DISTRIBUTIONALLY ROBUST OPTIMIZATION: A REVIEW ON THEORY AND APPLICATIONS

FENGMING LIN AND XIAOLEI FANG
Edward P. Fitts Department of Industrial and Systems Engineering
North Carolina State University, Raleigh, NC 27606, USA
ZHEMING GAO*
College of Information Science and Engineering
Northeastern University, Shenyang, Liaoning 110819, China

ABSTRACT. In this paper, we survey the primary research on the theory and applications of distributionally robust optimization (DRO). We start with reviewing the modeling power and computational attractiveness of DRO approaches, induced by the ambiguity sets structure and tractable robust counterpart reformulations. Next, we summarize the efficient solution methods, out-of-sample performance guarantee, and convergence analysis. Then, we illustrate some applications of DRO in machine learning and operations research, and finally, we discuss the future research directions.

1 Introduction. Optimization under uncertainty is a classical topic that has been studied extensively since the 1950s in the mathematical programming and industrial engineering community. The uncertainty was first depicted with random vectors following a given distribution in Dantzig’s work [45], who presented a two-stage linear programming model in which the demand in the second stage is uncertain and follows a preset distribution. The approach is to minimize the total expected cost associated with the random vectors in order to find the optimal solutions. Since proposed, this approach has been widely discussed in various fields [22, 65, 89, 95] and blossomed into an optimization branch named stochastic optimization (SO). SO concerns the uncertainty with randomness that the probability distributions of the random variables are exactly known. Optimization under uncertainty without any distribution assumption started to attract attention in the 1950s as the popularity of the worst case analysis and Wald’s maximin model [178]. In the worst case analysis, the decision maker pursues an optimal solution that performs the best in the worst case over an uncertainty set—that is—a minimax approach. The minimax approach gradually established its own discipline in the 1970s [165], which was known as robust optimization (RO) according to the work of Ben-Tal and Nemirovski at the end of the 20th century [8–10].

In 1958, Scarf et al. [145] applied the minimax approach to solving an uncertain newsvendor problem, in which the demand is a random vector with a known mean and standard deviation. It was the first work that used the min-max approach to

2020 Mathematics Subject Classification. Primary: 90-02, 90C17, 90C47; Secondary: 90C90.
Key words and phrases. Distributionally robust optimization, uncertain decision-making, tractable methods, machine learning, operations research.

* Corresponding author: Zheming Gao.
solve an uncertain optimization problem with partially known distribution information. Following this pioneer work, a series of minimax analysis on SO problems \([13, 55–57, 157, 159]\) were studied afterwards. Ye and Delage \([47]\) firstly named this type of problems as distributionally robust stochastic problems (DRSP), which were later well-known as distributionally robust optimization (DRO) problems. As a combination of RO and SO, DRO possesses the advantages of both of them. Utilizing the distribution information, DRO inherits the randomness of SO, which helps get over the strong conservativeness limitation of RO. Moreover, DRO can break the limitation of knowing the exact distribution or avoid the error of estimating the true distribution which are required in SO. In addition, DRO benefits from the tractability of robust counterpart reformulation. Therefore, a huge amount of research related to DRO has been conducted in the past decades. We use the Web of Science Core Collection database to get the citation report for references in topic “minimax stochastic” and “distributionally robust optimization” from 1950 to 2021. The total number of publications is 1346, and the sum of times cited is 17478. Figure 1 shows that the total number of publications per year dramatically increases recently and reaches about 200 publications per year. Figure 2 describes the fast increasing trend of the sum of times cited which measures up around 3000 citations per year.

