# DISTRIBUTIONALLY ROBUST OPTIMIZATION: A REVIEW

HAMED RAHIMIAN* AND SANJAY MEHROTRA†

**Abstract.** The concepts of risk-aversion, chance-constrained optimization, and robust optimization have developed significantly over the last decade. 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 its relationships with robust optimization, risk-aversion, chance-constrained optimization, and function regularization.

**Key words.** Distributionally robust optimization; Robust optimization; Stochastic optimization; Risk-averse optimization; Chance-constrained optimization; Statistical learning

**AMS subject classifications.** 90C15, 90C22, 90C25, 90C30, 90C34, 90C90, 68T37, 68T05

## 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. Stochastic optimization 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., Shapiro et al. [295], Birge and Louveaux [45]. The probability distribution of the random parameters is inferred from prior beliefs, expert opinions, errors in predictions based on the historical data (e.g., Kim and Mehrotra [171]), or a mixture of these. In robust optimization, 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., El Ghaoui and Lebret [99], El Ghaoui et al. [100], Ben-Tal and Nemirovski [19], Bertsimas and Sim [34], Ben-Tal and Nemirovski [20], Ben-Tal et al. [26]. The concept of robust optimization has a relationship with chance-constrained optimization, where in certain cases there is a direct relationship between a robust optimization model and a chance-constrained optimization model, see, e.g., Boyd and Vandenberghe [57, pp157–158].

We often have partial knowledge on the statistical properties of the model parameters. Specifically, the probability distribution quantifying the model parameter uncertainty is known ambiguously. 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 [311], or maximum entropy principle [126]. 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 [55], (2) stability analysis of a SO model with respect to a change in

*Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208 (hamed.rahimian@northwestern.edu).

†Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208 (mehrotra@northwestern.edu).

the probability distribution, see, e.g, Rachev [250], Römisch [263], or (3) input uncertainty analysis in stochastic simulation models, see, e.g., Lam [175] 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 [96], Schultz [272], Heitsch et al. [142], Rachev and Römisch [248], Pflug and Pichler [230]. While these approaches quantify the input uncertainty, they do not provide a systematic modeling 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 probability distribution by taking a worst-case approach. Scarf [271] 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 the seminal work of Scarf, and particularly in the past few years, significant research has been done on ambiguous stochastic optimization problems. This paper provides a review of the theoretical, modeling, and computational developments in this area. Moreover, we review the applications of the ambiguous stochastic optimization model that have been developed in the recent years. This paper also puts DRO 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 $\tilde{\boldsymbol{\xi}} : \Xi \mapsto \Omega \subseteq \mathbb{R}^d$, a random cost function $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) : \mathcal{X} \times \Xi \mapsto \mathbb{R}$, and a vector of random functions $\boldsymbol{g}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) : \mathcal{X} \times \Xi \mapsto \mathbb{R}^m$, i.e., $\boldsymbol{g}(\boldsymbol{x}, \cdot) := [g_1(\boldsymbol{x}, \cdot), \ldots, g_m(\boldsymbol{x}, \cdot)]^\top$. Given this setup, a general stochastic optimization problem has the form

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \left\{ \mathcal{R}_P \left[ h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right] \,\middle|\, \mathcal{R}_P \left[ \boldsymbol{g}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right] \leq \mathbf{0} \right\}, \tag{SO}$$

where $P$ denotes the (known) probability measure on $(\Xi, \mathcal{F})$ and $\mathcal{R}_P : \mathcal{Z} \mapsto \mathbb{R}$ denotes a (componentwise) 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, for a given fixed probability measure $P$. This setup represents a broad range of problems in statistics, optimization, and control, such as regression and classification models [106, 163], simulation-optimization [107, 227], stochastic optimal control [31], Markov decision processes [246], and stochastic programming [45, 295].

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

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \ \mathbb{E}_P \left[ h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right], \tag{1.1}$$

and

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \left\{ h(\boldsymbol{x}) \,\middle|\, \mathbb{E}_P \left[ \boldsymbol{g}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right] \leq \mathbf{0} \right\}, \tag{1.2}$$

where $\mathcal{R}_P[\cdot]$ is taken as the expected-value functional $\mathbb{E}_P[\cdot]$. Note that by taking $h(\boldsymbol{x},\cdot) := \mathbb{1}_{A(\boldsymbol{x})}(\cdot)$ in (1.1), where $\mathbb{1}_{A(\boldsymbol{x})}(\cdot)$ denotes an indicator function for an arbitrary set $A(\boldsymbol{x}) \subseteq \mathcal{B}(\mathbb{R}^d)$ (we define the indicator function and $\mathcal{B}(\mathbb{R}^d)$ precisely in Section 2), we obtain the class of problems with a probabilistic objective function of the form $P\{\tilde{\boldsymbol{\xi}} \in A(\boldsymbol{x})\}$, see, e.g., Prékopa [244]. The set $A(\boldsymbol{x})$ is called a *safe region* and may be of the form $\boldsymbol{a}(\boldsymbol{x})^\top\tilde{\boldsymbol{\xi}} \leq \boldsymbol{b}(\boldsymbol{x})$ or $\boldsymbol{a}(\tilde{\boldsymbol{\xi}})^\top\boldsymbol{x} \leq \boldsymbol{b}(\tilde{\boldsymbol{\xi}})$[1]. Similarly, by taking $h(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) := h(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x},\cdot) := [\mathbb{1}_{A_1(\boldsymbol{x})}(\cdot),\ldots,\mathbb{1}_{A_m(\boldsymbol{x})}(\cdot)]^\top$, for suitable indicator functions $\mathbb{1}_{A_j(\boldsymbol{x})}(\cdot)$, $j=1,\ldots,m$, (1.2) is in the form of probabilistic (i.e., chance) constraints $P\{\tilde{\boldsymbol{\xi}} \in A_j(\boldsymbol{x})\} \leq 0,\ j=1,\ldots,m$, see, e.g., Charnes et al. [68], Charnes and Cooper [67], Prékopa [243, 245], Dentcheva [86]. Note that the case where the event $\{\tilde{\boldsymbol{\xi}} \in A_j(\boldsymbol{x})\}$ is formed via several constraints is called *joint chance constraint* as compared to *individual chance constraint*, where the event $\{\tilde{\boldsymbol{\xi}} \in A_j(\boldsymbol{x})\}$ is formed via one constraint.

A robust optimization model is defined as

$$\text{(RO)} \qquad \inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\boldsymbol{\xi}\in\mathcal{U}} \left\{ h(\boldsymbol{x},\boldsymbol{\xi}) \,\middle|\, \sup_{\boldsymbol{\xi}\in\mathcal{U}} \boldsymbol{g}(\boldsymbol{x},\boldsymbol{\xi}) \leq \boldsymbol{0} \right\},$$

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

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\boldsymbol{\xi}\in\mathcal{U}} h(\boldsymbol{x},\boldsymbol{\xi}) \tag{1.3}$$

and

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \left\{ h(\boldsymbol{x}) \,\middle|\, \sup_{\boldsymbol{\xi}\in\mathcal{U}} \boldsymbol{g}(\boldsymbol{x},\boldsymbol{\xi}) \leq \boldsymbol{0} \right\} \tag{1.4}$$

are two special cases of (RO).

Problem (SO), as well as (1.1) and (1.2), require the knowledge of the underlying measure $P$, whereas (RO), as well as (1.3) and (1.4), ignore all distributional knowledge of $\tilde{\boldsymbol{\xi}}$, except for its support. An ambiguous version of (SO) is formulated as

$$\text{(DRO)} \qquad \inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{P\in\mathcal{P}} \left\{ \mathcal{R}_P\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \,\middle|\, \sup_{P\in\mathcal{P}} \mathcal{R}_P\left[\boldsymbol{g}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \leq \boldsymbol{0} \right\}.$$

Here, $\mathcal{P}$ denotes the *ambiguity set of probability measures*, i.e., a family of measures consistent with the prior knowledge about uncertainty. Note that if we consider the measurable space $(\Omega,\mathcal{B})$, where $\mathcal{B}$ denotes the Borel $\sigma$-field on $\Omega$, i.e., $\mathcal{B} = \Omega \cap \mathcal{B}(\mathbb{R}^d)$, then $\mathcal{P}$ can be viewed as an ambiguity set of probability distributions $\mathbb{P}$ defined on $(\Omega,\mathcal{B})$ and induced by $\tilde{\boldsymbol{\xi}}$[2].

As discussed before, (DRO) finds a decision that minimizes the worst-case of the functional $\mathcal{R}$ of the cost $h$ among all probability measures in the ambiguity set

[1] We say a safe region of the form $\boldsymbol{a}(\boldsymbol{x})^\top\tilde{\boldsymbol{\xi}} \leq \boldsymbol{b}(\boldsymbol{x})$ is bi-affine in $\boldsymbol{x}$ and $\boldsymbol{\xi}$ if $\boldsymbol{a}(\boldsymbol{x})$ and $\boldsymbol{b}(\boldsymbol{x})$ are both affine in $\boldsymbol{x}$. Similarly, we say a safe region of the form $\boldsymbol{a}(\tilde{\boldsymbol{\xi}})^\top\boldsymbol{x} \leq \boldsymbol{b}(\tilde{\boldsymbol{\xi}})$ is bi-affine in $\boldsymbol{x}$ and $\boldsymbol{\xi}$ if $\boldsymbol{a}(\tilde{\boldsymbol{\xi}})$ and $\boldsymbol{b}(\tilde{\boldsymbol{\xi}})$ are both affine in $\tilde{\boldsymbol{\xi}}$. Observe that a bi-affine safe region of the form $\boldsymbol{a}(\boldsymbol{x})^\top\tilde{\boldsymbol{\xi}} \leq \boldsymbol{b}(\boldsymbol{x})$ can be equivalently written as a bi-affine safe region of the form $\boldsymbol{a}(\tilde{\boldsymbol{\xi}})^\top\boldsymbol{x} \leq \boldsymbol{b}(\tilde{\boldsymbol{\xi}})$, and vice versa.

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

provided that the (componentwise) worst-case of the functional $\mathcal{R}$ of the function $\boldsymbol{g}$ is non-positive. The ambiguous versions of (1.1) and (1.2) are formulated as follows:

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{P\in\mathcal{P}} \mathbb{E}_P \left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right], \tag{1.5}$$

and

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \left\{h(\boldsymbol{x}) \,\middle|\, \sup_{P\in\mathcal{P}} \mathbb{E}_P \left[\boldsymbol{g}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \leq \mathbf{0}\right\}. \tag{1.6}$$

Models (1.5) and (1.6) 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 [271]; Žáčková [313] (a.k.a. Dupačová); Dupačová [95], Breton and El Hachem [58], Shapiro and Kleywegt [292], Shapiro and Ahmed [291]. Delage and Ye [82] 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.

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 $\tilde{\boldsymbol{\xi}}$, (DRO) reduces to (SO). On the other hand, if $\mathcal{P}$ contains all probability distributions on the support of the random vector $\tilde{\boldsymbol{\xi}}$, supported on $\mathcal{U}$, then, (DRO) reduces to (RO). Thus, a judicial 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 also Qian et al. [247]).

**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. While there are separate review papers on RO, see, e.g., [40, 108, 124], to the best of our knowledge, there are a few tutorials and survey papers on DRO within the operations research community. A tutorial on DRO, its connection to risk-averse optimization, and the use of $\phi$-divergence to construct the ambiguity set is presented in Bayraksan and Love [13]. Shapiro [290] provides a general tutorial on DRO and its connection to risk-averse optimization. Postek et al. [240] surveys different papers that address distributionally robust risk constraints, with a variety of risk functional and ambiguity sets. Similar to [13, 290, 240], in this paper, we show the connection between DRO and risk aversion. However, the current review is different from those in the literature from a number of 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 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. 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. In Section 2, we introduce the notation and the basic definitions. Section 3 reviews the connection of DRO to different concepts: game theory in Section 3.1, robust optimization in Section 3.2, risk-aversion and chance-constrained optimization with its relationship to robust optimization in Section 3.3, and regularization in statistical learning in Section 3.4. In Section 4, we review two main solution techniques to solve a DRO model by introducing tools in semi-infinite programming and duality. In Section 5, we discuss different models to construct the ambiguity set of distributions. This includes discrepancy-based models in Section 5.1, moment-based models in Section 5.2, shape-preserving-based models in Section 5.3, and kernel-based models in Section 5.4. In Section 6, we discuss the calibration of different parameters used in the ambiguity set of distributions. In Section 7, we discuss different functionals that amount for quantifying the uncertainty in the outcomes of a fixed decision. This includes regret functions in Section 7.1, risk measures in Section 7.2, and utility functions in Section 7.3. In Section 8, we introduce some modeling toolboxes for a DRO model.

**2. 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 in this section.

For a given space $\Xi$ and a $\sigma$-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 $\sigma$-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 $\mathfrak{M}_+(\cdot, \cdot)$ and $\mathfrak{M}(\cdot, \cdot)$ 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 $\nu_2$ is preferred over a measure $\nu_1$, 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 $\tilde{\boldsymbol{\xi}} : (\Xi, \mathcal{F}) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is always denoted with a tilde sign, while a realization of the random vector $\tilde{\boldsymbol{\xi}}$ is denoted by the same symbol without a tilde, i.e., $\boldsymbol{\xi}$. For a probability measure $Q \in \mathfrak{M}(\Xi, \mathcal{F})$, we define a probability space $(\Xi, \mathcal{F}, Q)$. We denote by $\mathbb{Q} := Q \circ \tilde{\boldsymbol{\xi}}^{-1}$ the probability distribution induced by a random vector $\tilde{\boldsymbol{\xi}}$ under $Q$, where $\tilde{\boldsymbol{\xi}}^{-1}$ denotes the inverse image of $\tilde{\boldsymbol{\xi}}$. 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 $\tilde{\boldsymbol{\xi}}$. 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 $\int_\Xi f(\tilde{\boldsymbol{\xi}}(s))Q(ds)$ equivalently as $\int_{\mathbb{R}^d} f(s)\mathbb{Q}(ds)$ with a change of measure. As we shall see later, we may denote $f(\tilde{\boldsymbol{\xi}}(s))$ 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 $\Xi$. A random variable $Z$ is $Q$-integrable if $\|Z\|_1 := \int_\Xi |Z| dQ$ 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 a finite space $\Xi$ 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 $q_i := q(s_i)$, $i \in \{1, \ldots, M\}$. A $\mathcal{F}$-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 $A : \mathcal{V} \mapsto \mathcal{W}$, we define the adjoint mapping $A^* : \mathcal{W}^* \mapsto \mathcal{V}^*$ by means of the equation $\langle w^*, Av \rangle = \langle A^* w^*, v \rangle$, $\forall v \in \mathcal{V}$. For two linear mappings, defined by finite dimensional matrices $A$ and $B$, $A \bullet B = Tr(A^T B)$ 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 $\ell_p$-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 $\Delta^d$ 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 $\mathbb{R}^d$. Let $(\cdot)_+$ denote $\max\{0, \cdot\}$.

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 $\mathcal{K}' := \{v^* \in \mathcal{V}^* \mid \langle v^*, v \rangle \geq 0,\ \forall v \in \mathcal{K}\}$. The negative of the dual cone is called polar cone and is denoted by $\mathcal{K}^{\mathrm{o}}$. 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 $\mathrm{conv}(\mathcal{K})$ and $\mathrm{int}\,(\mathcal{K})$ to denote the convex hull and the interior of $\mathcal{K}$, respectively.

Because we also review DRO papers in the context of statistical learning in this paper, we introduce some terminologies in statistical learning. For every approach

that uses a set of (training) data 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) data, independent from the training set. Consider a given set of (training) data $\{\boldsymbol{\xi}^i\}_{i=1}^N$. Suppose that $\mathbb{P}_N$ is the empirical probability distribution on $\{\boldsymbol{\xi}^i\}_{i=1}^N$. Data-driven approaches are interested in the performance of a data-driven solution (or, in-sample solution) $\hat{\boldsymbol{x}}_N$ that is constructed using $\{\boldsymbol{\xi}^i\}_{i=1}^N$. A primitive data-driven solution for a problem of the form (1.1) can be obtained by solving a *sample average approximation* (SAA) of that problem, where the underlying distribution is chosen to be $\mathbb{P}_N$ [295]. Assessing the quality of this solution is well-studied in the context of SO, see, e.g., Bayraksan and Morton [14, 15], Homem-de-Mello and Bayraksan [147]. Here, we introduce the analogous of such performance measure that are used to assess the quality of a solution in the context of a DRO model. Let us focus on a DRO problem of the form (1.5) for the ease of exposition. 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 (1.5), where the ambiguity set $\mathcal{P}$ is constructed using data, namely $\mathcal{P}_N$. The out-of-sample performance of $\boldsymbol{x}_N$ is defined as $\mathbb{E}_{\mathbb{P}^{\text{true}}}\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}\right]$, which is the expected cost of $\boldsymbol{x}_N$ given a new (test) sample that is independent of $\{\boldsymbol{\xi}^i\}_{i=1}^N$, drawn from an unknown true distribution $\mathbb{P}^{\text{true}} := P^{\text{true}} \circ \tilde{\boldsymbol{\xi}}^{-1}$. However, as $\mathbb{P}^{\text{true}}$ is unknown, one need to establish performance guarantees. One such guarantee, referred to as *finite-sample performance guarantee* or *generalization bound* is defined as

$$\mathbb{P}_N\left\{\mathbb{E}_{\mathbb{P}^{\text{true}}}\left[h(\boldsymbol{x}_N, \tilde{\boldsymbol{\xi}}\right] \leq \hat{V}_N\right\} \geq 1 - \alpha,$$

which guarantees that an (in-sample) *certificate* $\hat{V}_N$ provides a $(1-\alpha)$ confidence (with respect to the training sample) on the out-of-sample performance of $\boldsymbol{x}_N$. The certificate $\hat{V}_N$ 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 $\boldsymbol{x}_N$, see, e.g., [205]. The other guarantee, referred to as *asymptotic consistency*, guarantees that as $N$ increases, the certificate $\hat{V}_N$ and the data-driven solution $\boldsymbol{x}_N$ converges—in some sense—to the optimal value and an optimal solution of the true (unambiguous) problem of the form (1.1), see, e.g., [205].

## 3. Relationship with Game Theory, Risk-Aversion, Chance-Constrained Optimization, and Regularization.

**3.1. Relationship with Game Theory.** In this section, we present a game-theoretic interpretation of DRO. Indeed, a worst-case approach to SO may be viewed to have its roots in John von Neumann's game theory. For ease of exposition, let us consider a problem of the form (1.5).

The decision maker, the first player in this setup, makes a decision $\boldsymbol{x} \in \mathcal{X}$ whose consequences (i.e., cost $h$) depends on the outcome of the random vector $\tilde{\boldsymbol{\xi}}$. The decision maker assumes that $\tilde{\boldsymbol{\xi}}$ 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 $\tilde{\boldsymbol{\xi}}$. Thus, in one hand, the decision maker is looking for a decision that minimizes the maximum expected cost with respect to $\mathcal{P}$, 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 minmax theorem.

THEOREM 3.1. *(Sion [299, Theorem 3.4]) Suppose that*
*(i)* $\mathcal{X}$ *and* $\mathcal{P}$ *are convex and compact spaces,*
*(ii)* $\boldsymbol{x} \mapsto \mathcal{R}_{\mathbb{P}}\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$ *is upper semicontinuous and quasiconcave on* $\mathcal{P}$ *for all* $\boldsymbol{x} \in \mathcal{X}$*, and*
*(iii)* $\mathbb{P} \mapsto \mathcal{R}_{\mathbb{P}}\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$ *is lower semicontinuous and quasiconvex on* $\mathcal{X}$ *for all* $\mathbb{P} \in \mathcal{P}$.

*Then,*

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

According to the above theorem, under appropriate conditions, the exchange of the order between inf and sup will not change the optimal value to $\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$. We refer to Grünwald and Dawid [126] 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 [126]: a probability distribution $\mathbb{P}^*$ that maximizes the *generalized entropy* $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathcal{R}_{\mathbb{P}}\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$ over $\mathcal{P}$ has an associated decision $\boldsymbol{x}^*$, achieving $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathcal{R}_{\mathbb{P}^*}\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$, and it achieves $\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}} \mathcal{R}_{\mathbb{P}}\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$.

**3.2. Relationship between DRO and RO.** In Section 1, we mentioned that when the ambiguity set of probability distributions contains all probability distributions on the support of the uncertain parameters, DRO and RO are equivalent. In this section, we present a different perspective on the relationship between DRO and RO under the assumption that the sample space $\Xi$ is finite. For ease of exposition, we focus on (1.5). A similar argument follows for (DRO).

Suppose that $\Xi$ is a finite sample space with $M$ atoms, $\Xi = \{s_1, \ldots, s_M\}$. Then, for a fixed $x \in \mathcal{X}$, $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ has $M$ possible outcomes $\{h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}(s_1)), \ldots, h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}(s_M))\}$. For short, let us write these outcomes as a vector $\boldsymbol{h}(\boldsymbol{x}) \in \mathbb{R}^M$, where $h_m(\boldsymbol{x}) := h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}(s_m))$. In (1.5), $\mathcal{P}$ is a subset of all probability measures on $\tilde{\boldsymbol{\xi}}$. So, one can think of $\mathcal{P}$ as a subset of all discrete probability distributions $\mathbb{P}$ on $\mathbb{R}^d$ induced by $\tilde{\boldsymbol{\xi}}$. 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, (1.5) is written as

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\boldsymbol{p}\in\mathcal{P}} \boldsymbol{p}^\top \boldsymbol{h}(\boldsymbol{x}). \tag{3.1}$$

By defining $f(\boldsymbol{x}, \boldsymbol{p}) := \boldsymbol{p}^\top \boldsymbol{h}(\boldsymbol{x})$, we can rewrite the above problem as $\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\boldsymbol{p}\in\mathcal{P}} f(\boldsymbol{x}, \boldsymbol{p})$. This problem has the form of (1.3), where the probability vector $\boldsymbol{p}$ takes values in an "uncertainty set" $\mathcal{P}$. Techniques that are applicable for specifying the uncertainty set in a RO model may now be used to specify $\mathcal{P}$ in (3.1), see, e.g., Ben-Tal and Nemirovski [18, 20], Bertsimas et al. [37], Chen et al. [74]. We also refer to Bertsimas et al. [42] and Section 3.3.2. 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. [29]. However, as we shall see below, DRO has the richness that allows the use of techniques developed in the statistical literature to model the problem. Moreover, its framework allows $\Xi$ to be continuous. We also refer to Xu et al. [330] for a distributional interpretation of RO.

### 3.3. Relationship with Risk-Aversion.

**3.3.1. 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 (1.5) or (1.6) are equivalent to a *coherent* risk measure [7, 258, 270]. Furthermore, under mild conditions, the worst-case expectations given in (1.5) or (1.6) are equivalent to a *law invariant* risk measure [289]. These results imply that DRO models have an equivalent risk-averse optimization problem. In order to explain the relationship between (1.5) and (1.6) and risk-averse optimization more precisely, we present some definitions and fundamental results.

DEFINITION 3.2. *(Artzner et al. [7, Definition 2.4], Shapiro et al. [295, 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: $\rho\left(tZ + (1-t)Z'\right) \leq t\rho(Z) + (1-t)\rho(Z')$, for all $Z, Z' \in \mathcal{Z}$ and all $t \in [0, 1]$.*

*A risk measure $\rho$ is called convex if it satisfies all the above axioms besides the positive homogeneity condition.*

*Remark* 3.3. In Definition 3.2, the convexity axiom can be replaced with the *subadditivity* axiom: $\rho\left(Z + Z'\right) \leq \rho(Z) + \rho(Z')$, 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. [7, Definition 2.4] defines a coherent risk measure with the subadditivity axiom, whereas Shapiro et al. [295, Definition 6.4] defines a coherent risk measure with the convexity axiom.

DEFINITION 3.4. *(Shapiro [289, Definition 2.1]) A (real-valued) risk measure $\rho : \mathcal{Z} \mapsto \mathbb{R}$ is called law invariant if for all $Z, Z' \in \mathcal{Z}$, $Z \overset{d}{\sim} Z'$ implies that $\rho(Z) = \rho(Z')$.*

DEFINITION 3.5. *(Shapiro [289, Definition 2.2]) A set $\mathcal{M}$ is called law invariant 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 $\tilde{\boldsymbol{\xi}}$ to coherent risk measures, we adopt the following result from Shapiro et al. [295, Theorem 6.7], Shapiro [286, Theorem 3.1].

THEOREM 3.6. *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 $\int_\Xi |dP| < \infty$. 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[3]. Recall $\mathfrak{M}(\Xi, \mathcal{F})$ denotes the space of all probability measures on $(\Xi, \mathcal{F})$: $\mathfrak{M}(\Xi, \mathcal{F}) = \left\{P \in \mathcal{Z}^* \,\middle|\, \int_\Xi dP = 1,\ P \succcurlyeq 0\right\}$. Let $\rho : \mathcal{Z} \mapsto \overline{\mathbb{R}}$. Then, $\rho$ is a real-valued coherent risk measure if and only if there exists a convex compact set $\mathcal{M} \subseteq \mathfrak{M}(\Xi, \mathcal{F})$*

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

*(in the weakly* topology of $\mathcal{Z}^*$) such that*

$$\rho(Z) = \sup_{P\in\mathcal{M}} \mathbb{E}_P[Z], \ \forall Z \in \mathcal{Z}. \tag{3.2}$$

*Moreover, given a real-valued coherent risk measure, the set $\mathcal{M}$ in (3.2) 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.* 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. [295, Theorem 6.7], we can show that if $\rho$ is a proper and lower semicontinuous coherent risk measure, then (3.2) holds when $\mathcal{M}$ is equal to the subdifferential of $\rho$ at $0 \in \mathcal{Z}$, i.e., $\mathcal{M} = \partial\rho(0)$, where

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

Now, we show that $\rho$ 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 $Z \ge Z'$ if and only if $Z(s) \ge Z'(s)$ a.e. on $\Xi$. 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[4]. Thus, by Shapiro et al. [295, Theorem 7.91], we conclude that $\rho$ is continuous and subdifferentiable on the interior of its domain. This, in turn, implies that the lower semicontinuity of $\rho$ is automatically satisfied. Moreover, by Shapiro et al. [295, Theorem 7.85], the subdifferentials of $\rho$ 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 (3.2) holds with the set $\mathcal{M}$ being a convex and weakly* compact subset of $\mathfrak{M}(\Xi, \mathcal{F})$. Then, $\rho$ is a real-valued coherent risk measure.

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

Before we proceed, let us characterize the set $\mathcal{M}$, as described in Theorem 3.6, for three well-studied coherent risk measures, namely *conditional value-at-risk* (CVaR), see, e.g., Rockafellar and Uryasev [260, 261], Rockafellar [258], convex combination of expectation and CVaR, see, e.g., Zhang et al. [339], and *mean-upper-absolute semideviation*, see, e.g., Shapiro et al. [295]. CVaR at level $\beta$, $0 < \beta < 1$, denoted by $\mathrm{CVaR}^Q_\beta[\cdot]$, is defined as $\mathrm{CVaR}^Q_\beta[Z] := \frac{1}{1-\beta}\int_\beta^1 \mathrm{VaR}_\alpha[Z]\, d\alpha$, where $\mathrm{VaR}^Q_\alpha[Z] := \inf\{u \,|\, Q\{Z \le u\} \ge \alpha\}$ is the Value-at-Risk (VaR) at level $\alpha$. 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* 3.7. Consider a probability space $(\Xi, \mathcal{F}, Q)$ and $\mathcal{Z} = \mathcal{L}_\infty(\Xi, \mathcal{F}, Q)$. Suppose that $\Xi$ is a finite space with $M$ atoms. For a coherent risk measure $\rho$, we have $\rho(Z) = \sup_{\boldsymbol{p}\in\mathcal{M}} \left\{\sum_{m=1}^M Z_m p_m\right\}$, $\forall Z \in \mathcal{Z}$, where $\mathcal{M}$ is closed convex subset of

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

[4]It is said a partial order relation induces a *lattice structure* on $\mathcal{Z}$ if the least upper bound exists for any $Z, Z' \in \mathcal{Z}$ [295]. A Banach space $\mathcal{Z}$ with lattice structure is called *Banach lattice* if $Z, Z' \in \mathcal{Z}$ and $|Z| \ge |Z'|$ implies $\|Z\| \ge \|Z'\|$ [295].

