in the ambiguity set $\mathcal{P}^{\mathrm{DY}}$, defined in (53). In this section, we discuss how these parameters are chosen in a data-driven setting.

The nominal distribution P_0 in the discrepancy-based ambiguity sets is usually obtained by the maximal likelihood estimator of the true unknown distribution. In the discrete case, P_0 is typically chosen as the empirical distribution on data. Additionally, it is known that even in the continuous distribution case, one may choose P_0 to be the empirical distribution, in which case the statistical guarantees are obtained by using tools related to the empirical or profile likelihood, see, e.g., Blanchet et al. [64], Duchi et al. [123], Lam [230], Lam and Zhou [232, 233]. Alternatively, Jiang and Guan [214] and Zhao and Guan [442] propose to obtain P_0 with nonparametric kernel density estimation methods, see, e.g., Devroye and Gyorfi [117].

Delage and Ye [105] propose to estimate $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ by their empirical estimates. We discuss in Section 7.2 how this choice of nominal parameters, in conjunction with other assumptions, ensure that the constructed ambiguity set $\mathcal{P}^{\mathrm{DY}}$ contains the true unknown probability distribution with a high probability.

7.2 Choice of Robustness Parameters

All discrepancy-based ambiguity sets, studied in Section 6.1, and some of the moment-based ambiguity sets, studied in Section 6.2, rely on parameters that control the size of the ambiguity set. For instance, parameter ϵ in the ambiguity set $\mathcal{P}^{\mathrm{W}}(P_0; \epsilon)$, defined in (32), and parameters ϱ_1 and ϱ_2 in the ambiguity set $\mathcal{P}^{\mathrm{DY}}$, defined in (53), control the size of their corresponding ambiguity sets. A judicious choice of these parameters reduce the level of conservatism of the resulting DRO. A natural question is then how to choose appropriate values for these parameters.

In this section, we review different approaches to choose the level-of-robustness parameters. To have a structured review, we make a distinction between data-driven DROs with i.i.d. data (Section 7.2.1) and non-i.i.d data (Section 7.2.2). We end this section by describing *cross-validation* as a widely used approach in practice.

7.2.1 Data-Driven DROs with i.i.d. Data

Data-driven DROs usually propose a robustness parameter that is inversely proportional to the number of available data points. This construction is motivated from the asymptotic convergence of the optimal value of DRO to that of the corresponding model under the true unknown distribution, with an increasing number of data points, see, e.g., Bertsimas et al. [51], Delage and Ye [105], Pflug and Wozabal [302].

A common underlying assumption in data-driven methods is that data points are independently and identically distributed (i.i.d.) from the unknown distribution. Given this assumption, data-driven approaches for discrepancy-based ambiguity sets propose to choose the level of robustness by analyzing the discrepancy (with respect to some metric) between the empirical distribution and the true unknown distribution[24], asymptotically, see, e.g., Ben-Tal et al. [32], Shafieezadeh-Abadeh et al. [355], or with a finite sample, see, e.g., Pflug and Wozabal [302]. A direct consequence of such analysis is that it establishes a finite-sample probabilistic guarantee on the discrepancy between the empirical distribution and the true unknown distribution. Hence, it gives rise to a probabilistic guarantee on the inclusion of the unknown distribution in the constructed set, with respect to the empirical distribution. By construction, such an ambiguity set can be interpreted as a confidence set on the true unknown distribution. Moreover, such a construction implies a finite-sample guarantee on the out-of-sample performance, so that the current optimal value provides an upper bound on the out-of-sample performance of the current solution with a high probability. A similar idea is used in moment-based ambiguity sets, see, e.g., Goldfarb and Iyengar [161] and Delage and Ye [105]. In a recent work, Gotoh et al. [165] propose to choose the level of robustness by trading off between the mean and variance of the out-of-sample objective function value. We refer the readers to that paper for a review of calibration approaches in DRO.

Below, we review some theoretical results on choosing the level of robustness for a number of models introduced in Section 6. In this section, we suppose that a set $\{\boldsymbol{\xi}^i\}_{i=1}^N$ of i.i.d data, distributed according to $\mathbb{P}^{\mathrm{true}}$, is available. Moreover, $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{\xi}^i}$ denotes the empirical probability distribution of data, where $\delta_{\boldsymbol{\xi}^i}$ denotes the Dirac mass point on $\boldsymbol{\xi}^i$. Also, $\mathbb{P}^N$ is the sampling distribution of $\widehat{\mathbb{P}}_N$.

[24] Some probability metrics, such as Wasserstein metric, metrize the weak convergence Gibbs and Su [153]. That is, the convergence between two probability distributions, with respect to some metric, implies the convergence in probability.

7.2.1.1 Optimal Transport Discrepancy

When the ambiguity set contains all discrete distributions around the empirical distribution in the sense of the Wasserstein metric, Pflug and Wozabal [302] and Pflug et al. [303] propose to choose the level of robustness based on a probabilistic statement on the Wasserstein metric between the empirical and true distributions, due to Dudley [124], as $\epsilon = \frac{CN^{-\frac{1}{d}}}{\alpha}$. This choice of ϵ guarantees that $\mathbb{P}^N\{\mathfrak{d}_c^{\mathrm{W}}(\mathbb{P}, \widehat{\mathbb{P}}_N) \geq \epsilon\} \leq \alpha$. In addition to the confidence level $1-\alpha$ and the number of available data points N, the proposed level of robustness in Pflug et al. [303], Pflug and Wozabal [302] depends on the dimension of $\boldsymbol{\xi}$, d, and a constant C. For such a Wasserstein-based ambiguity set, one can also choose the size of the set by utilizing the probabilistic statement on the discrepancy between empirical distribution and the true unknown distribution, established in Fournier and Guillin [143].

Similarly, when the ambiguity set contains all (discrete and continuous) distributions around the empirical distribution in the sense of the Wasserstein metric, Mohajerin Esfahani and Kuhn [266] present a theoretical result on how to choose the level of robustness. This analysis is also based on the measure concentration property in Fournier and Guillin [143], assuming that the underlying distribution has an exponentially decaying tail. This assumption trivially holds if the support set Ω is compact.

▶ **Theorem 39** (Mohajerin Esfahani and Kuhn [266, Theorem 3.5]). *Suppose that there exists $a > 1$ and $A > 0$ such that $\mathbb{E}_{\mathbb{P}^{true}}[\exp\{c(\boldsymbol{\xi}, \boldsymbol{\xi}_0)\}] \leq A$ for some $\boldsymbol{\xi}_0 \in \mathbb{R}^d$, where $c(\,\cdot\,,\cdot\,)$ denotes the transportation cost in the definition of the optimal transport discrepancy* (31)*. Let $\alpha \in (0, 1]$ and define*

$$\epsilon_N(\alpha) := \begin{cases} \left(\frac{\log(c_1/\alpha)}{c_2 N}\right)^{1/\max\{d,2\}} & \text{if } N \geq \frac{\log(c_1/\alpha)}{c_2}, \\ \left(\frac{\log(c_1/\alpha)}{c_2 N}\right)^{1/a} & \text{if } N < \frac{\log(c_1/\alpha)}{c_2}, \end{cases}$$

where $d \neq 2$ and $c_1, c_2 > 0$ are constants that only depend on a, A, and d. Let $\boldsymbol{x}_N^*$ denote an optimal solution to (8) *with $\mathcal{P} = \mathcal{P}^W\big(\widehat{\mathbb{P}}_N; \epsilon\big)$, where $\epsilon \geq \epsilon_N(\alpha)$. Then,*

$$\mathbb{P}^N\left\{\mathbb{E}_{\mathbb{P}^{true}}[h_0(\boldsymbol{x}_N^*, \boldsymbol{\xi})] \leq \sup_{\mathbb{P} \in \mathcal{P}^W\left(\hat{\mathbb{P}}_N; \epsilon\right)} \mathbb{E}_{\mathbb{P}}[h_0(\boldsymbol{x}_N^*, \boldsymbol{\xi})]\right\} \geq 1 - \alpha.$$

Using the above result, Mohajerin Esfahani and Kuhn [266, Theorem 3.6] implies that under proper assumptions, if the significance level α_N converges to 0 at a judiciously chosen rate, e.g., $\alpha_N = \exp(-\sqrt{N})$, then the optimal value to (8) with $\mathcal{P} = \mathcal{P}^{\mathrm{W}}\big(\widehat{\mathbb{P}}_N; \epsilon_N(\alpha_N)\big)$ converges to that of (1) under the true distribution. A similar asymptotic consistency holds for the corresponding data-driven optimal solution.

Observe that because all the utilized probabilistic statements stated so far rely on some exogenous constants C, even if they can be computed, the size of the resulting ambiguity set calculated from the theoretical analysis may be very conservative; hence, such proposals are not practical. By acknowledging the issue raised above, some researchers propose to choose the level of robustness without relying on exogenous constants. For cases that the ambiguity set contains all discrete distributions, supported on a compact space and around the empirical distribution, Ji and Lejeune [211] derive a closed-form expression for computing the size of the Wasserstein-based ambiguity set.

▶ **Theorem 40** (Ji and Lejeune [211, Theorem 2]). *Suppose that the random vector $\boldsymbol{\xi}$ is supported on a finite Polish space (Ω, d), where $\Omega \subseteq \mathbb{R}^d$ and $d(\,\cdot\,,\cdot\,)$ is the ℓ_1-norm. Choose $c(\,\cdot\,,\cdot\,) = d(\,\cdot\,,\cdot\,)$ in the definition of the optimal transport discrepancy* (31)*. For some $\boldsymbol{\xi}_0$, assume that $\log \mathbb{E}_{\mathbb{P}^{true}}[\exp\{\lambda d(\boldsymbol{\xi}, \boldsymbol{\xi}_0)\}] < \infty, \forall\ \lambda > 0$. Let $\theta := \sup\{d(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) : \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \Omega\}$ be the diameter of Ω. Then,*

$$\mathbb{P}^N\{\mathfrak{d}_d^W(\mathbb{P}^{true}, \widehat{\mathbb{P}}_N) \leq \epsilon\} \geq 1 - \exp\left\{-N\left(\frac{\sqrt{4\epsilon(4\theta+3) + (4\theta+3)^2}}{4\theta+3} - 1\right)^2\right\}.$$

Moreover, if

$$\epsilon \geq \left(\theta + \frac{3}{4}\right)\left(-\frac{1}{N}\log\alpha + 2\sqrt{-\frac{1}{N}\log\alpha}\right),$$

then

$$\mathbb{P}^N\{\mathfrak{d}_d^W(\mathbb{P}^{true}, \widehat{\mathbb{P}}_N) \leq \epsilon\} \geq 1 - \alpha.$$

Unlike the result in Pflug and Wozabal [302], the proposed level of robustness in Ji and Lejeune [211], stated in Theorem 40, depends only on the confidence level α, the number of available data points, and the diameter of the compact support Ω. Ji and Lejeune [211] obtain this result by bounding the Wasserstein distance between two probability distributions from above, using the properties of the weighted total variation Bolley and Villani [68], and the weighted Csiszar–Kullback–Pinsker inequality Villani [403], and consequently applying Sanov's large deviation theorem Dembo and Zeitouni [108] to reach a probabilistic statement on the Wasserstein distance between two distributions. As stated in Theorem 40, such a result guarantees that the constructed set contains the unknown probability distribution with a high probability. Moreover, it implies a probabilistic guarantee on the true optimal value.

Another criticism of methods such as those proposed in Pflug and Wozabal [302] and Pflug et al. [303] is that they merely rely on the discrepancy between two probability distributions, and the optimization framework plays no role in the prescription. By making connection between the regularizer parameter and the size of the ambiguity for Wassersetin-based sets, Blanchet et al. [64] aim to optimally choose the regularization parameter. A key component of their analysis is a *robust Wasserstein profile* (RWP) function. At a given solution $\boldsymbol{x}$, this function calculates the minimum Wasserstein distance from the nominal distribution to the set of optimal probability distributions for the inner problem at $\boldsymbol{x}$. For any confidence level α, they show that the size of the ambiguity set should be chosen as $(1-\alpha)$-quantile of RWP at the optimal solution to the minimization problem under the true unknown distribution. Using this selection of ϵ, the optimal solution to the true problem belongs to the set of optimal solutions to the DRO problem, with $(1-\alpha)$ confidence for all $\mathbb{P} \in \mathcal{P}^{\mathrm{W}}(\mathbb{P}_N, \epsilon)$. As such a result is based on the true optimal solution, they study the asymptotic behavior of the RWP function and discuss how to use it to optimally choose the regularization parameter without cross validation. The work in Blanchet et al. [64] is extended in Blanchet and Kang [59, 61]. Blanchet and Kang [59] utilize the RWP function to introduce a data-driven (statistical) criterion for the optimal choice of the regularization parameter and study its asymptotic behavior. For a DRO approach to linear regression, Chen and Paschalidis [86] give guidance on the selection of the regularization parameter from the standpoint of a confidence region.

7.2.1.2 Goodness-of-Fit Test

Bertsimas et al. [51] propose to form the ambiguity set of distributions using the confidence set of the unknown distribution via goodness-of-fit tests. With such an approach, one chooses the level of robustness as the threshold value of the corresponding test, depending on the confidence level α, data, and the null hypothesis.

7.2.1.3 ϕ-Divergences

By noting that the class of ϕ-divergences can be used in statistical hypothesis tests, a similar approach to the one in Bertsimas et al. [51] can be used to choose the level of robustness for ϕ-divergence-based ambiguity sets. For the case that the distributional ambiguity in discrete distributions is modeled via ϕ-divergences, some papers propose to choose the level of robustness by relying on the asymptotic behavior of the discrepancy between the empirical distribution and true unknown distribution, see, e.g., Bayraksan and Love [17], Ben-Tal et al. [32], Yanıkoğlu and den Hertog [431].

Suppose that Ξ is finite sample space of size m and the ϕ-divergence function in (40) is twice continuously differentiable in a neighborhood of 1, with $\phi''(1) > 0$. It is shown in Pardo [293] that under the true distribution, the statistics $\frac{2N}{\phi''(1)}\mathfrak{d}^{\phi}(\mathbb{P}^{\text{true}}, \widehat{\mathbb{P}_N})$ converges in distribution to a χ^2_{m-1}-distribution, with $m-1$ degrees of freedom. Thus, at a given confidence level α, one can set the level of robustness to $\frac{\phi''(1)}{2N}\chi^2_{m-1,1-\alpha}$, where $\chi^2_{m-1,1-\alpha}$ is the $(1-\alpha)$-quantile of χ^2_{m-1}, to obtain an (approximate) confidence set on the true unknown distribution. Ben-Tal et al. [32] show that such a choice of the level of robustness gives a one-sided confidence interval with (asymptotically) inexact coverage on the true optimal value of $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{\text{true}}}[h_0(\boldsymbol{x}, \boldsymbol{\xi})]$. For corrections for small sample sizes, we refer readers to Pardo [293]. Lam and Zhou [232, 233] show that by using a χ^2-distribution with a smaller degree of freedom than $m-1$, one can obtain asymptotically valid coverage, including cases where the objective and constraints are both stochastic and observed from data. Additionally, Lam [230] shows that in general, one needs to use the supremum of a χ^2-process, instead of only a χ^2 random variable, to obtain asymptotically valid coverage of an expectation constraint in a non-conservative fashion.