![Figure 1. Total publications per year for DRO references.](image1)

![Figure 2. Sum of times cited per year for DRO references.](image2)

Significant progress has been made in the research of DRO, both in theory and in practice. In theory, some underlying works combining modeling and solving analysis in the context of DRO were proposed. For example, Goh and Sim \([77]\) designed
a general framework for linear DRO problems with flexible piecewise linear decision rules. Inspired by the benchmark research of RO [8, 18], Wiesemann et al. [186] proposed a framework that provides a canonical form of tractable DRO problems. Being reformulated as linear, second order conic, or semidefinite programs, the tractable DRO problems can be solved by using well-known polynomial-time algorithms. In addition, Bertsimas et al. [19] developed a modular and tractable framework for solving an adaptive distributionally robust linear optimization problem. With delightful theoretical properties and computational advantages, DRO has been applied in many research fields such as machine learning [6, 12, 27, 28, 36, 54, 66, 88, 108, 147, 161, 190], operations research [3, 47, 73, 96, 109, 112, 122, 134, 142, 173], and other fields [40, 52, 148, 180, 181, 194, 195].

1.1 Motivation and contributions. In this paper, we provide an overview on both the theory and applications of DRO from the 1950s up to now. Similarly to SO and RO, DRO plays an essential role in solving optimization problems under uncertainty. However, most of the existing review papers consider SO [22, 65, 89, 95] and RO [6, 11, 20, 67], but a few focuses on DRO.

Most of the DRO review papers focus on some specific topics. For example, Bayraksan and Love [4] surveyed the DRO problems over $\phi$-divergence-based ambiguity sets. Postek et al. [135] summarized and analyzed DRO problems with risk measures. Kun et al. [100] recently published a survey of Wasserstein DRO with theory and applications in machine learning. Shapiro [156] presented a tutorial on modeling and solving multistage SO problems via the distributionally robust and risk averse approaches. Rahimian and Mehrotra [138] provided a comprehensive review of the main concepts and contributions in DRO. Particularly, they focused on risk-aversion and chance-constrained optimization problems. Unlike the existing articles, this paper focuses on providing a comprehensive review on the theory and applications of DRO. Our main contributions are listed as follows.

- We summarize the connections among RO, SO, and DRO. We illustrate the theory of DRO following the logic of modeling problems, solving algorithms, and evaluating models. We also summarize the applications of DRO in machine learning, operations research, and other areas.
- We present a detailed discussion to guide the ambiguity sets construction of DRO. The tractable reformulations are illustrated over different types of ambiguity sets. Moreover, some classical examples are given to help readers capture the principles.
- We discuss how the DRO problems are solved efficiently via various methods including numerical methods on SIP, convex reformulations, and approximation algorithms. Furthermore, we provide the performance guarantee and convergence analysis for some DRO problems.

1.2 Organization. The rest of the paper is organized as follows. In Section 2, we introduce the basic SO, RO, and general DRO models, and the relationships among them. We also discuss the advantages of DRO over SO and RO. Section 3 presents the ambiguity set construction of DRO. Three main different types ambiguity sets and some special sets are illuminated. Additionally, the tractable reformulations over different ambiguity sets are presented in this section. We discuss the solution techniques and performance analysis in Section 4. Applications of DRO focusing on machine learning and operations research are introduced in Section 5. In Section 6, we provide some future research directions of DRO.
2 General DRO models. In this section, we introduce the general optimization models used in this paper. Let \( x \in \mathcal{X} \subseteq \mathbb{R}^n \) be the decision variable. The objective function is defined as \( f(\mathbf{x}) : \mathcal{X} \rightarrow \mathbb{R}; \ g(\mathbf{x}) : \mathcal{X} \rightarrow \mathbb{R}^m \) is defined as a vector of functions, where \( g(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}))^T; \ g_i(\mathbf{x}) : \mathcal{X} \rightarrow \mathbb{R}, i = 1, \ldots, m. \) A general optimization problem has the form

\[
\inf_{x \in \mathcal{X}} \{ f(\mathbf{x}) \mid g(\mathbf{x}) \leq 0 \}. \tag{1}
\]

When an uncertain parameter \( \xi \in \mathbb{R}^k \) is involved, problem (1) becomes an uncertain optimization problem. If \( \xi \) is a random vector following a known distribution, then stochastic optimization is typically used to solve the uncertain optimization problem. Furthermore, if we do not know the exact distribution of \( \xi \), but know a set of its possible distributions, DRO takes the responsibility. The notations introduced in this section will be used in the remaining of this paper.