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

- When $\rho(Z) = \mathrm{CVaR}_\beta^Q[Z]$, we have
$$\mathcal{M} = \left\{ \boldsymbol{p} \in \mathcal{D} \,\middle|\, p_m \in [0, \frac{q_m}{1-\beta}],\ m = 1, \ldots, M \right\}.$$
- 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
$$\mathcal{M} = \{ \boldsymbol{p} \in \mathcal{D} \,|\, p_m \in [q_m\gamma_1, q_m\gamma_2],\ m = 1, \ldots, M \}.$$
The above risk measure is also equivalent to $\gamma_1 \mathbb{E}_Q[Z] + (1-\gamma_1)\mathrm{CVaR}_\beta^Q[Z]$, 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 3.6 relates problems (1.5) and (1.6) to risk-averse optimization problems, involving the coherent risk-measure $\rho$. Consider a fixed $\boldsymbol{x} \in \mathcal{X}$. With an appropriate transformation of measure $\mathbb{P} = P \circ \tilde{\boldsymbol{\xi}}^{-1}$, we can write the inner problem $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$ in (1.5) as $\sup_{P \in \mathcal{P}} \mathbb{E}_P[h(\boldsymbol{x}, s)]$, where in the former, $\mathcal{P}$ is a set of probability distributions induced by $\tilde{\boldsymbol{\xi}}$, while in the latter, $\mathcal{P}$ is a set of probability measures on $(\Xi, \mathcal{F})$. Then, by applying Theorem 3.6 and setting $Z = h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$, $\sup_{P \in \mathcal{P}} \mathbb{E}_P[h(\boldsymbol{x}, s)]$ evaluates a (real-valued) coherent risk measure $\rho[h(\boldsymbol{x}, s)]$, provided that $\mathcal{P} \subset \mathfrak{M}(\Xi, \mathcal{F})$ is a convex compact set. It is easy to verify that such a function $\rho$ is coherent:

- *Translation Equivariance:* Consider $\boldsymbol{x} \in \mathcal{X}$ and $a \in \mathbb{R}$. Then, $\rho[h(\boldsymbol{x}, s) + a] = \sup_{P \in \mathcal{P}} \mathbb{E}_P[h(\boldsymbol{x}, s) + a] = \sup_{P \in \mathcal{P}} \mathbb{E}_P[h(\boldsymbol{x}, s)] + a = \rho[h(\boldsymbol{x}, s)] + a$.
- *Positive Homogeneity:* Consider $\boldsymbol{x} \in \mathcal{X}$ and $t \ge 0$. Then, $\rho[th(\boldsymbol{x}, s)] = \sup_{P \in \mathcal{P}} \mathbb{E}_P[th(\boldsymbol{x}, s)] = t \sup_{P \in \mathcal{P}} \mathbb{E}_P[h(\boldsymbol{x}, s)] = t\rho[h(\boldsymbol{x}, s)]$.
- *Monotonicity:* Consider $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ such that $h(\boldsymbol{x}, s) \ge h(\boldsymbol{x}', s)$. Thus, $\mathbb{E}_P[h(\boldsymbol{x}, s)] \ge \mathbb{E}_P[h(\boldsymbol{x}', s)]$ for any $P \in \mathcal{P}$, which implies $\rho[h(\boldsymbol{x}, s)] = \sup_{P \in \mathcal{P}} \mathbb{E}_P[h(\boldsymbol{x}, s)] \ge \sup_{P \in \mathcal{P}} \mathbb{E}_P[h(\boldsymbol{x}', s)] = \rho[h(\boldsymbol{x}', s)]$.
- *Convexity:* Consider $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and $t \in [0,1]$. Then, we have
$$\begin{aligned}
\rho[th(\boldsymbol{x}, s) + (1-t)h(\boldsymbol{x}', s)] &= \sup_{P \in \mathcal{P}} \mathbb{E}_P[th(\boldsymbol{x}, s) + (1-t)h(\boldsymbol{x}', s)] \\
&\le \sup_{P \in \mathcal{P}} \mathbb{E}_P[th(\boldsymbol{x}, s)] + \sup_{P \in \mathcal{P}} \mathbb{E}_P[(1-t)h(\boldsymbol{x}', s)] \\
&= t \sup_{P \in \mathcal{P}} \mathbb{E}_P[h(\boldsymbol{x}, s)] + (1-t) \sup_{P \in \mathcal{P}} \mathbb{E}_P[h(\boldsymbol{x}', s)] \\
&= t\rho[h(\boldsymbol{x}, s)] + (1-t)\rho[h(\boldsymbol{x}', s)],
\end{aligned}$$
where we used the translation equivariance property.

Consequently, (1.5) is equivalent to minimizing a coherent risk measure. Similarly, (1.6) 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 3.6 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 $h$ is a convex random function in (1.5), i.e., $h(\cdot, \boldsymbol{\xi})$ is convex in $\boldsymbol{x}$ for almost every $\boldsymbol{\xi}$, then, $\rho\left[h(\cdot, \tilde{\boldsymbol{\xi}})\right]$ is convex in $\boldsymbol{x}$. Convexity of $\boldsymbol{g}$ in (1.6) also implies the convexity of the region induced by the risk constraints $\rho\left[\boldsymbol{g}(\cdot, \tilde{\boldsymbol{\xi}})\right] \leq \mathbf{0}$. In our setup, neither $h(\cdot, \boldsymbol{\xi})$ nor $\boldsymbol{g}(\cdot, \boldsymbol{\xi})$ 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 $\tilde{\boldsymbol{\xi}}$ to law invariant risk measures.

THEOREM 3.8. *(Shapiro [289, Theorem 2.3]) Consider $\mathcal{Z}$ and $\mathcal{Z}^*$ as defined in Theorem 3.6. Also, consider $\rho : \mathcal{Z} \mapsto \mathbb{R}$, defined as $\rho(Z) = \sup_{P \in \mathcal{P}} \mathbb{E}_P[Z]$, $\forall Z \in \mathcal{Z}$, If the set $\mathcal{P}$ is law invariant, then the corresponding risk measure $\rho$ is law invariant. Conversely, if the risk measure $\rho$ 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 [286, 288, 290].

**3.3.2. 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 $\tilde{\boldsymbol{\xi}}$, as in RO.

Many approaches in RO construct the uncertainty set for the parameters $\tilde{\boldsymbol{\xi}}$ 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 $g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0$ for a fixed $\boldsymbol{x}$. Suppose that $\tilde{\boldsymbol{\xi}}$ belongs to a bounded uncertainty set $\mathcal{U} \subseteq \mathbb{R}^d$, i.e., $\mathcal{U}$ is the support of $\tilde{\boldsymbol{\xi}}$. The RO counterpart of this constraint then can be formulated as

$$g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0, \ \forall \boldsymbol{\xi} \in \mathcal{U}. \tag{3.3}$$

Two criticisms of (3.3) are that: (1) it treats all uncertain parameters $\boldsymbol{\xi} \in \mathcal{U}$ with equal weights and (2) all the parametrized 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 $\tilde{\boldsymbol{\xi}}$) instead of enforcing the constraint to be satisfied almost everywhere. Under the assumption that $\tilde{\boldsymbol{\xi}}$ is defined on a probability space $(\Xi, \mathcal{F}, P^{\text{true}})$, the chance constraint framework can be represented as follows:

$$P^{\text{true}}\{g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0\} \geq 1 - \epsilon, \tag{3.4}$$

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

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

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 (3.4), many approaches in RO construct an uncertainty set $\mathcal{U}_\epsilon$ such that a feasible solution to a problem of the form (3.3) will also be feasible with probability at least $1-\epsilon$ with respect to $P^{\text{true}}$. More precisely, for any fixed $\boldsymbol{x}$, these constructions guarantee that the following implication holds:

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

However, as we argued before, the probability measure $P^{\text{true}}$ 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[5] to (3.4)[6]: (1) scenario approximation scheme of (3.3) based on Monte Carlo sampling, see, e.g., Campi and Calafiore [63], Calafiore and Campi [60], Nemirovski and Shapiro [214], Campi and Garatti [62], Luedtke and Ahmed [198], Ben-Tal and Nemirovski [21], and (2) DRO approach to (3.4), see, e.g., Nemirovski and Shapiro [213], Erdoğan and Iyengar [102]. Research on scenario approximation of (3.3) focuses on providing probabilistic guarantee (with respect to the sample probability measure) that a solution to the sampled problem of (3.3) is feasible to (3.4) with a high probability.

The DRO approach, on the other hand, forms a version of (3.4) as follows:

$$P\{g(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \le 0\} \ge 1-\epsilon,\ \forall P \in \mathcal{P} \equiv \inf_{P\in\mathcal{P}}\ P\{g(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \le 0\} \ge 1-\epsilon. \tag{3.5}$$

Let $\bar{\mathcal{X}}_\epsilon$ denote the feasibility set induced by (3.5):

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

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 $\mathcal{X}_\epsilon$[7]. By leveraging a goodness-of-fit test, Bertsimas et al. [42] construct a $(1-\alpha)$-confidence region $\mathcal{P}(\alpha)$ for $P^{\text{true}}$. Such a construction leads to an uncertainty set $\mathcal{U}_\epsilon(\alpha)$ that guarantees the implication (C1) [42].

Let us now assume that the sample space $\Xi$ is finite. By the relationship between RO and DRO, discussed in Section 3.2, one may think the parameter $\boldsymbol{\xi}$ in (3.3) represents a probability distribution $\boldsymbol{p}$ on $\mathbb{R}^d$, which is random. That said, we may define $f(\boldsymbol{x},\boldsymbol{p}) := \mathcal{R}_{\boldsymbol{p}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]$. By leveraging the results in Bertsimas et al. [42], we aim to construct a data-driven ambiguity set $\mathcal{P}_\epsilon$ that guarantees the following implication:

$$\text{(C2)} \qquad \text{If } \mathcal{R}_{\boldsymbol{p}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \le 0,\ \forall \boldsymbol{p} \in \mathcal{P}_\epsilon,\ \text{then, } P^{\text{true}}\{\mathcal{R}_{\tilde{\boldsymbol{p}}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \le 0\} \ge 1-\epsilon.$$

[5] 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.

[6] There is another stream of research that approximates (3.4) by CVaR or its approximations, see, e.g., Chen et al. [74], Chen and Sim [71], Chen et al. [72] and references there in.

[7] One can in turn seek a safe approximation to (3.5). For example, one stream of such approximations includes using Chebyshev's inequality, see, e.g., Popescu [238], Bertsimas and Popescu [33], Bernstein's inequality, see, e.g., Nemirovski and Shapiro [213], or Hoeffding's inequality. We review such safe approximations to (3.5) in Section 5 in details.

THEOREM 3.9. *(Bertsimas et al. [42, Theorem 2]) Suppose that for any fixed* $\boldsymbol{x}$, $\mathcal{R}_{\boldsymbol{p}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]$ *is concave in* $\boldsymbol{p}$. *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* $\boldsymbol{y}^\top\tilde{\boldsymbol{p}}$ *at level* $1-\epsilon$ *over* $\mathcal{P}_\epsilon(\alpha)$, *i.e.,* $\sup_{P\in\mathcal{P}_\epsilon(\alpha)}\ \mathrm{VaR}^P_{1-\epsilon}\left[\boldsymbol{y}^\top\tilde{\boldsymbol{p}}\right]\le l_\epsilon(\boldsymbol{y};\alpha)$, $\boldsymbol{y}\in\mathbb{R}^d$. *Then, the closed, convex set* $\mathcal{P}_\epsilon(\alpha)$ *for which* $\delta^*\big(\boldsymbol{y}|\mathcal{P}_\epsilon(\alpha)\big)=l_\epsilon(\boldsymbol{y};\alpha)$ *guarantees the implication* (C2) *with probability at least* $(1-\alpha)$ *(with respect to the sample probability measure).*

As a byproduct of Theorem 3.9, $\delta^*\big(\boldsymbol{y}|\mathcal{P}_\epsilon(\alpha)\big)\le\boldsymbol{b}$ provides a safe approximation to $\sup_{P\in\mathcal{P}_\epsilon(\alpha)}\ P\{\boldsymbol{y}^\top\tilde{\boldsymbol{p}}\le\boldsymbol{b}\}\ge 1-\epsilon$. That is, there is a one-to-one correspondence between the ambiguity set $\mathcal{P}_\epsilon(\alpha)$ that satisfies the probabilistic guarantee (C2) and safe approximations to $\sup_{P\in\mathcal{P}_\epsilon(\alpha)}\ P\{\boldsymbol{y}^\top\tilde{\boldsymbol{p}}\le\boldsymbol{b}\}\ge 1-\epsilon$.

**3.4. 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.

**3.4.1. DRO and Regularization.** Some papers have shown that DRO problems via the *optimal transport discrepancy* and *$\phi$-divergences* are connected to regularization. When the optimal transport discrepancy is used, as shown in Shafieezadeh-Abadeh et al. [273], Blanchet et al. [50], Gao and Kleywegt [110], 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 [49], and we relegate the technical details and assumptions to Section 5. On the other hand, when $\phi$-divergences are used, DRO problem is connected to variance regularization, see, e.g., Duchi et al. [92], Namkoong and Duchi [209].

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 $P_2$:

$$\Pi(P_1,P_2)=\{\pi\in\mathfrak{M}(\Xi\times\Xi,\mathcal{F}\times\mathcal{F})\,|\,\pi(A\times\Xi)=P_1(A),\pi(\Xi\times A)=P_2(A)\forall A\in\mathcal{F}\}.$$

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[8]:

$$\mathfrak{d}_c^{\mathrm{W}}(P_1,P_2):=\inf_{\pi\in\Pi(P_1,P_2)}\int_{\Xi\times\Xi}c(s_1,s_2)\pi(ds_1\times ds_2).$$

THEOREM 3.10. *(Blanchet and Murthy [49, Remark 1]) Consider an ambiguity set of probability measures as*

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

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

*formed via the optimal transport discrepancy $\mathbb{W}_c(P, P_0)$, where $c$ is the transportation cost function, $\epsilon$ 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*

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

We can use Theorem 3.10 to explicitly state the connection between DRO and regularization. We adopt the following two theorems from Blanchet and Murthy [49], due to their generality. However, similar results are obtained in other papers, see, e.g., Shafieezadeh-Abadeh et al. [273], Gao and Kleywegt [110].

THEOREM 3.11. *(Blanchet et al. [50, Theorem 2–3]) Consider a given set of data $\{\boldsymbol{\xi}^i := (\boldsymbol{u}^i, y^i)\}_{i=1}^N$, 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 $\mathbb{P}_N$ 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 $h_1(\boldsymbol{x}, \boldsymbol{\xi}) := (y - \boldsymbol{x}^\top \boldsymbol{u})^2$, we have*

$$\inf_{\boldsymbol{x} \in \mathbb{R}^n} \sup_{\mathbb{P} \in \mathcal{P}^W(\mathbb{P}_N;\epsilon)} \mathbb{E}_{\mathbb{P}}\left[h_1(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] = \inf_{\boldsymbol{x} \in \mathbb{R}^n} \left\{ \epsilon^{\frac{1}{2}} \|\boldsymbol{x}\|_p + \left(\mathbb{E}_{\mathbb{P}_N}\left[h_1(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]\right)^{\frac{1}{2}} \right\}^2,$$

- *For a logistic regression model with cost function $h_2(\boldsymbol{x}, \boldsymbol{\xi}) := \log(1 + e^{-y\boldsymbol{x}^\top \boldsymbol{u}})$, we have*

$$\inf_{\boldsymbol{x} \in \mathbb{R}^n} \sup_{\mathbb{P} \in \mathcal{P}^W(\mathbb{P}_N;\epsilon)} \mathbb{E}_{\mathbb{P}}\left[h_2(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] = \inf_{\boldsymbol{x} \in \mathbb{R}^n} \left\{ \epsilon\|\boldsymbol{x}\|_p + \mathbb{E}_{\mathbb{P}_N}\left[h_2(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] \right\},$$

- *For a SVM with Hinge loss $h_3(\boldsymbol{x}, \boldsymbol{\xi}) := (1 - y\boldsymbol{x}^\top \boldsymbol{u})_+$, we have*

$$\inf_{\boldsymbol{x} \in \mathbb{R}^n} \sup_{\mathbb{P} \in \mathcal{P}^W(\mathbb{P}_N;\epsilon)} \mathbb{E}_{\mathbb{P}}\left[h_3(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] = \inf_{\boldsymbol{x} \in \mathbb{R}^n} \left\{ \epsilon\|\boldsymbol{x}\|_p + \mathbb{E}_{\mathbb{P}_N}\left[h_3(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] \right\}.$$

As stated in Theorem 3.11, 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 3.11 about the connection between DRO and regularization: (i) the shape of the transportation cost $c$ 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 judicially choose an appropriate regularization parameter for the problem in hand by using the DRO equivalent reformulation. We review the papers that draw this conclusion in Section 5.1.

Now, let us focus on DRO problems formulated via $\phi$-divergences. For two probability measures $P_1, P_2 \in \mathfrak{M}(\Xi, \mathcal{F})$, the $\phi$-divergence between $P_1$ and $P_2$ is defined as $\mathfrak{d}^\phi(P_1, P_2) := \int_\Xi \phi\left(\frac{dP_1}{dP_2}\right) dP_2$, where the $\phi$-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 $a > 0$[9].

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

THEOREM 3.12. *(Duchi et al. [92, Theorem 2]) Consider an ambiguity set of probability distributions as*

$$\mathcal{P}^{\phi}(\mathbb{P}_0;\epsilon) := \left\{\mathbb{P}\in\mathfrak{P}(\mathbb{R}^d,\mathfrak{B}(\mathbb{R}^d)) \,\middle|\, \mathfrak{d}^{\phi}(\mathbb{P},\mathbb{P}_0)\le\epsilon\right\},$$

*formed via the $\phi$-divergence $\mathfrak{d}^{\phi}(\mathbb{P},\mathbb{P}_0)$, where $\epsilon$ is the size of the ambiguity set and $\mathbb{P}_0$ 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 $\mathcal{X}$, $\mathbb{E}_{\mathbb{P}^{true}}\left[M(\tilde{\boldsymbol{\xi}})^2\right]<\infty$, and $\mathbb{E}_{\mathbb{P}^{true}}\left[|h(\boldsymbol{x}_0,\tilde{\boldsymbol{\xi}})|\right]<\infty$ for some $\boldsymbol{x}_0\in\mathcal{X}$. Then,*

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

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

As Theorem 3.12, we can rewrite the inner problem of a model of the form (1.5) with $\phi$-divergences as the expected cost plus a regularization term that accounts for the standard deviation of the cost, under the empirical distribution.

**4. General Solution Techniques to Solve DRO Models.** In this section, we discuss two approaches to solve (DRO). Let us first reformulate (DRO) as follows:

$$\inf_{\boldsymbol{x}\in\mathcal{X},\theta}\ \theta \tag{4.1a}$$

$$\text{s.t.}\quad \theta\ge\mathcal{R}_P\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right],\ \forall P\in\mathcal{P} \tag{4.1b}$$

$$\mathcal{R}_P\left[\boldsymbol{g}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]\le\mathbf{0},\ \forall P\in\mathcal{P}. \tag{4.1c}$$

Reformulation (4.1) is a semi-infinite program (SIP), and at a first glance, obtaining an optimal solution to this problem looks unreachable[10]. It is well-known that even convex SIPs cannot be solved directly with numerical 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 (4.1) 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 $h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$ and $\boldsymbol{g}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$. As we shall see in details in Section 5, 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 (4.1).

In the context of DRO, there are two main approaches to handle the semi-infinite quantifier $\forall P$ and to numerically solve (4.1). 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.

[10] The study of SIPs is pioneered by Haar [130], and followed up in Charnes et al. [64, 65, 66], which focus on linear SIPs. The first- and second-order optimality conditions of general SIP are also obtained in Hettich and Jongen [143, 144], Hettich and Still [146], Nuernberger [222], Nürnberger [223], Still [300]. For reviews of the theory and methods for SIPs, we refer the readers to Hettich and Kortanek [145], Reemtsen and Görner [255], López and Still [194].

**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 (4.1) 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., Mehrotra and Papp [202], Hettich and Kortanek [145]). We refer to Pflug and Wozabal [229], Rahimian et al. [251], Bansal et al. [9] 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 (4.1) with a finite number of probability distributions to optimally and (ii) generate an $\epsilon$-optimal solution[11] to a distribution separation subproblem [200].

THEOREM 4.1. *(Luo and Mehrotra [200, Theorem 3.2]) Suppose that $\mathcal{X} \times \mathcal{P}$ is compact, and $\mathcal{R}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$ and $\mathcal{R}_P\left[\boldsymbol{g}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$ 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 (4.1) for any finite set $\mathcal{P}_k \subseteq \mathcal{P}$, and an oracle that generates an $\epsilon$-optimal solution of the distribution generation subproblem*

$$\sup_{P \in \mathcal{P}} \max \left\{ \mathcal{R}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right], \mathcal{R}_P\left[g_1(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right], \ldots, \mathcal{R}_P\left[g_m(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] \right\}$$

*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 $\mathcal{R}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] \leq \theta_k + \frac{\epsilon}{2}$ and $\mathcal{R}_P\left[g_j(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] \leq \frac{\epsilon}{2}$, $j = 1, \ldots, m$, guarantee that an $\epsilon$-feasible solution[12] to problem (4.1), yielding an objective function value lower bounding the optimal value of (4.1), can be obtained in a finite number of iterations.*

It is worth noting that the distribution generation 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 convexfied without causing a change to the optimal value. 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 4.2. *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:*

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \sup_{P \in conv(\mathcal{P})} \left\{ \mathcal{R}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] \,\middle|\, \sup_{P \in conv(\mathcal{P})} \mathcal{R}_P\left[\boldsymbol{g}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] \leq \boldsymbol{0} \right\}. \tag{4.2}$$

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

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

*Proof.* Problems (DRO) and (4.2) can be reformulated, respectively, as $\min\{\theta \,|\, (x,\theta) \in \mathcal{G}\}$ and $\min\{\theta \,|\, (x,\theta) \in \mathcal{G}'\}$, where

$$\mathcal{G} := \left\{(x,\theta) \in \mathbb{R}^{n+1} \,\middle|\, \boldsymbol{x} \in \mathcal{X},\ \mathcal{R}_P\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \le \theta,\ \mathcal{R}_P\left[\boldsymbol{g}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \le \mathbf{0},\ \forall P \in \mathcal{P}\right\},$$

and

$$\mathcal{G}' := \left\{(x,\theta) \in \mathbb{R}^{n+1} \,\middle|\, \boldsymbol{x} \in \mathcal{X},\ \mathcal{R}_P\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \le \theta,\ \mathcal{R}_P\left[\boldsymbol{g}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \le \mathbf{0},\ \forall P \in \mathrm{conv}(\mathcal{P})\right\}.$$

Because $\mathcal{P} \subseteq \mathrm{conv}(P)$, we have $\mathcal{G}' \subseteq \mathcal{G}$, and thus, an optimal solution to (4.2) is optimal to (DRO). We now show that $\mathcal{G} \subseteq \mathcal{G}'$. Consider an arbitrary $(\boldsymbol{x},\theta) \in \mathcal{G}$. For an arbitrary $P \in \mathrm{conv}(\mathcal{P})$, there exists a collection $\{P^i\}_{i\in\mathcal{I}}$ such that $P = \sum_{i\in\mathcal{I}} \lambda^i P^i$, where $\sum_{i\in\mathcal{I}} \lambda^i = 1$, $P^i \in \mathcal{P}$, $\lambda^i \ge 0$, $i \in \mathcal{I}$. Now, by the convexity of $\mathcal{R}_P[\cdot]$ in $P$ on $\mathfrak{M}(\Xi,\mathcal{F})$, we have $\mathcal{R}_P\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \le \sum_{i\in\mathcal{I}} \lambda^i \mathcal{R}_{P^i}\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \le \theta$ and $\mathcal{R}_P\left[\boldsymbol{g}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \le \sum_{i\in\mathcal{I}} \lambda^i \mathcal{R}_{P^i}\left[\boldsymbol{g}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \le \mathbf{0}$. Thus, it follows that $(\boldsymbol{x},\theta) \in \mathcal{G}'$, and hence, $\mathcal{G} \subseteq \mathcal{G}'$. ∎

**4.2. Dual Method.** The second approach to solve (DRO) handles the quantifier $\forall P$ through the dualization of $\sup_{P\in\mathcal{P}} \mathcal{R}_P\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]$ and $\sup_{P\in\mathcal{P}} \mathcal{R}_P\left[\boldsymbol{g}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \le \mathbf{0}$. 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., Delage and Ye [82], Bertsimas et al. [39], Wiesemann et al. [317], Ben-Tal et al. [28]. A situation where one benefits from the application of the dual method to solve (DRO) arises in cases where the ambiguity set of probability distribution depends on decision $\boldsymbol{x}$ as formulated below, see, e.g., Luo and Mehrotra [199], Noyan et al. [221]:

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{P\in\mathcal{P}(\boldsymbol{x})} \left\{\mathcal{R}_P\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \,\middle|\, \sup_{P\in\mathcal{P}(\boldsymbol{x})} \mathcal{R}_P\left[\boldsymbol{g}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \le \mathbf{0}\right\}, \tag{4.3}$$

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. [16], Bertsekas [30], Ruszczyński [269], Rockafellar [257]. For detailed discussions of the duality theory in infinite-dimensional convex problems, we refer to Rockafellar [257], and we refer to Isii [162] and Shapiro [284] for duality theory in conic linear programs. Below, we briefly present the results from conic duality that are widely used in the dualization of DRO models.

THEOREM 4.3. *(Shapiro [284, 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*

$$\min_{v\in\mathcal{C}} \langle c, v\rangle \tag{4.4a}$$

$$s.t. \quad Av \succcurlyeq_{\mathcal{K}} b, \tag{4.4b}$$

*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 (4.4) is written as*

$$\max_{w^* \in \mathcal{K}'} \ \langle w^*, b\rangle \tag{4.5a}$$

$$s.t. \quad A^*w^* \preccurlyeq_{\mathcal{C}'} c. \tag{4.5b}$$

*Moreover, there is no duality gap between (4.4) and (4.5) 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 $\langle c - A^*w^*, v\rangle = 0$.*

It is worth noting that other numerical methods to solve a SIP, such as penalty methods, see, e.g., Lin et al. [190], Yang et al. [336], smooth approximation and projection methods, see, e.g., Xu et al. [332], and primal methods, see, e.g., Wang and Yuan [314], have not been popular in the DRO literature, although there are a few exceptions. Liu et al. [192] propose to discretize DRO by a min-max problem in a finite dimensional space, where the ambiguity set is replaced by a set of distributions on a discrete support set. Then, they consider lifting techniques to reformulate the discretized DRO as a saddle-point problem, if needed, and implement a primal-dual hybrid algorithm to solve the problem. They showcase this method for cases where the ambiguity set is formed via the moment constraints as in (5.23) or the Wasserstein metric, and they present the quantitative convergence of the optimal value and optimal solutions. Other iterative primal methods that have been proposed to solve a DRO model include Lam and Ghosh [177] for $\chi^2$-distance, and Ghosh et al. [115], Namkoong and Duchi [208], Ghosh and Lam [114] for general $\phi$-divergences.

**5. Choice of Ambiguity Set of Probability Distributions.** The ambiguity set of distribution 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 the 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., Lasserre [182], Bertsimas et al. [38], Bertsimas and Popescu [33], Popescu [238, 239], Gilboa and Schmeidler [117]. 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 choose the ambiguity set: (1) $\mathcal{P}$ should be chosen as small as possible, (2) $\mathcal{P}$ should contain the unknown true distribution with certainty (or at least, with a high confidence). Abiding by these two principles not only reduces the conservatism of the problem but it also robustifies the problem against the unknown true distribution. These two, in turn, give rise to two questions: (1) what should be the shape of the ambiguity set and (2) what should be the size of the ambiguity set. We discuss the latter in Section 6, and focus on the shape of the ambiguity set in this section.

Except for a few exceptions, the common practice in constructing the ambiguity set is that first, the shape of the set is determined by decision makers/modelers.

In this step, data does not directly affect the choice of the shape of the ambiguity set. Then, 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 size and shape of the ambiguity set might not necessarily be chosen separately. To make the transition between Section 5 and 6 somewhat smoother, we devote Section 5.4 to review those papers that address these two questions simultaneously.

When dealing with the question of the shape of the ambiguity set, 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: *moment-based* and *discrepancy-based* ambiguity sets. In short, moment-based ambiguity sets contain distributions whose moments satisfy certain properties, while discrepancy-based ambiguity sets contain distributions that are close to a nominal distribution in the sense of some *discrepancy* measure. Within these two groups, some specific ambiguity sets have been given names, see, e.g., Hanasusanto et al. [134]. 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 written as an infinite convex combination of extremal distributions of the set,
- *Mixture* ambiguity set contains all distributions that can be written 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 5.1),
- Moment-based ambiguity sets (Section 5.2),
- Shape-preserving ambiguity sets (Section 5.3),
- Kernel-based ambiguity sets (Section 5.4).

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 also contain distributions that are formed via a kernel and its parameters are close to the

parameters of a nominal kernel function. The above groups are not necessarily disjoint from a modeling perspective and there are some overlaps between them. However, we try to assign papers to these categories as close as possible to what the authors explicitly or implicitly might have stated in their work.

We review these four groups of ambiguity sets in Sections 5.1–5.4. Finally, we review the papers that are general and do not consider a specific form for the ambiguity set in Section 5.5.

**5.1. Discrepancy-Based Ambiguity Sets.** In many situations, 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:

$$\mathcal{P}^{\mathfrak{d}}(P_0;\epsilon) = \{P \in \mathfrak{M}(\Xi,\mathcal{F}) \,|\, \mathfrak{d}(P,P_0) \leq \epsilon\}, \tag{5.1}$$

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 $\epsilon$ is also referred to as the *level of robustness*.

A generic ambiguity set of the form (5.1) has been widely studied in the DRO literature. We relegate the discussion about $P_0$ and $\epsilon$ to Section 6. 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) *$\phi$-divergences*, (iii) *total variation metric*, (iv) *goodness-of-fit test*, (v) *Prohorov metric*, (vi) *$\ell_p$-norm*, (vii) *$\zeta$-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 [116], while *Kullback-Leibler* and $\chi^2$-distance from the family of $\phi$-divergences are not a probability metric. Thus, we refer to the models of the form (5.1) collectively as *discrepancy-based* ambiguity sets.

**5.1.1. 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 $P_2$:

$$\Pi(P_1,P_2) = \{\pi \in \mathfrak{M}(\Xi\times\Xi, \mathcal{F}\times\mathcal{F}) \,|\, \pi(A\times\Xi) = P_1(A), \pi(\Xi\times A) = P_2(A) \forall A \in \mathcal{F}\}.$$

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:

$$\mathfrak{d}_c^{\mathrm{W}}(P_1,P_2) := \inf_{\pi\in\Pi(P_1,P_2)} \int_{\Xi\times\Xi} c(s_1,s_2)\pi(ds_1\times ds_2). \tag{5.2}$$

If, in addition, function $c$ is symmetric (i.e., $c(s_1,s_2) = c(s_2,s_1)$) and $c^{\frac{1}{r}}(\cdot)$ satisfies a triangle inequality for some $1 \leq r < \infty$ (i.e., $c^{\frac{1}{r}}(s_1,s_2) \leq c^{\frac{1}{r}}(s_1,s_3) + c^{\frac{1}{r}}(s_3,s_2)$),

then, $\mathfrak{d}_{c^{\frac{1}{r}}}^{\mathrm{W}}(P_1, P_2)$ metricizes the weak convergence in $\mathfrak{M}(\Xi, \mathcal{F})$, see, e.g., Villani [312, Theorem 6.9]. If $\Xi$ is equipped with a metric $d$ and $c(\cdot) = d^r(\cdot)$, then $\mathfrak{d}_c^{\mathrm{W}}(P_1, P_2)$ is called *Wasserstein metric of order $r$ or $r$-Wasserstein metric*, for short[13].