By generalizing the empirical likelihood framework Owen [291] on a separable metric space (not necessarily finite), Duchi et al. [123] propose to choose the level of robustness ϵ such that a confidence interval $[l_N, u_N]$ on

the true optimal value of $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{\text{true}}}[h_0(\boldsymbol{x},\boldsymbol{\xi})]$ has an asymptotically exact coverage $1-\alpha$, i.e.,

$$\lim_{N\to\infty} \mathbb{P}^N \left\{ \inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{\text{true}}}[h_0(\boldsymbol{x},\boldsymbol{\xi})] \in [l_N, u_N] \right\} = 1-\alpha,$$

where

$$u_N := \inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}^{\phi}(\hat{\mathbb{P}}_N;\epsilon)} \mathbb{E}_{\mathbb{P}}[h_0(\boldsymbol{x},\boldsymbol{\xi})]$$

and

$$l_N := \inf_{\boldsymbol{x}\in\mathcal{X}} \inf_{\mathbb{P}\in\mathcal{P}^{\phi}(\hat{\mathbb{P}}_N;\epsilon)} \mathbb{E}_{\mathbb{P}}[h_0(\boldsymbol{x},\boldsymbol{\xi})].$$

▶ **Theorem 41** (Duchi et al. [123, Theorem 4])**.** *Suppose that the ϕ function is three time continuously differentiable in a neighborhood of 1, and normalized with $\phi(1)=\phi'(1)=0$[25] and $\phi''(1)=2$. Furthermore, suppose that $\mathcal{X}$ is compact, there exists a measurable function $M:\Omega\mapsto\mathbb{R}_+$ such that for all $\boldsymbol{\xi}\in\Omega$, $h(\cdot,\boldsymbol{\xi})$ is $M(\boldsymbol{\xi})$-Lipschitz with respect to some norm $\|\cdot\|$ on $\mathcal{X}$, $\mathbb{E}_{\mathbb{P}^{true}}\left[M(\boldsymbol{\xi})^2\right]<\infty$, and $\mathbb{E}_{\mathbb{P}^{true}}[|h_0(\boldsymbol{x}_0,\boldsymbol{\xi})|]<\infty$ for some $\boldsymbol{x}_0\in\mathcal{X}$. Additionally, suppose that $h(\cdot,\boldsymbol{\xi})$ is proper and lower semicontinuous for $\boldsymbol{\xi}$, $\mathbb{P}^{true}$-almost surely. If $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{true}}[h_0(\boldsymbol{x},\boldsymbol{\xi})]$ has a unique solution, then*

$$\lim_{n\to\infty} \mathbb{P}^N \left\{ \inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{true}}[h_0(\boldsymbol{x},\boldsymbol{\xi})] \le u_N \right\} = 1-\frac{1}{2}P(\chi_1^2 \ge N\epsilon)$$

and

$$\lim_{n\to\infty} \mathbb{P}^N \left\{ \inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{true}}[h_0(\boldsymbol{x},\boldsymbol{\xi})] \ge l_N \right\} = 1-\frac{1}{2}P(\chi_1^2 \ge N\epsilon).$$

According to Theorem 41, if $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{\text{true}}}[h_0(\boldsymbol{x},\boldsymbol{\xi})]$ has a unique solution, the desired asymptotic guarantee is achieved with the choice $\epsilon=\frac{\chi^2_{1,1-\alpha}}{N}$. Duchi et al. [123] also give rates at which $u_N-l_N\to 0$. Moreover, the upper confidence interval $(-\infty,u_N]$ is a one-sided confidence interval with an asymptotic exact coverage when $\epsilon=\chi^2_{1,1-2\alpha}$.

On another note, it can be seen from Table 3 that the ϕ-divergence function corresponding to the variation distance is not twice differentiable at 1. Hence, one cannot use the above result. However, by utilizing the first inequality in Lemma 26, i.e., the relationship between the variation distance and the Hellinger distance, Jiang and Guan [214] propose to set the level of robustness to $\sqrt{\frac{1}{N}\chi^2_{m-1,1-\alpha}}$ in order to obtain an (approximate) confidence set on the true unknown discrete distribution. The proposed choice of the level of robustness ensures that the unknown discrete distribution belongs to the ambiguity set with a high probability. For the case that $\boldsymbol{\xi}$ follows a continuous distribution, the proposed level of robustness in Jiang and Guan [214] depends on some constants that appear in the probabilistic statement of the discrepancy between the empirical distributions and the true distribution.

7.2.1.4 ℓ_p-Norm

For the case that ℓ_∞-norm is used to model the distributional ambiguity, Jiang and Guan [214] propose to choose the level of robustness based on a probabilistic statement on the discrepancy between the empirical distributions and the true distribution as $\epsilon=\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{N}}\max_{i=1}^m\sqrt{p_i(1-p_i)}$, where $z_{1-\frac{\alpha}{2}}$ represents the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. The proposed choice of the level of robustness ensures that the unknown discrete distribution belongs to the ambiguity set with a high probability. Similar to the ℓ_1-norm (i.e., the variation distance) case, when $\boldsymbol{\xi}$ follows a continuous distribution, the proposed level of robustness depends on some constants that appear in the probabilistic statement of the discrepancy between the empirical distributions and the true distribution.

[25] As in the definition of ϕ-divergence, the assumptions $\phi(1)=\phi'(1)=0$ are without loss of generality because the function $\psi(t)=\phi(t)-\phi'(1)(t-1)$ yields identical discrepancy measure to ϕ Pardo [293]

7.2.1.5 ζ-Structure

By exploiting the relationship between different metrics in the ζ-structure family, see, e.g., Lemma 30, Zhao and Guan [442] provide guidelines on how to choose the level of robustness for the ambiguity sets of the unknown discrete distribution formed via bounded Lipschitz, Kantorovich, and Fortet–Mourier metrics as follows.

▶ **Theorem 42.** *Suppose that the random vector $\boldsymbol{\xi}$ is supported on a bounded finite space Ω and θ denotes the diameter of Ω, as defined in Theorem 40.*

1. *if $\epsilon \geq \theta\sqrt{-2\frac{\log\alpha}{N}}$, then $\mathbb{P}^N\{\mathfrak{d}^K(\mathbb{P}^{true}, \widehat{\mathbb{P}}_N) \leq \epsilon\} \geq 1-\alpha$ and $\mathbb{P}^N\{\mathfrak{d}^{BL}(\mathbb{P}^{true}, \widehat{\mathbb{P}}_N) \leq \epsilon\} \geq 1-\alpha$.*
2. *if $\epsilon \geq \theta \max\{1, \theta^{q-1}\}\sqrt{-2\frac{\log\alpha}{N}}$, then $\mathbb{P}^N\{\mathfrak{d}^{FM}(\mathbb{P}^{true}, \widehat{\mathbb{P}}_N) \leq \epsilon\} \geq 1-\alpha$.*

Proof. See Appendix A. ◀

As it can be seen from Theorem 42, the proposed levels of robustness for the case that the unknown distribution is discrete depend on the diameter of Ω, the number of data points N, and the confidence level $1-\alpha$. However, the results in Zhao and Guan [442] for the continuous case suffer from similar practical issues as in Jiang and Guan [214], Pflug et al. [303], Pflug and Wozabal [302].

7.2.1.6 Chebyshev

A data-driven approach to construct a Chebyshev ambiguity set is proposed in Goldfarb and Iyengar [161]. Recall the linear model for the asset returns $\boldsymbol{\xi}$ in Goldfarb and Iyengar [161]: $\boldsymbol{\xi} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{f} + \boldsymbol{\epsilon}$, where $\boldsymbol{\mu}$ is the vector of mean returns, $\boldsymbol{f} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ is the vector of random returns that derives the market, $\boldsymbol{A}$ is the factor loading matrix, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{B})$ is the vector of residual returns with a diagonal matrix $\boldsymbol{B}$. Under the assumption that the covariance matrix $\boldsymbol{\Sigma}$ is known, Goldfarb and Iyengar [161] study three different models to form the uncertainty in $\boldsymbol{B}$, $\boldsymbol{A}$, and $\boldsymbol{\mu}$ as follows:

$$\mathcal{U}_{\boldsymbol{B}} = \left\{\boldsymbol{B} \,\middle|\, \boldsymbol{B} = \operatorname{diag}(\boldsymbol{b}),\ b_i \in [\underline{b}_i, \overline{b}_i],\ i = 1, \ldots, d\right\},$$
$$\mathcal{U}_{\boldsymbol{A}} = \left\{\boldsymbol{A} \,|\, \boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{C},\ \|\boldsymbol{c}_i\|_g \leq \rho_i,\ i = 1, \ldots, d\right\},$$
$$\mathcal{U}_{\boldsymbol{\mu}} = \left\{\boldsymbol{\mu} \,|\, \boldsymbol{\mu} = \boldsymbol{\mu}_0 + \boldsymbol{\zeta},\ |\zeta_i| \leq \gamma_i,\ i = 1, \ldots, d\right\},$$

where $\boldsymbol{c}_i$ denotes the i-th column of $\boldsymbol{C}$, and $\|\boldsymbol{c}_i\|_g = \sqrt{\boldsymbol{c}_i^\top \boldsymbol{G} \boldsymbol{c}_i^\top}$ denotes the elliptic norm of $\boldsymbol{c}_i$ with respect to a symmetric positive definite matrix $\boldsymbol{G}$. Calibrating the uncertainty sets $\mathcal{U}_{\boldsymbol{B}}$, $\mathcal{U}_{\boldsymbol{A}}$, and $\mathcal{U}_{\boldsymbol{\mu}}$ involves choosing parameters $\underline{b}_i$, $\overline{b}_i$, ρ_i, γ_i, $i = 1, \ldots, d$, vector $\boldsymbol{\mu}_0$, and matrices $\boldsymbol{A}_0$ and $\boldsymbol{G}$. Assuming that a set of data points is available on $\boldsymbol{\xi}$ and $\boldsymbol{f}$, by relying on the multivariate linear regression, Goldfarb and Iyengar [161] obtain least square estimates $(\boldsymbol{\mu}_0, \boldsymbol{A}_0)$ of $(\boldsymbol{\mu}, \boldsymbol{A})$, respectively, and construct a multidimensional confidence region of $(\boldsymbol{\mu}, \boldsymbol{A})$ around $(\boldsymbol{\mu}_0, \boldsymbol{A}_0)$. Now, projecting this confidence region along vector $\boldsymbol{\mu}$ and matrix $\boldsymbol{A}$ gives the corresponding uncertainty sets $\mathcal{U}_{\boldsymbol{\mu}}$ and $\mathcal{U}_{\boldsymbol{A}}$, respectively. To form the uncertainty set $\mathcal{U}_{\boldsymbol{B}}$, they propose to use a bootstrap confidence interval around the regression error of the residual.

7.2.1.7 Ellipsoid and Matrix Inequality

Data-driven methods to construct the ambiguity set $\mathcal{P}^{\mathrm{DY}}$ is proposed in Delage and Ye [105].

▶ **Theorem 43** (Delage and Ye [105, Corollary 4])**.** *Suppose that the random vector $\boldsymbol{\xi}$ is supported on a bounded space Ω. Consider the following parameters:*

$$\widehat{\boldsymbol{\mu}}_0 = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\xi}^i,$$
$$\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{N-1}\sum_{i=1}^{N} (\boldsymbol{\xi}^i - \widehat{\boldsymbol{\mu}}_0)(\boldsymbol{\xi}^i - \widehat{\boldsymbol{\mu}}_0)^\top,$$
$$\widehat{\theta} = \sup_{i=1}^{N} \left\| \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} (\boldsymbol{\xi}^i - \widehat{\boldsymbol{\mu}}_0) \right\|_2,$$

where $\widehat{\boldsymbol{\mu}}_0$, $\widehat{\boldsymbol{\Sigma}}_0$, *and* $\widehat{\theta}$ *are estimates of the mean, covariance, and diameter of the support of* $\boldsymbol{\xi}$, *respectively. Moreover, for a confidence level* $1-\alpha$, *let us define*

$$\bar{\theta} = \left(1 - (\widehat{\theta}^2+2)\frac{2+\sqrt{2\log(\frac{4}{\bar{\alpha}})}}{\sqrt{N}}\right)^{-\frac{1}{2}}\widehat{\theta},$$
$$\bar{\gamma}_1 = \frac{\bar{\theta}^2}{\sqrt{N}}\left(\sqrt{1-\frac{d}{\bar{\theta}^4}}+\sqrt{\log\left(\frac{4}{\bar{\alpha}}\right)}\right)$$
$$\bar{\gamma}_2 = \frac{\bar{\theta}^2}{N}\left(2+\sqrt{2\log\left(\frac{2}{\bar{\alpha}}\right)}\right),$$
$$\bar{\varrho}_1 = \frac{\bar{\gamma}_2}{1-\bar{\gamma}_1-\bar{\gamma}_2},$$
$$\bar{\varrho}_2 = \frac{1+\bar{\gamma}_2}{1-\bar{\gamma}_1-\bar{\gamma}_2},$$

where $\bar{\alpha} = 1-\sqrt{1-\alpha}$. *Let* $\mathcal{P}^{DY}(\Omega, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_0, \bar{\varrho}_1, \bar{\varrho}_2)$ *be the ambiguity set formed via* (53), *using parameters* $\widehat{\boldsymbol{\mu}}_0$, $\widehat{\boldsymbol{\Sigma}}_0$, $\bar{\varrho}_1$, *and* $\bar{\varrho}_2$. *Then, we have*

$$\mathbb{P}^N\left\{\mathbb{P}^{true} \in \mathcal{P}^{DY}(\Omega, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_0, \bar{\varrho}_1, \bar{\varrho}_2)\right\} \geq 1-\alpha.$$

7.2.2 Data-Driven DROs with non-i.i.d. Data

As mentioned before, data-driven DRO models typically assume that a set of i.i.d. sampled data is available from the unknown true distribution. In many situations, however, there is no guarantee that the future uncertainty is drawn from the same distribution, see, e.g., Glasserman and Yang [157], Lam [229] in the context of the model uncertainty, robust control and robust risk analysis. Recognizing this fact, some research is devoted to choosing the level of robustness in situations where the i.i.d. assumption is violated and data-driven methods to calibrate the level of robustness may be unsuitable. Rahimian et al. [328] use the notions of maximal effective subsets and prices of optimism/pessimism and nominal/worst-case regrets to calibrate the level of robustness in discrepancy-based DRO models. Price of optimism is defined as the loss by being too optimistic (i.e., using SO model with the nominal distribution)—and hence, implementing the corresponding solution—while DRO accurately represents the ambiguity in the distribution. Similarly, the price of pessimism is defined as the loss by being too pessimistic (i.e., using RO model with no distributional information except for the support of uncertainty). Nominal/worst-case regret is defined as the loss of being unnecessarily ambiguous/not being ambiguous enough—and hence, implementing the corresponding solution—while DRO is ill-calibrated. Rahimian et al. [328] suggest to balance the price of optimism and pessimism if the decision-maker is indifferent regarding the error from using too optimistic or pessimistic solutions. They refer to the smallest level of robustness for which such a balance happens as *indifferent-to-solution* level of robustness. On the other hand, Rahimian et al. [328] propose to balance the nominal and worst-case regrets if the decision-maker wants to be indifferent regarding the error from using an ill-calibrated DRO model in either the optimistic or the pessimistic scenarios. They refer to the smallest level of robustness for which such a balance happens as *indifferent-to-distribution* level of robustness.

7.3 Cross-Validation

In this section, we describe a practical approach to choose the level-of-robustness parameter for a DRO model. Suppose that ϵ denotes a generic level-of-robustness parameter. Recall the discussion in Section 2, and let $\boldsymbol{x}_N^*(\epsilon)$ be an optimal data-driven solution, obtained by solving a data-driven DRO model using parameter ϵ. With regards to (12), one would ideally select an optimal ϵ^* such that it would minimize $\widehat{V}_N\big(\boldsymbol{x}_N^*(\epsilon)\big) := \mathcal{R}_{\mathbb{P}^{\text{true}}}\left[h_0(\boldsymbol{x}_N^*(\epsilon), \boldsymbol{\xi})\right]$ over all $\epsilon \geq 0$. As $\mathbb{P}^{\text{true}}$ is unknown, it is not possible to calculate $\widehat{V}_N\big(\boldsymbol{x}_N^*(\epsilon)\big)$. However, one can estimate the best choice of ϵ^* (possibly from a finite number of candidates $\mathcal{E}$).

A commonly used approach to choose the level-of-robustness parameter is *cross-validation* Friedman et al. [144]. Below, we describe K-fold cross-validation, which is widely used in situation where there is not enough data and the use of a DRO model is justified:

- Split data into K roughly equal-sized subsets (at random),
- For each fold $k = 1, \ldots, K$,

- Use exactly one subset as the test and merge the remaining $K-1$ subsets into the training set,
- Use the training set to solve a data-driven DRO, formed via $\epsilon \in \mathcal{E}$, and obtain $\boldsymbol{x}_N^*(\epsilon)$,
- Use the test set to estimate the out-of-sample performance $\widehat{V}_N\big(\boldsymbol{x}_N^*(\epsilon)\big)$, $\epsilon \in \mathcal{E}$,
- Choose an ϵ_k that minimizes $\widehat{V}_N\big(\boldsymbol{x}_N^*(\epsilon)\big)$ over all choices in $\mathcal{E}$,

- Choose $\widehat{\epsilon} = \frac{1}{K}\sum_{k=1}^{K} \epsilon_k$ as an estimate of ϵ^*.

Now, using $\widehat{\epsilon}$, one can solve a data-driven DRO and obtain an optimal data-driven solution.

8 Modeling Toolboxes

In recent years, several open-source tools have been developed to handle RO/DRO problems. These tools support an algebraic modeling interface to enter the problem and uncertainty/ambiguity set. In addition, they provide tools to obtain an exact or approximate robust/distributional robust reformulation. They also connect to the existing open-source or commercial solvers to solve the resulting reformulation. Goh and Sim [160] develop a MATLAB-based algebraic modeling toolbox, named ROME (Robust Optimization Made Easy), for a class of DRO problems with conic-representable sets for the support and mean, known covariance matrix, and upper bounds on the directional deviations studied in Goh and Sim [159]. Goh and Sim [160] elucidate the practicability of this toolbox in the context several application domains. A C++-based algebraic modeling package, named ROC, is developed in Bertsimas et al. [53], to demonstrate the practicability and scalability of the studied adaptive DRO model. Some features of ROC include declaration of uncertain parameters and linear decision rules, transcriptions of ambiguity sets, and reformulation of DRO using the results obtained in Bertsimas et al. [53]. A brief introduction to ROC and some illustrative examples to declare the objects of a model, such as variables, constraints, ambiguity set, among others, are given in an early version of Bertsimas et al. [50]. XProg (`http://xprog.weebly.com`), is a MATLAB-based algebraic modeling package that also implements the proposed model in Bertsimas et al. [53]. Chen et al. [98] develop a MATLAB-based algebraic modeling package, named RSOME (Robust Stochastic Optimization Made Easy), to illustrate the modeling power of their proposed ambiguity set. RSOME supports more general ambiguity sets than ROME and is capable of handling static and multistage DRO problems. A Python version of RSOME is presented in Chen and Xiong [95].