2.1 Stochastic optimization models. Given an outcome space \( \Xi \) and its \( \sigma \)-algebra \( \mathcal{F} \subseteq 2^\Xi \), a function \( P : \mathcal{F} \rightarrow \mathbb{R} \) is a probability measure if 1) \( P(\Xi) = 1 \); 2) for all \( A \in \mathcal{F}, P(A) \geq 0 \); 3) \( P(A^c) = 1 - P(A) \); 4) \( P(\bigcup_i A_i) = \sum_i P(A_i) \) for all countable collections of disjoint sets \( \{A_i\} \), with \( A_i \in \mathcal{F} \) for all \( i \). Probability measures are also called probability distributions, or just distributions. The triplet \((\Xi, \mathcal{F}, P)\) is called a probability space. Denote \( \mathcal{B} \) as Borel \( \sigma \)-algebra. Let a \( \mathcal{F}/\mathcal{B}(\mathbb{R}) \)-measurable function \( \xi : \Xi \rightarrow \mathbb{R}^k \) be a random vector on the measure space \((\Xi, \mathcal{F})\).

The basic assumption of SO is that an accurate probabilistic description of the random variables is available, under the form of the probability distribution functions (pdf), cumulative density functions (cdf), or more generally, probability measures \( P \) [22].

Denote \( \mathcal{H}_P : \mathcal{E} \rightarrow \mathbb{R} \) a real-valued functional associated with \( P \) from a linear space \( \mathcal{E} \) of measurable functions on \((\Xi, \mathcal{F})\) to the real space [138]. \( \mathcal{H}_P \) qualifies the uncertainty, which can be chosen to different functional forms according to the real applications including statistical learning, stochastic optimal control, and so on. A general stochastic optimization model can be formulated as follows:

\[
\inf_{x \in \mathcal{X}} \{ \mathcal{H}_P[f(\mathbf{x}, \xi)] \mid \mathcal{H}_P[g(\mathbf{x}, \xi)] \leq 0 \}. \tag{2}
\]

When \( \mathcal{H}_P \) is defined as the expectation function in the form of \( \mathcal{H}_P(\cdot) = \mathbb{E}_P(\cdot) \), it becomes a classical SO model

\[
\inf_{x \in \mathcal{X}} \{ \mathbb{E}_P[f(\mathbf{x}, \xi)] \}, \tag{3}
\]

or

\[
\inf_{x \in \mathcal{X}} \{ f(\mathbf{x}) \mid \mathbb{E}_P[g(\mathbf{x}, \xi)] \leq 0 \}. \tag{4}
\]

Under the expected-value qualification of uncertainty, one can easily find that if let \( g_i(\mathbf{x}) = 1\{\xi \in S_i(\mathbf{x})\}, i = 1, \ldots, m \) be the indicator functions, then we have \( \mathbb{E}_P[g_i(\mathbf{x}, \xi)] = P(\xi \in S_i(\mathbf{x})) \). The SO model (4) becomes a typical chance-constrained problem

\[
\inf_{x \in \mathcal{X}} \{ f(\mathbf{x}) \mid P[\xi \in S_i(\mathbf{x})] \leq 0, i = 1, \ldots, m \}. \tag{5}
\]

Readers are recommended to read well-developed books and survey papers of SO [22, 65, 89, 95] for more information. SO is a powerful tool to solve the uncertain optimization problems with random parameters, however, it has two inherent drawbacks. First, the multi-dimensional integrals involved in the computation of
expectations are generally intractable. Second, a true distribution is always hard to be estimated. In addition, SO may yield over-fitted decisions when using a distribution estimated with bias, and this may lead to unpleasant out-of-sample performance [94]. Even a distribution estimated without bias can still lead to biased objective values, which is known as the optimizer’s curse [163]. In decision analysis with uncertainty, an optimal solution is determined when the highest expected value is achieved. The values are estimates that are subject to random error. Researchers have found that even if the value estimates are unbiased, the optimality results are highly biased because of the uncertainty in the estimates coupled with the optimization-based selection process [100, 163]. This phenomenon is so called “optimizer’s curse”, or “optimization bias”. The following is an example to illustrate the bias.