The optimal transport discrepancy (5.2) can be used to form an ambiguity set of probability measures as follows:

$$\mathcal{P}^{\mathrm{W}}(P_0; \epsilon) := \left\{ P \in \mathfrak{M}(\Xi, \mathcal{F}) \,\middle|\, \mathfrak{d}_c^{\mathrm{W}}(P, P_0) \leq \epsilon \right\}. \tag{5.3}$$

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., Pflug and Wozabal [229], Mehrotra and Zhang [203], Mohajerin Esfahani and Kuhn [205], Gao and Kleywegt [110], Chen et al. [76], Blanchet et al. [53], Lee and Mehrotra [183], Luo and Mehrotra [200], Shafieezadeh-Abadeh et al. [273], Sinha et al. [298], Lee and Raginsky [185], Shafieezadeh-Abadeh et al. [275], Singh and Póczos [297]. Pioneered by the work of Pflug and Wozabal [229], most of the literature has focused on the Wasserstein metric. Before we review these papers, we present a duality result on $\sup_{P \in \mathcal{P}^{\mathrm{W}}(P_0;\epsilon)} \mathbb{E}_P \left[ h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right]$, proved in a general form in Blanchet and Murthy [49].

Because the infimum in the defintion of (5.2) is attained for a lower semicontinuous function $c$ [312, 249], we can rewrite $\sup_{P \in \mathcal{P}^{\mathrm{W}}(P_0;\epsilon)} \mathbb{E}_P \left[ h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right]$ as follows:

$$\sup_{\pi \in \Phi_{P_0,\epsilon}} \int_{\Xi} h(\boldsymbol{x}, s) \pi(\Xi \times ds), \tag{5.4}$$

where

$$\Phi_{P_0,\epsilon} := \left\{ \pi \in \mathfrak{M}(\Xi \times \Xi, \mathcal{F} \times \mathcal{F}) \,\middle|\, \pi \in \cup_{P \in \mathfrak{M}(\Xi,\mathcal{F})} \Pi(P_0, P),\ \int_{\Xi \times \Xi} c(s_1, s_2) \pi(ds_1 \times ds_2) \leq \epsilon \right\}.$$

Recall that $\mathcal{S}(\Xi, \mathcal{F})$ is the collection of all $\mathcal{F}$-measurable functions $Z : (\Xi, \mathcal{F}) \mapsto (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$. With the primal problem (5.4), we have a dual problem

$$\inf_{(\lambda,\phi) \in \Lambda_{c,h(\boldsymbol{x},\cdot)}} \left\{ \lambda \epsilon + \int_{\Xi} \phi(s) P_0(ds) \right\}, \tag{5.5}$$

where

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

THEOREM 5.1. *(Blanchet and Murthy [49, Theorem 1]) For a fixed $\boldsymbol{x} \in \mathcal{X}$, suppose that $h(\boldsymbol{x}, \cdot)$ is upper semicontinuous and $P_0$-integrable, i.e., $\int_{\Xi} |h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}(s))| P_0(ds) < \infty$. Then,*

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

[13] Wasserstein metric of order 1 is sometimes referred to as *Kantorovich* metric. Wasserstein metric of order $\infty$ is defined as $\inf_{\pi \in \Pi(P_1,P_2)} \pi\text{-ess sup}\ d(s_1, s_2)$, where $\pi\text{-ess sup}_{\Xi \times \Xi}[\cdot]$ is the essential supremum with respect to measure $\pi$: $\pi\text{-ess sup}_{\Xi \times \Xi}\ d(s_1, s_2) = \inf\{a \in \mathbb{R} : \pi(s \in \Xi : \exists s' \in \Xi \text{ s.t. } d(s, s') > a) = 0\}$.

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

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

*if and only if*

$$h(\boldsymbol{x}, s_2) - \lambda^* c(s_1, s_2) = \sup_{s_3 \in \Xi} \{h(\boldsymbol{x}, s_3) - \lambda^* c(s_1, s_3)\},\ \pi^*\text{-almost surely}, \tag{5.6a}$$

$$\lambda^*\Big(\int_{\Xi\times\Xi} c(s_1, s_2)\pi(ds_1 \times ds_2) - \epsilon\Big) = 0. \tag{5.6b}$$

Corollary 5.2. *Suppose that* $h(\boldsymbol{x}, \cdot)$ *is upper semicontinuous and* $P_0$*-integrable. Then,*

$$\sup_{P \in \mathcal{P}^W(P_0;\epsilon)} \mathbb{E}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] = \inf_{\lambda \geq 0} \left\{\lambda\epsilon + \mathbb{E}_{P_0}\left[\sup_{s\in\Xi} \{h(\boldsymbol{x}, s) - \lambda c(\tilde{s}, s)\right]\right\}. \tag{5.7}$$

The importance of Theorem 5.1 and Corollary 5.2 is that (1) the transportion cost $c(\cdot,\cdot)$ is only known to be lower semicontinuous, (2) function $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is assumed to be upper semicontinuous and integrable, and (3) $\Xi$ is a general Polish space. In fact, there are only mild conditions on $h(\boldsymbol{x}, \cdot)$ and function $c$, and $P_0$ can be any probability measure. Moreover, $\sup_{P\in\mathcal{P}^W(P_0;\epsilon)} \mathbb{E}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$ can be obtained by solving a univariate reformulation of the dual problem (5.5), where it involves an expectation with respect to $P_0$ and a linear term in the level of robustness $\epsilon$. We shall shortly comment on similar results in the literature but under stronger assumptions. As shown in Section 3.4, by using Theorem 5.1 or its weaker forms, researchers have shown many mainstream machine learning algorithms, such as regularized logistic regression and LASSO, have a DRO representation, see, e.g., Blanchet et al. [50], Blanchet and Kang [48, 47], Gao et al. [112], Shafieezadeh-Abadeh et al. [273, 274].

While a strong duality result for DRO formed via the optimal transport discrepancy is provided in Blanchet and Murthy [49] under mild assumptions by utilizing Fenchel duality, Mohajerin Esfahani and Kuhn [205] and Gao and Kleywegt [110] are also among notable papers in this area. Below, we first highlight the main differences of Mohajerin Esfahani and Kuhn [205] and Gao and Kleywegt [110] with Blanchet and Murthy [49]. Then, we comment on their main contributions.

In Mohajerin Esfahani and Kuhn [205], it is assumed that the transportation cost $c(\cdot,\cdot)$ is a norm on $\mathbb{R}^n$, function $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ has specific structures, and the nominal probability measure $P_0$ is the empirical distribution of data supported on $\mathbb{R}^n$. On the other hand, Gao and Kleywegt [110] consider a more general setting than the one in Mohajerin Esfahani and Kuhn [205], but slightly more restricted than that of Blanchet et al. [50]. More precisely, in contrast to Blanchet et al. [50], it is assumed in Gao and Kleywegt [110] that the transportation cost $c(\cdot,\cdot)$ forms a metric on the underlying Polish space.

Mohajerin Esfahani and Kuhn [205] study data-driven DRO problems formed via 1-Wasserstein metric utilizing an arbitrary norm on $\mathbb{R}^n$. The main contribution of Mohajerin Esfahani and Kuhn [205] is in proving a strong duality result for the studied problem and to reformulate it as a finite-dimesnional convex program for different cost functions, including a pointwise maximum of finitely many concave functions,

convex functions, and sums of maxima of concave 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., [319, Theorem 6]. In addition, Mohajerin Esfahani and Kuhn [205] propose a procedure to construct an extremal distribution (respectively, a sequence of distributions) that attains the worst-case expectation precisely (or, asymptotically). They further show that their solutions enjoy finite-sample and asymptotic consistency guarantees. The results were applied to the mean-risk portfolio optimization and to the uncertainty quantification problems.

Gao and Kleywegt [110] study DRO problems formed via $p$-Wasserstein metric utilizing an arbitrary metric on a Polish space $\Xi$. Recognizing the fact that the ambiguity set should be chosen judicially for the application in hand, they 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 $\phi$-divergence-based sets, see Section 5.1.2. They prove a strong duality result for the studied problem by utilizing Lagrangian duality and approximate the worst-case distributions (or obtain a worst-case distribution, if it exists) explicitly via the first-order optimality conditions of the dual reformulation. Using this, they show data-driven DRO problems can be approximated by robust optimization problems.

In addition to the papers by Blanchet and Murthy [49], Mohajerin Esfahani and Kuhn [205], Gao and Kleywegt [110], 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. In the following review, we mention the properties of the transportation cost $c(\cdot,\cdot)$ in the definition of the optimal transport discrepancy, function $\boldsymbol{g}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$ or $h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$, and the nominal distribution $\mathbb{P}_0$ and its underlying space as studied in these papers. Zhao and Guan [343] study a data-driven distributionally robust two-stage stochastic linear program over a Wasserstein ambiguity set, with 1-Wasserstein metric utilizing $\ell_1$-norm. By developing a strong duality result, they reformulate the problem as a semi-infinite linear two-stage robust optimization problem. In addition, under mild conditions, 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. They also show the convergence of the problem to the corresponding stochastic program under the true unknown probability distribution as the data points increase.

Hanasusanto and Kuhn [132] derive conic programming reformulation to distributionally robust two-stage stochastic linear programs formed via $p$-Wasserstein metric utilizing an arbitrary norm. In particular, by relying on the strong duality result from Mohajerin Esfahani and Kuhn [205] and Gao and Kleywegt [110], they show that when the ambiguity set is formed via the 2-Wasserstein metric 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 that the two-stage distributionally robust stochastic linear program with nonrandom cost function in the second stage, where the ambiguity set is formed via the 1-Wasserstein metric around a discrete distribution is equivalent to a linear program. They further extend their result to a case where optimized certainty equivalent (OCE)

[22, 23] is used as a risk measure. As applications, they demonstrate their results for the least absolute deviations regression and multitask learning problems.

For random variables supported on a compact set and a bounded continuous function $h(\boldsymbol{x}, \cdot)$, Luo and Mehrotra [200] study (1.5) formed via the 1-Wasserstein metric utilizing an arbitrary norm, around the empirical distribution of data. They present an equivalent SIP reformulation of the problem by reformulating the inner problem as a conic linear program. In order to solve the resulting SIP, they propose a finitely convergent exchange method when the cost function $h$ is a general nonlinear function in $\boldsymbol{x}$, and a central cutting-surface method with a linear rate of convergence when the cost function $h(\cdot, \boldsymbol{\xi})$ is convex in $\boldsymbol{x}$ and $\mathcal{X}$ is convex. They investigate a logistic regression model to exemplify their algorithmic ideas, and the benefits of using 1-Wasserstein metric.

Pflug and Pichler [231] study a DRO approach to single- and two-stage stochastic programs formed via the $p$-Wasserstein metric utilizing an arbitrary norm. They assume that all probability distributions in the ambiguity set are supported on discrete, fixed atoms, while only the probabilities of atoms are changing in the ambiguity set. Hence, the ambiguity set can be represented as a subset of a finite-dimensional space. To solve the resulting problem, they apply the exchange method, proposed in Pflug and Wozabal [229]. Mehrotra and Zhang [203] study a distributionally robust ordinary least squares problem, where the ambiguity set of probability distribution is formed via 1-Wasserstein metric utilizing $\ell_1$-norm. Similar to Pflug and Pichler [231], they restrict the ambiguity set of distributions to all discrete distributions and show that the resulting problem can be solved by using an equivalent SOCP reformulation.

Unlike Pflug and Pichler [231] and Mehrotra and Zhang [203] that only allow varying the probabilities on atoms identical to those of the nominal distribution, the ambiguity set is allowed to contain an infinite-dimensional distribution in Wozabal [319]. Wozabal [319] 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 $\ell_1$-norm around the empirical distribution. Because such a model has a higher complexity than that of those considered in Pflug and Pichler [231] and Mehrotra and Zhang [203], they propose to reformulate the problem into an equivalent finite-dimensional, nonconvex saddle-point optimization problem, under appropriate conditions. The key ideas in Wozabal [319] 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 approximated via a probability distribution supported on a uniform number of atoms, and (ii) considering only the extremal distributions in the ambiguity set suffices to obtain the equivalent reformulation. Furthermore, for a portfolio selection problem complemented via a broad class of convex risk measures appearing in the constraints, they obtain an equivalent finite-dimensional, nonconvex, semidefinite saddle-point optimization problem. They propose to solve such a reformulated problem via the exchange method, proposed in Pflug and Wozabal [229].

Pichler and Xu [236] 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 a two-stage stochastic program with recourse.

A class of data-driven distributionally robust fractional optimization problems,

representing a reward-risk ratio, is studied in Ji and Lejeune [164] as follows:

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{P\in\mathcal{P}} \frac{\mathcal{R}_P^1\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]}{\mathcal{R}_P^2\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]}, \tag{5.8}$$

where $\mathcal{R}_P^1 : \mathcal{Z} \mapsto \mathbb{R}$ is a reward measure and $\mathcal{R}_P^2 : \mathcal{Z} \mapsto \mathbb{R}_+$ is a nonnegative risk measure. Assuming that the underlying distribution is discrete, Ji and Lejeune [164] model the ambiguity about discrete distributions using the 1-Wasserstein metric utilizing $\ell_1$-norm, around the empirical distribution. They provide a nonconvex reformulation for the resulting model and propose a bisection algorithm to obtain the optimal value by solving a sequence of convex programming problems. As in Postek et al. [240], the reformulation is obtained through investigating the support function of the ambiguity set and the convex conjugate of the ratio function. They further apply their results to portfolio optimization problem for the Sharpe ratio [296] and Omega ratio [170].

Motivated by the drawback of moment-based DRO problems, Gao and Kleywegt [111] study DRO 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. They obtain tractable reformulations of these models and apply their results to a portfolio optimization problem. Along the same lines as Gao and Kleywegt [111], Pflug and Pohl [232] 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. [221] study DRO model with decision-dependent ambiguity set, where the ambiguity set is formed via the $p$-Wasserstein metric utilizing $\ell_p$-norm. They consider two types of ambiguity sets: (1) *continuous* ambiguity set, where there is ambiguity in both probability distribution of $\tilde{\boldsymbol{\xi}}$ and its realizations, and (2) *discerte* ambiguity set, where there is only ambiguity in the probability distribution of $\tilde{\boldsymbol{\xi}}$, while the realizations are fixed. They apply their results to problems in machine scheduling and humanitarian logistics. Rujeerapaiboon et al. [268] study continuous and discrete scenario reduction [97, 139, 140, 141, 6], where $p$-Wasserstein metric utilizing $\ell_p$-norm is used as a measure of discrepancy between distributions.

**5.1.1.1. Discrete Problems.** We now review DRO models over Wasserstein ambiguity sets, with discrete decisions. Bansal et al. [9] study a distributionally robust integer program with pure binary first-stage and mixed-binary second-stage variables on a finite set of scenarios as follows:

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

where

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

They propose a decomposition-based L-shaped algorithm and a cutting surface algorithm to solve the resulting model. They investigate the conditions and the ambiguity

sets under which the proposed algorithm is finitely convergent. They show that the ambiguity set of distributions formed via 1-Wasserstein metric utilizing an arbitrary norm satisfy these conditions. Xu and Burer [327] study a mixed 0-1 linear program, where the coefficients of the objective functions are affinely dependent on the random vector $\tilde{\boldsymbol{\xi}}$. 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 $\ell_2$-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.

**5.1.1.2. Chance Constraints.** In this section, we review distributionally robust chance-constrained programs over Wasserstein ambiguity sets, see, e.g., Jiang and Guan [166], Chen et al. [76], Xie [322], Yang [333]. Ji and Lejeune [165] study a distributionally robust individual chance constraint, where the ambiguity set of distributions is formed via 1-Wasserstein metric utilizing $\ell_1$-norm, and $g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ in (1.6) is defined as

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

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., $g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) := \mathbb{1}_{[\boldsymbol{a}^\top \boldsymbol{x} \leq \tilde{\boldsymbol{\xi}}]}(\tilde{\boldsymbol{\xi}})$, and the linear random technology matrix case, i.e., $g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) := \mathbb{1}_{[\tilde{\boldsymbol{\xi}}^\top \boldsymbol{x} \leq \boldsymbol{b}]}(\tilde{\boldsymbol{\xi}})$, and provide techniques to strengthen the formulations. 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. [76] 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 $\ell_1$-norm or $\ell_\infty$-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.

**5.1.1.3. Statistical Learning.** DRO problems formed via the optimal transport discrepency has been widely studied in the context of statistical learning. We already mentioned Mehrotra and Zhang [203] as an example in this area. Below, we review the latest developments of DRO in the context of statistical learning. 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. [215]. 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 $\ell_1$-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. They apply their results to linear discriminant analysis, portfolio selection, and solar irradiation patterns inference problems.

Lee and Mehrotra [183] study a distributionally robust framework for finding support vector machines via the 1-Wasserstein metric. They provide SIP formulation of the resulting model and propose a cutting-plane algorithm to solve the problem. Lee and Raginsky [184, 185] study a distributionally robust statistical learning problem formed via the $p$-Wasserstein metric utilizing $\ell_p$-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 a generalization bound and generalization error guarantees for the problem.

Gao et al. [113] 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. [233], Pichler [235], Wozabal [320] draw the connection between robustification and regularization, where as in Theorem 3.11, 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 [235] 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 when the function $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is linear in $\tilde{\boldsymbol{\xi}}$, the worst-case value is the sum of the risk of $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ under the nominal distribution and a regularization term. Pflug et al. [233] and Wozabal [320] 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 $\ell_p$-norm around the empirical distribution, reduces to the sum of the nominal risk and a regularization term whenever the function $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is affine in $\tilde{\boldsymbol{\xi}}$.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. They apply their results to a portfolio selection problem. Important parts of the derivation of results in Pflug et al. [233], Pichler [235], Wozabal [320] are Kusuoka's representation of risk measures [173, 287] and Fenchel-Moreau theorem [262, 270].

In the context of statistical learning, the connection between DRO and regularization was first made in Shafieezadeh-Abadeh et al. [273], 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 the resulting problem admits an equivalent reformulation as a tractable convex program. As stated in Theorem 3.11, 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 linear programs.

Shafieezadeh-Abadeh et al. [274] extend the work of Shafieezadeh-Abadeh et al. [273] 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 [205] 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 [12, 276], 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).

Gao et al. [112] 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. [206] 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 [48] 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 [47]. This method enhances the general-

ization 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 [70] study a DRO approach to linear regression using an $\ell_1$-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. They study the application of the proposed model to outlier detection, arising in an abnormally high radiation exposure in CT exams, and show it achieves a higher performance than M-estimation [161].

**5.1.1.4. Choice of the Transportation Cost.** When forming a Wasserstein ambiguity set, the transportation cost function $c(\cdot, \cdot)$ should be chosen besides the nominal probability measure $P_0$ and the size of the ambiguity set $\epsilon$. Blanchet et al. [52] 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. They select the level of robustness using cross-validation. Blanchet et al. [51] 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. [52], and it allows to enhance the generalization properties of regularized estimators while reducing the variability in the out-of-sample performance error.

**5.1.1.5. Multistage Setting.** The single- and two-stage stochastic programs in Pflug and Pichler [231] are extended in Analui and Pflug [3] and Pflug and Pichler [231] to the multistage case, 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. Both papers further apply their results to a multiperiod production/inventory control problem. Built upon the above results, Glanzer et al. [118] 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. Glanzer et al. [118] also study a DRO framework formed via nested distance that allows for setting up bid and ask prices for acceptability pricing of contingent claims. Another study of multistage linear optimization can also be found in Bazier-Mattea and Delage [17].

TABLE 1
*Examples of $\phi$-divergence functions, their conjugates $\phi^*(a)$, and their DRO counterparts*

| Divergence | $\phi(t)$ | $\phi(t),\ t \geq 0$ | $\mathfrak{d}^\phi(P_1, P_2)$ | $\phi^*(a)$ | DRO Counterpart |
|---|---|---|---|---|---|
| Kullback-Leibler | $\phi_{\rm kl}(t)$ | $t\log t - t + 1$ | $\int_\Xi \log\left(\frac{dP_1}{dP_2}\right) dP_1$ | $e^a - 1$ | Convex program |
| Burg entropy | $\phi_{\rm b}(t)$ | $-\log t + t - 1$ | $\int_\Xi \log\left(\frac{dP_2}{dP_1}\right) dP_2$ | $-\log(1-a),\ a<1$ | Convex program |
| J-divergence | $\phi_{\rm j}(t)$ | $(t-1)\log t$ | $\int_\Xi \log\left(\frac{dP_1}{dP_2}\right)(dP_1 - dP_2)$ | No closed form | Convex program |
| $\chi^2$-distance | $\phi_{\rm c}(t)$ | $\frac{1}{t}(t-1)^2$ | $\int_\Xi \frac{(dP_1 - dP_2)^2}{dP_1}$ | $2 - 2\sqrt{1-a},\ a<1$ | SOCP |
| Modified $\chi^2$-distance | $\phi_{\rm mc}(t)$ | $(t-1)^2$ | $\int_\Xi \frac{(dP_1 - dP_2)^2}{dP_2}$ | $\begin{cases} -1 & a < -2 \\ a + \frac{a^2}{4} & a \geq -2 \end{cases}$ | SOCP |
| Hellinger distance | $\phi_{\rm h}(t)$ | $(\sqrt{t}-1)^2$ | $\int_\Xi (\sqrt{dP_1} - \sqrt{dP_2})^2$ | $\frac{a}{1-a},\ a<1$ | SOCP |
| $\chi$-divergence of order $\theta>1$ | $\phi_{{\rm ca}^\theta}(t)$ | $\lvert t-1\rvert^\theta$ | $\int_\Xi \lvert 1 - \frac{dP_1}{dP_2}\rvert^\theta dP_2$ | $a + (\theta-1)\left(\frac{\lvert a\rvert}{\theta}\right)^{\frac{\theta}{\theta-1}}$ | SOCP |
| Variation distance | $\phi_{\rm v}(t)$ | $\lvert t-1\rvert$ | $\int_\Xi \lvert dP_1 - dP_2\rvert$ | $\begin{cases} -1 & a \leq -1 \\ a & -1 \leq a \leq 1 \end{cases}$ | LP |
| Cressie-Read | $\phi_{{\rm cr}^\theta}(t)$ | $\frac{1-\theta+\theta t - t^\theta}{\theta(1-\theta)}$ | $\frac{1}{\theta(1-\theta)}\left(1 - \int_\Xi dP_1^\theta dP_2^{1-\theta}\right)$ | $\frac{1}{\theta}\left(1 - a(1-\theta)\right)^{\frac{\theta}{1-\theta}} - \frac{1}{\theta},\ a < \frac{1}{1-\theta}$ | SOCP |

**5.1.2. $\phi$-Divergences.** Another popular way to model the distributional ambiguity is to use *$\phi$-divergences*, a class of measures used in information theory. A $\phi$-divergence measures the discrepancy between two probability measures $P_1, P_2 \in \mathfrak{M}(\Xi, \mathcal{F})$ as $\mathfrak{d}^\phi(P_1, P_2) := \int_\Xi \phi\left(\frac{dP_1}{dP_2}\right) dP_2$[14], where the $\phi$-divergence function $\phi : \mathbb{R}_+ \to \mathbb{R}_+ \cup \{+\infty\}$ is convex, and satisfy the following properties: $\phi(1) = 0$[15], $0\phi\left(\frac{0}{0}\right) := 0$, and $a\phi\left(\frac{a}{0}\right) := a\lim_{t\to\infty} \frac{\phi(t)}{t}$ if $a > 0$. Note that a $\phi$-divergence does not necessarily induce a metric on the underlying space. For detailed information on $\phi$-divergences, we refer to Read and Cressie [254], Vajda [306], Pardo [226].

A $\phi$-divergence can be used to model the distributional ambiguity as follows:

$$\mathcal{P}^\phi(P_0; \epsilon) := \left\{P \in \mathfrak{M}(\Xi, \mathcal{F}) \,\middle|\, \mathfrak{d}^\phi(P, P_0) \leq \epsilon\right\}, \tag{5.9}$$

where as before $P_0$ is a nominal probability measure and $\epsilon$ controls the size of the ambiguity set. Table 1 presents a list of commonly used $\phi$-divergence functions in DRO and their conjugate functions $\phi^*$.

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

THEOREM 5.3. *Suppose that $\epsilon > 0$ in (5.9). Then, for a fixed $\boldsymbol{x} \in \mathcal{X}$, we have*

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

*where $\Lambda_{\phi,h(\boldsymbol{x},\cdot)} := \left\{(\lambda,\mu) \,\middle|\, \lambda \geq 0,\ h(\boldsymbol{x}, s) - \mu - \lambda\lim_{t\to\infty}\frac{\phi(t)}{t} \leq 0, \forall s \in \Xi\right\}$, 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 $a > 0$.*

The above result can be obtained by taking the Lagrangian dual of $\sup_{P\in\mathcal{P}^\phi(P_0;\epsilon)} \mathbb{E}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$, and we refer the readers to Ben-Tal et al. [28], Bayraksan and Love [13], Love and Bayraksan [196] for a detailed derivation.

The robust counterpart of linear and nonlinear optimization problems with an uncertainty set of parameters defined via general $\phi$-divergence is studied in Ben-Tal et al. [28]. As it is presented in Table 1, when the uncertain parameter is a

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

[15] 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 $\phi$ [226].

finite-dimensional probability vector, the robust counterpart is tractable for most of the choices of $\phi$-divergence function considered in the literature. The use of $\phi$-divergence to model the distributional ambiguity in DRO is systematically introduced in Bayraksan and Love [13] and Love and Bayraksan [197]. To elucidate the use of $\phi$-divergences for models with different sources of data and decision makers with different risk preferences, they present a classification of $\phi$-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 $\phi$-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 analyze the value of data and propose a decomposition algorithm to solve the dual of the resulting DRO model formed via a general $\phi$-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. [27] propose to minimize the expected cost under the nominal distribution while the maximum expected cost over an infinite nested family of ambiguity sets, parametrized by $\epsilon$, 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 $\epsilon$ is proportional to $\epsilon$. They refer to this approach as *soft robust optimization* and relate the feasibility region induced by this approach to the convex risk measures. They illustrate that the ambiguity sets formed via $\phi$-divergences are related to an optimized certainty equivalent risk measure formed via $\phi$-functions [23]. 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 $\epsilon$, 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. Furthermore, they study the downside probability guarantees implied by both the soft robust and standard robust approaches. They also apply their results to portfolio optimization and asset allocation problems.

A data-driven DRO approach to chance-constrained problems modeled via $\phi$-divergences is studied in Yanıkoğlu and den Hertog [337]. They propose safe approximations to these ambiguous chance constraints. Their approach is capable of handling joint chance constraints, dependent uncertain parameter, and a general nonlinear function $\boldsymbol{g}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$.

Hu et al. [157] and Jiang and Guan [166] show that distributionally robust chance-constrained programs formed via $\phi$-divergences can be transformed into a chance-constrained problem under the nominal distribution but with an adjusted risk level. For a general $\phi$-divergence, a bisection line search algorithm to obtain the perturbed risk level is proposed in Hu et al. [157], Jiang and Guan [166]. In addition, closed-form expressions for the adjusted risk level are obtained for the case of the variation distance (see, Hu et al. [157] and Jiang and Guan [166]), and Kullback-Leibler divergence and $\chi^2$-distance (see, Jiang and Guan [166]). For the ambiguous probabilistic programs formed via $\phi$-divergences, similar results to the chance-constrained programs are shown in Hu et al. [157]. Hu et al. [157] 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.

**5.1.2.1. Statistical Learning.** Hu et al. [155] study distributionally robust supervised learning, where the ambiguity set of distributions is formed via $\phi$-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 $\phi$-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. [92] 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 $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is lower-semicontinuous.

**5.1.2.2. Specific $\phi$-Divergences.** In this section, we review papers that consider specific $\phi$-divergences.

**Kullback-Leibler Divergence.** Calafiore [59] 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 [156] 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 $\tilde{\boldsymbol{\xi}}$[16] , and the nominal distribution belongs to the exponential families of distributions, both the nominal and worst-case distribution belong to the same distribution family.

Blanchet et al. [53] 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 Réyni divergences [226], that includes Kullback-Leibler as a special case[17]. Kullback-Leibler is also used for the DRO

[16] Recall the discussion following (1.1) and (1.2), where we gave a characterization of $A(\boldsymbol{x})$ as $\boldsymbol{a}(\boldsymbol{x})^\top \tilde{\boldsymbol{\xi}} \leq \boldsymbol{b}(\boldsymbol{x})$ and $\boldsymbol{a}(\tilde{\boldsymbol{\xi}})^\top \boldsymbol{x} \leq \boldsymbol{b}(\tilde{\boldsymbol{\xi}})$. A safe region characterized by a bi-affine expression in $\tilde{\boldsymbol{\xi}}$ and $\boldsymbol{x}$ means that both $\boldsymbol{a}(\boldsymbol{x})$ and $\boldsymbol{b}(\boldsymbol{x})$ are affine in $\boldsymbol{x}$ for the form $\boldsymbol{a}(\boldsymbol{x})^\top \tilde{\boldsymbol{\xi}} \leq \boldsymbol{b}(\boldsymbol{x})$, and both $\boldsymbol{a}(\tilde{\boldsymbol{\xi}})$ and $\boldsymbol{b}(\tilde{\boldsymbol{\xi}})$ are affine in $\tilde{\boldsymbol{\xi}}$ for the form $\boldsymbol{a}(\tilde{\boldsymbol{\xi}})^\top \boldsymbol{x} \leq \boldsymbol{b}(\tilde{\boldsymbol{\xi}})$.