There are also some other packages to handle RO problems. A C++-based algebraic modeling language, named ROC++, is presented in Vayanos et al. [401]. ROC++ supports static and multistage problems involving both exogenous and endogenous uncertain parameters. A Python package, named ROmodel, is also developed in Wiebe and Misener [408] to handle RO problems within Pyomo (Hart et al. [181]). We also refer the readers to Pyros (Isenberg et al. [208]), standing for Pyomo Robust Optimization Solver, PICOS (Sagnol and Stahlberg [349]), standing for Python Interface to Conic Optimization Solvers, and JuMPeR (Dunning [125]), standing for JuMP extension for Robust) in JuMP (Dunning et al. [126]).

9 Conclusion and Future Research Directions

This paper provided an overview over the modeling paradigm DRO. Starting from Scarf's seminal work Scarf [351] and till early 2000s, there had been several papers on this paradigm, mainly under the name of minimax or ambiguous stochastic optimization. However, there has been an increasing attention on this paradigm since late 2000s, mostly under the widely dominant term "distributionally robust optimization" (Calafiore and El Ghaoui [74], Delage and Ye [105]). DRO is often described as a modeling approach that seeks a trade-off between stochastic optimization, assuming full distributional information, and robust optimization, assuming no distributional information except for the support of the uncertain parameters. By modeling the distributional ambiguity in a way that respects statistical and/or structural properties of the underlying unknown distribution, the DRO approach seeks decisions that are immune with respect to the distributional parameter uncertainty due to limited observability of data, noisy measurements, implementation and prediction errors.

Recognizing the interest in DRO within operations research and machine learning communities, this paper provided a holistic view that connects DRO to other widely studied concepts in these communities. These include concepts such as game theory, risk aversion, chance constraint, robust optimization, and function regularization. We explained general solution techniques to solve DRO models, and discussed several models to express the distributional ambiguity as well as calibration of the resulting ambiguity sets. We also provided an overview of efforts to unify modeling approaches and to develop software packages with promising capabilities.

We believe that the success of DRO is made possible by recent advances in SDP, SOCP, integer programming, numerical optimization algorithms, among others. Despite extensive research on modeling, theoretical, and computational aspects of DRO, from reformulation, to customized solution approaches, to statistical properties of the resulting decisions, we envision several research directions that are to be tackled in the near future to make DRO a more appealing modeling paradigm for decision making and estimation:

- **Randomized policies.** Decision-making problems within the DRO literature usually aim for an optimal deterministic policies. As DRO problems may be interpreted as a minimax game between the decision maker and an adversary, randomized policies may have the potential to exhibit a better out-of-sample performance compared to commonly considered deterministic policies for several applications, as witnessed in Delage et al. [106], Delage and Saif [104].
- **Sequential decision-making.** Most of the DRO literature focuses on static and two-stage setting, with the exception of a few papers that study the multistage setting, see, e.g., Bertsimas et al. [55], Duque and Morton [131], Philpott et al. [304], Pichler and Shapiro [306], Rahimian et al. [329], Shapiro [368], Yu and Shen [435]. Distributional robustness is also studied in the context of reinforcement learning, see, e.g., Derman and Mannor [116], Smirnova et al. [379], Zhou et al. [446]. As most stochastic decision-making problems are dynamic in nature, with information being revealed over time and decisions made sequentially given the available information, further research on sequential decision-making under distributional ambiguity would make DRO more applicable in practical setting.
- **Decision-making with side information.** In statistical learning, DRO is usually performed with the hope of achieving a predictive model (between the independent and dependent variables) with reasonably acceptable generalization properties (Blanchet et al. [66]). On the other hand, decision-making models usually ignore the side information in the optimization framework and create an offline predictive model between the independent and dependent variables, whose outputs are to be used indirectly in the optimization framework. With the abundance of historical data and access to side information when decision-making, we envision the modeling framework of conditional stochastic optimization (Ban and Rudin [11], Bertsimas and Kallus [41], Kannan et al. [217]) and their distributionally robust versions (Bertsimas et al. [54], Bertsimas and Van Parys [44], Esteban-Pérez and Morales [139], Kannan et al. [218], Nguyen et al. [283]) receive an increasing attention in theory and practice.
- **Model-free distributional robustness.** Most of the DRO literature is model-based, in the sense that a "model" to express the distributional ambiguity is selected a priori and the parameters of the ambiguity sets are calibrated. This approach would raise the question of whether the selected ambiguity set is "appropriate" or there are other types of ambiguity sets that should be chosen. Even if an appropriate ambiguity set is selected, finding appropriate values for its parameters remains challenging. There are several efforts in the DRO literature that are model-free and instead, specify an acceptable target cost or regret compared to a baseline decision-maker, see, e.g., Bennouna and Van Parys [35], Sutter et al. [389], Van Parys [395], Van Parys et al. [399]. This research is close in spirits to Long et al. [250], Ramachandra et al. [330], Sim et al. [375], and we expect that it will continue to receive much attention in the years to come.
- **Globalized and soft robustness.** Two core characteristics of RO are its constraint-wise nature and the assumption that no constraint can be violated for any scenarios in the uncertainty set. To reduce conservatism of RO, these assumptions are relaxed by "light robustness" (Fischetti and Monaci [142]), "globalized robustness" (Ben-Tal et al. [29, 34]), and "soft robustness" (Ben-Tal et al. [31]). The idea of "globalization" is also extended to the DRO setting see, e.g., Ding et al. [120], Li and Xing [244], Liu et al. [247], and we envision these less conservative alternatives receive much attention in the future.
- **Emerging real-world applications.** To develop theoretical and computational results, most DRO papers focus on synthesized and stylized problems, although in different application domains. These include finance (Pflug and Wozabal [302]), environment and energy systems (Park and Bayraksan [294], Zhao and Jiang [444]), scheduling and project management (Jiang et al. [216], Natarajan et al. [276]), and network flow and transportation (Ahipaşaoğlu et al. [2], Carlsson et al. [77]). Resorting to real data, further development of data-driven DRO models in emerging real-world applications and public good areas, such as humanitarian logistics and operations management (Lu and Shen [257]), is interesting.

Appendix: Proofs and Further Discussions

All the emitted proofs are presented in Section A. We also describe kernel-based models in RO in Section B.

A Proofs

Proof of Proposition 3. We have that

$$\begin{aligned}\mathbb{E}_{\mathbb{P}^N}\left[\mathcal{R}_{\hat{\mathbb{P}}_N}\left[h_0(\boldsymbol{x}_N^*,\boldsymbol{\xi})\right]\right] &= \mathbb{E}_{\mathbb{P}^N}\left[\inf_{\boldsymbol{x}\in\mathcal{X}}\mathcal{R}_{\hat{\mathbb{P}}_N}\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right]\right] \leq \inf_{\boldsymbol{x}\in\mathcal{X}}\mathbb{E}_{\mathbb{P}^N}\left[\mathcal{R}_{\hat{\mathbb{P}}_N}\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right]\right]\\ &\leq \inf_{\boldsymbol{x}\in\mathcal{X}}\mathcal{R}_{\mathbb{P}^{\text{true}}}\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right] \leq \mathcal{R}_{\mathbb{P}^{\text{true}}}\left[h_0(\boldsymbol{x}_N^*,\boldsymbol{\xi})\right],\end{aligned}$$

where the second inequality is due to the hypothesis and the third inequality is due to the suboptimality of $\boldsymbol{x}_N^*$. ◀

Proof of Theorem 8. First note that $\mathcal{Z}$ is a Banach space, paired with the dual space $\mathcal{Z}^*$, which is also a Banach space. Then, by a similar proof to Shapiro et al. [374, Theorem 6.7], we can show that if ρ is a proper and lower semicontinuous coherent risk measure, then (13) holds when $\mathcal{M}$ is equal to the subdifferential of ρ at $0\in\mathcal{Z}$, i.e., $\mathcal{M}=\partial\rho(0)$, where

$$\partial\rho(Z) = \arg\max_{P\in\mathcal{M}}\mathbb{E}_P\left[Z\right].$$

Now, we show that ρ is a proper and lower semicontinuous coherent risk measure. Consider the cone $\mathcal{C}\subset\mathcal{Z}$ of nonnegative functions Z. This cone is closed, convex, and pointed, and it defines a partial order relation on $\mathcal{Z}$ that $Z\geq Z'$ if and only if $Z(s)\geq Z'(s)$ a.e. on Ξ. We let the least upper bound of Z,Z' be $Z\vee Z'$, where $(Z\vee Z')(s)=\max\{Z(s),Z'(s)\}$. It follows that $\mathcal{Z}$ with cone $\mathcal{C}$ forms a Banach lattice[26]. Thus, by Shapiro et al. [374, Theorem 7.91], we conclude that ρ is continuous and subdifferentiable on the interior of its domain. This, in turns, implies that the lower semicontinuity of ρ is automatically satisfied. Moreover, by Shapiro et al. [374, Theorem 7.85], the subdifferentials of ρ at any point form a nonempty, convex, and weakly* compact subset of $\mathcal{Z}^*$. In particular, $\mathcal{M}=\partial\rho(0)$ is a convex and weakly* compact set $\mathcal{M}\subseteq\mathfrak{M}(\Xi,\mathcal{F})$.

Conversely, suppose that (13) holds with the set $\mathcal{M}$ being a convex and weakly* compact subset of $\mathfrak{M}(\Xi,\mathcal{F})$. Then, ρ is a real-valued coherent risk measure.

To prove the last part notice that for any $Z\in\mathcal{Z}$, we have $\rho(Z)\geq\rho(0)+\mathbb{E}_P\left[Z-0\right]$, for all $P\in\partial\rho(0)$. Now, by the facts that $\mathcal{M}=\partial\rho(0)$ and $\rho(0)=0$, we conclude $\mathcal{M}=\{P\in\mathfrak{M}(\Xi,\mathcal{F})\,|\,\mathbb{E}_P\left[Z\right]\leq\rho(Z),\ \forall\ Z\in\mathcal{Z}\}$. ◀

Proof of Lemma 17. Problems (DRO) and (18) can be reformulated, respectively, as $\min\{\theta\,|\,(\boldsymbol{x},\theta)\in\mathcal{G}\}$ and $\min\{\theta\,|\,(\boldsymbol{x},\theta)\in\mathcal{G}'\}$, where

$$\mathcal{G}:=\left\{(\boldsymbol{x},\theta)\in\mathbb{R}^{n+1}\,\middle|\,\boldsymbol{x}\in\mathcal{X},\mathcal{R}_P\Big[h_0(\boldsymbol{x},\boldsymbol{\xi})\Big]\leq\theta,\mathcal{R}_P\Big[h_j(\boldsymbol{x},\boldsymbol{\xi})\Big]\leq 0,\forall\ P\in\mathcal{P},j\in[m]\right\},$$

and

$$\mathcal{G}':=\left\{(\boldsymbol{x},\theta)\in\mathbb{R}^{n+1}\,\middle|\,\boldsymbol{x}\in\mathcal{X},\mathcal{R}_P\Big[h_0(\boldsymbol{x},\boldsymbol{\xi})\Big]\leq\theta,\mathcal{R}_P\Big[h_j(\boldsymbol{x},\boldsymbol{\xi})\Big]\leq 0,\forall\ P\in\text{conv}(\mathcal{P}),j\in[m]\right\}.$$

Because $\mathcal{P}\subseteq\text{conv}(\mathcal{P})$, we have $\mathcal{G}'\subseteq\mathcal{G}$. We now show that $\mathcal{G}\subseteq\mathcal{G}'$. Consider an arbitrary $(\boldsymbol{x},\theta)\in\mathcal{G}$. For an arbitrary $P\in\text{conv}(\mathcal{P})$, there exists a collection $\{P^i\}_{i\in\mathcal{I}}$ such that $P=\sum_{i\in\mathcal{I}}\lambda^iP^i$, where $\sum_{i\in\mathcal{I}}\lambda^i=1$, $P^i\in\mathcal{P}$, $\lambda^i\geq 0$, $i\in\mathcal{I}$. Now, by the convexity of $\mathcal{R}_P[\cdot]$ in P on $\mathfrak{M}(\Xi,\mathcal{F})$, we have $\mathcal{R}_P\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right]\leq\sum_{i\in\mathcal{I}}\lambda^i\mathcal{R}_{P^i}\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right]\leq\theta$ and $\mathcal{R}_P\left[h_j(\boldsymbol{x},\boldsymbol{\xi})\right]\leq\sum_{i\in\mathcal{I}}\lambda^i\mathcal{R}_{P^i}\left[h_j(\boldsymbol{x},\boldsymbol{\xi})\right]\leq 0$, $j\in[m]$. Thus, it follows that $(\boldsymbol{x},\theta)\in\mathcal{G}'$, and hence, $\mathcal{G}\subseteq\mathcal{G}'$. ◀

Proof of Lemma 26. The first two inequalities in (41) can be found in e.g., Reiss [334, p. 99][27] and the last two inequalities can be found in e.g., Jiang et al. [215, Lemma 1]. Then, (42) follows from (41). ◀

Proof of Theorem 33. Using the conic duality results from Theorem 18, we write the dual of $\sup_{P\in\mathcal{P}^{\text{MM}}}\mathbb{E}_P\left[h_0(\boldsymbol{x},\boldsymbol{\xi})\right]$ as

$$\inf_{\boldsymbol{W},\boldsymbol{Y}}\sum_{i=1}^m\boldsymbol{W}_i\bullet\boldsymbol{U}_i-\sum_{i=1}^m\boldsymbol{Y}_i\bullet\boldsymbol{L}_i\quad\text{s.t.}\quad\begin{cases}\sum_{i=1}^m\boldsymbol{W}_i\bullet\boldsymbol{F}_i-\sum_{i=1}^m\boldsymbol{Y}_i\bullet\boldsymbol{F}_i\succcurlyeq_{\mathfrak{M}_+'(\Xi,\mathcal{F})}h_0(\boldsymbol{x},\cdot),\\ \boldsymbol{W},\boldsymbol{Y}\succcurlyeq\boldsymbol{0},\end{cases}$$

[26] It is said a partial order relation induces a *lattice structure* on $\mathcal{Z}$ if the least upper bound exists for any $Z,Z'\in\mathcal{Z}$ (Shapiro et al. [374]). A Banach space $\mathcal{Z}$ with lattice structure is called *Banach lattice* if $Z,Z'\in\mathcal{Z}$ and $|Z|\geq|Z'|$ implies $\|Z\|\geq\|Z'\|$ (Shapiro et al. [374]).