Example 2.1. Optimizer’s curse [100].
Consider a decision problem under uncertainty. Let \( l(\xi) : \Xi \rightarrow \mathbb{R} \) be a measurable extended real-valued loss function. Assume the random vector \( \xi \) captures all uncertain risk factors and follows an unknown distribution \( P \). Let \( L \) denote the feasible set of all available loss functions. Let \( \mathcal{V}(P, l) = \mathbb{E}_P[l(\xi)] \) be the risk of decision \( l \) which is the expected loss under \( P \). Then the optimal risk is

\[
\mathcal{V}(P, L) = \inf_{l \in L} \mathcal{V}(P, l). \tag{6}
\]

Let \( \hat{P} \) be an unbiased estimator of \( P \), that is, \( \mathbb{E}_{P^N}[\hat{P}] = P \), where \( P^N \) is the distribution derived from any \( N \) independent training samples generated by \( P \). Then the risk of the loss function \( l \) is

\[
\mathbb{E}_{P^N}[\mathcal{V}(\hat{P}, l)] = \mathbb{E}_{P^N}[\mathbb{E}_P[l(\xi)]] = \mathbb{E}_P[l(\xi)] = \mathcal{V}(P, l), \tag{7}
\]

which implies that \( \mathcal{V}(\hat{P}, l) \) is an unbiased estimator of \( \mathcal{V}(P, l) \). For the optimal risk, we have

\[
\mathbb{E}_{P^N}[\mathcal{V}(\hat{P}, \mathcal{L})] = \mathbb{E}_{P^N}\left[\inf_{l \in \mathcal{L}} \mathcal{V}(P, l)\right] \leq \inf_{l \in \mathcal{L}} \mathbb{E}_{P^N}[\mathcal{V}(P, l)] = \inf_{l \in \mathcal{L}} \mathcal{V}(P, l) = \mathcal{V}(P, \mathcal{L}), \tag{8}
\]

which implies that \( \mathcal{V}(\hat{P}, \mathcal{L}) \) is an optimistically biased estimator of \( \mathcal{V}(P, l) \). Therefore, even \( \hat{P} \) is an unbiased estimator of \( P \), the corresponding estimator \( \mathcal{V}(\hat{P}, \mathcal{L}) \) underestimates the true risk \( \mathcal{V}(P, \mathcal{L}) \) after the minimization process.

2.2 Robust optimization models. As a relatively new approach for optimization problems with uncertainty, RO considers an uncertainty model over a deterministic and set-based region rather than following a stochastic distribution. The uncertain parameter \( \xi \in \mathbb{R}^k \) is assumed to take arbitrary values in the uncertainty set \( \mathcal{U} \subseteq \mathbb{R}^k \), which is also referred to as ambiguity set. The goal is to compute the minimal cost solution \( x^* \) among all the feasible solutions for all the realizations of \( \xi \in \mathcal{U} \). To reach this goal, the following optimization problem is formulated:

\[
\inf_{x \in X} \{ f(x) \mid g(x, \xi) \leq 0, \ \forall \ \xi \in \mathcal{U} \}. \tag{9}
\]

If \( \mathcal{U} \) is a continuous set, then problem (9) will have infinite number of constraints, and it can be treated as a semi-infinite programming problem. Referring to literature [81, 86, 167], semi-infinite programming (SIP) problems are hard to solve. However, RO coincides with the idea that the decision maker is fully responsible for the consequences of the decisions to be made for any observation revealed from
Therefore, RO pursues an optimal solution that can minimize the cost even in the worst-case, that is, a robustest optimal solution. It leads to the following basic formulation of RO:

$$\inf_{x \in X} \sup_{\xi \in \mathcal{U}} \{f(x, \xi) \mid g(x, \xi) \leq 0\}. \quad (10)$$

Readers are referred to [11, 20, 67] for more details about the theory and applications of RO.

One of the biggest advantages of RO is its tractability, which depends on the baseline optimization problem and the ambiguity set. There are many RO problems that can be reformulated into linear programming (LP), second-order cone programming (SOCP), or semidefinite programming (SDP) problems. However, RO has disadvantages, either. In some cases, the deterministic uncertainty set may give a very rare worst case, such that the corresponding result is too conservative to be utilized in practice. Without losing the advantage of tractability, DRO provides a feasible approach to effectively address the conservativeness limitation of RO.