[17] The class of Réyni divergences is defined as $\mathfrak{d}_r^{\mathrm{R}}(P_1, P_2) := \frac{1}{1-r} \int_\Xi \left(\frac{dP_1}{dP_2}\right)^{r-1} dP_1$. This class is not a $\phi$-divergence, but $\mathfrak{d}_r^{\mathrm{R}}(P_1, P_2)$ can be rewritten as $h(\mathcal{D}_\phi(P_1, P_2))$, where $h(t) = \frac{1}{r-1} \log[(r-1)t + 1]$ and $\phi(t) = \frac{t^r - r(t-1) - 1}{r-1}$ [226].

approach to hypothesis testing in Levy [186], Gül and Zoubir [128], Gül [127].

**Burg Entropy.** Wang et al. [316] 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. To test the performance of the model, they apply it to the newsvendor problem and the portfolio selection problem.

Wiesemann et al. [317] 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 [176] 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.

**$\chi^2$-Distance.** Hanasusanto and Kuhn [137] 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 $\chi^2$-distance, reduces to a tractable conic program. They apply their results to problems arising in index tracking and wind energy commitment applications. Klabjan et al. [172] 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 $\chi^2$-distance. They show that the resulting model generalizes the Bayesian model, and it can be interpreted as minimizing demand-history-dependent risk measures.

**Modified $\chi^2$-Distance.** A *stochastic dual dynamic programming* (SDDP) approach to solve a distributionally robust multistage optimization model formed via the modified $\chi^2$-distance is porposed in Philpott et al. [234].

**Variation Distance.** Variation distance, or $\ell_1$-norm, as defined in Table 1, can be used to safely approximate several ambiguity sets formed via $\phi$-divergences, including $\chi$-divergence of order 2, $J$-divergence, Kullback-Leibler divergence, and Hellinger distance. The following lemma states the above result more formally.

LEMMA 5.4. *The following relationship holds between $\phi$-divergences, as defined in Table 1:*

$$\frac{1}{4}\big(\mathfrak{d}^{\phi_v}(P,P_0)\big)^2 \le \mathfrak{d}^{\phi_h}(P,P_0) \le \mathfrak{d}^{\phi_{kl}}(P,P_0) \le \mathfrak{d}^{\phi_j}(P,P_0) \le \mathfrak{d}^{\phi_{ca^2}}(P,P_0), \tag{5.10}$$

*which implies*

$$\mathcal{P}^{\phi_{ca^2}}(P_0;\epsilon) \subseteq \mathcal{P}^{\phi_j}(P_0;\epsilon) \subseteq \mathcal{P}^{\phi_{kl}}(P_0;\epsilon) \subseteq \mathcal{P}^{\phi_h}(P_0;\epsilon) \subseteq \mathcal{P}^{\phi_v}(P_0;2\epsilon^{\frac{1}{2}}). \tag{5.11}$$

*Proof.* The first two inequalities in (5.10) can be found in e.g., Reiss [256, p. 99][18] and the last two inequalities can be found in e.g., Jiang et al. [168, Lemma 1]. Then, (5.11) follows from (5.10). ∎

**5.1.3. 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 derivaties $f_1$ and $f_2$, respectively, then, $\mathfrak{d}^{\mathrm{TV}}(P_1, P_2) = \frac{1}{2}\int_\Xi |f_1(s) - f_2(s)|\nu(ds)$. Note that the total variation distance can be obtained from other classes of probability metrics: (1) it is a $\phi$-divergence with $\phi(t) = \frac{1}{2}|t-1|$, (2) it is half of the $\ell_1$-norm, and (3) it is obtained from the optimal transport discrepancy (5.2) with

$$c(s_1, s_2) = \begin{cases} 0, & \text{if } s_1 = s_2, \\ 1, & \text{if } s_1 \neq s_2. \end{cases} \tag{5.12}$$

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

$$\mathcal{P}^{\mathrm{TV}}(P_0; \epsilon) := \left\{ P \in \mathfrak{M}(\Xi, \mathcal{F}) \,\middle|\, \mathfrak{d}^{\mathrm{TV}}(P, P_0) \leq \epsilon \right\}, \tag{5.13}$$

where as before $P_0$ is a nominal probability measure and $\epsilon$ controls the size of the ambiguity set.

The total variation distance between $P_1$ and $P_2$ is also related to the *one-sided* variation distances $\frac{1}{2}\int_\Xi (f_1(s) - f_2(s))_+\nu(ds)$ and $\frac{1}{2}\int_\Xi (f_2(s) - f_1(s))_+\nu(ds)$ [251], which are $\phi$-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 $\sup_{P\in\mathcal{P}^{\mathrm{TV}}(P_0;\epsilon)} \mathbb{E}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$.

THEOREM 5.5. *(Jiang and Guan [167, Theorems 1–2], Rahimian et al. [251, Proposition 3], Shapiro [289]) For a fixed $\boldsymbol{x} \in \mathcal{X}$, we have*

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

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

*Remark* 5.6. (Rahimian et al. [251, Proposition 3], Shapiro [289]) 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 5.5.

[18] As shown for e.g., in Reiss [256] and [116], $\mathfrak{d}^{\phi_{\mathrm{h}}}(P, P_0) \leq \mathfrak{d}^{\phi_{\mathrm{kl}}}(P, P_0)$. However, in Jiang et al. [168, Lemma 1] this relationship has been shown incorrectly as $\mathfrak{d}^{\phi_{\mathrm{h}}}(P, P_0) \leq \left(\mathfrak{d}^{\phi_{\mathrm{kl}}}(P, P_0)\right)^{\frac{1}{2}}$.

Jiang and Guan [167] study distributionally robust two-stage stochastic programs 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. [251] 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 the effective and ineffective scenarios for the case that the distributional ambiguity is modeled via the total variation distance. Rahimian et al. [252] extends the work of Rahimian et al. [251] 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.

**5.1.4. Goodness-of-Fit Test.** Postek et al. [240] 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 $\phi$-divergences, Kolmogrov-Smirnov, Wasserstein, Anderson-Darling, Cramer-von Mises, Watson, and Kuiper. They exemplify the results in portfolio optimization and antenna array design problems. Bertsimas et al. [42] and Bertsimas et al. [43] 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 3.9). They consider the situation that (i) $\mathbb{P}^{\text{true}} = P^{\text{true}} \circ \tilde{\boldsymbol{\xi}}^{-1}$ may have continuous support, and the components of $\tilde{\boldsymbol{\xi}}$ are independent, (ii) $\mathbb{P}^{\text{true}}$ may have continuous support, and data are drawn from its marginal distributions asynchronously, and (iii) $\mathbb{P}^{\text{true}}$ may have continuous support, and data are drawn from its joint distribution. They also study a wide range of statistical hypothesis tests, including $\chi^2$, G, Kolmogrov-Smirnov, Kuiper, Cramer-von Mises, Watson, and Anderson-Darling goodness-of-fit tests, and they characterize the geometric shape of the corresponding ambiguity sets.

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

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

where $A^\gamma := \{s \in \Xi \,|\, \inf_{s' \in A} \; d(s, s') \leq \gamma\}$ [116]. The Prohorov metric takes values in $[0, 1]$ and can be used to model the distributional ambiguity as follows:

$$\mathcal{P}^{\text{P}}(P_0; \epsilon) := \{P \in \mathfrak{M}(\Xi, \mathcal{F}) \,|\, \mathfrak{d}^{\text{P}}(P, P_0) \leq \epsilon\}, \tag{5.14}$$

where as before $P_0$ is a nominal probability measure and $\epsilon$ controls the size of the ambiguity set. A specialization of the Prohorov metric to the univariate distributions is called *Levy* metric, which is defined as [116]

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

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

$$\mathcal{P}^{\text{L}}(P_0; \epsilon) := \left\{P \in \mathfrak{M}(\Xi, \mathcal{F}) \,\middle|\, \mathfrak{d}^{\text{L}}(P, P_0) \leq \epsilon\right\}. \tag{5.15}$$

Erdoğan and Iyengar [102] study an optimization problem subject to a set of parameterized convex constraints. Similar to the argument in Section 3.3.2, 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 [63, 60], 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.

**5.1.6. $\ell_p$-Norm.** Calafiore and El Ghaoui [61] 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 $\ell_1$- and $\ell_\infty$-norm[19]. Mevissen et al. [204] study distributionally robust polynomial optimization, where the distribution of the uncertain parameter is estimated using polynomial basis functions via the $\ell_p$-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. They apply their techniques to a water network optimization problem.

Jiang and Guan [167] study distributionally robust two-stage stochastic programs formed via $\ell_\infty$-norm. Huang et al. [158] study extend the work of Jiang and Guan [167] 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 minmax 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. They test their results on the hydrothermal scheduling problem.

**5.1.7. $\zeta$-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 $\zeta$-structure metric is defined as $\mathfrak{d}^{\mathcal{Z}}(P_1, P_2) := \sup_{z\in\mathcal{Z}} \left|\mathbb{E}_{P_1}\left[z(\tilde{\boldsymbol{\xi}})\right] - \mathbb{E}_{P_2}\left[z(\tilde{\boldsymbol{\xi}})\right]\right|$. A wide range of metrics in probability theory can be written as special cases of the above family of metrics [342, 236]. Let us introduce them below.

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

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

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

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

  where $L_1(z) =: \sup\{|z(\boldsymbol{u}) - z(\boldsymbol{v})|/d(\boldsymbol{u}, \boldsymbol{v}) \,|\, \boldsymbol{u} \neq \boldsymbol{v}\}$, is the Lipschitz modulus.

[19] Consider the sets $\mathcal{H}(\boldsymbol{A}, \boldsymbol{\xi}_0) := \{\boldsymbol{\xi} = \boldsymbol{\xi}_0 + \boldsymbol{A}\omega \,|\, \|\omega\|_\infty \leq 1\}$ and $\mathcal{E}(\boldsymbol{B}, \boldsymbol{\xi}_0) := \{\boldsymbol{\xi} = \boldsymbol{\xi}_0 + \boldsymbol{B}\omega \,|\, \|\omega\|_1 \leq 1\}$, where $\boldsymbol{A}$ is a diagonal positive-definite matrix and $\boldsymbol{B}$ is a positive-definite matrix. A random vector $\tilde{\boldsymbol{\xi}}$ has a probability distribution $P$ 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 $\tilde{\boldsymbol{\xi}} - \mathbb{E}_P\left[\tilde{\boldsymbol{\xi}}\right] = \boldsymbol{A}\omega$ (respectively, $\tilde{\boldsymbol{\xi}} - \mathbb{E}_P\left[\tilde{\boldsymbol{\xi}}\right] = \boldsymbol{B}\omega$), where $\omega$ is a random vector having the probability density $f_\omega$ such that $f_\omega(\omega) = t(\|\omega\|_\infty)$ for $\|\omega\|_\infty \leq 1$ and 0 otherwise (respectively, $f_\omega(\omega) = t(\|\omega\|_1)$ for $\|\omega\|_1 \leq 1$ and 0 otherwise) and $t(\cdot)$ is a nonincreasing function. The class of radially-symmetric distributions contains for example Gaussian, truncated Gaussian, uniform distribution on ellipsoidal support, and nonunimodal densities [61]

- *Kantorovich metric* $\mathfrak{d}^{\mathrm{K}}(P_1, P_2)$:

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

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

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

where

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

with $\|\cdot\|$ as the Euclidean norm. Note that when $q = 1$, Fortet-Mourier metric is the same as the Kantorovich metric.

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

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

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

$$\mathcal{P}^{\mathcal{Z}}(P_0; \epsilon) := \big\{P \in \mathfrak{M}(\Xi, \mathcal{F}) \,\big|\, \mathfrak{d}^{\mathcal{Z}}(P, P_0) \leq \epsilon\big\}, \tag{5.16}$$

where as before $P_0$ is a nominal probability measure and $\epsilon$ controls the size of the ambiguity set.

LEMMA 5.7. *Suppose that the support $\Omega$ of $\tilde{\boldsymbol{\xi}}$ is bounded with diameter $\theta$, 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 $\zeta$-structure metrics:*

$$\begin{aligned}
&\mathfrak{d}^{BL}(P, P_0) \leq \mathfrak{d}^{K}(P, P_0)\\
&\mathfrak{d}^{K}(P, P_0) \leq \mathfrak{d}^{TV}(P, P_0)\\
&\mathfrak{d}^{U}(P, P_0) \leq \mathfrak{d}^{TV}(P, P_0)\\
&\mathfrak{d}^{K}(P, P_0) \leq \mathfrak{d}^{FM}(P, P_0)\\
&\mathfrak{d}^{FM}(P, P_0) \leq \max\{1, \theta^{q-1}\}\mathfrak{d}^{K}(P, P_0).
\end{aligned}$$

*Proof.* The proof is immediate from Zhao and Guan [342, Lemmas 1–4]. ∎

Zhao and Guan [342] study distributionally robust two-stage stochastic programs via $\zeta$-structure metrics. They discuss how to construct the ambiguity set from historical data while utilizing a family of $\zeta$-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. They test their results on newsvendor and facility location problems.

Pichler and Xu [236] study a DRO model with a expectation as the risk measure and form the ambiguity set of distribution via $\zeta$-structure metric. They investigate how the variation of the ambiguity set would affect the optimal value and the optimal solution in the resulting optimization problem. They illustrate their results in the context of a two-stage stochastic program with recourse.

**5.1.8. Contamination Neighborhood.** The contamination neighborhood around a nominal probability measure $P_0$ is defined as

$$\mathcal{P}^{\mathrm{c}}(P_0;\epsilon) = \{P \in \mathfrak{M}(\Xi,\mathcal{F}) \,|\, P = (1-\epsilon)P_0 + \epsilon Q,\ Q \in \mathfrak{Q}\}, \tag{5.18}$$

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 [160], Huber and Ronchetti [161], and it has also been used in the economics literature, see, e.g., Nishimura and Ozaki [219, 220]. Bose and Daripa [56] study ambiguity aversion in a mechanism design problem using a maximin expected utility model of Gilboa and Schmeidler [117]. The contamination neighborhood is also used in the context of statistical learning, see, e.g., Duchi et al. [93] and hypothesis testing, see, e.g., Huber [159].

**5.1.9. General Discrepancy-Based Ambiguity Sets.** We devote this subsection to the papers that consider general discrepancy-based models. Postek et al. [240] review and derive tractable reformulations of distributionally robust risk constraints over discrete probability distributions and for function $\boldsymbol{g}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$ in $\tilde{\boldsymbol{\xi}}$. They provide a comprehensive list for risk measures and ambiguity sets, formed using statistical goodness-of-fit tests or probability metrics. They consider risk measures such as (1) expectation, (2) sum of expectation and standard deviation/variance, (3) variance, (4) mean absolute deviation from the median, (5) Sharpe ratio, (6) lower partial moments, (7) certainty equivalent, (8) optimized certainty equivalent, (9) shortfall risk, (10) VaR, (11) CVaR, (12) entropic VaR, (13) mean absolute deviation from the mean, (14) distortion risk measures, (15) coherent risk measures, and (16) spectral risk measures. They also consider (1) $\phi$-divergences, (2) Kolmogrov-Smirnov, (3) Wasserstein, (4) Anderson-Darling, (5) Cramer-von Mises, (6) Watson, and (7) Kuiper to model the distributional ambiguity. 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 (i.e., the convex conjugate of the indicator function of the ambiguity set). They exemplify the results in portfolio optimization and antenna array design problems.

A connection between DRO models formed via discrepancy-based ambiguity sets and law invariant risk measures is made in Shapiro [289] as described in Theorem 3.8. They specifically derive law invariant risk measures for cases when Wasserstein metric, $\phi$-divergences, and total variation distance is used to model the distributional ambiguity. They also propose a SAA approach to solve the corresponding dual of these problems, and establish the statistical properties of the optimal solutions and optimal value, similar to the results for the risk-neutral stochastic programs, see, e.g., Shapiro et al. [295], Shapiro [285].

**5.2. 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.

**5.2.1. Chebyshev.** Scarf [271] models the distributional ambiguity in a 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 the Scarf's seminal work, other researchers have investigated the Chebyshev ambiguity set in the context of the newsvendor model. Gallego and Moon [109] study

multiple extensions of the problem studied in Scarf [271]. 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.

Unlike the ambiguity sets studied in Scarf [271] and Gallego and Moon [109], the mean and covariance matrix can be unknown themselves and belong to some uncertainty sets. El Ghaoui et al. [101] study a distributionally robust one-period portfolio optimization, where the worst-case VaR over an ambiguity set of distributions with a known mean and covariance matrix is minimized. They show that this problem can be reformulated as a SOCP. Moreover, they 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. They extend their study to the case that the first two order 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 [131] for a similar result). Moreover, they show that in the presence of additional information on the distribution, besides the first two order 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. [101], Li [189] showcases the results in the context of a risk-averse portfolio optimization problem. Unlike El Ghaoui et al. [101] that considers polytopic and interval uncertainty sets for the mean and covariance, Lotfi and Zenios [195] 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 return constraint. They show that both problems can be reformulated as SOCPs.

Goldfarb and Iyengar [122] study a distributionally robust portfolio selection problem, where the asset returns $\tilde{\boldsymbol{\xi}}$ are formed by a linear factor model of the form $\tilde{\boldsymbol{\xi}} = \boldsymbol{\mu} + \boldsymbol{A}\tilde{\boldsymbol{f}} + \tilde{\boldsymbol{\epsilon}}$, where $\boldsymbol{\mu}$ is the vector of mean returns, $\tilde{\boldsymbol{f}} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ is the vector of random returns that derives the market, $\boldsymbol{A}$ is the factor loading matrix, and $\tilde{\boldsymbol{\epsilon}} \sim N(\boldsymbol{0}, \boldsymbol{B})$ is the vector of residual returns with a diagonal matrix $\boldsymbol{B}$. It is assumed that $\tilde{\boldsymbol{\epsilon}}$ is independent of $\tilde{\boldsymbol{f}}, \boldsymbol{F}$, and $\boldsymbol{B}$. Thus, $\tilde{\boldsymbol{\xi}} \sim N(\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top + \boldsymbol{B})$; hence, the uncertainty in the mean is independent of the uncertainty in the covariance matrix of the returns. Under the assumption that the covariance matrix $\boldsymbol{\Sigma}$ is known, Goldfarb and Iyengar [122] study three different models to form the uncertainty in $\boldsymbol{B}$, $\boldsymbol{A}$, and $\boldsymbol{\mu}$ as follows:

$$\mathcal{U}_{\boldsymbol{B}} = \big\{\boldsymbol{B} \,\big|\, \boldsymbol{B} = \operatorname{diag}(\boldsymbol{b}),\ b_i \in [\underline{b}_i, \overline{b}_i],\ i = 1, \ldots, d\big\}, \tag{5.19}$$

$$\mathcal{U}_{\boldsymbol{A}} = \{\boldsymbol{A} \,|\, \boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{C},\ \|\boldsymbol{c}_i\|_g \le \rho_i,\ i = 1, \ldots, d\}, \tag{5.20}$$

$$\mathcal{U}_{\boldsymbol{\mu}} = \{\boldsymbol{\mu} \,|\, \boldsymbol{\mu} = \boldsymbol{\mu}_0 + \boldsymbol{\zeta},\ |\zeta_i| \le \gamma_i,\ i = 1, \ldots, d\}, \tag{5.21}$$

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 $\underline{d}_i$, $\overline{d}_i$, $\rho_i$, $\gamma_i$, $i = 1, \ldots, d$, vector $\boldsymbol{\mu}_0$, and matrices $\boldsymbol{A}_0$ and $\boldsymbol{G}$. Given this setup, Goldfarb and Iyengar [122] study a DRO approach to different portfolio optimization problems for the return $\tilde{\boldsymbol{\xi}}^\top \boldsymbol{x}$ on the portfolio $\boldsymbol{x}$, where $\sum_{i=1}^n x_i = 1$. This includes: (1) minimum variance,

$\mathrm{Var}\,[\cdot]$, subject to a minimum expected return constraint

$$\min_{\boldsymbol{x}\geq\mathbf{0}}\ \max_{\boldsymbol{A}\in\mathcal{U}_{\boldsymbol{A}},\boldsymbol{B}\in\mathcal{U}_{\boldsymbol{B}}}\left\{\mathrm{Var}\left[\tilde{\boldsymbol{\xi}}^{\top}\boldsymbol{x}\right]\,\middle|\,\min_{\boldsymbol{\mu}\in\mathcal{U}_{\boldsymbol{\mu}}}\mathbb{E}\left[\tilde{\boldsymbol{\xi}}^{\top}\boldsymbol{x}\right]\geq\alpha,\ \sum_{i=1}^{n}x_i=1\right\},$$

(2) maximum expected return subject to a maximum variance constraint

$$\max_{\boldsymbol{x}\geq\mathbf{0}}\ \min_{\boldsymbol{\mu}\in\mathcal{U}_{\boldsymbol{\mu}}}\left\{\mathbb{E}\left[\tilde{\boldsymbol{\xi}}^{\top}\boldsymbol{x}\right]\,\middle|\,\max_{\boldsymbol{A}\in\mathcal{U}_{\boldsymbol{A}},\boldsymbol{B}\in\mathcal{U}_{\boldsymbol{B}}}\mathrm{Var}\left[\tilde{\boldsymbol{\xi}}^{\top}\boldsymbol{x}\right]\leq\lambda,\ \sum_{i=1}^{n}x_i=1\right\},$$

(3) maximum Sharpe ratio

$$\max_{\boldsymbol{x}\geq\mathbf{0}}\ \min_{\boldsymbol{\mu}\in\mathcal{U}_{\boldsymbol{\mu}},\boldsymbol{A}\in\mathcal{U}_{\boldsymbol{A}},\boldsymbol{B}\in\mathcal{U}_{\boldsymbol{B}}}\left\{\frac{\mathbb{E}\left[\tilde{\boldsymbol{\xi}}^{\top}\boldsymbol{x}\right]-\boldsymbol{\xi}_0^{\top}\boldsymbol{x}}{\sqrt{\mathrm{Var}\left[\tilde{\boldsymbol{\xi}}^{\top}\boldsymbol{x}\right]}}\,\middle|\,\sum_{i=1}^{n}x_i=1\right\},$$

where $\boldsymbol{\xi}_0$ is a risk-free return rate, and (4) maximum expected return subject to a maximum VaR constraint

$$\max_{\boldsymbol{x}\geq\mathbf{0}}\ \min_{\boldsymbol{\mu}\in\mathcal{U}_{\boldsymbol{\mu}}}\left\{\mathbb{E}\left[\tilde{\boldsymbol{\xi}}^{\top}\boldsymbol{x}\right]\,\middle|\,\max_{\boldsymbol{\mu}\in\mathcal{U}_{\boldsymbol{\mu}},\boldsymbol{A}\in\mathcal{U}_{\boldsymbol{A}},\boldsymbol{B}\in\mathcal{U}_{\boldsymbol{B}}}\mathrm{VaR}_{\beta}\left[\tilde{\boldsymbol{\xi}}^{\top}\boldsymbol{x}\right]\geq\alpha,\ \sum_{i=1}^{n}x_i=1\right\}.$$

Note that the constraint $\mathrm{VaR}_{\beta}\left[\tilde{\boldsymbol{\xi}}^{\top}\boldsymbol{x}\right]\geq\alpha$ is equivalent to $P\{\tilde{\boldsymbol{\xi}}^{\top}\boldsymbol{x}\leq\alpha\}\leq\beta$. They show that all the above four classes of problems can be reformulated as SOCPs. They further assume the covariance matrix $\boldsymbol{\Sigma}$ or its inverse are unknown and belong to ellipsoidal uncertainty sets, and show that the above problems can be reformulated as SOCPs. El Ghaoui et al. [101] study a similar linear factor model as the one in Goldfarb and Iyengar [122], but 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.

Li and Kwon [188] 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 variance 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.

Grünwald and Dawid [126] 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.

Rujeerapaiboon et al. [267] 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.

**5.2.2. Delage and Ye.** Unlike the ambiguity sets studied in Scarf [271] and Gallego and Moon [109], Delage and Ye [82] allow the mean and covariance matrix to be unknown themselves. This ambiguity set is defined as follows [82]:

$$\mathcal{P}^{DY} := \left\{ P \in \mathfrak{M}(\Xi, \mathcal{F}) \;\middle|\; \begin{array}{l} P\{\tilde{\boldsymbol{\xi}} \in \Omega\} = 1, \\ \left(\mathbb{E}_P\left[\tilde{\boldsymbol{\xi}}\right] - \boldsymbol{\mu}_0\right)^\top \boldsymbol{\Sigma}_0^{-1} \left(\mathbb{E}_P\left[\tilde{\boldsymbol{\xi}}\right] - \boldsymbol{\mu}_0\right) \leq \varrho_1, \\ \mathbb{E}_P\left[(\tilde{\boldsymbol{\xi}} - \boldsymbol{\mu}_0)(\xi - \boldsymbol{\mu}_0)^\top\right] \preccurlyeq \varrho_2 \boldsymbol{\Sigma}_0 \end{array} \right\}. \tag{5.22}$$

The first constraint denotes the smallest closed convex set $\Omega \subseteq \mathbb{R}^d$ that contains $\tilde{\boldsymbol{\xi}}$ with probability one (w.p. 1), i.e., $\Omega$ is the support of $\mathbb{P} = P \circ \tilde{\boldsymbol{\xi}}^{-1}$ w.p. 1. The second constraint ensures that the mean of $\tilde{\boldsymbol{\xi}}$ lies in an ellipsoid of size $\varrho_1$ and centered around the nominal mean estimate $\boldsymbol{\mu}_0$. Note that we can equivalently write this constraint as

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

The third constraint defines the second central-moment matrix of $\tilde{\boldsymbol{\xi}}$ by a matrix inequality. The parameters $\varrho_1$ and $\varrho_2$ 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 (5.22), with $\varrho_1 = 0$ and $\varrho_2 = 1$. Delage and Ye [82] propose data-driven methods to form confidence regions for the mean and the covariance matrix of the random vector $\tilde{\boldsymbol{\xi}}$ using the concentration inequalities of McDiarmid [201], 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. A conic generalization of the ambiguity set $\mathcal{P}^{\text{DY}}$, beyond the first and second moment information is also studied in Delage [80]. Below, we present a duality result for $\sup_{\mathbb{P}\in\mathcal{P}^{\text{DY}}} \mathbb{E}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$ given a fixed $\boldsymbol{x} \in \mathcal{X}$, due to Delage and Ye [82].

THEOREM 5.8. *(Delage and Ye [82, 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 $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is $P$-integrable for all $P \in \mathcal{P}^{DY}$. Then, $\sup_{P\in\mathcal{P}^{DY}} \mathbb{E}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$ is equal to the optimal value of the following semi-infinite convex conic optimization problem:*

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

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

The reformulated problem in Theorem 5.8 is polynomial-time solvable under the following assumptions [82]:

- The sets $\mathcal{X}$ and $\Omega$ are convex and compact, and are both equipped with oracles that confirm the feasibility of a point $\boldsymbol{x}$ and $\tilde{\boldsymbol{\xi}}$, or provide a hyperplane that separates the infeasible point from its corresponding feasible set in time polynomial in the dimension of the set.
- Function $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) := \max_{k\in\{1,\ldots,K\}} h_k(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is piecewise and is such that for each $k$, $h_k(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is convex in $\boldsymbol{x}$ and concave in $\tilde{\boldsymbol{\xi}}$. In addition, for any given pair $(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$, one can evaluate $h_k(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$, find a supergradient of $h_k(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ in $\tilde{\boldsymbol{\xi}}$, and find a subgradient of $h_k(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ in $\boldsymbol{x}$, in time polynomial in the dimension of $\mathcal{X}$ and $\Omega$.

As a special case where $\Omega$ is an ellipsoid, the resulting reformulation in Theorem 5.8 reduces to a SDP of finite size. Motivated by the computational challenges of solving a semidefinite reformulation of (1.5) formed via (5.22), Cheng et al. [79] 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.

Popescu [239] study a class of stochastic optimization problems, where the objective function is characterized with one- or two-point support functions. They show that when the ambiguity set of distributions is formed with all distributions with known mean and covaraince, the problem reduces to a deterministic parametric quadratic program. In particular, this result holds for increasing concave utilities with convex or concave-convex derivatives.

Goh and Sim [120] study a DRO approach to a stochastic linear optimization problem with expectation constraints, where the support and mean of the random parameters belong to a conic-representable set, while the covariance matrix is assumed to be known.

**5.2.2.1. Discrete Problems.** Under the assumption that the mean and covariance are known, Natarajan and Teo [210] investigate the worst-case expected value of the maximum of a linear function of random variables as follows:

$$\sup_{P\in\mathcal{P}} \ \mathbb{E}_P\left[Z(\tilde{\boldsymbol{\xi}})\right],$$

where $Z(\tilde{\boldsymbol{\xi}}) = \max\left\{\tilde{\boldsymbol{\xi}}^\top \boldsymbol{x} \,\middle|\, \boldsymbol{x}\in\mathcal{X}\right\}$. The set $\mathcal{X}$ is specified with either a finite number of points or a bounded feasible region to a mixed-integer LP. To obtain an upper bound, they approximate the copostive programming reformulation of the problem, presented in Natarajan et al. [211, Theorem 3.3], with a SDP. They show that the complexity of computing this bound is closely related to characterizing the convex hull of the quadratic forms of the points in the feasible region.

Xie and Ahmed [323] study a DRO approach to a two-stage stochastic program with a simple integer round-up recourse function, defined as follows:

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

where

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

The ambiguity set is formed by the product of one-dimensional ambiguity sets for each component of the random parameter $\tilde{\boldsymbol{\xi}}$, 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.

Ahipasaoğlu et al. [2] study distributionally robust project crashing problems. They assume the underlying joint probability distribution of the activity durations lies in an ambiguity set of distributions with the given mean, standard deviation, and correlation information. The goal is to select the means and standard deviations to minimize the worst-case expected makespan for the project network with respect to the ambiguity set of distributions. Unlike the typical use of the SDP solvers to directly solve the problem, they exploit the problem structure to reformulate it as a convex-concave saddle point problem over the first two moment variables in order to solve the formulation in polynominal time.