[27] As shown for e.g., in Reiss [334] and Gibbs and Su [153], $\mathfrak{d}^{\phi_{\text{h}}}(P,P_0)\leq\mathfrak{d}^{\phi_{\text{kl}}}(P,P_0)$. However, in Jiang et al. [215, Lemma 1] this relationship has been shown as $\mathfrak{d}^{\phi_{\text{h}}}(P,P_0)\leq\left(\mathfrak{d}^{\phi_{\text{kl}}}(P,P_0)\right)^{\frac{1}{2}}$.

where $\mathfrak{M}'_+(\Xi,\mathcal{F})$ is the dual cone of $\mathfrak{M}_+(\Xi,\mathcal{F})$:

$$\mathfrak{M}'_+(\Xi,\mathcal{F}) = \left\{ Z \in \mathcal{S}(\Xi,\mathcal{F}) \,\middle|\, \int_\Xi Z(s)P(\mathrm{d}s) \geq 0,\ \forall\, P \in \mathfrak{M}_+(\Xi,\mathcal{F}) \right\}.$$

Thus, we can write the first constraint above as

$$\sum_{i=1}^m \boldsymbol{W}_i \bullet \int_\Xi \boldsymbol{F}_i(s)P(\mathrm{d}s) - \sum_{i=1}^m \boldsymbol{Y}_i \bullet \int_\Xi \boldsymbol{F}_i(s)P(\mathrm{d}s) \geq \int_\Xi h_0(\boldsymbol{x},\boldsymbol{\xi}(s))P(\mathrm{d}s),\ \forall\, P \in \mathfrak{M}_+(\Xi,\mathcal{F}).$$

The Slater-type condition ensures that the strong duality holds (Shapiro [363]). ◀

Proof of Theorem 42. The proof is immediate from the relationship between ζ-structure metrics, stated in Lemma 30, and the fact that $\mathbb{P}^N\{\mathfrak{d}^{\mathrm{K}}(\mathbb{P}^{\mathrm{true}},\mathbb{P}_N) \leq \epsilon\} \geq 1-\exp\{-\frac{\epsilon^2 N}{2\theta^2}\}$ due to Zhao and Guan [442, Proposition 3]. ◀

B Kernel-based Models in Robust Optimization

The papers mentioned in Section 6.4 incorporate machine learning into the optimization framework using kernel functions in the virtue of SO and DRO. There are also papers that study this integration in the sense of RO, and mainly in order to learn the uncertainty, see, e.g., Tulabandhula and Rudin [391, 393, 392]. This is an important question to investigate, particularly, when facing high-dimensional uncertain parameters. In fact, in these situations, it may not be practical to fix the form of uncertainty set a priori; this is even more complicated with the calibration of different parameters describing the set. An alternative practice is to learn the form of the uncertainty set by using unsupervised learning algorithms on the historical data. Different from Bertsimas and Kallus [41], Tulabandhula and Rudin [391] study a framework that simultaneously seeks a best statistical model and a corresponding decision policy. In their framework, in addition to $\{(\boldsymbol{u}^i,\boldsymbol{\xi}^i)\}_{i=1}^N$, a new set of unlabeled data is available that in conjunction with the statistical model affects the cost. The minimum of such a cost function over the set of possible decisions is cast by a regularization term in the objective function of the learning algorithm. Tulabandhula and Rudin [391] show that under some conditions this problem is equivalent to a RO model, where the uncertainty set of the statistical model contains all models that are within ϵ-optimality from the predictive model describing $\{(\boldsymbol{u}^i,\boldsymbol{\xi}^i)\}_{i=1}^N$. Similar to Tulabandhula and Rudin [391], Tulabandhula and Rudin [392] use a new set of unlabeled data in addition to $\{(\boldsymbol{u}^i,\boldsymbol{\xi}^i)\}_{i=1}^N$ in order to combine machine learning and decision making. Their idea to form the uncertainty set of $\boldsymbol{\xi}$ is to consider a class of "good" predictive models with low training error on the data set $\{(\boldsymbol{u}^i,\boldsymbol{\xi}^i)\}_{i=1}^N$. Recognizing that the uncertainty can be decomposed into the predictive model uncertainty and residual uncertainty, they form the uncertainty by the Minkowski sum of two sets: (1) predictions of the new data set with the class of "good" predictive models, and (2) residuals of the new data set with the class of "good" predictive models. To form the class of "good" predictive models, one can use loss functions such as least squares and hing loss.

Kernel density estimation (KDE) (Devroye and Gyorfi [117]) in combination with *principal component analysis* (PCA) is also used in the RO literature to construct the uncertainty set (Ning and You [285]). PCA captures the correlation between uncertain parameters and transforms data into their corresponding uncorrelated principal components. KDE, then, captures the distributional information of the transformed, uncorrelated uncertain parameters along the principal components, by using kernel smoothing methods. Ning and You [285] propose to use a Gaussian kernel K defined between the latent uncertainty along the principal component k, w_k, and the projected data along the principal component k, t_k. By incorporating forward and backward deviations to allow for asymmetry (Chen et al. [91]), Ning and You [285] propose the following polytopic uncertainty set that resembles the intersection of a box, with the so-called *budget*, and polyhedral uncertainty sets:

$$\mathcal{U} = \left\{ \boldsymbol{u} \,\middle|\, \begin{array}{l} \boldsymbol{u} = \boldsymbol{\mu}_0 + \boldsymbol{V}\boldsymbol{w},\ \boldsymbol{w} = \underline{\boldsymbol{w}} \odot \boldsymbol{z}^- + \overline{\boldsymbol{w}} \odot \boldsymbol{z}^+, \\ \boldsymbol{0} \leq \boldsymbol{z}^-, \boldsymbol{z}^+ \leq \boldsymbol{1},\ \boldsymbol{z}^- + \boldsymbol{z}^+ \leq \boldsymbol{1},\ \boldsymbol{1}^\top(\boldsymbol{z}^- + \boldsymbol{z}^+) \leq \Gamma, \\ \underline{\boldsymbol{w}} = [F_1^{-1}(\alpha),\ldots,F_m^{-1}(\alpha)]^\top, \\ \overline{\boldsymbol{w}} = [F_1^{-1}(1-\alpha),\ldots,F_m^{-1}(1-\alpha)]^\top \end{array} \right\}.$$

Let us define $\boldsymbol{U} = [\boldsymbol{u}^1,\ldots,\boldsymbol{u}^N]^\top$. Above $\boldsymbol{\mu}_0 = \frac{1}{N}\sum_{i=1}^N \boldsymbol{u}^i$, and $\boldsymbol{V}$ is a square matrix consists of all m eigenvvectors (i.e., principal components) obtained from the eignevalue decomposition of the sample covariance matrix

$\boldsymbol{S} = \frac{1}{N-1}(\boldsymbol{U} - \mathbf{1}\boldsymbol{\mu}_0^\top)^\top(\boldsymbol{U} - \mathbf{1}\boldsymbol{\mu}_0^\top)$. Moreover, $\boldsymbol{z}^-$ is a backward deviation, $\boldsymbol{z}^+$ is a forward deviation vector, and Γ is the uncertainty budget. In addition, $F_k^{-1} := \min\{w_k | F_k(w_k) \geq \alpha\}$, $k \in [m]$, where $F_k(w_k)$ is the cdf of w_k, with the density function is obtained using KDE as follows: $f_k(w_k) = \frac{1}{N}\sum_{i=1}^n K_b(w_k, t_k^i)$. Ning and You [285] further extend their approach to the data-driven static and adaptive robust optimization.

In the context of RO, *support vector clustering* (SVC) is proposed to form the uncertainty set, which seeks for a sphere with the smallest radius that encloses all data mapped in the covariate space (Shang et al. [362]). In SVC, to avoid overfitting, the violations of the data outside the sphere is penalized by a regularization term as follows:

$$\min_{\delta,\boldsymbol{s},\boldsymbol{c}} \delta^2 + \frac{1}{N\gamma}\sum_{i=1}^N s_i \quad \text{s.t.} \begin{cases} \|\Phi(\boldsymbol{u}^i) - \boldsymbol{c}\|_2^2 \leq \delta^2 + s_i, \ i = 1, \ldots, N, \\ \boldsymbol{s} \geq \mathbf{0}. \end{cases}$$

Dualizing the problem of finding the smallest sphere using dual multipliers $\boldsymbol{\pi}$ results in a quadratic problem where the kernel function appears in the objective function. It is shown that commonly used kernel functions in SVC, such as polynomial, radial basis function, sigmoid function kernel, lead to an intractable robust counterpart problem for the corresponding uncertainty set. Hence, Shang et al. [362] propose to use a piecewise linear kernel, referred to as a *weighted generalized intersection kernel*, defined as follows:

$$K(\boldsymbol{u}, \boldsymbol{v}) = \sum_{k=1}^m l_k - \|\boldsymbol{Q}(\boldsymbol{u} - \boldsymbol{v})\|_1, \tag{63}$$

where $\boldsymbol{Q} = \boldsymbol{S}^{-\frac{1}{2}}$ and $\boldsymbol{S} = \frac{1}{N-1}\sum_{i=1}^N \left[\boldsymbol{u}^i(\boldsymbol{u}^i)^\top - \left(\sum_{i=1}^N \boldsymbol{u}^i\right)\left(\sum_{i=1}^N \boldsymbol{u}^i\right)^\top\right]$, and l_k, $k \in [m]$, is chosen such that $l_k > \max_{i=1}^N \boldsymbol{Q}_{\cdot k}^\top \boldsymbol{u}^i - \min_{i=1}^N \boldsymbol{Q}_{\cdot k}^\top \boldsymbol{u}^i$. Such a kernel not only incorporates covariance information, but also gives rise to the following results.

▶ **Theorem 44** (Shang et al. [362, Propositions 1, Propositions 3–4])**.** *Suppose that the kernel function is constructed as in* (63)*. Then,*

1. *The kernel matrix induced by the kernel K is positive definite.*
2. *The constructed uncertainty set*

$$\mathcal{U} = \left\{ \boldsymbol{u} \;\middle|\; \begin{array}{l} \exists\ \boldsymbol{v}_i,\ i \in \mathcal{S} s.t. \\ \sum_{i\in\mathcal{S}} \pi_i \boldsymbol{v}_i^\top \mathbf{1} \leq \epsilon, \\ -\boldsymbol{v}_i \leq \boldsymbol{Q}(\boldsymbol{u} - \boldsymbol{u}^i) \leq \boldsymbol{v}_i,\ i \in \mathcal{S} \end{array} \right\},$$

   where $\mathcal{S} := \{i \,|\, \pi_i > 0\}$, $\epsilon = \sum_{i\in\mathcal{S}} \pi_i \|\boldsymbol{Q}(\boldsymbol{u}^j - \boldsymbol{u}^i)\|_1$, $j \in \mathcal{B}$, and $\mathcal{B} := \left\{i \,\middle|\, 0 < \pi_i < \frac{1}{N\gamma}\right\}$, is a polytope; hence, the robust counterpart $\max_{\boldsymbol{u}\in\mathcal{U}} \boldsymbol{u}^\top \boldsymbol{x} \leq b$ has the same complexity as the deterministic problem.
3. *The regularization parameter γ gives an upper bound on the fraction of the outliers; hence, a feasible solution $\boldsymbol{x}$ in the robust counterpart $\max_{\boldsymbol{u}\in\mathcal{U}} \boldsymbol{u}^\top \boldsymbol{x} \leq b$ is also feasible to a SAA-based chance-constrained problem $P\{\boldsymbol{u}^\top \boldsymbol{x} \leq b\} \geq 1 - \gamma$.*
4. *As the number of data points increases, the fraction of outliers converges to the regularization parameter γ with probability one.*
5. *The regularization parameter γ gives a lower bound on the fraction of the support vectors.*

Shang and You [361] further propose to calibrate the radius of the uncertainty set and provide a probabilistic guarantee of the proposed uncertainty set. Shang and You [359] use PCA in combination with SVC to construct the uncertainty set. By employing PCA, the data space is decomposed into the principal subspace and residual subspace. Then, they utilize the uncertainty set formed in Shang et al. [362] to explain the variation in the principal subspace, and utilize a polyhedral set to explain noise in the residual subspace. The proposed uncertainty set is then the intersection of the above two sets.

References

1 Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *J. Bank. Financ.*, 26 (7):1505–1518, 2002.

2 Selin Damla Ahipaşaoğlu, Uğur Arıkan, and Karthik Natarajan. Distributionally robust Markovian traffic equilibrium. *Transport. Sci.*, 53(6):1546–1562, 2019.

3 Amir Ahmadi-Javid. An information-theoretic approach to constructing coherent risk measures. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2125–2127. IEEE, 2011.

4 Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *J. Optim. Theory Appl.*, 155(3):1105–1123, 2012.

5 Bita Analui and Georg Ch. Pflug. On distributionally robust multiperiod stochastic optimization. *Comput. Manag. Sci.*, 11(3):197–220, 2014.

6 Amir Ardestani-Jaafari and Erick Delage. Robust Optimization of Sums of Piecewise Linear Functions with Application to Inventory Problems. *Oper. Res.*, 64(2):474–494, 2016.

7 Benjamin Armbruster and Erick Delage. Decision making under uncertainty when preference information is incomplete. *Manage. Sci.*, 61(1):111–128, 2015.

8 Benjamin Armbruster and James R. Luedtke. Models and formulations for multivariate dominance-constrained stochastic programs. *IIE Trans.*, 47(1):1–14, 2015.

9 Sebastián Arpón, Tito Homem-de-Mello, and Bernardo Pagnoncelli. Scenario reduction for stochastic programs with Conditional Value-at-Risk. *Math. Program.*, 170(1):327–356, 2018.

10 Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent Measures of Risk. *Math. Financ.*, 9(3):203–228, 1999.

11 Gah-Yi Ban and Cynthia Rudin. The Big Data Newsvendor: Practical Insights from Machine Learning. *Oper. Res.*, 67(1):90–108, 2019.

12 Gah-Yi Ban, Jérémie Gallien, and Adam J. Mersereau. Dynamic procurement of new products with covariate information: The residual tree method. *Manuf. Serv. Oper. Management*, 21(4):798–815, 2019.

13 Manish Bansal and Sanjay Mehrotra. On Solving Two-Stage Distributionally Robust Disjunctive Programs with a General Ambiguity Set. *Eur. J. Oper. Res.*, 279(2):296–307, 2019.

14 Manish Bansal and Yingqiu Zhang. Scenario-based cuts for structured two-stage stochastic and distributionally robust p-order conic mixed integer programs. *J. Glob. Optim.*, 81(2):391–433, 2021.

15 Manish Bansal, Kuo-Ling Huang, and Sanjay Mehrotra. Decomposition algorithms for two-stage distributionally robust mixed binary programs. *SIAM J. Optim.*, 28(3):2360–2383, 2018.

16 Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(Nov):463–482, 2002.

17 Güzin Bayraksan and David K. Love. Data-Driven Stochastic Programming Using Phi-Divergences. In *The Operations Research Revolution*, pages 1–19. INFORMS TutORials in Operations Research, 2015.

18 Güzin Bayraksan and David P. Morton. Assessing solution quality in stochastic programs. *Math. Program.*, 108 (2-3):495–514, 2006.

19 Güzin Bayraksan and David P. Morton. Assessing solution quality in stochastic programs via sampling. In *Decision Technologies and Applications*, pages 102–122. INFORMS TutORials in Operations Research, 2009.

20 Mokhtar S. Bazaraa, Hanif D. Sherali, and Chitharanjan M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, 3rd edition, 2006.

21 Aharon Ben-Tal and Eithan Hochman. More bounds on the expectation of a convex function of a random variable. *J. Appl. Probab.*, 9(4):803–812, 1972.

22 Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Math. Oper. Res.*, 23(4):769–805, 1998.

23 Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Math. Program.*, 88(3):411–424, 2000.

24 Aharon Ben-Tal and Arkadi Nemirovski. On safe tractable approximations of chance-constrained linear matrix inequalities. *Math. Oper. Res.*, 34(1):1–25, 2009.

25 Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: Analysis, Algorithms, Engineering Applications*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2019.

26 Aharon Ben-Tal and Marc Teboulle. Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Manage. Sci.*, 32(11):1445–1466, 1986.

27 Aharon Ben-Tal and Marc Teboulle. An Old-New Concept of Convex Risk Measures: The Optimized Certainty Equivalent. *Math. Financ.*, 17(3):449–476, 2007.

28 Aharon Ben-Tal, Alexander Goryashko, Elana Guslitzer, and Arkadi Nemirovski. Adjustable robust solutions of uncertain linear programs. *Math. Program.*, 99(2):351–376, 2004.

29 Aharon Ben-Tal, Stephen Boyd, and Arkadi Nemirovski. Extending scope of robust optimization: Comprehensive robust counterparts of uncertain problems. *Math. Program.*, 107(1-2):63–89, 2006.

30 Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

31 Aharon Ben-Tal, Dimitris Bertsimas, and David B. Brown. A soft robust model for optimization under ambiguity. *Oper. Res.*, 58(4, Part 2):1220–1234, 2010.

32 Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Manage. Sci.*, 59(2):341–357, 2013.

33 Aharon Ben-Tal, Dick Den Hertog, and Jean-Philippe Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Math. Program.*, 149(1-2):265–299, 2015.

34 Aharon Ben-Tal, Ruud Brekelmans, Dick Den Hertog, and Jean-Philippe Vial. Globalized robust optimization for nonlinear uncertain inequalities. *INFORMS J. Comput.*, 29(2):350–366, 2017.

35 M. Bennouna and Bart P. G. Van Parys. Learning and Decision-Making with Data: Optimal Formulations and Phase Transitions. `https://arxiv.org/abs/2109.06911`, 2021.

36 Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 3rd edition, 2016.

37 Dimitri P. Bertsekas. *Dynamic programming and optimal control*. Athena Scientific, 4th edition, 2017.

38 Dimitris Bertsimas and David B. Brown. Constructing uncertainty sets for robust linear optimization. *Oper. Res.*, 57(6):1483–1495, 2009.

39 Dimitris Bertsimas and Constantine Caramanis. Finite adaptability in multistage linear optimization. *IEEE Trans. Autom. Control*, 55(12):2751–2766, 2010.

40 Dimitris Bertsimas and Iain R. Dunning. Relative robust and adaptive optimization. *INFORMS J. Comput.*, 32(2): 408–427, 2020.

41 Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Manage. Sci.*, 66(3):1025–1044, 2020.

42 Dimitris Bertsimas and Ioana Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM J. Optim.*, 15(3):780–804, 2005.

43 Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Oper. Res.*, 52(1):35–53, 2004.

44 Dimitris Bertsimas and Bart P. G. Van Parys. Bootstrap robust prescriptive analytics. *Math. Program.*, 2021. doi: 10.1007/s10107-021-01679-2. `https://doi.org/10.1007/s10107-021-01679-2`.

45 Dimitris Bertsimas, Karthik Natarajan, and Chung-Piaw Teo. Probabilistic combinatorial optimization: Moments, semidefinite programming, and asymptotic bounds. *SIAM J. Optim.*, 15(1):185–209, 2004.