### 2.3 Distributionally robust optimization models

Distributionally robust optimization can be regarded as a complementary approach provided by robust optimization and stochastic optimization. It adopts the worst-case approach in RO and the probability information utility in SO. DRO aims to find an optimal solution that can perform well even in the worst case over an ambiguity set whose elements are possible distributions. This distributionally ambiguity set helps to overcome the difficulty of estimating an underlying distribution and the error derived by naive reliance on a single probabilistic model in SO, as well as the conservativeness of ignoring statistical information in RO. By adopting a worst-case approach, DRO regularizes the optimization problem and thereby mitigates the optimizers curse characteristic attached with SO. It can also inherit counterpart reformulation techniques of RO. Some problems formulated in SO may not be solved efficiently, even when the data-generating distribution is known. However, they can be solved in polynomial time with DRO formulations in some cases [61].

Denote $\mathcal{P} \subseteq \mathcal{M}(\Xi, \mathcal{F})$ as the ambiguity set consisting of all possible distributions of the random vector $\xi \in \Xi$. Given the objective function $f(x, \xi)$ with decision variable $x$ and the restriction functions $g(x, \xi)$, a general DRO model can be formulated as follows:

$$\inf_{x \in X} \sup_{P \in \mathcal{P}} \{\mathcal{H}_P[f(x, \xi)] \mid \sup_{P \in \mathcal{P}} \mathcal{H}_P[g(x, \xi)] \leq 0\}. \quad (11)$$

Similarly, the classical models are formulated as follows when $\mathcal{H}_P$ takes the expectation function with respect to $P$:

$$\inf_{x \in X} \sup_{P \in \mathcal{P}} \{E_P[f(x, \xi)]\}. \quad (12)$$

If the true distribution is given, that is, $\mathcal{P}$ is a singleton, then DRO reduces to SO. In addition, when an optimal distribution can be obtained by solving the inner worst-case problem over $\mathcal{P}$, the outer minimization problem reduces to an SO problem (see Section 4.3).

Moreover, if we consider ambiguity sets of the form $\mathcal{P} = \{P \subseteq \mathcal{M}(\Xi, \mathcal{F}) \mid P(\xi \in \Xi) = 1\}$, then the distributionally robust counterpart recovers the classical robust counterpart. DRO therefore constitutes a true generalization of the classical RO paradigm. Some specific DRO problems can also be proved to be equivalent to a RO problem. Dimitris Bertsimas et al. [19] proved that the adaptive distributionally
robust linear optimization problem can be formulated as a classical RO problem. Xu et al. [191] built up a connection between RO and DRO. They showed that RO can be equivalently reformulated as a DRO problem with respect to a special class of distributions supported in the general Euclidean space. This general equivalence is of significance in the DRO interpretation of RO.

The ambiguity set $\mathcal{P}$ will essentially decide the tractability of the DRO models and the quality of the final results. The following two sections focus on the analysis of different types of ambiguity sets and their associated tractable reformulations. Solution methods and model performance are also discussed.

3 Ambiguity sets and tractability. In this section, we first review different types of ambiguity sets utilized in the literature. The ambiguity set $\mathcal{P}$ is a distribution family constructed by limited distribution information of the random vector $\xi$.

The probability measure $P$ on the measurable space $(\Xi, \mathcal{F})$ induced by a random vector $\xi$ is uniquely determined by its distribution function [136]. It indicates that the exact distribution can be regarded as the “complete” data information of a random vector. While the distribution $P$ is often unknown, some specific properties of $P$ are still able to be extracted from existing support set information or statistical analysis, for example, estimation of means and covariances from historical data.

The shape of distributions can provide a good approach to model the statistical properties of a population [166]. A moment is a specific quantitative measure of the shape of a function in mathematics and statistics. The first moment is known as expectation that provides information of the location of a dataset. The second moment related to variance describes the variability or the spread dispersion. Furthermore, the third moment is used to define the skewness of a distribution, which is the measure of the symmetry of the shape. The fourth moment defines the kurtosis of a distribution, which measures the flatness or peakness of a distribution. Therefore, an ambiguity set can be constructed by sharing known moment information. This type of ambiguity set $\mathcal{P}_M$ is called “moment-based ambiguity sets”, which will be discussed in section 3.1.