A distributionally robust approach to an individual chance constraint with binary decisions is studied in Zhang et al. [340]. They consider the following individual chance constraints with $g_j(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$, $j = 1,\ldots,m$, in (1.6) is defined as

$$g_j(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) := \mathbb{1}_{[\tilde{\boldsymbol{\xi}}^\top \boldsymbol{x} \leq \boldsymbol{b}]}(\tilde{\boldsymbol{\xi}}),$$

where $\boldsymbol{x} \in \{0,1\}^n$. They form the ambiguity set of distributions by all joint distributions whose marginal means and covarinces satisfy the constraints in (5.22). They reformulate the chance constraints as binary second-order conic (SOC) constraints.

**5.2.2.2. Risk and Chance Constraints.** Risk-based DRO models formed via the ambiguity set (5.22) are also studied in the literature. Bertsimas et al. [39] study a risk-averse distributionally robust two-stage stochastic linear optimization problem where 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 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. An explicit construction of the worst-case distributions is also given. The results are applied to the production-transportation problem and a single facility minimax distance problem. Li [189] 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.

Zymler et al. [346] extend the work of El Ghaoui et al. [101] 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. [101], Zymler et al. [346] 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. [345], Rujeerapaiboon et al. [266] 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. [72] 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 [61] 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. [345] 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 $\tilde{\boldsymbol{\xi}}$, 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 $\tilde{\boldsymbol{\xi}}$ by relying on nonlinear Farkas lemma and $\mathcal{S}$-lemma, see, e.g., Pólik and Terlaky [237].

Chen et al. [72] 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. [345] show that the CVaR approximation is exact for joint chance constraints whose constraint functions depend linearly on $\tilde{\boldsymbol{\xi}}$. They evaluate the performance of their approximation for joint chance constraint in the context of a water reservoir control problem for hydro power generation and show it outperforms the Bonferroni approximation and the method of Chen et al. [72].

Motivated by the fact that chance constraints do not take into account the magnitude of the violation, Xu et al. [330] study a probabilistic envelope constraint. This approach can be interpreted as a continum of chance constraints with nondecreasing target values and probabilities. They show that when the first two order moments are known, an ambigious probabilistic envelope constraint is equivalent to a deterministic SIP, which is called as a *comprehensive robust optimization* problem [25, 27]. 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 [335] for an extension of the work in Xu et al. [330] to

the nonlinear inequalities.

**5.2.2.3. Statistical Learning.** Lanckriet et al. [179] 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.

**5.2.2.4. Multistage Setting.** Xin and Goldberg [326] study a multistage distributionally robust newvendor problem 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 [288], and provide sufficient conditions for time consistency. Unlike Xin and Goldberg [326] that suppose the demand process is stage-wise independent, Xin and Goldberg [325] 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 [334] and Van Parys et al. [307] study a stochastic optimal control model to minimize the worst-case probability that a system remains in a safe region for all stages. Yang [334] 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. [307] form the ambiguity set by all distributions with a known mean and covariance.

**5.2.3. 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:

$$\mathcal{P}^{MM} := \left\{ P \in \mathfrak{M}_+(\Xi, \mathcal{F}) \,\middle|\, \nu_1 \preceq P \preceq \nu_2,\ \int_\Xi \boldsymbol{f} dP \in [\boldsymbol{l}, \boldsymbol{u}] \right\}, \tag{5.23}$$

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 (5.23) enforces a preference relationship between probability 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 [291] propose this framework, and special cases of it appear in Popescu [238],

Bertsimas and Popescu [33], Perakis and Roels [228], Mehrotra and Papp [202], among others. Note that if the first constraint in (5.23) 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 [180]. Using this unified set, one can impose bounds on the standard moments, by setting the $i$th entry of $\boldsymbol{f}$ to have the form: $f_i(\tilde{\boldsymbol{\xi}}) := (\xi_1)^{k_{i1}} \cdot (\xi_2)^{k_{i2}} \cdots (\xi_d)^{k_{id}}$, where $k_{ij}$ is a nonnegative integer indicating the power of $\xi_j$ for the $i$th 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 [238] and Perakis and Roels [228] to model the unimodality and symmetry of distributions within this framework (see also Section 5.3).

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

THEOREM 5.9. *(Shapiro and Ahmed [291, Proposition 2.1]) For a fixed $\boldsymbol{x} \in \mathcal{X}$, suppose that $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is $\nu_2$-integrable, i.e., $\int_\Xi |h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})| d\nu_2 < \infty$, as defined in (5.23). Moreover, suppose that $\boldsymbol{f}$ is $\nu_2$-integrable, and there exists $\nu_1 \preceq P \preceq \nu_2$ such that $\int_\Xi \boldsymbol{f} dP \in (\boldsymbol{l}, \boldsymbol{u})$. If $\sup_{P \in \mathcal{P}^{MM}} \mathbb{E}_P \left[ h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right]$ is finite, then, it can be written as the optimal value of the following problem:*

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

*where*

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

Shapiro and Ahmed [291] focus on a special case of (5.23), 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 $h$ and two-stage stochastic programs. Popescu [238], Bertsimas and Popescu [33], Mehrotra and Papp [202] study the classical problem of moments, i.e., ambiguity set is formed via only the second constraints in (5.23). When $\boldsymbol{f}$ are moment functions, Mehrotra and Papp [202] show that under mild conditions (continuous function $h$ and compact support $\Omega$), the optimal value of a sequence of problems of the form (1.5), 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 (1.1) under the reference distribution. Moreover, using the SIP reformulation of (1.5), Mehrotra and Papp [202] propose a cutting surface method to solve a convex (1.5). This method can be applied to problems where bounds of moments are of arbitrary order, and possibly, bounds on nonpolynomial moments are available.

Royset and Wets [265] study a DRO model with a decision-dependent ambiguity set, where the ambiguity set has the form of (5.23), without the second set of constraints, and the first constraint is formed via the decision-dependent cumulative distribution functions (cdf). They establish the convergence properties of the solutions to this problem by exploiting and refining results in variational analysis.

Besides Shapiro and Ahmed [291], there are other studies that focus on special types of cost function $h$. Two-stage stochastic programs have received much attention

in this class. Chen et al. [73] 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 (5.23) 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 $\Xi$. 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 [4] 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 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 linear program.

**5.2.3.1. Discrete Problems.** Bansal et al. [9] study a (two-stage) distributionally robust integer program with pure binary first-stage and mixed-binary second stage decisions 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 (5.23) without the first constraint, satisfy these conditions. Hanasusanto et al. [135] study a finite adaptability scheme to approximate the following two-stage distributionally robust linear program, with binary recourse decisions and optimized certainty equivalent as a risk measure:

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

where

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

and $\mathcal{R}_P\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]$ is an optimized certainty equivalent risk measure corresponding to the utility function $u$: $\mathcal{R}_P\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] = \inf_{\eta\in\mathbb{R}}\eta + \mathbb{E}_P\left[u\big(h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-\eta\big)\right]$ [22, 23]. 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 (5.23) but without the first constraint, where the support is assumed to be a polytope and functions $f_i$ are also convex piecewise linear in $\tilde{\boldsymbol{\xi}}$. 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 linear programs. Postek et al. [242] study a two-stage stochastic integer program, where the second-stage problem is a mixed-integer program. 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. They apply their results to a surgery block allocation problem.

**5.2.3.2. Risk and Chance Constraints.** Bertsimas and Popescu [33] study the worst-case bound on the probability of a multivariate random vector falling outside a semialgebreic confidence region (i.e., a set described via polynomial inequalities) over an ambiguity set of the form (5.23), where functions $\boldsymbol{f}$ are represented by all polynomials of up to $k$th-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 [33] appears in Vandenberghe et al. [310], 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. [73], Liu et al. [191] 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. The results are applied to a portfolio optimization problem.

Nemirovski and Shapiro [213] study a convex approximation, referred to as *Bernstein* approximation, to an ambiguous joint chance-constrained problem of the form

$$\min_{\boldsymbol{x}\in\mathcal{X}} \; h(\boldsymbol{x}) \tag{5.24a}$$

$$\text{s.t.} \quad \inf_{P\in\mathcal{P}} \; P\left\{\tilde{\boldsymbol{\xi}} : g_{i0}(\boldsymbol{x}) + \sum_{j=1}^{d} \tilde{\xi}_j g_{ij}(\boldsymbol{x}) \leq 0, \; i = 1,\ldots,m\right\} \geq 1-\epsilon. \tag{5.24b}$$

THEOREM 5.10. *(Nemirovski and Shapiro [213, Theorem 6.2]) Suppose that the ambiguous joint chance-constrained problem (5.24) is such that (i) the components of the random vector $\tilde{\boldsymbol{\xi}}$ are independent of each other, with finite-valued moment generating functions, (ii) function $h(\boldsymbol{x})$ and all functions $g_{ij}$, $i = 1,\ldots,m$, $j = 0,\ldots,d$, are convex and well defined on $\mathcal{X}$, and (iii) the ambiguity set of probability distributions $\mathcal{P}$ forms a convex set. Let $\epsilon_i$, $i = 1,\ldots,m$, be positive real values such that $\sum_{i=1}^{m} \epsilon_i \leq \epsilon$. Then, the problem*

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathcal{X}} \; h(\boldsymbol{x}) \\ &s.t. \quad \inf_{t>0} \; \left[g_{i0}(\boldsymbol{x}) + t\hat{\Psi}(t^{-1}\boldsymbol{z}^i[\boldsymbol{x}]) - t\log\epsilon_i\right] \leq 0, \; i = 1,\ldots,m, \end{aligned}$$

[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 $\gamma > 0$ [191].

*where* $z^i(\boldsymbol{x}) = \big(g_{i1}(\boldsymbol{x}), \ldots, g_{id}(\boldsymbol{x})\big)$ *and*

$$\hat{\Psi}(\boldsymbol{z}) := \max_{Q_1 \times \ldots \times Q_d \in \mathcal{P}} \sum_{j=1}^{d} \log\Big(\int_{\Xi} \exp\{z_j s\} dQ_j(s)\Big),$$

*is a conservative approximation of problem* (5.24)*, i.e., every feasible solution to the approximation is feasible for the chance-constrained problem* (5.24)*. This approximation is a convex program and is efficiently solvable, provided that all* $g_{ij}$ *and* $\hat{\Psi}$ *are efficiently computable, and* $\mathcal{X}$ *is computationally tractable.*

Hanasusanto et al. [136] study a distributionally robust joint chance constrained stochastic program where each chance constraint is linear in $\tilde{\boldsymbol{\xi}}$, and the technology matrix and right hand-side are affine in $\boldsymbol{x}$. They form the ambiguity set of distributions as in (5.23) 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 $\boldsymbol{x}$. They apply their results to problems in project management and image reconstruction. While their formulation is exact for the distributionally robust chance constrained project crashing problem, the size of the formulation grows in the number of paths in the network. For other research in chance-constrained optimization problem, we refer to Xie et al. [324], Xie and Ahmed [321].

**5.2.3.3. Statistical Learning.** Fathony et al. [104] study a distributonally robust approach to graphical models for leveraging the graphical structure among the variables. The proposed model in Fathony et al. [104] seeks a predictor to make a probabilistic prediction $\hat{P}(\hat{y}|\boldsymbol{u})$ over all possible label assignments so that it minimizes the worst-case conditional expectation of the prediction loss $l(\hat{y}, \bar{y})$ with respect to $\bar{P}(\bar{y}|\boldsymbol{u})$ as follows:

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

where $\Phi(\boldsymbol{U}, Y)$ is a given feature function and $\breve{\Phi} = \mathbb{E}_{(\boldsymbol{U},Y)\sim\breve{P}}[\Phi(\boldsymbol{U}, Y)]$. 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. [104] show that the DRO approach enjoys the consistency guarantees of probabilistic graphical models, see, e.g., Lafferty et al. [174], and has the advantage of incorporating customized loss metrics during the training as in large margin models, see, e.g., Tsochantaridis et al. [302].

**5.2.4. Moment Matrix Inequalities.** In this section we review an ambiguity set that generalizes both the ambiguity set $\mathcal{P}^{\text{DY}}$ (5.22) and the ambiguity set $\mathcal{P}^{\text{MM}}$ (5.23) as follows:

$$\mathcal{P}^{MMI} := \left\{ P \in \mathfrak{M}_+(\Xi, \mathcal{F}) \;\middle|\; \boldsymbol{L} \preccurlyeq \int_{\Xi} \boldsymbol{F} dP \preccurlyeq \boldsymbol{U} \right\}, \tag{5.25}$$

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 scalars. As in (5.23), 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. [331], 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 (5.23). Moreover, by setting $\boldsymbol{F}_2$ to be a matrix as $\begin{pmatrix} -\boldsymbol{\Sigma}_0 & \boldsymbol{\mu}_0 - \tilde{\boldsymbol{\xi}} \\ (\boldsymbol{\mu}_0 - \tilde{\boldsymbol{\xi}})^\top & -\varrho_1 \end{pmatrix}$, $\boldsymbol{F}_3$ to be a matrix as $(\tilde{\boldsymbol{\xi}} - \boldsymbol{\mu}_0)(\xi - \boldsymbol{\mu}_0)^\top$, $\boldsymbol{L}_2 = -\boldsymbol{\infty}$, $\boldsymbol{U}_2 = \boldsymbol{L}_3 = \boldsymbol{0}$, and $\boldsymbol{U}_3 = \varrho_2 \boldsymbol{\Sigma}_0$, we can recover (5.22).

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

THEOREM 5.11. *For a fixed $\boldsymbol{x} \in \mathcal{X}$, suppose that $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ and $\boldsymbol{F}$ are integrable for all $P \in \mathcal{P}^{MMI}$. In addition, suppose that the following Slater-type condition holds:*

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

*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 $\sup_{P\in\mathcal{P}^{MM}} \mathbb{E}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$ is finite, then, it can be written as the optimal value of the following problem:*

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

*Proof.* Using the conic duality results from Theorem 4.3, we write the dual of $\sup_{P\in\mathcal{P}^{\text{MM}}} \mathbb{E}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$ as

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

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

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

Thus, we can write the first constraint above as

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

The Slater-type condition ensures that the strong duality holds [284]. ∎

Suppose that every finite subset of $\Xi$ is $\mathcal{F}$-measurable, i.e., for every $s \in \Xi$, the corresponding Dirac measure $\delta(s)$ (of mass one at point $s$) belongs to $\mathfrak{M}_+(\Xi, \mathcal{F})$. Then, the first constraint in Theorem 5.11 can be written as follows [284]:

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

Motivated by the difficulty in verifying the Slater-type conditions to guarantee strong duality for $\sup_{P \in \mathcal{P}^{\text{MMI}}} \mathbb{E}_P \left[ h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right]$ and its dual, Xu et al. [331] investigate the duality conditions from the perspective of lower semicontinuity of the optimal value function inner maximization problem, with a perturbed ambiguity set. While these conditions are restrictive in general, they show that they are satisfied in the case of compact $\Xi$ or bounded $\boldsymbol{F}_i$. Xu et al. [331] 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. They illustrate their results for the portfolio optimization and multiproduct newsvendor problems.

**5.2.5. Cross-Moment or Nested Moment.** In an attempt to unify modeling and solving DRO models, Wiesemann et al. [318] propose a framework for modeling the ambiguity set of probability distributions as follows:

(5.26)
$$\mathcal{P}^{\text{WKS}} := \left\{ \mathbb{P} \in \mathfrak{P}\left(\mathbb{R}^d \times \mathbb{R}^r, \mathfrak{B}(\mathbb{R}^d) \times \mathfrak{B}(\mathbb{R}^r)\right) \,\middle|\, \begin{array}{l} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{A}\tilde{\boldsymbol{\xi}} + \boldsymbol{B}\tilde{\boldsymbol{u}}\right] = \boldsymbol{b}, \\ P\{(\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{u}}) \in \mathcal{C}_i\} \in [\underline{p}_i, \overline{p}_i], \ i \in \mathcal{I} \end{array} \right\},$$

where $\mathbb{P}$ represents a joint probability distribution of $\tilde{\boldsymbol{\xi}}$ and some auxiliary random vector $\tilde{\boldsymbol{u}} \in \mathbb{R}^r$. 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

(5.27)
$$\mathcal{C}_i := \left\{ (\boldsymbol{\xi}, \boldsymbol{u}) \in \mathbb{R}^d \times \mathbb{R}^r \,\middle|\, \boldsymbol{C}_i \boldsymbol{\xi} + \boldsymbol{D}_i \boldsymbol{u} \preccurlyeq_{\mathcal{K}_i} \boldsymbol{c}_i \right\},$$

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 $(\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{u}})$. This set contains all distributions with prescribed conic-representable confidence sets and with mean values residing on an affine manifold. An important aspect of (5.26) is that the inclusion of an auxiliary random vector $\tilde{\boldsymbol{u}}$ gives the flexibility to model a rich variety of structural information about the marginal distribution of $\tilde{\boldsymbol{\xi}}$ in a unified manner. Using this framework, Wiesemann et al. [318] show that many ambiguity sets studied in the literature can be represented by a projection of the ambiguity set

(5.26) on the space of $\tilde{\boldsymbol{\xi}}$. 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 5.12. *(Wiesemann et al. [318, 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:*

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

*as well as the lifted ambiguity set*

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

*which involves the auxiliary random vector $\tilde{\boldsymbol{u}} \in \mathbb{R}^N$. We have that (i) $\mathcal{P}'$ is the union of all marginal distributions of $\tilde{\boldsymbol{\xi}}$ 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 (5.26).*

Using Theorem 5.12, Wiesemann et al. [318] show how an ambiguity set of the form $\mathcal{P}^{\text{WKS}}$, defined in (5.26), with conic-representable expectation constraints and a collection of conic-representable confidence sets, can represent ambiguity sets formed via (1) $\phi$-divergences, (2) mean, (3) mean and upper bound on the covariance matrix (i.e., a special case of the ambiguity set (5.22)), (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 (5.26) 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 (5.26) 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 [78].

Wiesemann et al. [318] determine conditions under which distributionally robust expectation constraints, formed via the proposed ambiguity set (5.26), can be solved in polynomial time as follows: (i) the cost function $g_j$, $j = 1, \ldots, m$, is convex and piecewise affine in $\boldsymbol{x}$ and $\tilde{\boldsymbol{\xi}}$ (i.e., $g_j(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) := \max_{k \in \{1,\ldots,K\}} g_{jk}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ with $g_{jk}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) := s_{jk}(\tilde{\boldsymbol{\xi}})\boldsymbol{x} + t_{jk}(\tilde{\boldsymbol{\xi}})$ such that $s_{jk}(\tilde{\boldsymbol{\xi}})$ and $t_{jk}(\tilde{\boldsymbol{\xi}})$ are affine in $\tilde{\boldsymbol{\xi}}$) 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 5.13. *(Wiesemann et al. [318, Theorem 1]) Consider a fixed $\boldsymbol{x} \in \mathcal{X}$. Then, under suitable regularity conditions, $\sup_{\mathbb{P} \in \mathcal{P}^{WKS}} \mathbb{E}_{\mathbb{P}}\left[g_j(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] \leq 0$, $j = 1, \ldots, m$, 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 $k \in \{1, \ldots, K\}$, that satisfy the following systems:*

$$\begin{aligned} &\boldsymbol{b}^\top \boldsymbol{\beta} + \sum_{i \in \mathcal{I}} (\overline{p}_i \kappa_i - \underline{p}_i \lambda_i) \leq 0, \\ &\boldsymbol{c}_i^\top \boldsymbol{\alpha}_{ik} + \boldsymbol{s}_k^\top \boldsymbol{x} + \boldsymbol{t}_k \leq \sum_{i' \in \{i\} \cup \mathcal{A}(i)} (\kappa_{i'} - \lambda_{i'}), \quad \forall i \in \mathcal{I},\ k \in \{1, \ldots, K\}, \\ &\boldsymbol{C}_i^\top \boldsymbol{\alpha}_{ik} + \boldsymbol{A}^\top \boldsymbol{\beta} = \boldsymbol{S}_k^\top \boldsymbol{x} + \boldsymbol{t}_k, \quad \forall i \in \mathcal{I},\ k \in \{1, \ldots, K\}, \end{aligned}$$

$$\boldsymbol{D}_i^\top \boldsymbol{\alpha}_{ik} + \boldsymbol{B}^\top \boldsymbol{\beta} = 0, \quad \forall i \in \mathcal{I},\ k \in \{1, \ldots, K\},$$

*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 5.13 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. [318] also provide tight tractable conservative approximations for problems that violate the nesting condition by proposing an outer approximation of (5.26). They discuss several mild modifications of the conditions on $\boldsymbol{g}$.

There are several papers that use the ambiguity set (5.26) and consider its generalization or special cases. Chen et al. [78] 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. [318]. However, unlike Wiesemann et al. [318], 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 $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ over both sets are equal, provided that the objective function $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is tractable and conic-representable in $\tilde{\boldsymbol{\xi}}$ 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. [78] 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. When 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 [82] 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. [241] derive exact reformulation of the worst-case expected constraints when function $g(\boldsymbol{x}, \cdot)$ is convex in $\tilde{\boldsymbol{\xi}}$, and the ambiguity set of distributions consists of all distributions of componentwise independent $\tilde{\boldsymbol{\xi}}$ with known support, mean, and mean-asboulute deviation information. They also obtain exact reformulation of the resulting model when $g(\boldsymbol{x}, \cdot)$ is concave in $\tilde{\boldsymbol{\xi}}$ and there is additional information on the probability that a component is greater than or equal to its mean. These reformulations involve a number of terms that are exponential in the dimension of $\tilde{\boldsymbol{\xi}}$. They show how upper bounds can be constructed that alleviate the independence restriction, and require only a linear number of terms, by exploiting models in which random variables are linearly aggregated and function $g(\boldsymbol{x}, \cdot)$ is convex. Under the assumption of independent random variables, they use the above results for the worst-case expected constraints to derive safe approximations to the corresponding individual chance constrained problems.

To reduce the conservatism of the robust optimization due to its constraint-wise approach and the assumption that all constraints are hard for all scenarios in the uncertainty set, Roos and den Hertog [264] 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 $\tilde{\boldsymbol{\xi}}$ with known support, mean, and mean-asboulute deviation information. When the right-hand side is uncertain, they use the results in Postek et al. [241] 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. [241] to derive tractable reformulations. We also refer to Sun et al. [301] for a two-stage quadratic stochastic optimization problem and DeMiguel and Nogales [84] for a portfolio optimization problem.

Bertsimas et al. [44] develop a modular and tractable framework for solving an adaptive distributionally robust two-stage linear optimization problem with recourse of the form

$$\min_{\boldsymbol{x}}\left\{\boldsymbol{c}^\top\boldsymbol{x}+\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]\,\middle|\,\boldsymbol{x}\in\mathcal{X}\right\},$$

where

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

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-representbale ambiguity set (5.26). 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. [318] 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. [44] discuss extension to the conic-representbale ambiguity set (5.26) and multistage problems. They also apply their results to medical appointment scheduling and single-item multiperiod newsvendor problems.

Following the approach in Bertsimas et al. [44], Zhen et al. [344] 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 uncertainty sets, there exists convex two-piecewise affine functions that are optimal for the adjustable variables. By studying the medical appointment scheduling considered in Bertsimas et al. [44], they show that their approach improves the solutions obtained in Bertsimas et al. [44].

[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 [74, 75, 24]. Whether or not the linear decision rules are optimal depends on the problem [293].

**5.2.5.1. Statistical Learning.** Gong et al. [123] 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 (5.26). They reformulate the resulting model as a conic optimization problem, based on the results in Wiesemann et al. [318].

**5.2.5.2. Multistage Setting..** A Markov decision process with unknown distribution for the transition probabilities and rewards for each state is studied in Xu and Mannor [329, 328]. 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 [338] extends the work in Xu and Mannor [329, 328] by forming the ambiguity set of distributions as in (5.26).

**5.2.6. Marginals (Fréchet).** All the moment-based ambiguity sets discussed so far, study the ambiguity of the joint probability distribution of the random vector $\tilde{\boldsymbol{\xi}}$. Papers reviewed in this section assume that additional information on the marginal distributions is available. We refer to the class of joint distributions with fixed marginal distributions as the *Fréchet* class of distributions [91].

**5.2.6.1. Discrete problems.** Chen et al. [69] study a problem of the form (1.5), where the cost function $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ 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. [36], to a class of integral polytopes. They discuss their results on four polynomial time solvable instances, arising in the appointment scheduling problem, max flow problem with random arc capacities, ranking problem with random utilities, and project scheduling problems with irregular random starting time costs.

**5.2.6.2. Risk and Chance Constraints.** Dhara et al. [89] 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. [91] study a DRO model of the form (1.5) with a convex piecewise linear objective function in $\tilde{\boldsymbol{\xi}}$ 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 linear programming. In particular, they develop a tight linear programming 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.

Natarajan et al. [212] 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:

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

where $h(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) = -\tilde{\boldsymbol{\xi}}^{\top}\boldsymbol{x} + \max_{\boldsymbol{y}\in\{0,1\}^{q_1}} \tilde{\boldsymbol{\xi}}^{\top}\boldsymbol{y}$ and

$$\mathrm{WCVaR}_{\alpha}^{P}\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] = \sup_{P\in\mathcal{P}} \mathrm{CVaR}_{\alpha}^{P}\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right].$$

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 (5.23). 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. They show the maximum weight subset selection problem is polynomially solvable under (i) and (ii). They illustrate their results on subset selection and the shortest path problems.

Zhang et al. [341] 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. They present a branch-and-price algorithm to solve the problem, and apply their results to a surgery scheduling problem for operating rooms.

**5.2.6.3. Statistical Learning.** Farnia and Tse [103] 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 [126] 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 [119] for the case of Shannon entropy. Farnia and Tse [103] 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 [103] 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. [98] 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. [253] study a DRO approach to

the binary classification problem, with an ambiguity set similar to that of Eban et al. [98], 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.

**5.2.7. 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* [169]. For example, a mixture model can be defined as the set of all mixtures of normal distributions with mean $\mu$ and standard deviation $\sigma$ 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:

$$P(B) = \int_{\mathcal{A}} Q_{\boldsymbol{a}}(B)\ M(d\boldsymbol{a}), \quad B \in \mathcal{F}, \tag{5.28}$$

where $M$ is any probability distribution on $\mathcal{A}$ [181]. Hence, modeling the ambiguity in the mixture probabilities may give rise to a DRO model over the *resultant or barycenter* $P$ of $M$ [238].

**5.2.7.1. Risk and Chance Constraints.** Lasserre and Weisser [181] 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. [169] 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 linear program, and in the latter case, they reformulate the problem as a SOCP.

Hanasusanto et al. [133] 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 (5.22). 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. [133] 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 $\chi^2$-distance. For this case, they also obtain exact SDP reformulation as well as a conservative SDP approximation.

**5.3. 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 [238] 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:

$$\mathcal{P}^{SP} := \left\{ P \in \mathfrak{M}_+(\Xi, \mathcal{F}) \,\middle|\, \int_{\Xi} \boldsymbol{f} dP = \boldsymbol{a} \right\} \cap \{P \text{ satisfies structural properties}\}. \tag{5.29}$$

Popescu [238] 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 (5.29). Popescu [238] uses conic duality to evaluate such lower and upper bounds via SDPs. The key to the development in Popescu [238] 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 (5.28). 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 [238] 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 [238] further derives generalized Chebyshev's inequality for symmetric and unimodal distributions of multivariate random variables. A related notion to unimodality is $\alpha$-unmiodality, which is defined as follows:

Definition 5.14. *Dharmadhikari and Joag-Dev [90] For $\alpha > 0$, a distribution $\mathbb{P} \in \mathfrak{P}(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$ is called $\alpha$-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. [308] further extend the work of Popescu [238] to obtain worst-case probability bounds over $\alpha$-unimodal multivariate distributions with the same mode and within the class of distributions in $\mathcal{P}^{\mathrm{DY}}$, defined in (5.22), 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 [238], Van Parys et al. [308] derive semidefinite representations for worst-case probability bounds using Choquet representation of the ambiguity set. They demonstrate that classical generalized Chebyshev and Guass 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 $\alpha$-multimodal multivariate distributions, defined via a mixture distribution.

By relying on information from classical statistics as well as robust statistics, Hanasusanto et al. [134] propose a unifying canonical ambiguity set that contains

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

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. [318], and a structural ambiguity set on the marginal distributions, representing information such as symmetry and $\alpha$-unimodality. As in [238], the key to the development in Hanasusanto et al. [134] 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 $\tilde{\boldsymbol{\xi}}$ 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. [134, 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 [238], Vandenberghe et al. [310]. Similarly, they recover the generalized Gauss inequality, discovered in Van Parys et al. [308], 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.

Li et al. [187] study distributionally robust chance- and CVaR-constrained stochastic programs, where the ambiguity set contains all $\alpha$-unimodal distributions with the same first two order moments, and the safe region is bi-affine in both $\tilde{\boldsymbol{\xi}}$ 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. [309].

Hu et al. [154] 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. [155] 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. [155] 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.

To overcome the difficulty in evaluating extremal performance due to the lack of data, Lam and Mottet [178] 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.

**5.4. Kernel-Based Models.** In Sections 5.1–5.3, we discussed different sets to model the distributional ambiguity. In all the papers we reviewed in those sections, the form of ambiguity set is endogenously chosen by decision makers. However, when facing high-dimensional uncertain parameters, it may not be practical to fix the form of ambiguity set a priori, being even more complicated with the calibration of different parameters describing the set (see Section 6). An alternative practice is to learn the form of the ambiguity set by using unsupervised learning algorithms on the historical data. 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 $\boldsymbol{\xi}^i \in \mathbb{R}^d$. Let $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ be a *kernel* function.