46 Dimitris Bertsimas, Dessislava Pachamanova, and Melvyn Sim. Robust linear optimization under general norms. *Oper. Res. Lett.*, 32(6):510–516, 2004.

47 Dimitris Bertsimas, Karthik Natarajan, and Chung-Piaw Teo. Persistence in discrete optimization under data uncertainty. *Math. Program.*, 108(2-3):251–274, 2006.

48 Dimitris Bertsimas, Xuan Vinh Doan, Karthik Natarajan, and Chung-Piaw Teo. Models for minimax stochastic linear optimization problems with risk aversion. *Math. Oper. Res.*, 35(3):580–602, 2010.

49 Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Rev.*, 53(3):464–501, 2011.

50 Dimitris Bertsimas, Melvyn Sim, and Meilin Zhang. A practicable framework for distributionally robust linear optimization, 2014. Optimization Online `www.optimization-online.org/DB_FILE/2013/07/3954.html`.

51 Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Math. Program.*, 167(2): 235–292, 2018.

52 Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Robust sample average approximation. *Math. Program.*, 171 (1):217–282, 2018.

53 Dimitris Bertsimas, Melvyn Sim, and Meilin Zhang. Adaptive distributionally robust optimization. *Manage. Sci.*, 65(2):604–618, 2019.

54 Dimitris Bertsimas, Christopher McCord, and Bradley Sturt. Dynamic optimization with side information. *Eur. J. Oper. Res.*, 2022. `https://doi.org/10.1016/j.ejor.2022.03.030`.

55 Dimitris Bertsimas, Shimrit Shtern, and Bradley Sturt. A data-driven approach to multistage stochastic linear optimization. *Manage. Sci.*, 2022. `https://doi.org/10.1287/mnsc.2022.4352`.

56 Dimitris Bertsimas, Shimrit Shtern, and Bradley Sturt. Two-stage sample robust optimization. *Oper. Res.*, 70(1): 624–640, 2022.

57 J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, 2nd edition, 2011.

58 David A. Blackwell and Meyer A. Girshick. *Theory of games and statistical decisions*. Dover Publications, 1979.

59 Jose Blanchet and Yang Kang. Distributionally robust groupwise regularization estimator. In *Asian Conference on Machine Learning*, pages 97–112. Proceedings of Machine Learning Research, 2017.

60 Jose Blanchet and Yang Kang. Semi-supervised Learning Based on Distributionally Robust Optimization. In Andreas Makrides, Alex Karagrigoriou, and Christos H Skiadas, editors, *Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods*, volume 5, pages 1–33. John Wiley & Sons, 2020.

61 Jose Blanchet and Yang Kang. Sample out-of-sample inference based on Wasserstein distance. *Oper. Res.*, 69(3): 985–1013, 2021.

62 Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Math. Oper. Res.*, 44(2):565–600, 2019.

63 Jose Blanchet, Yang Kang, Fan Zhang, Fei He, and Zhangyi Hu. Doubly Robust Data-driven Distributionally Robust Optimization. In Yannis Dimotikalis, Alex Karagrigoriou, Christina Parpoula, and Christos H Skiadas, editors, *Applied Modeling Techniques and Data Analysis 1*, pages 75–90. John Wiley & Sons. doi: 10.1002/9781119821588.ch4.

64 Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.*, 56(3):830–857, 2019.

65 Jose Blanchet, Yang Kang, Karthyek Murthy, and Fan Zhang. Data-driven optimal transport cost selection for distributionally robust optimization. In *Proceedings of the 2019 Winter Simulation Conference (WSC '19)*, pages 3740–3751, 2019.

66 Jose Blanchet, Karthyek Murthy, and Viet Anh Nguyen. Statistical Analysis of Wasserstein Distributionally Robust Estimators. In *Emerging Optimization Methods and Modeling Techniques with Applications*, pages 227–254. INFORMS TutORials in Operations Research, 2021.

67 Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal Transport-Based Distributionally Robust Optimization: Structural Properties and Iterative Schemes. *Math. Oper. Res.*, 2021.

68 François Bolley and Cédric Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. volume 14 of *6*, pages 331–352, 2005.

69 Frédéric Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer, 2013.

70 Subir Bose and Arup Daripa. A dynamic mechanism and surplus extraction under ambiguity. *J. Econ. Theory*, 144 (5):2084–2114, 2009.

71 Michèle Breton and Saeb El Hachem. Algorithms for the solution of stochastic dynamic minimax problems. *Comput. Optim. Appl.*, 4(4):317–345, 1995.

72 Giuseppe C. Calafiore. Ambiguous risk measures and optimal robust portfolios. *SIAM J. Optim.*, 18(3):853–877, 2007.

73 Giuseppe C. Calafiore and Marco C. Campi. Uncertain convex programs: randomized solutions and confidence levels. *Math. Program.*, 102(1):25–46, 2005.

74 Giuseppe C. Calafiore and Laurent El Ghaoui. On distributionally robust chance-constrained linear programs. *J. Optim. Theory Appl.*, 130(1):1–22, 2006.

75 Marco C. Campi and Giuseppe C. Calafiore. Decision making in an uncertain environment: the scenario-based optimization approach. In *Multiple Participant Decision Making*, pages 99–111. Advanced Knowledge International, 2004.

76 Marco C. Campi and Simone Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM J. Optim.*, 19(3):1211–1230, 2008.

77 John Gunnar Carlsson, Mehdi Behroozi, and Kresimir Mihic. Wasserstein distance and the distributionally robust TSP. *Oper. Res.*, 66(6):1603–1624, 2018.

78 Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

79 Abraham Charnes and William W. Cooper. Chance-constrained programming. *Manage. Sci.*, 6(1):73–79, 1959.

80 Abraham Charnes and William W. Cooper. Deterministic equivalents for optimizing and satisficing under chance constraints. *Oper. Res.*, 11(1):18–39, 1963.

81 Abraham Charnes, William W. Cooper, and Gifford H. Symonds. Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil. *Manage. Sci.*, 4(3):235–263, 1958.

82 Abraham Charnes, William W. Cooper, and Kenneth O. Kortanek. Duality, Haar programs, and finite sequence spaces. *Proc. Natl. Acad. Sci. USA*, 48(5):783–786, 1962.

83 Abraham Charnes, William W. Cooper, and Kenneth O. Kortanek. Duality in semi-infinite programs and some works of Haar and Carathéodory. *Manage. Sci.*, 9(2):209–228, 1963.

84 Abraham Charnes, William W. Cooper, and Kenneth O. Kortanek. On the theory of semi-infinite programming and a generalization of the Kuhn-Tucker saddle point theorem for arbitrary convex functions. *Nav. Res. Logist. Q.*, 16(1):41–52, 1969.

85 Louis Chen, Will Ma, Karthik Natarajan, David Simchi-Levi, and Zhenzhen Yan. Distributionally robust linear and discrete optimization with marginals. *Oper. Res.*, 2022. `https://doi.org/10.1287/opre.2021.2243`.

86 Ruidi Chen and Ioannis Ch. Paschalidis. A Robust Learning Approach for Regression Models Based on Distributionally Robust Optimization. *J. Mach. Learn. Res.*, 19(13):1–48, 2018.

87 Wenqing Chen and Melvyn Sim. Goal-driven optimization. *Oper. Res.*, 57(2):342–357, 2009.

88 Wenqing Chen, Melvyn Sim, Jie Sun, and Chung-Piaw Teo. From CVaR to uncertainty set: Implications in joint chance-constrained optimization. *Oper. Res.*, 58(2):470–485, 2010.

89 Xiaojun Chen, Hailin Sun, and Huifu Xu. Discrete approximation of two-stage stochastic and distributionally robust linear complementarity problems. *Math. Program.*, 177(1):255–289, 2019.

90 Xin Chen and Yuhan Zhang. Uncertain linear programs: Extended affinely adjustable robust counterparts. *Oper. Res.*, 57(6):1469–1482, 2009.

91 Xin Chen, Melvyn Sim, and Peng Sun. A robust optimization perspective on stochastic programming. *Oper. Res.*, 55(6):1058–1071, 2007.

92 Xin Chen, Melvyn Sim, Peng Sun, and Jiawei Zhang. A linear decision-based approximation approach to stochastic programming. *Oper. Res.*, 56(2):344–357, 2008.

93 Yannan Chen, Hailin Sun, and Huifu Xu. Decomposition and discrete approximation methods for solving two-stage distributionally robust optimization problems. *Comput. Optim. Appl.*, 78(1):205–238, 2021.

94 Zhi Chen and Weijun Xie. Regret in the newsvendor model with demand and yield randomness. *Prod. Oper. Manage.*, 30(11):4176–4197, 2021.

95 Zhi Chen and Peng Xiong. RSOME in Python: An Open-Source Package for Robust Stochastic Optimization Made Easy, 2021. Optimization Online `http://www.optimization-online.org/DB_HTML/2021/06/8443.html`.

96 Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. Data-Driven Chance Constrained Programs over Wasserstein Balls. `https://arxiv.org/abs/1809.00210`, 2018.

97 Zhi Chen, Melvyn Sim, and Huan Xu. Distributionally robust optimization with infinitely constrained ambiguity sets. *Oper. Res.*, 67(5):1328–1344, 2019.

98 Zhi Chen, Melvyn Sim, and Peng Xiong. Robust stochastic optimization made easy with RSOME. *Manage. Sci.*, 66 (8):3329–3339, 2020.

99 Jianqiang Cheng, Richard Li-Yang Chen, Habib N. Najm, Ali Pinar, Cosmin Safta, and Jean-Paul Watson. Distributionally Robust Optimization with Principal Component Analysis. *SIAM J. Optim.*, 28(2):1817–1841, 2018.

100 Andreas Christmann and Ingo Steinwart. On robustness properties of convex risk minimization methods for pattern recognition. *J. Mach. Learn. Res.*, 5(Aug):1007–1034, 2004.

101 Andreas Christmann, Ingo Steinwart, et al. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.

102 Erick Delage. *Distributionally robust optimization in context of data-driven problems*. Ph.D. dissertation, Stanford University, 2009.

103 Erick Delage and Jonathan Y. Li. Minimizing risk exposure when the choice of a risk measure is ambiguous. *Manage. Sci.*, 64(1):327–344, 2018.

104 Erick Delage and Ahmed Saif. The value of randomized solutions in mixed-integer distributionally robust optimization problems. *INFORMS J. Comput.*, 34(1):333–353, 2022.

105 Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.*, 58(3):595–612, 2010.

106 Erick Delage, Daniel Kuhn, and Wolfram Wiesemann. "Dice"sion-Making Under Uncertainty: When Can a Random Decision Reduce Risk? *Manage. Sci.*, 65(7):3282–3301, 2019.

107 Erick Delage, Shaoyan Guo, and Huifu Xu. Shortfall Risk Models When Information on Loss Function Is Incomplete. *Oper. Res.*, 2022.

108 Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer, 1998.

109 Victor DeMiguel and Francisco J. Nogales. Portfolio selection with robust estimation. *Oper. Res.*, 57(3):560–577, 2009.

110 Yunxiao Deng and Suvrajeet Sen. Learning Enabled Optimization: Towards a Fusion of Statistical Learning and Stochastic Optimization, 2018. Optimization Online `http://www.optimization-online.org/DB_HTML/2017/03/5904.html`.

111 Darinka Dentcheva. Optimization Models with Probabilistic Constraints. In Giuseppe C. Calafiore and Fabrizio Dabbene, editors, *Probabilistic and Randomized Methods for Design under Uncertainty*, pages 49–97. Springer, 2006.

112 Darinka Dentcheva and Andrzej Ruszczynski. Optimization with stochastic dominance constraints. *SIAM J. Optim.*, 14(2):548–566, 2003.

113 Darinka Dentcheva and Andrzej Ruszczyński. Optimality and duality theory for stochastic optimization problems with nonlinear dominance constraints. *Math. Program.*, 99(2):329–350, 2004.

114 Darinka Dentcheva and Andrzej Ruszczyński. Optimization with multivariate stochastic dominance constraints. *Math. Program.*, 117(1-2):111–127, 2009.

115 Darinka Dentcheva and Andrzej Ruszczyński. Robust stochastic dominance and its application to risk-averse optimization. *Math. Program.*, 123(1):85–100, 2010.

116 Esther Derman and Shie Mannor. Distributional robustness and regularization in reinforcement learning. `https://arxiv.org/abs/2003.02894`, 2020.

117 Luc Devroye and Laszlo Gyorfi. *Nonparametric density estimation: The L1 View*. John Wiley & Sons, 1985.

118 Anulekha Dhara, Bikramjit Das, and Karthik Natarajan. Worst-case expected shortfall with univariate and bivariate marginals. *INFORMS J. Comput.*, 33(1):370–389, 2021.

119 Sudhakar Dharmadhikari and Kumar Joag-Dev. *Unimodality, convexity, and applications*. Academic Press Inc., 1988.

120 Ke-wei Ding, Nan-jing Huang, and Lei Wang. Globalized distributionally robust optimization problems under the moment-based framework. `https://arxiv.org/abs/2008.08256`, 2020.

121 Xuan Vinh Doan, Xiaobo Li, and Karthik Natarajan. Robustness to dependency in portfolio optimization using overlapping marginals. *Oper. Res.*, 63(6):1468–1488, 2015.

122 John C. Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally Robust Losses Against Mixture

Covariate Shifts. https://arxiv.org/abs/2007.13982, 2019.

123 John C. Duchi, Peter W. Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Math. Oper. Res.*, 46(3):946–969, 2021.

124 Richard Mansfield Dudley. The speed of mean Glivenko-Cantelli convergence. *Ann. Math. Stat.*, 40(1):40–50, 1969.

125 Iain R. Dunning. *Advances in robust and adaptive optimization: algorithms, software, and insights*. PhD thesis, Massachusetts Institute of Technology, 2016.

126 Iain R. Dunning, Joey Huchette, and Miles Lubin. JuMP: A Modeling Language for Mathematical Optimization. *SIAM Rev.*, 59(2):295–320, 2017. doi: 10.1137/15M1020575.

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

128 Jitka Dupačová. Stability and sensitivity-analysis for stochastic programming. *Ann. Oper. Res.*, 27(1):115–142, 1990.

129 Jitka Dupačová, Nicole Gröwe-Kuska, and Werner Römisch. Scenario reduction in stochastic programming. *Math. Program.*, 95(3):493–511, 2003.

130 Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Petr Plechác. Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *SIAM/ASA J. Uncertain. Quantif.*, 4(1): 80–111, 2016.

131 Daniel Duque and David P. Morton. Distributionally robust stochastic dual dynamic programming. *SIAM J. Optim.*, 30(4):2841–2865, 2020.

132 Elad Eban, Elad Mezuman, and Amir Globerson. Discrete Chebyshev classifiers. In *31st International Conference on Machine Learning*, pages 1233–1241. Proceedings of Machine Learning Research, 2014.

133 Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM J. Matrix Anal. Appl.*, 18(4):1035–1064, 1997.

134 Laurent El Ghaoui, Francois Oustry, and Hervé Lebret. Robust solutions to uncertain semidefinite programs. *SIAM J. Optim.*, 9(1):33–52, 1998.

135 Laurent El Ghaoui, Maksim Oks, and Francois Oustry. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Oper. Res.*, 51(4):543–556, 2003.

136 Paul Embrechts and Giovanni Puccetti. Aggregating risk capital, with an application to operational risk. *Geneva Risk Insur. Rev.*, 31(2):71–90, 2006.

137 Paul Embrechts and Giovanni Puccetti. Bounds for functions of multivariate risks. *J. Multivariate Anal.*, 97(2): 526–547, 2006.

138 Emre Erdoğan and Garud Iyengar. Ambiguous chance constrained problems and robust optimization. *Math. Program.*, 107(1-2):37–61, 2006.

139 Adrián Esteban-Pérez and Juan M. Morales. Distributionally robust stochastic programs with side information based on trimmings. *Math. Program.*, 2021. https://doi.org/10.1007/s10107-021-01724-0.

140 Farzan Farnia and David Tse. A Minimax Approach to Supervised Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4240–4248. Curran Associates, Inc., 2016.

141 Rizal Fathony, Ashkan Rezaei, Mohammad Ali Bashiri, Xinhua Zhang, and Brian Ziebart. Distributionally Robust Graphical Models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8354–8365. Curran Associates, Inc., 2018.

142 Matteo Fischetti and Michele Monaci. Light robustness. In Ravindra K Ahuja, Rolf H Möhring, and Christos D Zaroliagis, editors, *Robust and online large-scale optimization: models and techniques for transportation systems*, pages 61–84. Springer, 2009.

143 Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Relat. Fields*, 162(3-4):707–738, 2015.

144 Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer Series in Statistics. Springer, 2nd edition, 2016.

145 Michael C. Fu. Handbook of simulation optimization. In Camille C. Price, editor, *International Series in Operations Research & Management Science*. Springer, 2016.

146 Virginie Gabrel, Cécile Murat, and Aurélie Thiele. Recent advances in robust optimization: An overview. *Eur. J. Oper. Res.*, 235(3):471–483, 2014. ISSN 0377-2217.

147 Guillermo Gallego and Ilkyeong Moon. The distribution free newsboy problem: review and extensions. *J. Oper. Res. Soc.*, 44(8):825–834, 1993.