The main downside of the moment-based ambiguity sets is the possibility of containing some unrealistic distributions that consequently leads to overly pessimistic. Recent work has declared a way to weaken the pessimism by restricting $\mathcal{P}_M$ to contain possible distributions under some structural requirements [134, 172]. The further structural information includes symmetry, unimodality, monotonicity and so on. We will introduce the “structural ambiguity set” $\mathcal{P}_{St}$ in Section 3.2.

Moments have their limitations to capture enough distribution information. If a nominal or baseline distribution $\hat{P}$ is available, a better way to construct the ambiguity set is to form a neighborhood of $\hat{P}$ that contains the true distribution with a high probability. The neighborhood consists of three main elements: the center $\hat{P}$, the distance measure between two distributions, and a radius or a perturbation size $\epsilon_D$. We will discuss the “distance-based ambiguity sets” $\mathcal{P}_D$ in Section 3.3. Compared with the moment-based ambiguity set, the distance-based ambiguity set can be more accurate in depicting the profile of the ambiguous probability distribution and can provide a less conservative model.

The analysis of ambiguity sets focuses on the following aspects.
Clarify the partial information of the problem. Researchers should figure out which level of distribution information is preferred according to the research interest, data property, and problem settings.

Construct an ambiguity set \( \mathcal{P} \). If qualified and certificated moment information can be given, see Section 3.1; if structural properties can be described properly, see Section 3.2; otherwise, distance-based ambiguity sets in Section 3.3 will be a good choice. Section 3.4 can guide the construction of some special ambiguity sets.

Evaluate the ambiguity set and parameter design. An effective ambiguity set should follow the principles [61,84]:

1. \( \mathcal{P} \) contains the true distribution \( P \) at least with a high probability \( 1 - \delta \), that is, \( P(P \in \mathcal{P}) > 1 - \delta \), where \( \delta \) is the confidence level.
2. A favorable \( \mathcal{P} \) should facilitate a tractable (approximated) reformulation of the DRO counterpart. Examples of classical optimization models and statistic problems are given in each subsection to illustrate the reformulation tractability.
3. If \( \mathcal{P} \) satisfies the principles 1 and 2, its size should be chosen as small as possible. This rule helps exclude pathological distributions and avoid overly conservative decisions.
4. It is easy to collect parameters for designing an ambiguity set. If the problem is data-driven, for example, the mean and variance should be easily calculated from the historical data.

3.1 Moment-based ambiguity sets. Moment-based ambiguity sets contain all distributions that satisfy certain moment constraints, which appear to display better tractability properties. There is growing evidence that distributionally robust models with moment-based ambiguity sets are more tractable than the corresponding stochastic models because the intractable high-dimensional integrals in the objective function are replaced with tractable (generalized) moment problems.

Xu et al. [192] provided a general discussion about the DRO models under moment-based ambiguity sets. Consider a general moment-based ambiguity set

\[
\mathcal{P}_{MG} = \left\{ P \in \mathcal{M}(\Xi, \mathcal{F}) \mid \begin{array}{c}
\mathbb{E}_P[\Psi_i(\xi)] = 0, \ i = 1, \ldots, p \\
\mathbb{E}_P[\Psi_i(\xi)] \preceq 0, \ i = p + 1, \ldots, q \\
\mathbb{E}_P[1] = 1
\end{array} \right\},
\]

where \( \Psi_i : \Xi \to \mathbb{R}_+^{n_i \times n_i}, i = 1, \ldots, q \) is a symmetric matrix or a scalar with measurable random components. As we will see in the later discussion, \( \Psi_i \) can take some special forms to derive different types of moment-based ambiguity sets in the literature.