Bertsimas and Kallus [32] 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 $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ in anticipation of the future, given the observation $\boldsymbol{u}$. However, the conditional distribution of $\tilde{\boldsymbol{\xi}}$ given $\boldsymbol{u}$ is not known and should be learned from data. Given $\{(\boldsymbol{u}^i, \boldsymbol{\xi}^i)\}_{i=1}^N$, they suggest to find a data-driven predictive prescription that minimizes $\sum_{i=1}^k w_k^i(\boldsymbol{u}) h(\boldsymbol{x}, \boldsymbol{\xi}^i)$ over $\mathcal{X}$. Functions $w_k^i(\boldsymbol{u})$ are weights learned locally from the data, in a way that predictions are made based on the mean or mode of the past observations that are in some way similar to the one at hand. Bertsimas and Kallus [32] 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 $\mathbb{E}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \middle| \boldsymbol{u}\right]$ using the Nadaraya-Watson kernel regression is obtained as

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

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: $K(a) = (1 - \|a\|^3)^3 \mathbb{1}_{[\|a\| \le 1]}$,
- Guassian 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 [32] resembles SAA. They show that under mild conditions, the problem is polynomially solvable and the resulting predictive prescription is asymptotically optimal and consistent. However, it is worth noting that Bertsimas and Kallus [32] illustrate that 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 [32] is investigated in other papers, see, e.g., Hannah et al. [138], Deng and Sen [85], Ban and Rudin [8], Pang Ho and Hanasusanto [225], to incorporate machine learning into decision making. Deng and Sen [85] use regression models such as $k$-nearest-neighbors regression to learn the conditional distribution of $\tilde{\boldsymbol{\xi}}$ 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 Hannah et al. [138], Ban and Rudin [8], Pang Ho and Hanasusanto [225], the weights are obtained by the Nadaraya-Watson kernel regression method.

For a newsvendor problem, Ban and Rudin [8] 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. Ban and Rudin [8] apply their study to the staffing levels of nurses for a hospital emergency room.

Tulabandhula and Rudin [303] incorporate machine learning for the decision making. But, different from Bertsimas and Kallus [32], they 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 [303] show that under some conditions this problem is equivalent to a robust optimization model, where the uncertainty set of the statistical model contains all models that are within $\epsilon$-optimality from the predictive model describing $\{(\boldsymbol{u}^i, \boldsymbol{\xi}^i)\}_{i=1}^N$. They illustrate the form of the uncertainty set for different loss functions used in the predictive statistical model, including least squares, 0-1, logistic, exponential, ramp, and hing losses. Tulabandhula and Rudin [305] study the application of the framework studied in Tulabandhula and Rudin [303] to a travelling repairman problem, where a repair crew is seeking for an optimal route to repair the nodes on a graph while the failure probabilities are unknown.

Similar to Tulabandhula and Rudin [303], Tulabandhula and Rudin [304] 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. However, unlike Bertsimas and Kallus [32], Deng and Sen [85], Tulabandhula and Rudin [303], and Tulabandhula and Rudin [305], Tulabandhula and Rudin [304] study a robust optimization framework. Their idea to form the uncertainty set of $\tilde{\boldsymbol{\xi}}$ 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.

Similar to Bertsimas and Kallus [32], Bertsimas and Van Parys [35] 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 [32], which safeguards against overfitting and lead to an improved out-of-sample performance. Both resulting prescriptive methods reduce to tractable convex optimization problems.

Kernel density estimation (KDE) [88] in combination with *principal component analysis* (PCA) is also used in the RO literature to construct the uncertainty set [217]. PCA captures the correlation between uncertain parameters and transfoms 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 [217] 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

$k$, $t_k$[23]. By incorporating forward and backward deviations to allow for asymmetry [74], Ning and You [217] propose the following polytopic uncertainty set that resembles the intersection of a box, with the so-called *budget*, and polyhedral uncertainty sets:

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

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 $\boldsymbol{S} = \frac{1}{N-1}(\boldsymbol{U} - \boldsymbol{1}\boldsymbol{\mu}_0^\top)^\top(\boldsymbol{U} - \boldsymbol{1}\boldsymbol{\mu}_0^\top)$. Moreover, $\boldsymbol{z}^-$ is a backward deviation, $\boldsymbol{z}^+$ is a forward deviation vector, and $\Gamma$ is the uncertainty budget. In addition, $F_k^{-1} := \min\{w_k | F_k(w_k) \ge \alpha\}$, $k = 1, \ldots, m$, 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 [217] 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 [282]. In SVC, to avoid overfitting, the violations of the data outside the sphere is penalized by a regularization term as follows:

$$\begin{array}{ll} \min\limits_{\delta, \boldsymbol{s}, \boldsymbol{c}} & \delta^2 + \dfrac{1}{N\gamma}\displaystyle\sum_{i=1}^N s_i \\ \text{s.t.} & \|\Phi(\boldsymbol{u}^i) - \boldsymbol{c}\|_2^2 \le \delta^2 + s_i,\ i = 1, \ldots, N, \\ & \boldsymbol{s} \ge \boldsymbol{0}. \end{array}$$

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. [282] propose to use a piecewise linear kernel, referred to as a *weighted generalized intersection kernel*, defined as follows:

$$K(\boldsymbol{u}, \boldsymbol{v}) = \sum_{k=1}^m l_k - \|\boldsymbol{Q}(\boldsymbol{u} - \boldsymbol{v})\|_1, \tag{5.30}$$

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 $l_k$, $k = 1, \ldots, m$, 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.

[23]It is known that for any positive definite symmetric kernel $K$, there is a mapping $\Phi$ from the covariates space to a higher-dimensional space $\mathbb{H}$ such that $K(\xi_k, t_k)$ is equal to the inner product between $\Phi(\xi_k)$ and $\Phi(t_k)$, see, e.g., Mohri et al. [207, Theorem 5.2]. Such a space $\mathbb{H}$ is called *reproducing kernel Hilbert space*. A kernel is said to be positive definite symmetric if the induced kernel matrix is symmetric positive semidefinite.

THEOREM 5.15. *(Shang et al. [282, Propositions 1, Propositions 3–4]) Suppose that the kernel function is constructed as in (5.30). Then,*

*(i) The kernal matrix induced by the kernel $K$ is positive definite.*

*(ii) The constructed uncertainty set*

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

*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 $\max_{\boldsymbol{u}\in\mathcal{U}} \boldsymbol{u}^\top \boldsymbol{x} \le b$ has the same complexity as the deterministic problem.*

*(iii) The regularization parameter $\gamma$ gives an upper bound on the fraction of the outliers; hence, a feasible solution $\boldsymbol{x}$ in the robust counterpart $\max_{\boldsymbol{u}\in\mathcal{U}} \boldsymbol{u}^\top \boldsymbol{x} \le b$ is also feasible to a SAA-based chance-constrained problem $P\{\tilde{\boldsymbol{u}}^\top \boldsymbol{x} \le b\} \ge 1 - \gamma$.*

*(iv) As the number of data points increases, the fraction of outliers converges to the regularization parameter $\gamma$ with probability one.*

*(v) The regularization parameter $\gamma$ gives a lower bound on the fraction of the support vectors.*

Shang and You [280] further propose to calibrate the radius of the uncertainty set and provide a probabilistic guarantee of the proposed uncertainty set. Shang and You [278] 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. [282] 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. Shang and You [279] adopt the ambiguity set proposed in Wiesemann et al. [318], and propose to use 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.

Applications of the proposed method in Ning and You [217] are studied in production scheduling [217] and in process network planning [217, 218, 216]. The proposed method in Shang et al. [282] is used in different application domains to construct the uncertainty set, see, e.g., control of irrigation system [283] and chemical process network planning [282]. Applications of the proposed method in Shang and You [279] are studied in production scheduling [279, 281] and in process network planning [279, 277].

**5.5. General Ambiguity Sets.** In Sections 5.1–5.4, we reviewed papers with specific distributional and structural properties for the random parameters, captured via discrepancy-based, moment-based, shape-preserving, and kernel-based ambiguity sets. In this section, we review papers that either do not consider any specific form for the ambiguity set or provide some general results for a broad class of ambiguty sets.

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. [77]. It is shown that the ambiguity sets formed via generalized moments, mixture distribution, Wasserstein metric, $\phi$-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. [318]. 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. [77], 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. [318], and a broad class of shape-preserving ambiguity sets is proposed in Hanasusanto et al. [134].

Luo and Mehrotra [199] study DRO problem where the ambiguity sets of probability distributions can depend on the decision variables. They consider a wide range of moment- and discrepancy-based ambiguity sets formed, such as (1) measure and moment inequalities (see Section 5.2.3), (2) bounds on moment constraints (see Section 5.2.1), (3) 1-Wasserstein metric utilizing $\ell_1$-norm, (4) $\phi$-divergences, and (5) Kolmogorov-Smirnov test. They present equivalent reformulations for these problems by relying on duality results.

Pflug and Wozabal [229] study a DRO problem, where the ambiguity exists in both the objective function and constraints as in (DRO). To solve the model, they propose an exchange method to successively generate a finite inner approximation of the ambiguity set of distributions. They show that when the ambiguity set is compact and convex, and the risk measure is jointly continuous in both $\boldsymbol{x}$ and $\mathbb{P}$, then the proposed algorithm is finitely convergent.

Bansal and Zhang [11] introduce two-stage stochastic integer programs in which the second-stage problem have $p$-order conic constraints as well as integer variables. They present sufficient conditions under which the addition of parametric (non)linear cutting planes along with the linear relaxation of the integrality constraints provides a convex programming equivalent for the second-stage problem. They show that this result is also valid for the distributionally robust counterpart of this problem. This paper generalizes the results on two-stage mixed-binary linear programs studied in Bansal et al. [9].

Bansal and Mehrotra [10] introduce two-stage distributionally robust disjunctive programs with disjunctive constraints in both stages and a general ambiguity set for the probability distributions. To solve the resulting model, they develop decomposition algorithms, which utilize Balas' linear programming equivalent for deterministic disjunctive programs or his sequential convexification approach within the L-shaped method. They demonstrate that the proposed algorithms are finitely convergent if a distribution separation subproblem can be solved in a finite number of iterations, as in sets formed via $\mathcal{P}^{\mathrm{MM}}$, defined in (5.23), 1-Wasserstein metric utilizing an arbitrary norm, and the total variation distance. These algorithms generalize the distribution-

ally robust integer L-shaped algorithm of Bansal et al. [9] for two-stage mixed binary linear programs.

Wang et al. [315] 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 $\tilde{\boldsymbol{\xi}}$, 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 substrcture 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. Wang et al. [315] apply their results to an operating room scheduling problem.

Guo et al. [129] 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 distrubutionally 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 measure, including Kolmogorov and the total variation distance. They apply their convergence results to the ambiguity sets formed via (5.23) and Kullback-Leibler divergence.

Delage and Saif [81] study the value of using a randomized policy, as compared to a deterministic policy, for mixed-integer DRO problems. They show that the value of randomization for such DRO models with a convex cost function $h$ 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. They also develop a column generation algorithm for solving a two-stage mixed-integer linear DRO problem, formed via (5.23) and 1-Wasserstein metric utilizing an arbitrary norm. They test their results on assignment problem, and on uncapacitated and capacitated facility location problems.

Long and Qi [193] study a distributionally robust binary stochastic program to minimize the entropic VaR, also known as Bernstein approximation for the chance constraint. They propose an approximation algorithm to solve the problem via solving a sequence of problems. They showcase their results for ambiguity set formed as in (5.23) for a stochastic shortest path problem.

Shapiro et al. [294] 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 stochastic dual dynamic programming (SDDP) method to solve this problem.

## 6. Calibration of the Ambiguity Set of Probability Distributions.

**6.1. Choice of the Nominal Parameters.** All discrepancy-based ambiguity sets, studied in Section 5.1, and some of the moment-based ambiguity sets, studied in Section 5.2, 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 (5.3), and parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ in the ambiguity set $\mathcal{P}^{\mathrm{DY}}$, defined in (5.22). 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. In the case that the true unknown distribution is continuous, Jiang and Guan [167] and Zhao and Guan [342] propose to obtain $P_0$ with nonparametric kernel density estimation methods, see, e.g., Devroye and Gyorfi [88].

Delage and Ye [82] propose to estimate $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ by their empirical estimates (see Section 6.2 for more details on how this choice of nominal parameters, in conjuction with other assumptions, ensure that the constructed ambiguity set $\mathcal{P}^{\mathrm{DY}}$ contains the true unknown probability distribution with a high probability).

**6.2. Choice of Robustness Parameters.** In Section 5, we reviewed different approaches to form the ambiguity set of distributions. All discrepancy-based ambiguity sets, studied in Section 5.1, and some of the moment-based ambiguity sets, studied in Section 5.2, rely on parameters that control the size of the ambiguity set. For instance, parameter $\epsilon$ in the ambiguity set $\mathcal{P}^{\mathrm{W}}(P_0;\epsilon)$, defined in (5.3), and parameters $\varrho_1$ and $\varrho_2$ in the ambiguity set $\mathcal{P}^{\mathrm{DY}}$, defined in (5.22), control the size of their corresponding ambiguity sets. A judicial 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 DRO and non-data-driven DRO.

**6.2.1. Data-Driven DROs.** 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., [229, 82, 42].

An 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. [28], Shafieezadeh-Abadeh et al. [273], or with a finite sample, see, e.g., Pflug and Wozabal [229]. 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 [122] and Delage and Ye [82]. In a recent work, Gotoh et al. [125] propose to choose

[24]Some probability metrics, such as Wasserstein metric, metrize the weak convergence [116]. That is, the convergence between two probability distributions, with respect to some metric, implies the convergence in probability.

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 the data-driven approaches to choose the level of robustness in more details. In this section, we suppose that a set $\{\boldsymbol{\xi}^i\}_{i=1}^N$ of i.i.d data, distributed according to $\mathbb{P}^{\text{true}}$, is available, where $\mathbb{P}_N$ denotes the empirical probability distribution of data.

**6.2.1.1. 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 [229] and Pflug et al. [233] 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 [94], as $\epsilon = \frac{CN^{-\frac{1}{d}}}{\alpha}$. This choice of $\epsilon$ guarantees that $\mathbb{P}\{\mathfrak{d}_c^{\mathrm{W}}(\mathbb{P}, \mathbb{P}_N) \geq \epsilon\} \leq \alpha$. In addition to the confidence level $1-\alpha$ and the number of available data points $N$, the proposed level of robustness in [229, 233] depends on the dimension of $\tilde{\boldsymbol{\xi}}$, $d$, 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 [105]. Nevertheless, because all the utilized probabilistic statements rely on the exogenous constant $C$, the size of the 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 [164] derive a closed-form expression for computing the size of the Wasserstein-based ambiguity set.

THEOREM 6.1. *(Ji and Lejeune [164, Theorem 2]) Suppose that the random vector $\tilde{\boldsymbol{\xi}}$ is supported on a finite Polish space $(\Omega, d)$, where $\Omega \subseteq \mathbb{R}^d$ and $d$ is the $\ell_1$-norm. Choose $c(\cdot,\cdot) = d(\cdot,\cdot)$ in the definition of the optimal transport discrepancy (5.2). Assume that*

$$\log \int_\Omega e^{\lambda d(\boldsymbol{\xi},\boldsymbol{\xi}_0)} \mathbb{P}^{true}(d\boldsymbol{\xi}) < \infty, \quad \forall \lambda > 0,$$

*for some $\boldsymbol{\xi}_0$. Let $\theta := \sup\{d(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) : \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \Omega\}$ be the diameter of $\Omega$. Then,*

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

*Moreover, if*

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

*then*

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

Unlike the result in Pflug and Wozabal [229], the proposed level of robustness in Ji and Lejeune [164], stated in Theorem 6.1, depends only on the confidence level $\alpha$, the number of available data points, and the diameter of the compact support $\Omega$. Ji and Lejeune [164] obtain this result by bounding the Wasserstein distance between two probability distributions from above, using the properties of the weighted total variation [54], and the weighted Csiszar-Kullback-Pinsker inequality [312], and

consequently applying Sanov's large deviation theorem [83] to reach a probabilistic statement on the Wasserstein distance between two distributions. As stated in Theorem 6.1, 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 [229] and Pflug et al. [233] 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. [50] 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 $\alpha$, 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 $\epsilon$, 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. [50] is extended in Blanchet and Kang [48, 46]. Blanchet and Kang [48] 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 [70] give guidance on the selection of the regularization parameter from the standpoint of a confidence region.

**6.2.1.2. Goodness-of-Fit Test.** Bertsimas et al. [42] 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 $\alpha$, data, and the null hypothesis.

**6.2.1.3. $\phi$-Divergences.** By noting that the class of $\phi$-divergences can be used in statistical hypothesis tests, a similar approach to the one in Bertsimas et al. [42] can be used to choose the level of robustness for $\phi$-divergence-based ambiguity sets. For the case that the distributional ambiguity in discrete distributions is modeled via $\phi$-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., Ben-Tal et al. [28], Bayraksan and Love [13], Yanıkoğlu and den Hertog [337].

Suppose that $\Xi$ is finite sample space of size $m$ and the $\phi$-divergence function in (5.9) is twice continuously differentiable in a neighborhood of 1, with $\phi''(1) > 0$. Then, it is shown in Pardo [226] that under the true distribution, the statistics $\frac{2N}{\phi''(1)}\mathcal{D}_\phi(\mathbb{P}^{\text{true}}, \mathbb{P}_0)$ converges in distribution to a $\chi^2_{m-1}$-distribution, with $m-1$ degrees of freedom. Thus, at a given confidence level $\alpha$, 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 $\chi^2_{m-1}$, to obtain an (approximate) confidence set on the true unknown distribution. Ben-Tal et al. [28] 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 $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{\text{true}}}\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]$. For corrections for small sample sizes, we refer readers to Pardo [226].

By generalizing the empirical likelihood framework [224] on a separable metric space (not necessarily finite), Duchi et al. [92] propose to choose the level of robustness $\epsilon$ such that a confidence interval $[l_N, u_N]$ on the true optimal value of $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{\text{true}}}\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]$ has an asymptotically exact coverage $1-\alpha$, i.e., $\lim_{N\to\infty} \mathbb{P}_N\{\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{\text{true}}}\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \in [l_N, u_N]\} = 1-\alpha$, where

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

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

and

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

THEOREM 6.2. *(Duchi et al. [92, Theorem 4]) Suppose that the $\phi$ 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 $\mathcal{X}$, $\mathbb{E}_{\mathbb{P}^{true}}\left[M(\tilde{\boldsymbol{\xi}})^2\right]<\infty$, and $\mathbb{E}_{\mathbb{P}^{true}}\left[|h(\boldsymbol{x}_0,\tilde{\boldsymbol{\xi}})|\right]<\infty$ 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 $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{true}}\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]$ has a unique solution, then*

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

*and*

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

According to Theorem 6.2, if $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^{\text{true}}}\left[h(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]$ has a unique solution, the desired asymptotic guarantee is achieved with the choice $\epsilon=\frac{\chi^2_{1,1-\alpha}}{N}$. Duchi et al. [92] 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 1 that the $\phi$-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 5.4, i.e., the relationship between the variation distance and the Hellinger distance, Jiang and Guan [167] 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 $\tilde{\boldsymbol{\xi}}$

[25] As in the definition of $\phi$-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 $\phi$ [226]

follows a continuous distribution, the proposed level of robustness in [167] depends on some constants that appear in the probabilistic statement of the discrepancy between the empirical distributions and the true distribution.

**6.2.1.4. $\ell_p$-Norm.** For the case that $\ell_\infty$-norm is used to model the distributional ambiguity, Jiang and Guan [167] propose to choose the level of robustness based on a probabilistic statement on the discrepancy between the empirical distributions and the true distribution as $\epsilon = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{N}} \max_{i=1}^{m} \sqrt{p_0^i(1-p_0^i)}$, where $z_{1-\frac{\alpha}{2}}$ represents the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution, and $\boldsymbol{p}_0 := [p_0^1, \ldots, p_0^m]$ denotes the empirical distribution of data. 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 $\ell_1$-norm (i.e., the variation distance) case, when $\tilde{\boldsymbol{\xi}}$ 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.

**6.2.1.5. $\zeta$-Structure.** By exploiting the relationship between different metrics in the $\zeta$-structure family, see, e.g., Lemma 5.7, Zhao and Guan [342] 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 6.3. *Suppose that the random vector $\tilde{\boldsymbol{\xi}}$ is supported on a bounded finite space $\Omega$ and $\theta$ denotes the diameter of $\Omega$, as defined in Theorem 6.1.*

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

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

*Proof.* The proof is immediate from the relationship between $\zeta$-structure metrics, stated in Lemma 5.7, 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 [342, Proposition 3]. □

As it can be seen from Theorem 6.3, the proposed levels of robustness for the case that the unknown distribution is discrete depend on the diameter of $\Omega$, the number of data points $N$, and the confidence level $1-\alpha$. However, the results in Zhao and Guan [342] for the continuous case suffer from similar practical issues as in [229, 233, 167].

**6.2.1.6. Chebyshev.** A data-driven approach to construct a Chebyshev ambiguity set is proposed in Goldfarb and Iyengar [122]. Recall the linear model for the asset returns $\tilde{\boldsymbol{\xi}}$ in Goldfarb and Iyengar [122]: $\tilde{\boldsymbol{\xi}} = \boldsymbol{\mu} + \boldsymbol{A}\tilde{\boldsymbol{f}} + \tilde{\boldsymbol{\epsilon}}$, where $\boldsymbol{\mu}$ is the vector of mean returns, $\tilde{\boldsymbol{f}} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ is the vector of random returns that derives the market, $\boldsymbol{A}$ is the factor loading matrix, and $\tilde{\boldsymbol{\epsilon}} \sim N(\boldsymbol{0}, \boldsymbol{B})$ is the vector of residual returns with a diagonal matrix $\boldsymbol{D}$. Under the assumption that the covariance matrix $\boldsymbol{\Sigma}$ is known, recall that Goldfarb and Iyengar [122] study three different models to form the uncertainty in $\boldsymbol{B}$, $\boldsymbol{A}$, and $\boldsymbol{\mu}$ as follows:

$$\begin{aligned}
\mathcal{U}_{\boldsymbol{B}} &= \big\{\boldsymbol{B} \,\big|\, \boldsymbol{B} = \mathrm{diag}(\boldsymbol{b}),\ b_i \in [\underline{b}_i, \bar{b}_i],\ i = 1, \ldots, d\big\},\\
\mathcal{U}_{\boldsymbol{A}} &= \{\boldsymbol{A} \,|\, \boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{C},\ \|\boldsymbol{c}_i\|_g \leq \rho_i,\ i = 1, \ldots, d\},\\
\mathcal{U}_{\boldsymbol{\mu}} &= \{\boldsymbol{\mu} \,|\, \boldsymbol{\mu} = \boldsymbol{\mu}_0 + \boldsymbol{\zeta},\ |\zeta_i| \leq \gamma_i,\ i = 1, \ldots, d\},
\end{aligned}$$

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 $\underline{d}_i$, $\overline{d}_i$, $\rho_i$, $\gamma_i$, $i = 1, \ldots, d$, vector $\boldsymbol{\mu}_0$, and matrices $\boldsymbol{A}_0$ and $\boldsymbol{G}$. Assuming that a set of data points is available on $\tilde{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{f}}$, by relying on the multivariate linear regression, Goldfarb and Iyengar [122] obtain least square estimates $(\boldsymbol{\mu}_0, \boldsymbol{A}_0)$ of $(\boldsymbol{\mu}, \boldsymbol{A})$, respectively, and construct a multdimensional confidence region of $(\boldsymbol{\mu}, \boldsymbol{A})$ around $(\boldsymbol{\mu}_0, \boldsymbol{A}_0)$. Now, projecting this confidence region along vector $\boldsymbol{A}$ and matrix $\boldsymbol{\mu}$ gives the corresponding uncertainty sets $\mathcal{U}_{\boldsymbol{A}}$ and $\mathcal{U}_{\boldsymbol{\mu}}$, 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.

**6.2.1.7. Delage and Ye.** Data-driven methods to construct the ambiguity set $\mathcal{P}^{\mathrm{DY}}$ is proposed in Delage and Ye [82].

THEOREM 6.4. *(Delage and Ye [82, Corollary 4]) Suppose that the random vector $\tilde{\boldsymbol{\xi}}$ is supported on a bounded space $\Omega$. Consider the following parameters:*

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_0 &= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\xi}^i, \\
\hat{\boldsymbol{\Sigma}}_0 &= \frac{1}{N-1}\sum_{i=1}^{N} (\boldsymbol{\xi}^i - \hat{\boldsymbol{\mu}}_0)(\boldsymbol{\xi}^i - \hat{\boldsymbol{\mu}}_0)^\top, \\
\hat{\theta} &= \sup_{i=1}^{N} \|\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}(\boldsymbol{\xi}^i - \hat{\boldsymbol{\mu}}_0)\|_2,
\end{aligned}$$

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

$$\begin{aligned}
\bar{\theta} &= \Big(1 - (\hat{\theta}^2 + 2)\frac{2 + \sqrt{2\log(\frac{4}{\bar{\alpha}})}}{\sqrt{N}}\Big)^{-\frac{1}{2}}\hat{\theta}, \\
\bar{\gamma}_1 &= \frac{\bar{\theta}^2}{\sqrt{N}}\Big(\sqrt{1 - \frac{d}{\bar{\theta}^4}} + \sqrt{\log(\frac{4}{\bar{\alpha}})}\Big) \\
\bar{\gamma}_2 &= \frac{\bar{\theta}^2}{N}\Big(2 + \sqrt{2\log(\frac{2}{\bar{\alpha}})}\Big), \\
\bar{\varrho}_1 &= \frac{\bar{\gamma}_2}{1 - \bar{\gamma}_1 - \bar{\gamma}_2}, \\
\bar{\varrho}_2 &= \frac{1 + \bar{\gamma}_2}{1 - \bar{\gamma}_1 - \bar{\gamma}_2},
\end{aligned}$$

*where $\bar{\alpha} = 1 - \sqrt{1-\alpha}$. Let $\mathcal{P}^{DY}(\Omega, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0, \bar{\varrho}_1, \bar{\varrho}_2)$ be the ambiguity set formed via (5.22), using parameters $\hat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{\Sigma}}_0$, $\bar{\varrho}_1$, and $\bar{\varrho}_2$. Then, we have*

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

**6.2.2. Non-Data-Driven DROs.** As mentioned before, data-driven DROs 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. 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. [252] 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 pessimism 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. [252] 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. [252] 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. 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 (1.1) and (1.2), 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 [23] and [259] 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.

**7.1. Regret Function.** 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:

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

In other words, $\mathcal{V}_P\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]$ calculates the expected additional loss that could have been avoided. This definition of regret function is used in Natarajan et al. [212] and Hu et al. [151] in the context of combinatorial optimization and multicriteria decision-making, respectively. Another way for formulating a regret function may be as

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

This type of regret function is used in Perakis and Roels [228] in the context of the newsvendor problem. Perakis and Roels [228] obtain closed form solutions to distributionally robust single-item newsvendor problems that minimize the worst-case expected regret of acting optimally, 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 $\mathcal{P}^{\mathrm{MM}}$, defined in (5.23). Perakis and Roels [228] 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.

**7.2. Risk Measure.** As introduced in Section 3.3.1, a functional that quantifies the uncertainty in the outcomes of a decision is a risk measure Artzner et al. [7], Acerbi [1], Kusuoka [173], Shapiro [287]. A risk measure $\rho_P$ 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. [233], Pichler [235], Wozabal [320], Pichler and Xu [236] for spectral and distortion risk measures, Calafiore [59] for variance, Calafiore [59] for mean absolute-deviation, Hanasusanto et al. [135], Wiesemann et al. [318] for optimized certainty equivalent, Hanasusanto et al. [133] for CVaR, and Postek et al. [240] for a variety of risk measures.

**7.3. Utility Function.** An alternative to using risk measures to compare random variables is to evaluate their expected utility Gilboa and Schmeidler [117]. 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 $\mathbb{E}_P[u(Z_1)] \geq \mathbb{E}_P[u(Z_2)]$ for a given univariate utility function $u$[26]. 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 $\zeta$, 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 $\zeta$. The notion of *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 $\zeta$ for a given family of cdf. Consequently, to address the problem of ambiguity in decision maker's utility or equivalently, cdf of the random variable $\zeta$, one can study

$$\min_{\boldsymbol{x} \in \mathcal{X}} \max_{\zeta \in \mathcal{U}} P\{h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \geq \zeta\}, \tag{7.1}$$

and

$$\min_{\boldsymbol{x} \in \mathcal{X}} \max_{\zeta \in \mathcal{U}} \left\{ h(\boldsymbol{x}) \,\middle|\, \max_{\zeta \in \mathcal{U}} P\{\boldsymbol{g}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \geq \zeta\} \leq \mathbf{0} \right\}, \tag{7.2}$$

where $\mathcal{U}$ denotes a given family of normalized and nondecreasing utility functions, or equivalently, a given family of cdf. Note that problems (7.1) and (7.2) have the form of problems (1.5) and (1.6), respectively. Hu and Mehrotra [150] study problem of the form (7.1), where $\mathcal{U}$ is further restricted to include concave utility functions

[26] For definitions in a multivariate case, we refer to Hu et al. [152, 153].

or equivalently, cdf, and satisfy functional bounds on the utility and marginal utility functions (cdf and pdf of $\zeta$) as in (5.23). 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 et al. [154], in Hu et al. [155], no shape restrictions on the utility function is assumed and only functional bounds on the utility function are enforced. Hu et al. [155] show that an SAA approach to the Lagrangian dual of the resulting problem can be used while solving a mixed-integer LP. They study the convergence properties of this SAA problem, and illustrate their results using examples in portfolio optimization and a streaming bandwidth allocation problem. Bertsimas et al. [39] study a DRO model of the form (1.5), where a convex nondecreasing disutility function is used to quantify the uncertainty in decision. A utility function is closely related to risk measures [150]. 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 [23], 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. [318] study a DRO model formed via (5.26), 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.

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 [5]. 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 [87] for unrestricted sets, Homem-de-Mello and Mehrotra [148], Hu et al. [151], Hu and Mehrotra [149] for polyhedral sets, and Hu et al. [152] for convex sets. For instance, Hu et al. [151] 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 minmax 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. [151] 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 [149] and Hu et al. [153]. For instance, Hu and Mehrotra [149] 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, $\ell_1$-norm, and $\ell_\infty$-norm, and (2) a conic-representable set, including those induced by a single cone (e.g., $\ell_p$-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 (5.22), they obtain a reformulation of the problem using the result in Delage and Ye [82]. 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. They illustrate the ideas set forth in the paper using examples from disaster planning and agriculture revenue management problems.