148 Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. https://arxiv.org/abs/1604.02199v2, 2016.

149 Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with dependence structure. https://arxiv.org/abs/1701.04200, 2017.

150 Rui Gao, Xi Chen, and Anton J. Kleywegt. Wasserstein distributional robustness and regularization in statistical

learning. https://arxiv.org/abs/1712.06050, 2017.

151 Rui Gao, Liyan Xie, Yao Xie, and Huan Xu. Robust Hypothesis Testing Using Wasserstein Uncertainty Sets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

152 Angelos Georghiou, Daniel Kuhn, and Wolfram Wiesemann. The decision rule approach to optimization under uncertainty: methodology and applications. *Comput. Manag. Sci.*, 16(4):545–576, 2019.

153 Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *Int. Stat. Rev.*, 70(3): 419–435, 2002.

154 Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *J. Math. Econ.*, 18(2):141–153, 1989. ISSN 0304-4068.

155 Martin Glanzer, Georg Ch. Pflug, and Alois Pichler. Incorporating statistical model error into the calculation of acceptability prices of contingent claims. *Math. Program.*, 174(1-2):499–524, 2019.

156 Paul Glasserman and Xingbo Xu. Robust risk measurement and model risk. *Quant. Finance*, 14(1):29–58, 2014.

157 Paul Glasserman and Linan Yang. Bounding Wrong-Way Risk in CVA Calculation. *Math. Financ.*, 28(1):268–305, 2018.

158 Amir Globerson and Naftali Tishby. The minimum information principle for discriminative learning. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 193–200. AUAI Press, 2004.

159 Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Oper. Res.*, 58(4, Part 1):902–917, 2010.

160 Joel Goh and Melvyn Sim. Robust Optimization Made Easy with ROME. *Oper. Res.*, 59(4):973–985, 2011.

161 Donald Goldfarb and Garud Iyengar. Robust portfolio selection problems. *Math. Oper. Res.*, 28(1):1–38, 2003.

162 Zhaohua Gong, Chongyang Liu, Jie Sun, and Kok Lay Teo. Distributionally robust L1-estimation in multiple linear regression. *Optim. Lett.*, 13(4):935–947, 2019.

163 Bram L. Gorissen, İhsan Yanıkoğlu, and Dick den Hertog. A practical guide to robust optimization. *Omega*, 53: 124–137, 2015.

164 Jun-ya Gotoh, Michael Jong Kim, and Andrew E. B. Lim. Robust empirical optimization is almost the same as mean–variance optimization. *Oper. Res. Lett.*, 46(4):448–452, 2018.

165 Jun-ya Gotoh, Michael Jong Kim, and Andrew E. B. Lim. Calibration of distributionally robust empirical optimization models. *Oper. Res.*, 69(5):1630–1650, 2021.

166 Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Ann. Stat.*, 32(4):1367–1433, 2004.

167 Gökhan Gül. Asymptotically Minimax Robust Hypothesis Testing. https://arxiv.org/abs/1711.07680, 2017.

168 Gökhan Gül and Abdelhak M. Zoubir. Minimax robust hypothesis testing. *IEEE Trans. Inf. Theory*, 63(9):5572–5587, 2017.

169 Shaoyan Guo, Huifu Xu, and Liwei Zhang. Convergence analysis for mathematical programs with distributionally robust chance constraint. *SIAM J. Optim.*, 27(2):784–816, 2017.

170 A. Haar. Über linear Ungleichungen. *Acta Sci. Math.*, 2, 1924.

171 Bjarni V. Halldórsson and Reha H. Tütüncü. An interior-point method for a class of saddle-point problems. *J. Optim. Theory Appl.*, 116(3):559–590, 2003.

172 Qiaoming Han, Donglei Du, and Luis F. Zuluaga. Technical Note-A Risk- and Ambiguity-Averse Extension of the Max-Min Newsvendor Order Formula. *Oper. Res.*, 62(3):535–542, 2014.

173 Grani A. Hanasusanto and Daniel Kuhn. Robust Data-Driven Dynamic Programming. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 827–835. Curran Associates, Inc., 2013.

174 Grani A. Hanasusanto and Daniel Kuhn. Conic Programming Reformulations of Two-Stage Distributionally Robust Linear Programs over Wasserstein Balls. *Oper. Res.*, 66(3):849–869, 2018.

175 Grani A. Hanasusanto, Daniel Kuhn, Stein W. Wallace, and Steve Zymler. Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Math. Program.*, 152(1-2):1–32, 2015.

176 Grani A. Hanasusanto, Daniel Kuhn, and Wolfram Wiesemann. K-adaptability in two-stage robust binary programming. *Oper. Res.*, 63(4):877–891, 2015.

177 Grani A. Hanasusanto, Vladimir Roitch, Daniel Kuhn, and Wolfram Wiesemann. A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Math. Program.*, 151(1):35–62, 2015.

178 Grani A. Hanasusanto, Daniel Kuhn, and Wolfram Wiesemann. K-adaptability in two-stage distributionally robust binary programming. *Oper. Res. Lett.*, 44(1):6–11, 2016.

179 Grani A. Hanasusanto, Vladimir Roitch, Daniel Kuhn, and Wolfram Wiesemann. Ambiguous joint chance constraints under mean and dispersion information. *Oper. Res.*, 65(3):751–767, 2017.

180 Lauren Hannah, Warren Powell, and David M. Blei. Nonparametric Density Estimation for Stochastic Optimization with an Observable State Variable. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta,

editors, *Advances in Neural Information Processing Systems 23*, pages 820–828. Curran Associates, Inc., 2010.

181 William E. Hart, Carl D. Laird, Jean-Paul Watson, David L. Woodruff, Gabriel A. Hackebeil, Bethany L. Nicholson, John D. Siirola, et al. *Pyomo-optimization modeling in Python*. Springer, 2017.

182 Holger Heitsch and Werner Römisch. Scenario reduction algorithms in stochastic programming. *Comput. Optim. Appl.*, 24(2-3):187–206, 2003.

183 Holger Heitsch and Werner Römisch. Scenario tree modeling for multistage stochastic programs. *Math. Program.*, 118(2):371–406, 2009.

184 Holger Heitsch and Werner Römisch. Scenario tree reduction for multistage stochastic programs. *Comput. Manag. Sci.*, 6(2):117–133, 2009.

185 Holger Heitsch, Werner Römisch, and Cyrille Strugarek. Stability of multistage stochastic programs. *SIAM J. Optim.*, 17(2):511–525, 2006.

186 Rainer Hettich and H. T. Jongen. On first and second order conditions for local optima for optimization problems in finite dimensions. *Methods Oper. Res.*, 23:82–97, 1977.

187 Rainer Hettich and H. T. Jongen. Semi-infinite programming: conditions of optimality and applications. In J Stoer, editor, *Optimization Techniques, Lecture Notes in Control and Information Science*, pages 82–97. Springer, 1978.

188 Rainer Hettich and Kenneth O. Kortanek. Semi-infinite programming: theory, methods, and applications. *SIAM Rev.*, 35(3):380–429, 1993.

189 Rainer Hettich and Georg Still. Second order optimality conditions for generalized semi-infinite programming problems. *Optimization*, 34(3):195–211, 1995.

190 Nam Ho-Nguyen, Fatma Kılınç-Karzan, Simge Küçükyavuz, and Dabeen Lee. Strong formulations for distributionally robust chance-constrained programs with left-hand side uncertainty under Wasserstein ambiguity. `https://arxiv.org/abs/2007.06750`, 2020.

191 Nam Ho-Nguyen, Fatma Kılınç-Karzan, Simge Küçükyavuz, and Dabeen Lee. Distributionally robust chance-constrained programs with right-hand side uncertainty under Wasserstein ambiguity. *Math. Program.*, 2021. `https://doi.org/10.1007/s10107-020-01605-y`.

192 Tito Homem-de-Mello and Güzin Bayraksan. Monte Carlo sampling-based methods for stochastic optimization. *Surv. Oper. Res. Manage. Sci.*, 19(1):56–85, 2014.

193 Tito Homem-de-Mello and Sanjay Mehrotra. A cutting-surface method for uncertain linear programs with polyhedral stochastic dominance constraints. *SIAM J. Optim.*, 20(3):1250–1273, 2009.

194 Jian Hu and Sanjay Mehrotra. Robust and stochastically weighted multiobjective optimization models and reformulations. *Oper. Res.*, 60(4):936–953, 2012.

195 Jian Hu and Sanjay Mehrotra. Robust decision making over a set of random targets or risk-averse utilities with an application to portfolio optimization. *IIE Trans.*, 47(4):358–372, 2015.

196 Jian Hu, Tito Homem-de-Mello, and Sanjay Mehrotra. Risk-adjusted budget allocation models with application in homeland security. *IIE Trans.*, 43(12):819–839, 2011.

197 Jian Hu, Tito Homem-de-Mello, and Sanjay Mehrotra. Sample average approximation of stochastic dominance constrained programs. *Math. Program.*, 133(1-2):171–201, 2012.

198 Jian Hu, Tito Homem-de-Mello, and Sanjay Mehrotra. Stochastically weighted stochastic dominance concepts with an application in capital budgeting. *Eur. J. Oper. Res.*, 232(3):572–583, 2014.

199 Jian Hu, Junxuan Li, and Sanjay Mehrotra. A data-driven functionally robust approach for simultaneous pricing and order quantity decisions with unknown demand function. *Oper. Res.*, 67(6):1564–1585, 2019.

200 Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does Distributionally Robust Supervised Learning Give Robust Classifiers? In *35th International Conference on Machine Learning*, pages 2034–2042. Proceedings of Machine Learning Research, 2018.

201 Zhaolin Hu and L. Jeff Hong. Kullback-Leibler divergence constrained distributionally robust optimization, 2012. Optimization Online `http://www.optimization-online.org/DB_HTML/2012/11/3677.html`.

202 Zhaolin Hu, L. Jeff Hong, and Anthony Man Cho So. Ambiguous probabilistic programs, 2013. Optimization Online `http://www.optimization-online.org/DB_HTML/2013/09/4039.html`.

203 Jianqiu Huang, Kezhuo Zhou, and Yongpei Guan. A Study of Distributionally Robust Multistage Stochastic Optimization. `https://arxiv.org/abs/1708.07930`, 2017.

204 Peter J. Huber. A robust version of the probability ratio test. *Ann. Math. Stat.*, pages 1753–1758, 1965.

205 Peter J. Huber. The use of Choquet capacities in statistics. *B. Int. Statist. Inst.*, 45(4):181–191, 1973.

206 Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. John Wiley & Sons, 2nd edition, 2009.

207 Leonid Hurwicz. The generalized Bayes minimax principle: a criterion for decision making uncer uncertainty. *Cowles Comm. Discuss. Paper: Stat.*, 1951.

208 Natalie Isenberg, John D. Siirola, and Chrysanthos Gounaris. Pyros: A Pyomo Robust Optimization Solver for Robust Process Design. In *2020 Virtual AIChE Annual Meeting*. AIChE, 2020.

209 Keiiti Isii. On sharpness of Tchebycheff-type inequalities. *Ann. Inst. Stat. Math.*, 14(1):185–197, 1962.

210 Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning.*

Springer, 2013.

211 Ran Ji and Miguel A. Lejeune. Data-driven optimization of reward-risk ratio measures. *INFORMS J. Comput.*, 33 (3):1120–1137, 2021.

212 Ran Ji and Miguel A. Lejeune. Data-driven distributionally robust chance-constrained optimization with Wasserstein metric. *J. Glob. Optim.*, 79(4):779–811, 2021.

213 Ruiwei Jiang and Yongpei Guan. Data-driven chance constrained stochastic program. *Math. Program.*, 158(1-2): 291–327, 2016.

214 Ruiwei Jiang and Yongpei Guan. Risk-averse two-stage stochastic program with distributional ambiguity. *Oper. Res.*, 66(5):1390–1405, 2018.

215 Ruiwei Jiang, Yongpei Guan, and Jean-Paul Watson. Risk-averse stochastic unit commitment with incomplete information. *IIE Trans.*, 48(9):838–854, 2016.

216 Ruiwei Jiang, Siqian Shen, and Yiling Zhang. Integer Programming Approaches for Appointment Scheduling with Random No-Shows and Service Durations. *Oper. Res.*, 65(6):1638–1656, 2017.

217 Rohit Kannan, Güzin Bayraksan, and James R. Luedtke. Data-driven sample average approximation with covariate information, 2020. Optimization Online `http:/www.optimization-online.org/DB_HTML/2020/07/7932.html`.

218 Rohit Kannan, Güzin Bayraksan, and James R. Luedtke. Residuals-based distributionally robust optimization with covariate information, 2020. `https://arxiv.org/abs/2012.01088`.

219 Michalis Kapsos, Nicos Christofides, and Berç Rustem. Worst-case robust Omega ratio. *Eur. J. Oper. Res.*, 234(2): 499–507, 2014.

220 Kibaek Kim and Sanjay Mehrotra. A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management. *Oper. Res.*, 63(6):1431–1451, 2015.

221 Diego Klabjan, David Simchi-Levi, and Miao Song. Robust Stochastic Lot-Sizing by Means of Histograms. *Prod. Oper. Manage.*, 22(3):691–710, 2013.

222 Frank Hyneman Knight. *Risk, uncertainty and profit*. Houghton Mifflin, 1921.

223 Daniel Kuhn, Wolfram Wiesemann, and Angelos Georghiou. Primal and dual linear decision rules in stochastic and robust optimization. *Math. Program.*, 130(1):177–209, 2011.

224 Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS TutORials in Operations Research, 2019.

225 Shigeo Kusuoka. On law invariant coherent risk measures. In Shigeo Kusuoka and Toru Maruyama, editors, *Advances in Mathematical Economics*, pages 83–95. Springer, 2001.

226 John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*, pages 282–289. Morgan Kaufmann Publishers Inc., 2001.

227 Henry Lam. Advanced tutorial: Input uncertainty and robust analysis in stochastic simulation. In *Proceedings of the 2016 Winter Simulation Conference (WSC '16)*, pages 178–192. IEEE, 2016.

228 Henry Lam. Robust sensitivity analysis for stochastic systems. *Math. Oper. Res.*, 41(4):1248–1275, 2016.

229 Henry Lam. Sensitivity to serial dependency of input processes: A robust approach. *Manage. Sci.*, 64(3):1311–1327, 2018.

230 Henry Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Oper. Res.*, 67(4):1090–1105, 2019.

231 Henry Lam and Clementine Mottet. Tail analysis without parametric models: A worst-case perspective. *Oper. Res.*, 65(6):1696–1711, 2017.

232 Henry Lam and Enlu Zhou. Quantifying uncertainty in sample average approximation. In *Proceedings of the 2015 Winter Simulation Conference (WSC '15)*, pages 3846–3857, 2015.

233 Henry Lam and Enlu Zhou. The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Oper. Res. Lett.*, 45(4):301–307, 2017.

234 Gert R. G. Lanckriet, Laurent El Ghaoui, Chiranjib Bhattacharyya, and Michael I. Jordan. A robust minimax approach to classification. *J. Mach. Learn. Res.*, 3:555–582, 2002.

235 Henry J. Landau, editor. *Moments in mathematics*, volume 37 of *Proceeding of Symposia in Applied Mathematics*. American Mathematical Society, 1987.

236 Jean B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM J. Optim.*, 11(3): 796–817, 2001.

237 Jean B. Lasserre and Tillmann Weisser. Distributionally robust polynomial chance-constraints under mixture ambiguity sets. *Math. Program.*, 185(1):409–453, 2021.

238 Changhyeok Lee and Sanjay Mehrotra. A distributionally-robust approach for finding support vector machines, 2015. Optimization Online `http://www.optimization-online.org/DB_HTML/2015/06/4965.html`.

239 Jaeho Lee and Maxim Raginsky. Minimax Statistical Learning with Wasserstein distances. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing*

Systems 31, pages 2692–2701. Curran Associates, Inc., 2018.

240 Bernard C. Levy. Robust Hypothesis Testing With a Relative Entropy Tolerance. *IEEE Trans. Inf. Theory*, 55(1): 413–421, 2009.

241 Bowen Li, Ruiwei Jiang, and Johanna L. Mathieu. Ambiguous risk constraints with moment and unimodality information. *Math. Program.*, 173(1-2):151–192, 2019.

242 Jonathan Y. Li. Technical Note-Closed-form solutions for worst-case law invariant risk measures with application to robust portfolio optimization. *Oper. Res.*, 66(6):1533–1541, 2018.

243 Jonathan Y. Li and Roy H. Kwon. Portfolio selection under model uncertainty: a penalized moment-based optimization approach. *J. Glob. Optim.*, 56(1):131–164, 2013.

244 Yueyao Li and Wenxun Xing. Globalized distributionally robust optimization based on samples. `https://arxiv.org/abs/2205.02994`, 2022.

245 Andrew E. B. Lim, George J. Shanthikumar, and Max Z. J. Shen. Model uncertainty, robust optimization, and learning. In *Models, Methods, and Applications for Innovative Decision Making*, pages 66–94. INFORMS, 2006.