**8. Modeling Toolboxes.** Goh and Sim [121] develop a MATLAB-based algebraic modeling toolbox, named ROME, 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 [120]. Goh and Sim [121] elucidate the practicability of this toolbox in the context of (1) a service-constrained inventory management problem, (2) a project-crashing problem, and (3) a portfolio optimization problem. A C++-based algebraic modeling package, named ROC, is developed in Bertsimas et al. [44], 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. [44]. 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. [41]. XProg (http://xprog.weebly.com), is a MATLAB-based algebraic modeling package that also implements the proposed model in Bertsimas et al. [44]. Chen et al. [77] develop an algebraic modeling package, AROMA, to illustrate the modeling power of their proposed ambiguity set.

**References.**


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

[2] S. D. Ahipasaoğlu, K. Natarajan, and D. Shi, *Distributionally robust project crashing with partial or no correlation information*, Networks, 74 (2019), pp. 79–106.

[3] B. Analui and G. C. Pflug, *On distributionally robust multiperiod stochastic optimization*, Comput. Management Sci., 11 (2014), pp. 197–220.

[4] A. Ardestani-Jaafari and E. Delage, *Robust optimization of sums of piecewise linear functions with application to inventory problems*, Oper. Res., 64 (2016), pp. 474–494.

[5] B. Armbruster and J. Luedtke, *Models and formulations for multivariate dominance-constrained stochastic programs*, IIE Trans., 47 (2015), pp. 1–14.

[6] S. Arpón, T. Homem-de-Mello, and B. Pagnoncelli, *Scenario reduction for stochastic programs with Conditional Value-at-Risk*, Math. Program., 170 (2018), pp. 327–356.

[7] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent measures of risk*, Math. Financ., 9 (1999), pp. 203–228.

[8] G.-Y. Ban and C. Rudin, *The big data newsvendor: Practical insights from machine learning*, Oper. Res., 67 (2019), pp. 90–108.

[9] M. Bansal, K. Huang, and S. Mehrotra, *Decomposition algorithms for two-stage distributionally robust mixed binary programs*, SIAM J. Optim., 28 (2018), pp. 2360–2383.

[10] M. Bansal and S. Mehrotra, *On solving two-stage distributionally robust disjunctive programs with a general ambiguity set*, Eur. J. Oper. Res., 279 (2019), pp. 296–307.

[11] M. Bansal and Y. Zhang, *Two-stage stochastic (and distributionally robust) p-order conic mixed integer programs: Tight second stage formulations*, 2018. Optimization Online http://www.optimization-online.org/DB_HTML/2018/05/6630.html.

[12] P. L. Bartlett and S. Mendelson, *Rademacher and gaussian complexities: Risk bounds and structural results*, J Mach Learn Res, 3 (2002), pp. 463–482.

[13] G. Bayraksan and D. K. Love. *Data-driven stochastic programming using phi-divergences.* in The Operations Research Revolution, pp. 1–19, INFORMS TutORials in Operations Research, 2015.

[14] G. Bayraksan and D. P. Morton, *Assessing solution quality in stochastic programs*, Math. Program., 108 (2006), pp. 495–514.

[15] G. Bayraksan and D. P. Morton. *Assessing solution quality in stochastic programs via sampling.* in Decision Technologies and Applications, pp. 102–122, INFORMS TutORials in Operations Research, 2009.

[16] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, 3rd ed., 2006.

[17] T. Bazier-Mattea and E. Delage, *Generalization bounds for regularized portfolio selection with market side information*, 2018. Optimization Online http://www.optimization-online.org/DB_HTML/2018/02/6476.html.

[18] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization*, MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, 2001.

[19] A. Ben-Tal and A. Nemirovski, *Robust convex optimization*, Math Oper Res, 23 (1998), pp. 769–805.

[20] A. Ben-Tal and A. Nemirovski, *Robust solutions of linear programming problems contaminated with uncertain data*, Math. Program., 88 (2000), pp. 411–424.

[21] A. Ben-Tal and A. Nemirovski, *On safe tractable approximations of chance-constrained linear matrix inequalities*, Math Oper Res, 34 (2009), pp. 1–25.

[22] A. Ben-Tal and M. Teboulle, *Expected utility, penalty functions, and duality in stochastic nonlinear programming*, Management Sci., 32 (1986), pp. 1445–1466.

[23] A. Ben-Tal and M. Teboulle, *An old-new concept of convex risk measures: The optimized certainty equivalent*, Math. Financ., 17 (2007), pp. 449–476.

[24] A. Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski, *Adjustable robust solutions of uncertain linear programs*, Math. Program., 99 (2004), pp. 351–376.

[25] A. Ben-Tal, S. Boyd, and A. Nemirovski, *Extending scope of robust optimization: Comprehensive robust counterparts of uncertain problems*, Math. Program., 107 (2006), pp. 63–89.

[26] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust optimization*, vol. 28, Princeton University Press, 2009.

[27] A. Ben-Tal, D. Bertsimas, and D. B. Brown, *A soft robust model for optimization under ambiguity*, Oper. Res., 58 (2010), pp. 1220–1234.

[28] A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen, *Robust solutions of optimization problems affected by uncertain probabilities*, Management Sci., 59 (2013), pp. 341–357.
[29] A. Ben-Tal, D. Den Hertog, and J.-P. Vial, *Deriving robust counterparts of nonlinear uncertain inequalities*, Math. Program., 149 (2015), pp. 265–299.
[30] D. P. Bertsekas, *Nonlinear Programming*, Athena scientific, 3rd ed., 2016.
[31] D. P. Bertsekas, *Dynamic programming and optimal control*, vol. 1, Athena scientific Belmont, MA, 4th ed., 2017.
[32] D. Bertsimas and N. Kallus, *From predictive to prescriptive analytics*, 2018. `arXiv preprint arXiv:1402.5481 [stat.ML]`.
[33] D. Bertsimas and I. Popescu, *Optimal inequalities in probability theory: A convex optimization approach*, SIAM J. Optim., 15 (2005), pp. 780–804.
[34] D. Bertsimas and M. Sim, *The price of robustness*, Oper. Res., 52 (2004), pp. 35–53.
[35] D. Bertsimas and B. Van Parys, *Bootstrap robust prescriptive analytics*, 2017. `arXiv preprint arXiv:1711.09974 [math.OC]`.
[36] D. Bertsimas, K. Natarajan, and C.-P. Teo, *Probabilistic combinatorial optimization: Moments, semidefinite programming, and asymptotic bounds*, SIAM J. Optim., 15 (2004), pp. 185–209.
[37] D. Bertsimas, D. Pachamanova, and M. Sim, *Robust linear optimization under general norms*, Oper. Res. Lett., 32 (2004), pp. 510–516.
[38] D. Bertsimas, K. Natarajan, and C.-P. Teo, *Persistence in discrete optimization under data uncertainty*, Math. Program., 108 (2006), pp. 251–274.
[39] D. Bertsimas, X. V. Doan, K. Natarajan, and C.-P. Teo, *Models for minimax stochastic linear optimization problems with risk aversion*, Math Oper Res, 35 (2010), pp. 580–602.
[40] D. Bertsimas, D. B. Brown, and C. Caramanis, *Theory and applications of robust optimization*, Siam Rev, 53 (2011), pp. 464–501.
[41] D. Bertsimas, M. Sim, and M. Zhang, *A practicable framework for distributionally robust linear optimization*, 2014. Optimization Online www.optimization-online.org/DB_FILE/2013/07/3954.html.
[42] D. Bertsimas, V. Gupta, and N. Kallus, *Data-driven robust optimization*, Math. Program., 167 (2018), pp. 235–292.
[43] D. Bertsimas, V. Gupta, and N. Kallus, *Robust sample average approximation*, Math. Program., 171 (2018), pp. 217–282.
[44] D. Bertsimas, M. Sim, and M. Zhang, *Adaptive distributionally robust optimization*, Management Sci., 65 (2018), pp. 604–618.
[45] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer, New York, 2nd ed., 2011.
[46] J. Blanchet and Y. Kang, *Sample out-of-sample inference based on Wasserstein distance*, 2016. `arXiv preprint arXiv:1605.01340 [math.ST]`.
[47] J. Blanchet and Y. Kang, *Semi-supervised learning based on distributionally robust optimization*, 2017. `arXiv preprint arXiv:1702.08848 [stat.ML]`.
[48] J. Blanchet and Y. Kang, *Distributionally robust groupwise regularization estimator*, 2017. `arXiv preprint arXiv:1705.04241 [math.ST]`.
[49] J. Blanchet and K. R. Murthy, *Quantifying distributional model risk via optimal transport*, 2017. `arXiv preprint arXiv:1604.01446 [math.PR]`.
[50] J. Blanchet, Y. Kang, and K. Murthy, *Robust Wasserstein profile inference and applications to machine learning*, 2016. `arXiv:1610.05627 [math.ST]`.
[51] J. Blanchet, Y. Kang, F. Zhang, F. He, and Z. Hu, *Doubly robust data-driven distributionally robust optimization*, 2017. `arXiv preprint arXiv:1705.07168 [stat.ML]`.
[52] J. Blanchet, Y. Kang, F. Zhang, and K. Murthy, *Data-driven optimal transport cost selection for distributionally robust optimization*, 2017. `arXiv preprint arXiv:1705.07152 [stat.ML]`.
[53] J. Blanchet, K. Murthy, and F. Zhang, *Optimal transport based distributionally robust optimization: Structural properties and iterative schemes*, 2018. `arXiv preprint arXiv:1810.02403 [math.OC]`.
[54] F. Bolley and C. Villani. *Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities.* in Annales de la faculté des sciences de Toulouse: Mathématiques, vol. 14, Université Paul Sabatier, Université Paul Sabatier, 2005.
[55] J. F. Bonnans and A. Shapiro, *Perturbation analysis of optimization problems*, Springer Science & Business Media, 2013.

[56] S. Bose and A. Daripa, *A dynamic mechanism and surplus extraction under ambiguity*, J Econ Theory, 144 (2009), pp. 2084–2114.
[57] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.
[58] M. Breton and S. El Hachem, *Algorithms for the solution of stochastic dynamic minimax problems*, Comput. Optim. Appl., 4 (1995), pp. 317–345.
[59] G. Calafiore, *Ambiguous risk measures and optimal robust portfolios*, SIAM J. Optim., 18 (2007), pp. 853–877.
[60] G. Calafiore and M. C. Campi, *Uncertain convex programs: randomized solutions and confidence levels*, Math. Program., 102 (2005), pp. 25–46.
[61] G. C. Calafiore and L. El Ghaoui, *On distributionally robust chance-constrained linear programs*, J Optimiz Theory App, 130 (2006), pp. 1–22.
[62] M. C. Campi and S. Garatti, *The exact feasibility of randomized solutions of uncertain convex programs*, SIAM J. Optim., 19 (2008), pp. 1211–1230.
[63] M. Campi and G. Calafiore, *Decision making in an uncertain environment: the scenario-based optimization approach*, Multiple Participant Decision Making, (2004), pp. 99–111.
[64] A. Charnes, W. Cooper, and K. Kortanek, *Duality, haar programs, and finite sequence spaces*, Proceedings of the National Academy of Sciences, 48 (1962), pp. 783–786.
[65] A. Charnes, W. Cooper, and K. Kortanek, *Duality in semi-infinite programs and some works of haar and carathéodory*, Management Sci., 9 (1963), pp. 209–228.
[66] A. Charnes, W. Cooper, and K. 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 (1969), pp. 41–52.
[67] A. Charnes and W. W. Cooper, *Chance-constrained programming*, Management Sci., 6 (1959), pp. 73–79.
[68] A. Charnes, W. W. Cooper, and G. H. Symonds, *Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil*, Management Sci., 4 (1958), pp. 235–263.
[69] L. Chen, W. Ma, K. Natarajan, D. Simchi-Levi, and Z. Yan, *Distributionally robust linear and discrete optimization with marginals*, 2018. Optimization Online http://www.optimization-online.org/DB_HTML/2018/04/6570.html.
[70] R. Chen and I. C. Paschalidis, *A robust learning approach for regression models based on distributionally robust optimization*, J Mach Learn Res, 19 (2018), pp. 1–48.
[71] W. Chen and M. Sim, *Goal-driven optimization*, Oper. Res., 57 (2009), pp. 342–357.
[72] W. Chen, M. Sim, J. Sun, and C.-P. Teo, *From cvar to uncertainty set: Implications in joint chance-constrained optimization*, Oper. Res., 58 (2010), pp. 470–485.
[73] X. Chen, H. Sun, and H. Xu, *Discrete approximation of two-stage stochastic and distributionally robust linear complementarity problems*, Math. Program., (2018), https://doi.org/10.1007/s10107-018-1266-4.
[74] X. Chen, M. Sim, and P. Sun, *A robust optimization perspective on stochastic programming*, Oper. Res., 55 (2007), pp. 1058–1071.
[75] X. Chen, M. Sim, P. Sun, and J. Zhang, *A linear decision-based approximation approach to stochastic programming*, Oper. Res., 56 (2008), pp. 344–357.
[76] Z. Chen, D. Kuhn, and W. Wiesemann, *Data-driven chance constrained programs over Wasserstein balls*, 2018. `arXiv preprint arXiv:1809.00210 [math.OC]`.
[77] Z. Chen, M. Sim, and P. Xiong, *Adaptive robust optimization with scenario-wise ambiguity sets*, 2018. Optimization Online http://www.optimization-online.org/DB_HTML/2017/06/6055.html.
[78] Z. Chen, M. Sim, and H. Xu, *Distributionally robust optimization with infinitely constrained ambiguity sets*, Operations Research, (2019), https://doi.org/10.1287/opre.2018.1799.
[79] J. Cheng, R. Li-Yang Chen, H. N. Najm, A. Pinar, C. Safta, and J.-P. Watson, *Distributionally robust optimization with principal component analysis*, SIAM J. Optim., 28 (2018), pp. 1817–1841.
[80] E. Delage, *Distributionally robust optimization in context of data-driven problems*, Ph.D. dissertation, Stanford University, Stanford, California, 2009.
[81] E. Delage and A. Saif, *The value of randomized solutions in mixed-integer distributionally robust optimization problems*, 2018. Optimization Online http://www.optimization-online.org/DB_HTML/2018/06/6668.html.

[82] E. Delage and Y. Ye, *Distributionally robust optimization under moment uncertainty with application to data-driven problems*, Oper. Res., 58 (2010), pp. 595–612.
[83] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, vol. 38 of Stochastic Modelling and Applied Probability, Springer, 1998.
[84] V. DeMiguel and F. J. Nogales, *Portfolio selection with robust estimation*, Oper. Res., 57 (2009), pp. 560–577.
[85] Y. Deng and S. 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.
[86] D. Dentcheva, *Optimization models with probabilistic constraints*, in Probabilistic and Randomized Methods for Design under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer, London, 2006, pp. 49–97.
[87] D. Dentcheva and A. Ruszczyński, *Optimization with multivariate stochastic dominance constraints*, Math. Program., 117 (2009), pp. 111–127.
[88] L. Devroye and L. Gyorfi, *Nonparametric density estimation: The L1 View*, John Wiley, 1985.
[89] A. Dhara, B. Das, and K. Natarajan, *Worst-case expected shortfall with univariate and bivariate marginals*, 2017. `arXiv preprint arXiv:1701.04167 [q-fin.RM]`.
[90] S. Dharmadhikari and K. Joag-Dev, *Unimodality, convexity, and applications*, Academic Press, 1988.
[91] X. V. Doan, X. Li, and K. Natarajan, *Robustness to dependency in portfolio optimization using overlapping marginals*, Oper. Res., 63 (2015), pp. 1468–1488.
[92] J. Duchi, P. Glynn, and H. Namkoong, *Statistics of robust optimization: A generalized empirical likelihood approach*, 2016. `arXiv preprint arXiv:1610.03425 [stat.ML]`.
[93] J. C. Duchi, T. Hashimoto, and H. Namkoong, *Distributionally robust losses against mixture covariate shifts*, 2019.
[94] R. Dudley, *The speed of mean Glivenko-Cantelli convergence*, The Annals of Mathematical Statistics, 40 (1969), pp. 40–50.
[95] J. Dupačová, *The minimax approach to stochastic programming and an illustrative application*, Stochastics: An International Journal of Probability and Stochastic Processes, 20 (1987), pp. 73–88.
[96] J. Dupačová, *Stability and sensitivity-analysis for stochastic programming*, Ann Oper Res, 27 (1990), pp. 115–142.
[97] J. Dupačová, N. Gröwe-Kuska, and W. Römisch, *Scenario reduction in stochastic programming*, Math. Program., 95 (2003), pp. 493–511.
[98] E. Eban, E. Mezuman, and A. Globerson. *Discrete Chebyshev classifiers.* in 31st International Conference on Machine Learning, 2014.
[99] L. El Ghaoui and H. Lebret, *Robust solutions to least-squares problems with uncertain data*, Siam J Matrix Anal A, 18 (1997), pp. 1035–1064.
[100] L. El Ghaoui, F. Oustry, and H. Lebret, *Robust solutions to uncertain semidefinite programs*, SIAM J. Optim., 9 (1998), pp. 33–52.
[101] L. El Ghaoui, M. Oks, and F. Oustry, *Worst-case value-at-risk and robust portfolio optimization: A conic programming approach*, Oper. Res., 51 (2003), pp. 543–556.
[102] E. Erdoğan and G. Iyengar, *Ambiguous chance constrained problems and robust optimization*, Math. Program., 107 (2006), pp. 37–61.
[103] F. Farnia and D. Tse. *A minimax approach to supervised learning.* in Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., pp. 4240–4248, Curran Associates, Inc., 2016, http://papers.nips.cc/paper/6247-a-minimax-approach-to-supervised-learning.pdf.
[104] R. Fathony, A. Rezaei, M. A. Bashiri, X. Zhang, and B. Ziebart. *Distributionally robust graphical models.* in Advances in Neural Information Processing Systems 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., Curran Associates, Inc., Curran Associates, Inc., 2018, http://papers.nips.cc/paper/8055-distributionally-robust-graphical-models.pdf.
[105] N. Fournier and A. Guillin, *On the rate of convergence in Wasserstein distance of the empirical measure*, Probab Theory Rel, 162 (2015), pp. 707–738.
[106] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1, Springer series in statistics New York, 2nd ed., 2016.
[107] M. C. Fu. *Handbook of simulation optimization.* in International Series in Operations Research & Management Science, C. C. Price, ed., vol. 216, Springer, 2016.
[108] V. Gabrel, C. Murat, and A. Thiele, *Recent advances in robust optimization: An*

*overview*, Eur. J. Oper. Res., 235 (2014), pp. 471 – 483.
[109] G. Gallego and I. Moon, *The distribution free newsboy problem: review and extensions*, J. Oper. Res. Soc., 44 (1993), pp. 825–834.
[110] R. Gao and A. J. Kleywegt, *Distributionally robust stochastic optimization with Wasserstein distance*, 2016. arXiv preprint arXiv:1604.02199v2 [math.OC].
[111] R. Gao and A. J. Kleywegt, *Distributionally robust stochastic optimization with dependence structure*, 2017. arXiv preprint arXiv:1701.04200 [math.OC.
[112] R. Gao, X. Chen, and A. J. Kleywegt, *Wasserstein distributional robustness and regularization in statistical learning*, 2017. arXiv preprint arXiv:1712.06050 [math.OC].
[113] R. Gao, L. Xie, Y. Xie, and H. Xu, *Robust hypothesis testing using Wasserstein uncertainty sets*, 2018. arXiv preprint arXiv:1805.10611 [stat.ML.
[114] S. Ghosh and H. Lam, *Robust analysis in stochastic simulation: Computation and performance guarantees*, 2018. arXiv preprint arXiv:1507.05609 [math.PR].
[115] S. Ghosh, M. Squillante, and E. Wollega, *Efficient stochastic gradient descent for distributionally robust learning*, 2018. arXiv preprint arXiv:1805.08728 [stat.ML].
[116] A. L. Gibbs and F. E. Su, *On choosing and bounding probability metrics*, Int Stat Rev, 70 (2002), pp. 419–435.
[117] I. Gilboa and D. Schmeidler, *Maxmin expected utility with non-unique prior*, J Math Econ, 18 (1989), pp. 141 – 153.
[118] M. Glanzer, G. C. Pflug, and A. Pichler, *Incorporating statistical model error into the calculation of acceptability prices of contingent claims*, Math. Program., 174 (2019), pp. 499–524.
[119] A. Globerson and N. Tishby. *The minimum information principle for discriminative learning.* in Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUAI Press, AUAI Press, 2004.
[120] J. Goh and M. Sim, *Distributionally robust optimization and its tractable approximations*, Oper. Res., 58 (2010), pp. 902–917.
[121] J. Goh and M. Sim, *Robust optimization made easy with rome*, Oper. Res., 59 (2011), pp. 973–985.
[122] D. Goldfarb and G. Iyengar, *Robust portfolio selection problems*, Math Oper Res, 28 (2003), pp. 1–38.
[123] Z. Gong, C. Liu, J. Sun, and K. L. Teo, *Distributionally robust L1-estimation in multiple linear regression*, Optim Lett, 13 (2019), pp. 935–947.
[124] B. L. Gorissen, İ. Yanikoğlu, and D. den Hertog, *A practical guide to robust optimization*, Omega, 53 (2015), pp. 124 – 137.
[125] J.-y. Gotoh, M. J. Kim, and A. E. B. Lim, *Calibration of distributionally robust empirical optimization models*, 2017. arXiv:1711.06565 [stat.ML].
[126] P. D. Grünwald and A. P. Dawid, *Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory*, the Annals of Statistics, 32 (2004), pp. 1367–1433.
[127] G. Gül, *Asymptotically minimax robust hypothesis testing*, 2017. arXiv preprint arXiv:1711.07680 [cs.IT].
[128] G. Gül and A. M. Zoubir, *Minimax robust hypothesis testing*, Ieee T Inform Theory, 63 (2017), pp. 5572–5587.
[129] S. Guo, H. Xu, and L. Zhang, *Convergence analysis for mathematical programs with distributionally robust chance constraint*, SIAM J. Optim., 27 (2017), pp. 784–816.
[130] A. Haar, *Über linear Ungleichungen*, Acta Mathematica Szeged, 2 (1924).
[131] B. V. Halldórsson and R. H. Tütüncü, *An interior-point method for a class of saddle-point problems*, J Optimiz Theory App, 116 (2003), pp. 559–590.
[132] G. A. Hanasusanto and D. Kuhn, *Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls*, Oper. Res., 66 (2018), pp. 849–869.
[133] G. A. Hanasusanto, D. Kuhn, S. W. Wallace, and S. Zymler, *Distributionally robust multi-item newsvendor problems with multimodal demand distributions*, Math. Program., 152 (2015), pp. 1–32.
[134] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann, *A distributionally robust perspective on uncertainty quantification and chance constrained programming*, Math. Program., 151 (2015), pp. 35–62.
[135] G. A. Hanasusanto, D. Kuhn, and W. Wiesemann, *K-adaptability in two-stage distributionally robust binary programming*, Oper. Res. Lett., 44 (2016), pp. 6 – 11.

[136] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann, *Ambiguous joint chance constraints under mean and dispersion information*, Oper. Res., 65 (2017), pp. 751–767.
[137] G. A. Hanasusanto and D. Kuhn. *Robust data-driven dynamic programming.* in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., pp. 827–835, Curran Associates, Inc., 2013, http://papers.nips.cc/paper/5123-robust-data-driven-dynamic-programming.pdf.
[138] L. Hannah, W. Powell, and D. M. Blei. *Nonparametric density estimation for stochastic optimization with an observable state variable.* in Advances in Neural Information Processing Systems 23, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, eds., pp. 820–828, Curran Associates, Inc., 2010, http://papers.nips.cc/paper/4098-nonparametric-density-estimation-for-stochastic-optimization-with-an-observable-state-variable.pdf.
[139] H. Heitsch and W. Römisch, *Scenario reduction algorithms in stochastic programming*, Comput. Optim. Appl., 24 (2003), pp. 187–206.
[140] H. Heitsch and W. Römisch, *Scenario tree modeling for multistage stochastic programs*, Math. Program., 118 (2009), pp. 371–406.
[141] H. Heitsch and W. Römisch, *Scenario tree reduction for multistage stochastic programs*, Comput. Management Sci., 6 (2009), pp. 117–133.
[142] H. Heitsch, W. Römisch, and C. Strugarek, *Stability of multistage stochastic programs*, SIAM J. Optim., 17 (2006), pp. 511–525.
[143] R. Hettich and H. T. Jongen, *On first and second order conditions for local optima for optimization problems in finite dimensions*, Methods Oper. Res., 23 (1977), pp. 82–97.
[144] R. Hettich and H. T. Jongen. *Semi-infinite programming: conditions of optimality and applications.* in Optimization Techniques, Lecture Notes in Control and Information Science, J. Stoer, ed., ch. 7, pp. 82–97, Springer-Verlag, Berlin, Heidelberg, New York, 1978.
[145] R. Hettich and K. O. Kortanek, *Semi-infinite programming: theory, methods, and applications*, Siam Rev, 35 (1993), pp. 380–429.
[146] R. Hettich and G. Still, *Second order optimality conditions for generalized semi-infinite programming problems*, Lect Notes Econ Math, 34 (1995), pp. 195–211.
[147] T. Homem-de-Mello and G. Bayraksan, *Monte carlo sampling-based methods for stochastic optimization*, Surv. Oper. Res. Manage. Sci., 19 (2014), pp. 56–85.
[148] T. Homem-de-Mello and S. Mehrotra, *A cutting-surface method for uncertain linear programs with polyhedral stochastic dominance constraints*, SIAM J. Optim., 20 (2009), pp. 1250–1273.
[149] J. Hu and S. Mehrotra, *Robust and stochastically weighted multiobjective optimization models and reformulations*, Oper. Res., 60 (2012), pp. 936–953.
[150] J. Hu and S. Mehrotra, *Robust decision making over a set of random targets or risk-averse utilities with an application to portfolio optimization*, IIE Trans., 47 (2015), pp. 358–372.
[151] J. Hu, T. Homem-de-Mello, and S. Mehrotra, *Risk-adjusted budget allocation models with application in homeland security*, IIE Trans., 43 (2011), pp. 819–839.
[152] J. Hu, T. Homem-de-Mello, and S. Mehrotra, *Sample average approximation of stochastic dominance constrained programs*, Math. Program., 133 (2012), pp. 171–201.
[153] J. Hu, T. Homem-de-Mello, and S. Mehrotra, *Stochastically weighted stochastic dominance concepts with an application in capital budgeting*, Eur. J. Oper. Res., 232 (2014), pp. 572–583.
[154] J. Hu, J. Li, and S. Mehrotra, *A data driven functionally robust approach for coordinating pricing and order quantity decisions with unknown demand function*, To appear in Operations Research, (2019). Optimization Online http://www.optimization-online.org/DB_HTML/2015/07/5016.html.
[155] W. Hu, G. Niu, I. Sato, and M. Sugiyama. *Does distributionally robust supervised learning give robust classifiers?* in 35th International Conference on Machine Learning, 2018.
[156] Z. Hu and L. J. Hong, *Kullback-Leibler divergence constrained distributionally robust optimization*, 2012. Optimization Online http://www.optimization-online.org/DB_HTML/2012/11/3677.html.
[157] Z. Hu, L. J. Hong, and A. M. C. So, *Ambiguous probabilistic programs*, 2013. Optimization Online http://www.optimization-online.org/DB_HTML/2013/09/4039.html.
[158] J. Huang, K. Zhou, and Y. Guan, *A study of distributionally robust multistage*

*stochastic optimization*, 2017. arXiv preprint arXiv:1708.07930 [math.OC].
[159] P. J. Huber, *A robust version of the probability ratio test*, The Annals of Mathematical Statistics, (1965), pp. 1753–1758.
[160] P. J. Huber, *The use of choquet capacities in statistics*, B Int Statist Inst, 45 (1973), pp. 181–191.
[161] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, John Wiley & Sons, 2nd ed., 2009.
[162] K. Isii, *On sharpness of tchebycheff-type inequalities*, Ann I Stat Math, 14 (1962), pp. 185–197.
[163] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112, Springer, 2013.
[164] R. Ji and M. Lejeune, *Data-driven optimization of reward-risk ratio measures*, 2017. Optimization Online http://www.optimization-online.org/DB_HTML/2017/01/5819.html.
[165] R. Ji and M. Lejeune, *Data-driven distributionally robust chance-constrained programming with Wasserstein metric*, 2018. Optimization Online http://www.optimization-online.org/DB_HTML/2018/07/6697.html.
[166] R. Jiang and Y. Guan, *Data-driven chance constrained stochastic program*, Math. Program., 158 (2016), pp. 291–327.
[167] R. Jiang and Y. Guan, *Risk-averse two-stage stochastic program with distributional ambiguity*, Oper. Res., 66 (2018), pp. 1390–1405.
[168] R. Jiang, Y. Guan, and J.-P. Watson, *Risk-averse stochastic unit commitment with incomplete information*, IIE Trans., 48 (2016), pp. 838–854.
[169] M. Kapsos, N. Christofides, and B. Rustem, *Worst-case robust omega ratio*, Eur. J. Oper. Res., 234 (2014), pp. 499–507.
[170] C. Keating and W. F. Shadwick, *A universal performance measure*, Journal of performance measurement, 6 (2002), pp. 59–84.
[171] K. Kim and S. Mehrotra, *A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management*, Oper. Res., 63 (2015), pp. 1431–1451.
[172] D. Klabjan, D. Simchi-Levi, and M. Song, *Robust stochastic lot-sizing by means of histograms*, Prod. Oper. Management, 22 (2013), pp. 691–710.
[173] S. Kusuoka, *On law invariant coherent risk measures*, in Advances in Mathematical Economics, S. Kusuoka and T. Maruyama, eds., vol. 3, Springer Japan, Tokyo, 2001, pp. 83–95.
[174] J. D. Lafferty, A. McCallum, and F. 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, pp. 282–289, San Francisco, CA, USA, 2001, Morgan Kaufmann Publishers Inc.
[175] H. Lam. *Advanced tutorial: Input uncertainty and robust analysis in stochastic simulation.* in Proceedings of the 2016 Winter Simulation Conference, WSC '16, pp. 178–192, Piscataway, NJ, USA, 2016, IEEE Press.
[176] H. Lam, *Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization*, 2016. arXiv preprint arXiv:1605.09349 [math.OC].
[177] H. Lam and S. Ghosh. *Iterative methods for robust estimation under bivariate distributional uncertainty.* in Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World, A. T. R. H. R. Pasupathy, S.-H. Kim and M. E. Kuhl, eds., WSC '13, pp. 193–204, Piscataway, NJ, USA, 2013, IEEE Press.
[178] H. Lam and C. Mottet, *Tail analysis without parametric models: A worst-case perspective*, Oper. Res., 65 (2017), pp. 1696–1711.
[179] G. R. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan, *A robust minimax approach to classification*, J Mach Learn Res, 3 (2002), pp. 555–582.
[180] H. J. Landau, ed., *Moments in mathematics*, vol. 37 of Proceeding of Symposia in Applied Mathematics, American Mathematical Society, Providence, RI, 1987.
[181] J. Lasserre and T. Weisser, *Representation of distributionally robust chance-constraints*, 2018. arXiv preprint arXiv:1803.11500 [math.OC].
[182] J. B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.
[183] C. Lee and S. Mehrotra, *A distributionally-robust approach for finding support vector machines*, 2015. Optimization Online http://www.optimization-online.org/DB_HTML/2015/06/4965.html.
[184] J. Lee and M. Raginsky, *Minimax statistical learning and domain adaptation with*