246 Qun Lin, Ryan Loxton, Kok Lay Teo, Yong Hong Wu, and Changjun Yu. A new exact penalty method for semi-infinite programming problems. *J. Comput. Appl. Math.*, 261:271–286, 2014.

247 Feng Liu, Zhi Chen, and Shuming Wang. Globalized Distributionally Robust Counterpart: Model, Reformulation, and Applications, 2021. Optimization Online `http://www.optimization-online.org/DB_HTML/2021/11/8663.html`.

248 Yongchao Liu, Rudabeh Meskarian, and Huifu Xu. Distributionally Robust Reward-Risk Ratio Optimization with Moment Constraints. *SIAM J. Optim.*, 27(2):957–985, 2017.

249 Yongchao Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Primal–dual hybrid gradient method for distributionally robust optimization problems. *Oper. Res. Lett.*, 45(6):625–630, 2017.

250 Daniel Long, Melvyn Sim, and Minglong Zhou. The Dao of Robustness: Achieving Robustness in Prescriptive Analytics, 2020. available at SSRN 3478930.

251 Daniel Zhuoyu Long and Jin Qi. Distributionally robust discrete optimization with Entropic Value-at-Risk. *Oper. Res. Lett.*, 42(8):532–538, 2014.

252 Daniel Zhuoyu Long, Melvyn Sim, and Minglong Zhou. Robust satisficing. *Oper. Res.*, 2022. `https://doi.org/10.1287/opre.2021.2238`.

253 Marco López and Georg Still. Semi-infinite programming. *Eur. J. Oper. Res.*, 180(2):491–518, 2007.

254 Somayyeh Lotfi and Stavros A Zenios. Robust VaR and CVaR optimization under joint ambiguity in distributions, means, and covariances. *Eur. J. Oper. Res.*, 269(2):556–576, 2018.

255 David K. Love and Güzin Bayraksan. Two-stage likelihood robust linear program with application to water allocation under uncertainty. In *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World (WSC '13)*, pages 77–88. IEEE, 2013.

256 David K. Love and Güzin Bayraksan. Phi-divergence constrained ambiguous stochastic programs for data-driven optimization, 2016. Optimization Online `http://www.optimization-online.org/DB_HTML/2016/03/5350.html`.

257 Mengshi Lu and Zuo-Jun Max Shen. A review of robust operations management under model uncertainty. *Prod. Oper. Manage.*, 30(6):1927–1943, 2021.

258 James R. Luedtke and Shabbir Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM J. Optim.*, 19(2):674–699, 2008.

259 Fengqiao Luo and Sanjay Mehrotra. Decomposition Algorithm for Distributionally Robust Optimization using Wasserstein Metric with an Application to a Class of Regression Models. *Eur. J. Oper. Res.*, 278(1):20–35, 2019.

260 Fengqiao Luo and Sanjay Mehrotra. Distributionally robust optimization with decision dependent ambiguity sets. *Optim. Lett.*, 14(8):2565–2594, 2020.

261 Colin McDiarmid. Concentration. In Michel Habib, Colin McDiarmid, Jorge Ramirez-Alfonsin, and Bruce Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, 1998.

262 Sanjay Mehrotra and Dávid Papp. A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. *SIAM J. Optim.*, 24(4):1670–1697, 2014.

263 Sanjay Mehrotra and He Zhang. Models and algorithms for distributionally robust least squares problems. *Math. Program.*, 146(1-2):123–141, 2014.

264 Yu Mei, Jia Liu, and Zhiping Chen. Distributionally Robust Second-Order Stochastic Dominance Constrained Optimization with Wasserstein Ball. *SIAM J. Optim.*, 32(2):715–738, 2022.

265 Martin Mevissen, Emanuele Ragnoli, and Jia Yuan Yu. Data-driven Distributionally Robust Polynomial Optimization. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 37–45. Curran Associates, Inc., 2013.

266 Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Math. Program.*, 171(1):115–166, 2018.

267 Peyman Mohajerin Esfahani, Soroosh Shafieezadeh-Abadeh, Grani A. Hanasusanto, and Daniel Kuhn. Data-driven inverse optimization with imperfect information. *Math. Program.*, 167(1):191–234, 2018.

268 Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, 2018.

269 Alfred Müller. Integral probability metrics and their generating classes of functions. *Adv. Appl. Probab.*, pages 429–443, 1997.

270 John M. Mulvey, Robert J. Vanderbei, and Stavros A. Zenios. Robust optimization of large-scale systems. *Oper. Res.*, 43(2):264–281, 1995.

271 Hongseok Namkoong and John C. Duchi. Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

272 Hongseok Namkoong and John C. Duchi. Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2208–2216. Curran Associates, Inc., 2016.

273 Hongseok Namkoong and John C. Duchi. Variance-based Regularization with Convex Objectives. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2971–2980. Curran Associates, Inc., 2017.

274 Karthik Natarajan and Chung-Piaw Teo. On reduced semidefinite programs for second order moment bounds with applications. *Math. Program.*, 161(1-2):487–518, 2017.

275 Karthik Natarajan, Dessislava Pachamanova, and Melvyn Sim. Constructing risk measures from uncertainty sets. *Oper. Res.*, 57(5):1129–1141, 2009.

276 Karthik Natarajan, Chung-Piaw Teo, and Zhichao Zheng. Mixed 0-1 linear programs under objective uncertainty: A completely positive representation. *Oper. Res.*, 59(3):713–728, 2011.

277 Karthik Natarajan, Chung-Piaw Teo, and Zhichao Zheng. Mixed 0-1 linear programs under objective uncertainty: A completely positive representation. *Oper. Res.*, 59(3):713–728, 2011.

278 Karthik Natarajan, Dongjian Shi, and Kim-Chuan Toh. A Probabilistic Model for Minmax Regret in Combinatorial Optimization. *Oper. Res.*, 62(1):160–181, 2014.

279 Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM J. Optim.*, 17(4):969–996, 2006.

280 Arkadi Nemirovski and Alexander Shapiro. Scenario Approximations of Chance Constraints. In Giuseppe C. Calafiore and Fabrizio Dabbene, editors, *Probabilistic and Randomized Methods for Design under Uncertainty*, pages 3–47. Springer, 2006.

281 Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

282 David Newton, Farzad Yousefian, and Raghu Pasupathy. Stochastic gradient descent: Recent trends. In *Recent Advances in Optimization and Modeling of Contemporary Problems*, pages 193–220. INFORMS TutORials in Operations Research, 2018.

283 Viet Anh Nguyen, Fan Zhang, Jose Blanchet, Erick Delage, and Yinyu Ye. Robustifying conditional portfolio decisions via optimal transport, 2021. `https://arxiv.org/abs/2103.16451`.

284 Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *Oper. Res.*, 70(1):490–515, 2022.

285 Chao Ning and Fengqi You. Data-driven decision making under uncertainty integrating robust optimization with principal component analysis and kernel smoothing methods. *Comput. Chem. Eng.*, 112:190–210, 2018.

286 Kiyohiko G. Nishimura and Hiroyuki Ozaki. Search and Knightian uncertainty. *J. Econ. Theory*, 119(2):299–333, 2004.

287 Kiyohiko G. Nishimura and Hiroyuki Ozaki. An Axiomatic Approach to ϵ-Contamination. *J. Econ. Theory*, 27(2): 333–340, 2006.

288 Nilay Noyan, Gábor Rudolf, and Miguel A. Lejeune. Distributionally Robust Optimization Under a Decision-Dependent Ambiguity Set with Applications to Machine Scheduling and Humanitarian Logistics. *INFORMS J. Comput.*, 34(2):729–751, 2022.

289 Günther Nürnberger. Global unicity in optimization and approximation. *Z. Angew. Math. Mech.*, 65(5):T319–T321, 1985.

290 Günther Nürnberger. Global unicity in semi-infinite optimization. *Numer. Funct. Anal. Optim.*, 8:173–191, 1985.

291 Art B. Owen. *Empirical likelihood*. Chapman & Hall/CRC, 2001.

292 Chin Pang Ho and Grani A. Hanasusanto. On Data-Driven Prescriptive Analytics with Side Information: A Regularized Nadaraya-Watson Approach, 2019. Optimziation Online `http://www.optimization-online.org/DB_HTML/2019/01/7043.html`.

293 Leandro Pardo. *Statistical inference based on divergence measures*. Chapman & Hall/CRC, 2005.

294 Jangho Park and Güzin Bayraksan. A Multistage Distributionally Robust Optimization Approach to Water Allocation under Climate Uncertainty. `https://arxiv.org/abs/2005.07811`, 2020.

295 Raghu Pasupathy and Soumyadip Ghosh. Simulation optimization: A concise overview and implementation guide. In *Theory Driven by Influential Applications*, pages 122–150. INFORMS TutORials in Operations Research, 2013.

296 Chun Peng and Erick Delage. Data-driven optimization with distributionally robust second-order stochastic

dominance constraints, 2020. Optimization Online `http://www.optimization-online.org/DB_HTML/2020/12/8173.html`.

297 Georgia Perakis and Guillaume Roels. Regret in the newsvendor model with partial information. *Oper. Res.*, 56(1): 188–203, 2008.

298 Ian R. Petersen, Matthew R. James, and Paul Dupuis. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Trans. Autom. Control*, 45(3):398–412, 2000.

299 Georg Ch. Pflug and Alois Pichler. A distance for multistage stochastic optimization models. *SIAM J. Optim.*, 22 (1):1–23, 2012.

300 Georg Ch. Pflug and Alois Pichler. The problem of ambiguity in stochastic optimization. In *Multistage Stochastic Optimization*, pages 229–255. Springer, 2014.

301 Georg Ch. Pflug and Mathias Pohl. A Review on Ambiguity in Stochastic Portfolio Optimization. *Set-Valued Var. Anal.*, 26(4):733–757, 2018.

302 Georg Ch. Pflug and David Wozabal. Ambiguity in portfolio selection. *Quant. Finance*, 7(4):435–442, 2007.

303 Georg Ch. Pflug, Alois Pichler, and David Wozabal. The 1/N investment strategy is optimal under high model ambiguity. *J. Bank. Financ.*, 36(2):410–417, 2012.

304 A. B. Philpott, V. L. de Matos, and Lea Kapelevich. Distributionally robust SDDP. *Comput. Manag. Sci.*, 15(3-4): 431–454, 2018.

305 Alois Pichler. Evaluations of Risk Measures for Different Probability Measures. *SIAM J. Optim.*, 23(1):530–551, 2013.

306 Alois Pichler and Alexander Shapiro. Mathematical foundations of distributionally robust multistage optimization. *SIAM J. Optim.*, 31(4):3044–3067, 2021.

307 Alois Pichler and Huifu Xu. Quantitative stability analysis for minimax distributionally robust risk optimization. *Math. Program.*, 191:47–77, 2022.

308 Imre Pólik and Tamás Terlaky. A survey of the S-lemma. *SIAM Rev.*, 49(3):371–418, 2007.

309 Ioana Popescu. A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Math. Oper. Res.*, 30(3):632–657, 2005.

310 Ioana Popescu. Robust mean-covariance solutions for stochastic optimization. *Oper. Res.*, 55(1):98–112, 2007.

311 Krzysztof Postek, Dick den Hertog, and Bertrand Melenberg. Computationally Tractable Counterparts of Distributionally Robust Constraints on Risk Measures. *SIAM Rev.*, 58(4):603–650, 2016.

312 Krzysztof Postek, Aharon Ben-Tal, Dick den Hertog, and Bertrand Melenberg. Robust Optimization with Ambiguous Stochastic Constraints Under Mean and Dispersion Information. *Oper. Res.*, 66(3):814–833, 2018.

313 Krzysztof Postek, Ward Romeijnders, Dick den Hertog, and Maarten H. van der Vlerk. An approximation framework for two-stage ambiguous stochastic integer programs under mean-MAD information. *Eur. J. Oper. Res.*, 274(2): 432–444, 2019.

314 Mehran Poursoltani and Erick Delage. Adjustable robust optimization reformulations of two-stage worst-case regret minimization problems. *Oper. Res.*, 2021. `https://doi.org/10.1287/opre.2021.2159`.

315 Andras Prékopa. On probabilistic constrained programming. In *Proceedings of the Princeton symposium on mathematical programming*, page 138. Princeton University Press, 1970.

316 Andras Prékopa. Programming under probabilistic constraints with a random technology matrix. *Statistics*, 5(2): 109–116, 1974.

317 Andras Prékopa. Probabilistic Programming. In Andrzej Ruszczyński and Alexander Shapiro, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*. Elsevier, 2003.

318 Giovanni Puccetti and Ludger Rüschendorf. Computation of sharp bounds on the distribution of a function of dependent risks. *J. Comput. Appl. Math.*, 236(7):1833–1840, 2012.

319 Giovanni Puccetti and Ludger Rüschendorf. Sharp bounds for sums of dependent risks. *J. Appl. Probab.*, 50(1): 42–53, 2013.

320 Giovanni Puccetti, Ludger Rüschendorf, et al. Bounds for joint portfolios of dependent risks. *Stat. Risk Model.*, 29 (2):107–132, 2012.

321 Martin L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2005.

322 Peng-Yu Qian, Zi-Zhuo Wang, and Zai-Wen Wen. A Composite Risk Measure Framework for Decision Making Under Uncertainty. *J. Oper. Res. Soc. China*, 7(1):43–68, 2019.

323 Svetlozar T. Rachev. *Probability metrics and the stability of stochastic models*. John Wiley & Son Ltd, 1991.

324 Svetlozar T. Rachev and Werner Römisch. Quantitative stability in stochastic programming: The method of probability metrics. *Math. Oper. Res.*, 27(4):792–818, 2002.

325 Svetlozar T. Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume I: Theory*. Springer, 1998.

326 Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. `https://arxiv.org/abs/1908.05659`, 2019.

327 Hamed Rahimian, Güzin Bayraksan, and Tito Homem-de-Mello. Identifying effective scenarios in distributionally

robust stochastic programs with total variation distance. *Math. Program.*, 173(1–2):393–430, 2019.

328 Hamed Rahimian, Güzin Bayraksan, and Tito Homem-de-Mello. Controlling Risk and Demand Ambiguity in Newsvendor Models. *Eur. J. Oper. Res.*, 279(3):854–868, 2019.

329 Hamed Rahimian, Güzin Bayraksan, and Tito Homem-de-Mello. Effective Scenarios in Multistage Distributionally Robust Optimization with a Focus on Total Variation Distance, 2021. to appear in *SIAM J. Optim.*, available on Optimization Online `http://www.optimization-online.org/DB_HTML/2021/09/8588.html`.

330 Arjun Ramachandra, Napat Rujeerapaiboon, and Melvyn Sim. Robust Conic Satisficing, 2021. `https://arxiv.org/abs/2107.06714`.

331 Meisam Razaviyayn, Farzan Farnia, and David Tse. Discrete Rényi Classifiers. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3276–3284. Curran Associates, Inc., 2015.

332 Timothy R. C. Read and Noel A. C. Cressie. *Goodness-of-fit statistics for discrete multivariate data*. Springer, 1988.

333 Rembert Reemtsen and Stephan Görner. Numerical methods for semi-infinite programming: A survey. In R. Reemtsen and J. J. Rückmann, editors, *Semi-infinite Programming, Nonconvex Optimization and Its Applications*, pages 195–275. Kluwer Academic Publishers, 1998.

334 Rolf-Dieter Reiss. *Approximate distributions of order statistics: with applications to nonparametric statistics*. Springer, 1989.

335 Tyrrell R. Rockafellar. *Conjugate Duality and Optimization*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1974.

336 Tyrrell R. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, 1997.

337 Tyrrell R. Rockafellar. Coherent approaches to risk in optimization under uncertainty. In *OR Tools and Applications: Glimpses of Future Technologies*, pages 38–61. INFORMS TutORials in Operations Research, 2007.

338 Tyrrell R. Rockafellar and Johannes O. Royset. Measures of residual risk with connections to regression, risk tracking, surrogate models, and ambiguity. *SIAM J. Optim.*, 25(2):1179–1208, 2015.

339 Tyrrell R. Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *J. Risk*, 2:21–42, 2000.

340 Tyrrell R. Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *J. Bank. Financ.*, 26(7):1443–1471, 2002.

341 Werner Römisch. Stability of Stochastic Programming Problems. In Andrzej Ruszczyński and Alexander Shapiro, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pages 483–554. Elsevier, 2003.

342 Ernst Roos and Dick den Hertog. Reducing conservatism in robust optimization. *INFORMS J. Comput.*, 32(4): 1109–1127, 2020.

343 Johannes O. Royset and Roger J.-B. Wets. Variational theory for optimization under stochastic ambiguity. *SIAM J. Optim.*, 27(2):1118–1149, 2017.

344 Napat Rujeerapaiboon, Daniel Kuhn, and Wolfram Wiesemann. Robust Growth-Optimal Portfolios. *Manage. Sci.*, 62(7):2090–2109, 2016.

345 Napat Rujeerapaiboon, Daniel Kuhn, and Wolfram Wiesemann. Chebyshev Inequalities for Products of Random Variables. *Math. Oper. Res.*, 43(3):887–918, 2018.

346 Napat Rujeerapaiboon, Kilian Schindler, Daniel Kuhn, and Wolfram Wiesemann. Scenario reduction revisited: fundamental limits and guarantees. *Math. Program.*, 191:207–242, 2022.

347 Andrzej Ruszczyński. *Nonlinear optimization*. Princeton University Press, 2006.

348 Andrzej Ruszczyński and Alexander Shapiro. Optimization of convex risk functions. *Math. Oper. Res.*, 31(3): 433–452, 2006.

349 Guillaume Sagnol and Maximilian Stahlberg. PICOS: A Python interface to conic optimization solvers. *J. Open Source Softw.*, 7(70):3915, 2022.

350 Leonard J. Savage. The theory of statistical decision. *J. Am. Stat. Assoc.*, 46(253):55–67, 1951.

351 Herbert Scarf. A min-max solution of an inventory problem. In Herbert Scarf, KJ Arrow, and S Karlin, editors, *Studies in the mathematical theory of inventory and production*, pages 201–209. Stanford University Press, Stanford, CA, 1958.

352 Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

353 Rüdiger Schultz. Some aspects of stability in stochastic programming. *Ann. Oper. Res.*, 100(1-4):55–84, 2000.

354 Chuen-Teck See and Melvyn Sim. Robust approximation to multiperiod inventory management. *Oper. Res.*, 58(3): 583–594, 2010.

355 Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally Robust Logistic Regression. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1576–1584. Curran Associates, Inc., 2015.

356 Soroosh Shafieezadeh-Abadeh, Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Wasserstein Distributionally Robust Kalman Filtering. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi,

and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 8474–8483. Curran Associates, Inc., 2018.

357 Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via Mass Transportation. *J. Mach. Learn. Res.*, 20(103):1–68, 2019.

358 Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

359 Chao Shang and Fengqi You. Robust Optimization in High-Dimensional Data Space with Support Vector Clustering. *IFAC-PapersOnLine*, 51(18):19–24, 2018.

360 Chao Shang and Fengqi You. Distributionally robust optimization for planning and scheduling under uncertainty. *Comput. Chem. Eng.*, 110:53–68, 2018.

361 Chao Shang and Fengqi You. A data-driven robust optimization approach to scenario-based stochastic model predictive control. *J. Process Control*, 75:24–39, 2019.

362 Chao Shang, Xiaolin Huang, and Fengqi You. Data-driven robust optimization based on kernel learning. *Comput. Chem. Eng.*, 106:464–479, 2017.

363 Alexander Shapiro. On Duality Theory of Conic Linear Problems. In Miguel Á. Goberna and Marco A. López, editors, *Semi-Infinite Programming: Recent Advances*, pages 135–165. Springer, 2001.

364 Alexander Shapiro. Minimax and risk averse multistage stochastic programming. *Eur. J. Oper. Res.*, 219(3):719–726, 2012.

365 Alexander Shapiro. On Kusuoka representation of law invariant risk measures. *Math. Oper. Res.*, 38(1):142–152, 2013.

366 Alexander Shapiro. Rectangular sets of probability measures. *Oper. Res.*, 64(2):528–541, 2016.

367 Alexander Shapiro. Distributionally robust stochastic programming. *SIAM J. Optim.*, 27(4):2258–2275, 2017.

368 Alexander Shapiro. Tutorial on risk neutral, distributionally robust and risk averse multistage stochastic programming. *Eur. J. Oper. Res.*, 288(1):1–13, 2021.

369 Alexander Shapiro and Shabbir Ahmed. On a class of minimax stochastic programs. *SIAM J. Optim.*, 14(4): 1237–1249, 2004.

370 Alexander Shapiro and Anton J. Kleywegt. Minimax analysis of stochastic problems. *Optim. Methods Softw.*, 17(3): 523–542, 2002.

371 Alexander Shapiro and Arkadi Nemirovski. On Complexity of Stochastic Programming Problems. In Vaithilingam Jeyakumar and Alexander Rubinov, editors, *Continuous Optimization: Current Trends and Modern Applications*, pages 111–146. Springer, 2005.

372 Alexander Shapiro and Arkadi Nemirovski. On complexity of stochastic programming problems. In Vaithilingam Jeyakumar and Alexander Rubinov, editors, *Continuous Optimization: Current Trends and Modern Applications*, pages 111–146. Springer, 2005.

373 Alexander Shapiro, Wajdi Tekaya, Murilo Pereira Soares, and Joari Paulo da Costa. Worst-case-expectation approach to optimization under uncertainty. *Oper. Res.*, 61(6):1435–1449, 2013.

374 Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2nd edition, 2014.

375 Melvyn Sim, Qinshen Tang, Minglong Zhou, and Taozeng Zhu. The analytics of robust satisficing, 2021. available at SSRN 3829562.

376 Shashank Singh and Barnabás Póczos. Minimax Distribution Estimation in Wasserstein Distance, 2018. `https://arxiv.org/abs/1802.08855`.

377 Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John C. Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training, 2018. `https://arxiv.org/abs/1710.10571`.

378 Maurice Sion. On general minimax theorems. *Pac. J. Math.*, 8(1):171–176, 1958.

379 Elena Smirnova, Elvis Dohmatob, and Jérémie Mary. Distributionally robust reinforcement learning, 2019. `https://arxiv.org/abs/1902.08708`.

380 James E Smith. Generalized Chebychev inequalities: theory and applications in decision analysis. *Oper. Res.*, 43(5): 807–825, 1995.

381 Alexander J. Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, editors, *18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

382 Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electron. J. Stat.*, 6:1550–1599, 2012.

383 Matthew Staib and Stefanie Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, volume 1, 2017.

384 Matthew Staib and Stefanie Jegelka. Distributionally Robust Optimization and Generalization in Kernel Methods. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

385 Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2(Nov):67–93, 2001.

386 Georg Still. Generalized semi-infinite programming: theory and methods. *Eur. J. Oper. Res.*, 119:301–313, 1999.

387 Anirudh Subramanyam, Chrysanthos Gounaris, and Wolfram Wiesemann. K-adaptability in two-stage mixed-integer robust optimization. *Math. Program. Comput.*, 12(2):193–224, 2020.

388 Jie Sun, Li-Zhi Liao, and Brian Rodrigues. Quadratic two-stage stochastic optimization with coherent measures of risk. *Math. Program.*, 168(1-2):599–613, 2018.

389 Tobias Sutter, Bart P. G. Van Parys, and Daniel Kuhn. A general framework for optimal data-driven optimization, 2020. `https://arxiv.org/abs/2010.06606`.

390 Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6(Sep):1453–1484, 2005.

391 Theja Tulabandhula and Cynthia Rudin. Machine learning with operational costs. *J. Mach. Learn. Res.*, 14(1): 1989–2028, 2013.

392 Theja Tulabandhula and Cynthia Rudin. Robust optimization using machine learning for uncertainty sets, 2014. `https://arxiv.org/abs/1407.1097`.

393 Theja Tulabandhula and Cynthia Rudin. On combining machine learning with decision making. *Mach. Learn.*, 97 (1-2):33–64, 2014.

394 Igor Vajda. *Theory of statistical inference and information*. Kluwer Academic Publishers, 1989.

395 Bart P. G. Van Parys. Efficient Data-Driven Optimization with Noisy Data, 2021. `https://arxiv.org/abs/2102.04363`.

396 Bart P. G. Van Parys, Paul J. Goulart, and Daniel Kuhn. Generalized Gauss inequalities via semidefinite programming. *Math. Program.*, 156(1):271–302, 2016.

397 Bart P. G. Van Parys, Daniel Kuhn, Paul J. Goulart, and Manfred Morari. Distributionally Robust Control of Constrained Stochastic Systems. *IEEE Trans. Autom. Control*, 61(2):430–442, 2016.

398 Bart P. G. Van Parys, Paul J. Goulart, and Manfred Morari. Distributionally robust expectation inequalities for structured distributions. *Math. Program.*, 173(1-2):251–280, 2019.

399 Bart P. G. Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From data to decisions: Distributionally robust optimization is optimal. *Manage. Sci.*, 67(6):3387–3402, 2021.

400 Lieven Vandenberghe, Stephen Boyd, and Katherine Comanor. Generalized Chebyshev bounds via semidefinite programming. *SIAM Rev.*, 49(1):52–64, 2007.

401 Phebe Vayanos, Qing Jin, and George Elissaios. ROC++: Robust Optimization in C++, 2020. `https://arxiv.org/abs/2006.08741`.

402 Anand N. Vidyashankar and Jie Xu. Stochastic Optimization Using Hellinger Distance. In *Proceedings of the 2015 Winter Simulation Conference (WSC '15)*, pages 3702–3713, 2015.

403 Cédric Villani. *Optimal transport: old and new*. Springer, 2008.

404 Bin Wang and Ruodu Wang. The complete mixability and convex minimization problems with monotone marginal densities. *J. Multivariate Anal.*, 102(10):1344–1360, 2011.

405 S. Wang and Y. Yuan. Feasible method for semi-infinite programs. *SIAM J. Optim.*, 25(4):2537–2560, 2015.

406 Shanshan Wang, Jinlin Li, and Sanjay Mehrotra. A Solution Approach to Distributionally Robust Joint-Chance-Constrained Assignment Problems. *INFORMS J. Optim.*, 2022. `https://doi.org/10.1287/ijoo.2021.0060`.

407 Zi-Zhuo Wang, Peter W. Glynn, and Yinyu Ye. Likelihood robust optimization for data-driven problems. *Comput. Manag. Sci.*, 13(2):241–261, 2016.

408 Johannes Wiebe and Ruth Misener. ROmodel: modeling robust optimization problems in Pyomo. *Optim. Eng.*, 2021. `https://doi.org/10.1007/s11081-021-09703-2`.

409 Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov Decision Processes. *Math. Oper. Res.*, 38(1): 153–183, 2013.

410 Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Oper. Res.*, 62 (6):1358–1376, 2014.

411 David Wozabal. A framework for optimization under ambiguity. *Ann. Oper. Res.*, 193(1):21–47, 2012.

412 David Wozabal. Robustifying Convex Risk Measures for Linear Portfolios: A Nonparametric Approach. *Oper. Res.*, 62(6):1302–1315, 2014.

413 Weijun Xie. Tractable reformulations of two-stage distributionally robust linear programs over the type-∞ Wasserstein ball. *Oper. Res. Lett.*, 48(4):513–523, 2020.

414 Weijun Xie. On distributionally robust chance constrained programs with Wasserstein distance. *Math. Program.*, 186(1):115–155, 2021.

415 Weijun Xie and Shabbir Ahmed. Distributionally robust simple integer recourse. *Comput. Manag. Sci.*, 15(3): 351–367, 2018.

416 Weijun Xie and Shabbir Ahmed. On Deterministic Reformulations of Distributionally Robust Joint Chance Constrained Optimization Problems. *SIAM J. Optim.*, 28(2):1151–1182, 2018.

417 Weijun Xie, Shabbir Ahmed, and Ruiwei Jiang. Optimized Bonferroni approximations of distributionally robust joint chance constraints. *Math. Program.*, 191:79–112, 2022.

418 Linwei Xin and David A. Goldberg. Time (in) consistency of multistage distributionally robust inventory models with moment constraints. *Eur. J. Oper. Res.*, pages 1127–1141, 2021.

419 Linwei Xin and David A. Goldberg. Distributionally robust inventory control when demand is a martingale. *Math. Oper. Res.*, 2022. `https://doi.org/10.1287/moor.2021.1213`.

420 Guanglin Xu and Samuel Burer. A data-driven distributionally robust bound on the expected optimal value of uncertain mixed 0-1 linear programming. *Comput. Manag. Sci.*, 15(1):111–134, 2018.

421 Huan Xu and Shie Mannor. Distributionally Robust Markov Decision Processes. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2505–2513. Curran Associates, Inc., 2010.

422 Huan Xu and Shie Mannor. Distributionally Robust Markov Decision Processes. *Math. Oper. Res.*, 37(2):288–300, 2012.

423 Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *J. Mach. Learn. Res.*, 10(Jul):1485–1510, 2009.

424 Huan Xu, Constantine Caramanis, and Shie Mannor. Optimization under probabilistic envelope constraints. *Oper. Res.*, 60(3):682–699, 2012.

425 Huifu Xu, Yongchao Liu, and Hailin Sun. Distributionally robust optimization with matrix moment constraints: Lagrange duality and cutting plane methods. *Math. Program.*, 169(2):489–529, 2018.

426 Mengwei Xu, Soon-Yi Wu, and J. Ye Jane. Solving semi-infinite programs by smoothing projected gradient method. *Comput. Math. Appl.*, 59(3):591–616, 2014.

427 Insoon Yang. A dynamic game approach to distributionally robust safety specifications for stochastic systems. *Automatica*, 94:94–101, 2018.

428 Insoon Yang. Wasserstein Distributionally Robust Stochastic Control: A Data-Driven Approach. *IEEE Trans. Autom. Control*, 2020.

429 Wenzhuo Yang and Huan Xu. Distributionally robust chance constraints for non-linear uncertainties. *Math. Program.*, 155(1-2):231–265, 2016.

430 Xiaoqi Yang, Zhangyou Chen, and Jinchuan Zhou. Optimality conditions for semi-infinite and generalized semi-infinite programs via lower order exact penalty functions. *J. Optim. Theory Appl.*, 169(3):984–1012, 2016.

431 İhsan Yanıkoğlu and Dick den Hertog. Safe approximations of ambiguous chance constraints using historical data. *INFORMS J. Comput.*, 25(4):666–681, 2012.

432 İhsan Yanıkoğlu, Bram L. Gorissen, and Dick den Hertog. A survey of adjustable robust optimization. *Eur. J. Oper. Res.*, 277(3):799–813, 2019.

433 Hui Yu, Jia Zhai, and Guang-Ya Chen. Robust Optimization for the Loss-Averse Newsvendor Problem. *J. Optim. Theory Appl.*, 171(3):1008–1032, 2016.

434 Pengqian Yu and Huan Xu. Distributionally robust counterpart in Markov decision processes. *IEEE Trans. Autom. Control*, 61(9):2538–2543, 2016.

435 Xian Yu and Siqian Shen. Multistage Distributionally Robust Mixed-Integer Programming with Decision-Dependent Moment-Based Ambiguity Sets. *Math. Program.*, 2020. doi: 10.1007/s10107-020-01580-4.

436 Jinfeng Yue, Bintong Chen, and Min-Chiang Wang. Expected value of distribution information for the newsvendor problem. *Oper. Res.*, 54(6):1128–1136, 2006.

437 Jitka Žáčková. On minimax solutions of stochastic linear programming problems. *Časopis pro pěstováná matematiky*, 091(4):423–430, 1966.

438 Jie Zhang, Huifu Xu, and Liwei Zhang. Quantitative Stability Analysis for Distributionally Robust Optimization with Moment Constraints. *SIAM J. Optim.*, 26(3):1855–1882, 2016.

439 Yiling Zhang, Ruiwei Jiang, and Siqian Shen. Ambiguous Chance-Constrained Binary Programs under Mean-Covariance Information. *SIAM J. Optim.*, 28(4):2922–2944, 2018.

440 Zhe Zhang, Shabbir Ahmed, and Guanghui Lan. Efficient Algorithms for Distributionally Robust Stochastic Optimization with Discrete Scenario Support. *SIAM J. Optim.*, 31(3):1690–1721, 2021.

441 Zheng Zhang, Brian T. Denton, and Xiaolan Xie. Branch and price for chance-constrained bin packing. *INFORMS J. Comput.*, 32(3):547–564, 2020.

442 Chaoyue Zhao and Yongpei Guan. Data-Driven Risk-Averse Two-Stage Stochastic Program with ζ-Structure Probability Metrics, 2015. Optimization Online `http://www.optimization-online.org/DB_HTML/2015/07/5014.html`.

443 Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Oper. Res. Lett.*, 46(2):262–267, 2018.

444 Chaoyue Zhao and Ruiwei Jiang. Distributionally robust contingency-constrained unit commitment. *IEEE Trans. Power Syst.*, 33(1):94–102, 2018.

445 Jianzhe Zhen, Dick den Hertog, and Melvyn Sim. Adjustable Robust Optimization via Fourier-Motzkin Elimination.

Oper. Res., 66(4):1086–1100, 2018.

446 Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter W. Glynn. Finite-Sample Regret Bound for Distributionally Robust Offline Tabular Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. Proceedings of Machine Learning Research, 2021.

447 Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf. Kernel Distributionally Robust Optimization: Generalized Duality Theorem and Stochastic Approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 280–288. Proceedings of Machine Learning Research, 2021.

448 Steve Zymler, Daniel Kuhn, and Berç Rustem. Distributionally robust joint chance constraints with second-order moment information. *Math. Program.*, 137(1):167–198, 2013.

449 Steve Zymler, Daniel Kuhn, and Berç Rustem. Worst-Case Value at Risk of Nonlinear Portfolios. *Manage. Sci.*, 59 (1):172–188, 2013.