*Wasserstein distances*, 2017. arXiv preprint arXiv:1705.07815 [cs.LG].
[185] J. Lee and M. Raginsky. *Minimax statistical learning with Wasserstein distances.* in Advances in Neural Information Processing Systems 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., pp. 2692–2701, Curran Associates, Inc., 2018, http://papers.nips.cc/paper/7534-minimax-statistical-learning-with-{W}asserstein-distances.pdf.
[186] B. C. Levy, *Robust hypothesis testing with a relative entropy tolerance*, Ieee T Inform Theory, 55 (2009), pp. 413–421.
[187] B. Li, R. Jiang, and J. L. Mathieu, *Ambiguous risk constraints with moment and unimodality information*, Math. Program., 173 (2019), pp. 151–192.
[188] J. Y. Li and R. H. Kwon, *Portfolio selection under model uncertainty: a penalized moment-based optimization approach*, J Global Optim, 56 (2013), pp. 131–164.
[189] J. Y.-M. Li, *Closed-form solutions for worst-case law invariant risk measures with application to robust portfolio optimization*, 2016. arXiv preprint arXiv:1609.04065 [q-fin.RM].
[190] Q. Lin, R. Loxton, K. L. Teo, Y. H. Wu, and C. Yu, *A new exact penalty method for semi-infinite programming problems*, J Comput Appl Math, 261 (2014), pp. 271–286.
[191] Y. Liu, R. Meskarian, and H. Xu, *Distributionally robust reward-risk ratio optimization with moment constraints*, SIAM J. Optim., 27 (2017), pp. 957–985.
[192] Y. Liu, X. Yuan, S. Zeng, and J. Zhang, *Primal–dual hybrid gradient method for distributionally robust optimization problems*, Oper. Res. Lett., 45 (2017), pp. 625–630.
[193] D. Z. Long and J. Qi, *Distributionally robust discrete optimization with entropic value-at-risk*, Oper. Res. Lett., 42 (2014), pp. 532 – 538.
[194] M. López and G. Still, *Semi-infinite programming*, Eur. J. Oper. Res., 180 (2007), pp. 491–518.
[195] S. Lotfi and S. A. Zenios, *Robust var and cvar optimization under joint ambiguity in distributions, means, and covariances*, Eur. J. Oper. Res., 269 (2018), pp. 556–576.
[196] D. Love and G. Bayraksan. *Two-stage likelihood robust linear program with application to water allocation under uncertainty.* in Simulation Conference (WSC), 2013 Winter, IEEE, IEEE, 2013.
[197] D. K. Love and G. 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.
[198] J. Luedtke and S. Ahmed, *A sample approximation approach for optimization with probabilistic constraints*, SIAM J. Optim., 19 (2008), pp. 674–699.
[199] F. Luo and S. Mehrotra, *Distributionally robust optimization with decision dependent ambiguity sets*, 2018. arXiv:1806.09215 [math.OC].
[200] F. Luo and S. Mehrotra, *Decomposition algorithm for distributionally robust optimization using Wasserstein metric with an application to a class of regression models*, Eur. J. Oper. Res., 278 (2019), pp. 20–35.
[201] C. McDiarmid, *Concentration*, in Probabilistic Methods for Algorithmic Discrete Mathematics, M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, eds., Springer Berlin Heidelberg, Berlin, Heidelberg, 1998, pp. 195–248.
[202] S. Mehrotra and D. Papp, *A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization*, SIAM J. Optim., 24 (2014), pp. 1670–1697.
[203] S. Mehrotra and H. Zhang, *Models and algorithms for distributionally robust least squares problems*, Math. Program., 146 (2014), pp. 123–141.
[204] M. Mevissen, E. Ragnoli, and J. Y. Yu. *Data-driven distributionally robust polynomial optimization.* in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., pp. 37–45, Curran Associates, Inc., 2013, http://papers.nips.cc/paper/4943-data-driven-distributionally-robust-polynomial-optimization.pdf.
[205] P. Mohajerin Esfahani and D. Kuhn, *Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations*, Math. Program., 171 (2018), pp. 115–166.
[206] P. Mohajerin Esfahani, S. Shafieezadeh-Abadeh, G. A. Hanasusanto, and D. Kuhn, *Data-driven inverse optimization with imperfect information*, Math. Program., 167 (2018), pp. 191–234.
[207] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*, MIT press, 2018.

[208] H. Namkoong and J. C. Duchi. *Stochastic gradient methods for distributionally robust optimization with f-divergences.* in Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., pp. 2208–2216, Curran Associates, Inc., 2016, http://papers.nips.cc/paper/6040-stochastic-gradient-methods-for-distributionally-robust-optimization-with-f-divergences.pdf.

[209] H. Namkoong and J. C. Duchi. *Variance-based regularization with convex objectives.* in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., pp. 2971–2980, Curran Associates, Inc., 2017, http://papers.nips.cc/paper/6890-variance-based-regularization-with-convex-objectives.pdf.

[210] K. Natarajan and C.-P. Teo, *On reduced semidefinite programs for second order moment bounds with applications*, Math. Program., 161 (2017), pp. 487–518.

[211] K. Natarajan, C. P. Teo, and Z. Zheng, *Mixed 0-1 linear programs under objective uncertainty: A completely positive representation*, Oper. Res., 59 (2011), pp. 713–728.

[212] K. Natarajan, D. Shi, and K.-C. Toh, *A probabilistic model for minmax regret in combinatorial optimization*, Oper. Res., 62 (2014), pp. 160–181.

[213] A. Nemirovski and A. Shapiro, *Convex approximations of chance constrained programs*, SIAM J. Optim., 17 (2006), pp. 969–996.

[214] A. Nemirovski and A. Shapiro, *Scenario approximations of chance constraints*, in Probabilistic and Randomized Methods for Design under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer, 2006, pp. 3–47.

[215] V. A. Nguyen, D. Kuhn, and P. M. Esfahani, *Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator*, 2018. `arXiv preprint arXiv:1805.07194 [math.OC]`.

[216] C. Ning and F. You. *Data-driven adaptive robust optimization framework based on principal component analysis.* in 2018 Annual American Control Conference (ACC), IEEE, IEEE, 2018.

[217] C. Ning and F. You, *Data-driven decision making under uncertainty integrating robust optimization with principal component analysis and kernel smoothing methods*, Computers & Chemical Engineering, 112 (2018), pp. 190–210.

[218] C. Ning, D. J. Garcia, and F. You, *Hedging against uncertainty in biomass processing network design using a data-driven approach*, Chemical Engineering Transactions, 70 (2018), pp. 1837–1842.

[219] K. G. Nishimura and H. Ozaki, *Search and Knightian uncertainty*, J. Econ. Theory, 119 (2004), pp. 299–333.

[220] K. G. Nishimura and H. Ozaki, *An axiomatic approach to $\epsilon$-contamination*, J. Econ. Theory, 27 (2006), pp. 333–340.

[221] N. Noyan, G. Rudolf, and M. Lejeune, *Distributionally robust optimization with decision-dependent ambiguity set*, 2018. Optimization Online http://www.optimization-online.org/DB_HTML/2018/09/6821.html.

[222] G. Nuernberger, *Global unicity in semi-infinite optimization*, Numer. Funct. Anal. Optim., 8 (1985), pp. 173–191.

[223] G. Nürnberger, *Global unicity in optimization and approximation*, Zeitschrift für angewandte Mathematik und Mechanik: ZAMM, 65 (1985), pp. T319–T321.

[224] A. B. Owen, *Empirical likelihood*, Chapman and Hall/CRC, 2001.

[225] C. Pang Ho and G. 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.

[226] L. Pardo, *Statistical inference based on divergence measures*, Chapman & Hall/CRC Press, 2005.

[227] R. Pasupathy and S. Ghosh. *Simulation optimization: A concise overview and implementation guide.* in Theory Driven by Influential Applications, vol. 7, pp. 122–150, INFORMS TutORials in Operations Research, 2013.

[228] G. Perakis and G. Roels, *Regret in the newsvendor model with partial information*, Oper. Res., 56 (2008), pp. 188–203.

[229] G. Pflug and D. Wozabal, *Ambiguity in portfolio selection*, Quant. Financ., 7 (2007), pp. 435–442.

[230] G. C. Pflug and A. Pichler, *A distance for multistage stochastic optimization models*, SIAM J. Optim., 22 (2012), pp. 1–23.

[231] G. C. Pflug and A. Pichler. *The problem of ambiguity in stochastic optimization.*

in Multistage Stochastic Optimization, pp. 229–255, Springer, 2014.
[232] G. C. Pflug and M. Pohl, *A review on ambiguity in stochastic portfolio optimization*, Set-Valued and Variational Analysis, 26 (2018), pp. 733–757.
[233] G. C. Pflug, A. Pichler, and D. Wozabal, *The 1/n investment strategy is optimal under high model ambiguity*, J. Bank. Financ., 36 (2012), pp. 410–417.
[234] A. Philpott, V. de Matos, and L. Kapelevich, *Distributionally robust SDDP*, Comput. Management Sci., 15 (2018), pp. 431–454.
[235] A. Pichler, *Evaluations of risk measures for different probability measures*, SIAM J. Optim., 23 (2013), pp. 530–551.
[236] A. Pichler and H. Xu, *Quantitative stability analysis for minimax distributionally robust risk optimization*, Math. Program., (2017), pp. 1–31, https://doi.org/10.1007/s10107-018-1347-4.
[237] I. Pólik and T. Terlaky, *A survey of the S-lemma*, Siam Rev, 49 (2007), pp. 371–418.
[238] I. Popescu, *A semidefinite programming approach to optimal-moment bounds for convex classes of distributions*, Math Oper Res, 30 (2005), pp. 632–657.
[239] I. Popescu, *Robust mean-covariance solutions for stochastic optimization*, Oper. Res., 55 (2007), pp. 98–112.
[240] K. Postek, D. den Hertog, and B. Melenberg, *Computationally tractable counterparts of distributionally robust constraints on risk measures*, Siam Rev, 58 (2016), pp. 603–650.
[241] K. Postek, A. Ben-Tal, D. den Hertog, and B. Melenberg, *Robust optimization with ambiguous stochastic constraints under mean and dispersion information*, Oper. Res., 66 (2018), pp. 814–833.
[242] K. Postek, W. Romeijnders, D. den Hertog, and M. H. van der Vlerk, *An approximation framework for two-stage ambiguous stochastic integer programs under mean-mad information*, Eur. J. Oper. Res., 274 (2019), pp. 432–444.
[243] A. Prékopa. *On probabilistic constrained programming.* in Proceedings of the Princeton symposium on mathematical programming, vol. 113, Princeton, NJ, Princeton, NJ, 1970.
[244] A. Prékopa. *Probabilistic programming.* vol. 10 of Handbooks in Operations Research and Management Science, Elsevier, 2003.
[245] A. Prékopa, *Programming under probabilistic constraints with a random technology matrix*, Statistics: A Journal of Theoretical and Applied Statistics, 5 (1974), pp. 109–116.
[246] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, 2005.
[247] P.-Y. Qian, Z.-Z. Wang, and Z.-W. Wen, *A composite risk measure framework for decision making under uncertainty*, Journal of the Operations Research Society of China, 7 (2019), pp. 43–68.
[248] S. T. Rachev and W. Römisch, *Quantitative stability in stochastic programming: The method of probability metrics*, Math Oper Res, 27 (2002), pp. 792–818.
[249] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems: Volume I: Theory*, vol. 1, Springer Science & Business Media, 1998.
[250] S. T. Rachev, *Probability metrics and the stability of stochastic models*, vol. 269, John Wiley & Son Ltd, 1991.
[251] H. Rahimian, G. Bayraksan, and T. Homem-de-Mello, *Identifying effective scenarios in distributionally robust stochastic programs with total variation distance*, Math. Program., 173 (2019), pp. 393–430.
[252] H. Rahimian, G. Bayraksan, and T. Homem-de-Mello, *Controlling risk and demand ambiguity in newsvendor models*, Eur. J. Oper. Res., 279 (2019), pp. 854–868.
[253] M. Razaviyayn, F. Farnia, and D. Tse. *Discrete Rényi classifiers.* in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds., pp. 3276–3284, Curran Associates, Inc., 2015, http://papers.nips.cc/paper/5698-discrete-renyi-classifiers.pdf.
[254] T. R. Read and N. A. Cressie, *Goodness-of-fit statistics for discrete multivariate data*, Springer-Verlag, 1988.
[255] R. Reemtsen and S. Görner. *Numerical methods for semi-infinite programming: A survey.* in Semi-infinite Programming, Nonconvex Optimization and Its Applications, R. Reemtsen and J. J. Rückmann, eds., ch. 25, pp. 195–275, Kluwer Boston, Boston, 1998.
[256] R.-D. Reiss, *Approximate distributions of order statistics: with applications to nonparametric statistics*, Springer science & business media, 1989.
[257] R. Rockafellar, *Conjugate Duality and Optimization*, CBMS-NSF Regional Con-

ference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, 1974.
[258] R. T. Rockafellar. *Coherent approaches to risk in optimization under uncertainty.* in OR Tools and Applications: Glimpses of Future Technologies, pp. 38–61, INFORMS TutORials in Operations Research, 2007.
[259] R. T. Rockafellar and J. O. Royset, *Measures of residual risk with connections to regression, risk tracking, surrogate models, and ambiguity*, SIAM J. Optim., 25 (2015), pp. 1179–1208.
[260] R. T. Rockafellar and S. Uryasev, *Optimization of conditional value-at-risk*, J Risk, 2 (2000), pp. 21–42.
[261] R. T. Rockafellar and S. Uryasev, *Conditional value-at-risk for general loss distributions*, J. Bank. Financ., 26 (2002), pp. 1443–1471.
[262] R. Rockafellar, *Convex Analysis*, Princeton landmarks in mathematics and physics, Princeton University Press, 1997.
[263] W. Römisch. *Stability of stochastic programming problems.* in Stochastic Programming, A. Ruszczyński and A. Shapiro, eds., vol. 10 of Handbooks in Operations Research and Management Science, pp. 483 – 554, Elsevier, 2003.
[264] E. Roos and D. den Hertog, *Reducing conservatism in robust optimization*, 2018. Optimziation Online http://www.optimization-online.org/DB_HTML/2017/12/6383.html.
[265] J. O. Royset and R. J.-B. Wets, *Variational theory for optimization under stochastic ambiguity*, SIAM J. Optim., 27 (2017), pp. 1118–1149.
[266] N. Rujeerapaiboon, D. Kuhn, and W. Wiesemann, *Robust growth-optimal portfolios*, Management Sci., 62 (2016), pp. 2090–2109.
[267] N. Rujeerapaiboon, D. Kuhn, and W. Wiesemann, *Chebyshev inequalities for products of random variables*, Math Oper Res, 43 (2018), pp. 887–918.
[268] N. Rujeerapaiboon, K. Schindler, D. Kuhn, and W. Wiesemann, *Scenario reduction revisited: fundamental limits and guarantees*, Math. Program., (2018), https://doi.org/10.1007/s10107-018-1269-1.
[269] A. Ruszczyński, *Nonlinear optimization*, Princeton university press, 2006.
[270] A. Ruszczyński and A. Shapiro, *Optimization of convex risk functions*, Math Oper Res, 31 (2006), pp. 433–452.
[271] H. Scarf. *A min-max solution of an inventory problem.* in Studies in the mathematical theory of inventory and production, H. Scarf, K. Arrow, and S. Karlin, eds., vol. 10, pp. 201–209, Stanford University Press, Stanford, CA, 1958.
[272] R. Schultz, *Some aspects of stability in stochastic programming*, Ann Oper Res, 100 (2000), pp. 55–84.
[273] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. *Distributionally robust logistic regression.* in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds., pp. 1576–1584, Curran Associates, Inc., 2015, http://papers.nips.cc/paper/5745-distributionally-robust-logistic-regression.pdf.
[274] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, *Regularization via mass transportation*, 2017. `arXiv preprint arXiv:1710.10016 [math.OC]`.
[275] S. Shafieezadeh-Abadeh, V. A. Nguyen, D. Kuhn, and P. M. Esfahani, *Wasserstein distributionally robust kalman filtering*, 2018. `arXiv preprint arXiv:1809.08830 [math.OC]`.
[276] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014.
[277] C. Shang and F. You, *Data-driven process network planning: A distributionally robust optimization approach*, IFAC-PapersOnLine, 51 (2018), pp. 150–155.
[278] C. Shang and F. You, *Robust optimization in high-dimensional data space with support vector clustering*, IFAC-PapersOnLine, 51 (2018), pp. 19–24.
[279] C. Shang and F. You, *Distributionally robust optimization for planning and scheduling under uncertainty*, Computers & Chemical Engineering, 110 (2018), pp. 53–68.
[280] C. Shang and F. You, *A data-driven robust optimization approach to stochastic model predictive control*, 2018. `arXiv preprint arXiv:1807.05146 [math.OC]`.
[281] C. Shang and F. You. *Process scheduling under ambiguity uncertainty probability distribution.* in Computer Aided Chemical Engineering, A. Friedl, J. J. Klemeš, S. Radl, P. S. Varbanov, and T. Wallek, eds., vol. 43, pp. 919–924, Elsevier, 2018.
[282] C. Shang, X. Huang, and F. You, *Data-driven robust optimization based on kernel learning*, Computers & Chemical Engineering, 106 (2017), pp. 464–479.

[283] C. Shang, W.-H. Chen, A. D. Stroock, and F. You, *Robust model predictive control of irrigation systems with active uncertainty learning and data analytics*, 2018. arXiv preprint arXiv:1810.05947 [cs.SY].
[284] A. Shapiro, *On duality theory of conic linear problems*, in Semi-Infinite Programming: Recent Advances, M. Á. Goberna and M. A. López, eds., Springer US, Boston, MA, 2001, pp. 135–165.
[285] A. Shapiro. *Monte carlo sampling methods.* in Handbooks in Operations Research and Management Science, A. Ruszczyński and A. Shapiro, eds., vol. 10, Elsevier, 2003.
[286] A. Shapiro, *Minimax and risk averse multistage stochastic programming*, Eur. J. Oper. Res., 219 (2012), pp. 719–726.
[287] A. Shapiro, *On kusuoka representation of law invariant risk measures*, Math Oper Res, 38 (2013), pp. 142–152.
[288] A. Shapiro, *Rectangular sets of probability measures*, Oper. Res., 64 (2016), pp. 528–541.
[289] A. Shapiro, *Distributionally robust stochastic programming*, SIAM J. Optim., 27 (2017), pp. 2258–2275.
[290] A. Shapiro, *Tutorial on risk neutral, distributionally robust and risk averse multistage stochastic programming*, 2018. Optimization Online http://www.optimization-online.org/DB_HTML/2018/02/6455.html.
[291] A. Shapiro and S. Ahmed, *On a class of minimax stochastic programs*, SIAM J. Optim., 14 (2004), pp. 1237–1249.
[292] A. Shapiro and A. Kleywegt, *Minimax analysis of stochastic problems*, Optim. Method. Softw., 17 (2002), pp. 523–542.
[293] A. Shapiro and A. Nemirovski. *On complexity of stochastic programming problems.* in Continuous Optimization: Current Trends and Modern Applications, V. Jeyakumar and A. Rubinov, eds., pp. 111–146, Springer, 2005.
[294] A. Shapiro, W. Tekaya, M. P. Soares, and J. P. da Costa, *Worst-case-expectation approach to optimization under uncertainty*, Oper. Res., 61 (2013), pp. 1435–1449.
[295] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*, MPS-SIAM series on optimization, Society for Industrial and Applied Mathematics, Philadelphia, USA, 2nd ed., 2014.
[296] W. F. Sharpe, *Mutual fund performance*, The Journal of business, 39 (1966), pp. 119–138.
[297] S. Singh and B. Póczos, *Minimax distribution estimation in Wasserstein distance*, 2018. arXiv preprint arXiv:1802.08855 [math.ST].
[298] A. Sinha, H. Namkoong, and J. Duchi, *Certifying some distributional robustness with principled adversarial training*, 2018. arXiv preprint arXiv:1710.10571 [stat.ML].
[299] M. Sion, *On general minimax theorems*, Pac J Math, 8 (1958), pp. 171–176.
[300] G. Still, *Generalized semi-infinite programming: theory and methods*, Eur. J. Oper. Res., 119 (1999), pp. 301–313.
[301] J. Sun, L.-Z. Liao, and B. Rodrigues, *Quadratic two-stage stochastic optimization with coherent measures of risk*, Math. Program., 168 (2018), pp. 599–613.
[302] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, *Large margin methods for structured and interdependent output variables*, J Mach Learn Res, 6 (2005), pp. 1453–1484.
[303] T. Tulabandhula and C. Rudin, *Machine learning with operational costs*, The Journal of Machine Learning Research, 14 (2013), pp. 1989–2028.
[304] T. Tulabandhula and C. Rudin, *Robust optimization using machine learning for uncertainty sets*, 2014. arXiv preprint arXiv:1407.1097[math.OC].
[305] T. Tulabandhula and C. Rudin, *On combining machine learning with decision making*, Mach Learn, 97 (2014), pp. 33–64.
[306] I. Vajda, *Theory of statistical inference and information*, vol. 11, Kluwer Academic Pub, 1989.
[307] B. P. G. Van Parys, D. Kuhn, P. J. Goulart, and M. Morari, *Distributionally robust control of constrained stochastic systems*, Ieee T Automat Contr, 61 (2016), pp. 430–442.
[308] B. P. G. Van Parys, P. J. Goulart, and D. Kuhn, *Generalized gauss inequalities via semidefinite programming*, Math. Program., 156 (2016), pp. 271–302.
[309] B. P. G. Van Parys, P. J. Goulart, and M. Morari, *Distributionally robust expectation inequalities for structured distributions*, Math. Program., 173 (2019), pp. 251–

280.
[310] L. Vandenberghe, S. Boyd, and K. Comanor, *Generalized Chebyshev bounds via semidefinite programming*, Siam Rev, 49 (2007), pp. 52–64.
[311] A. N. Vidyashankar and J. Xu. *Stochastic optimization using Hellinger distance.* in Proceedings of the 2015 Winter Simulation Conference, WSC '15, pp. 3702–3713, Piscataway, NJ, USA, 2015, IEEE Press.
[312] C. Villani, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.
[313] J. Žáčková, *On minimax solutions of stochastic linear programming problems*, Časopis pro pěstováná matematiky, 091 (1966), pp. 423–430, http://eudml.org/doc/20949.
[314] S. Wang and Y. Yuan, *Feasible method for semi-infinite programs*, SIAM J. Optim., 25 (2015), pp. 2537–2560.
[315] S. Wang, J. Li, and S. Mehrotra, *Distributionally robust chance-constrained assignment problem with an application to operating room planning*, 2019. Optimziation Online http://www.optimization-online.org/DB_HTML/2019/05/7207.html.
[316] Z. Wang, P. W. Glynn, and Y. Ye, *Likelihood robust optimization for data-driven problems*, Comput. Management Sci., 13 (2016), pp. 241–261.
[317] W. Wiesemann, D. Kuhn, and B. Rustem, *Robust markov decision processes*, Math Oper Res, 38 (2013), pp. 153–183.
[318] W. Wiesemann, D. Kuhn, and M. Sim, *Distributionally robust convex optimization*, Oper. Res., 62 (2014), pp. 1358–1376.
[319] D. Wozabal, *A framework for optimization under ambiguity*, Ann Oper Res, 193 (2012), pp. 21–47.
[320] D. Wozabal, *Robustifying convex risk measures for linear portfolios: A nonparametric approach*, Oper. Res., 62 (2014), pp. 1302–1315.
[321] W. Xie and S. Ahmed, *On deterministic reformulations of distributionally robust joint chance constrained optimization problems*, SIAM J. Optim., 28 (2018), pp. 1151–1182.
[322] W. Xie, *On distributionally robust chance constrained program with Wasserstein distance*, 2018. `arXiv preprint arXiv:1806.07418 [math.OC]`.
[323] W. Xie and S. Ahmed, *Distributionally robust simple integer recourse*, Comput. Management Sci., 15 (2018), pp. 351–367.
[324] W. Xie, S. Ahmed, and R. Jiang, *Optimized Bonferroni approximations of distributionally robust joint chance constraints*, 2017. Optimization Online http://www.optimization-online.org/DB_HTML/2017/02/5860.html.
[325] L. Xin and D. A. Goldberg, *Distributionally robust inventory control when demand is a martingale*, 2018. `arXiv preprint arXiv:1511.09437 [math.OC]`.
[326] L. Xin and D. A. Goldberg, *Time (in) consistency of multistage distributionally robust inventory models with moment constraints*, 2018. `arXiv preprint arXiv:1304.3074 [math.OC]`.
[327] G. Xu and S. Burer, *A data-driven distributionally robust bound on the expected optimal value of uncertain mixed 0-1 linear programming*, Comput. Management Sci., 15 (2018), pp. 111–134.
[328] H. Xu and S. Mannor. *Distributionally robust markov decision processes.* in Advances in Neural Information Processing Systems 23, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, eds., pp. 2505–2513, Curran Associates, Inc., 2010, http://papers.nips.cc/paper/3927-distributionally-robust-markov-decision-processes.pdf.
[329] H. Xu and S. Mannor, *Distributionally robust markov decision processes*, Math Oper Res, 37 (2012), pp. 288–300.
[330] H. Xu, C. Caramanis, and S. Mannor, *Optimization under probabilistic envelope constraints*, Oper. Res., 60 (2012), pp. 682–699.
[331] H. Xu, Y. Liu, and H. Sun, *Distributionally robust optimization with matrix moment constraints: Lagrange duality and cutting plane methods*, Math. Program., 169 (2018), pp. 489–529.
[332] M. Xu, S.-Y. Wu, and J. Y. Jane, *Solving semi-infinite programs by smoothing projected gradient method*, Comput. Optim. Appl., 59 (2014), pp. 591–616.
[333] I. Yang, *Wasserstein distributionally robust stochastic control: A data-driven approach*, 2018. `arXiv preprint arXiv:1812.09808 [math.OC]`.
[334] I. Yang, *A dynamic game approach to distributionally robust safety specifications for stochastic systems*, Automatica, 94 (2018), pp. 94–101.
[335] W. Yang and H. Xu, *Distributionally robust chance constraints for non-linear uncertainties*, Math. Program., 155 (2016), pp. 231–265.

[336] X. Yang, Z. Chen, and J. Zhou, *Optimality conditions for semi-infinite and generalized semi-infinite programs via lower order exact penalty functions*, J Optimiz Theory App, 169 (2016), pp. 984–1012.

[337] İ. Yanikoğlu and D. den Hertog, *Safe approximations of ambiguous chance constraints using historical data*, INFORMS J. Comput., 25 (2012), pp. 666–681.

[338] P. Yu and H. Xu, *Distributionally robust counterpart in markov decision processes*, Ieee T Automat Contr, 61 (2016), pp. 2538–2543.

[339] J. Zhang, H. Xu, and L. Zhang, *Quantitative stability analysis for distributionally robust optimization with moment constraints*, SIAM J. Optim., 26 (2016), pp. 1855–1882.

[340] Y. Zhang, R. Jiang, and S. Shen, *Ambiguous chance-constrained binary programs under mean-covariance information*, SIAM J. Optim., 28 (2018), pp. 2922–2944.

[341] Z. Zhang, B. Denton, and X. Xie, *Branch and price for chance constrained bin packing*, 2015. Optimization Online http://www.optimization-online.org/DB_HTML/2015/11/5217.html.

[342] C. Zhao and Y. Guan, *Data-driven risk-averse two-stage stochastic program with $\zeta$-structure probability metrics*, 2015. Optimization Online http://www.optimization-online.org/DB_HTML/2015/07/5014.html.

[343] C. Zhao and Y. Guan, *Data-driven risk-averse stochastic optimization with Wasserstein metric*, Oper. Res. Lett., 46 (2018), pp. 262–267.

[344] J. Zhen, D. den Hertog, and M. Sim, *Adjustable robust optimization via Fourier-Motzkin elimination*, Oper. Res., 66 (2018), pp. 1086–1100.

[345] S. Zymler, D. Kuhn, and B. Rustem, *Distributionally robust joint chance constraints with second-order moment information*, Math. Program., 137 (2013), pp. 167–198.

[346] S. Zymler, D. Kuhn, and B. Rustem, *Worst-case value at risk of nonlinear portfolios*, Management Sci., 59 (2013), pp. 172–188.