Consider the inner maximization problem of (12) when \( \mathcal{P} = \mathcal{P}_{MG} \),

\[
\sup_{P \in \mathcal{U}_+} \mathbb{E}_P[f(x, \xi)] \\
\text{s.t.} \quad \mathbb{E}_P[\Psi_i(\xi)] = 0, \ i = 1, \ldots, p \\
\mathbb{E}_P[\Psi_i(\xi)] \preceq 0, \ i = p + 1, \ldots, q \\
\mathbb{E}_P[1] = 1,
\]

where \( \mathcal{U}_+ \) denotes the positive linear space of all signed measures generated by \( \mathcal{M}(\Xi, \mathcal{F}) \). Its Lagrangian dual problem is

\[
\inf_{\lambda_0, \Lambda_1, \ldots, \Lambda_q} \lambda_0 \\
\text{s.t.} \quad f(x, \xi) - \lambda_0 - \sum_{i=1}^q \Lambda_i \cdot \Psi_i(\xi) \leq 0, \forall \xi \in \Xi \\
\lambda_0 \in \mathbb{R} \\
\Lambda_i \geq 0, \ i = p + 1, \ldots, q.
\]
The dual problem (15) is a semi-infinite problem. Readers may refer to review papers and books of SIP for more information [76, 86, 154]. Hettich et al. [85] also provided useful numerical methods including exchange methods, discretization methods, and reduction methods to solve SIP. Furthermore, if Ξ is well-defined in some special forms, problem (15) could be a semidefinite programming (SDP) problem, which is tractable in many situations with the existence of polynomial algorithms with efficient implementations. People may refer to the review articles of SDP [116, 174] to learn the tractable algorithms. The reformulation models and their corresponding solution methods that are widely used in solving various DRO problems will be summarized in Section 4.

One of the key steps to solve (12) over $P_{MG}$ by taking advantage of the tractable optimization models is to evaluate the duality gap between (14) and (15). Literature shows that there is no duality gap if the support set Ξ is compact and $Ψ_i(\cdot)$ is continuous [158], or if the moment problem satisfies Slater type condition [152]. The following sections introduce different types of moment-based ambiguity sets in DRO study. The general set $P_{MG}$ in [192] covers most of them.

### 3.1.1 Exact moment information.

Markov ambiguity sets are known as the moment-based ambiguity sets with exact known means and variances. Scarf [145] assumed that the mean $μ_0$ and the standard variance $Σ_0$ are exactly known. The ambiguity set is denoted as $P_{MS}$.

$$P_{MS} = \left\{ P \in \mathcal{M}(Ξ, F) \mid \mathbb{E}_P[ξ] = μ_0, \mathbb{E}_P[(ξ - μ_0)(ξ - μ_0)^T] = Σ_0 \right\}. \quad (16)$$

One can easily see that $P_{MS}$ is a special case of $P_{MG}$.

For the newsvendor problem with uncertain demand, Scarf provided an explicit optimal solution for the optimal order quantity and proved that there exists a two point distribution achieving the worst case. Gallego and Moon [68] then extended this problem to the resource cases in which a second purchasing opportunity is allowed. Based on Scarf’s work [145], Yue et al. [199] considered different objective functions while the uncertainty information remains the same. Zhu et al. [203] also studied the Newsvedor problem by minimizing the regret under the ambiguity set constructed as a known mean and standard deviation. The above works gave a closed-form solution of the worst case as a point distribution. Popescu [134] explored the same ambiguity set based on the known mean and covariance information while they first used a projection to reduce the the optimization problem with multivariate distributions to a univariate mean-variance DRO problem. Natarajan et al. [122] derived exact and approximate optimal trading strategies for a distributionally robust utility model when the first and second moment information is given. In addition, they considered a box-type ambiguity set in the mean and variance matrix. Assuming information about their first two moments were given, Rujeerapaiboon et al. [143] derived probability bounds on the tails of a product of symmetric nonnegative random variables.

The strong reserve information of $P_{MS}$ yields a more tractable optimization model. Wagner [177] considered the binary ambiguous LP problem when certain moment information was known exactly. The linear DRO model in the work turned out to be equivalent to a deterministic integer linear program if the first moment information was given and an SOCP problem if the second moment information was given. Zymler et al. [205] provided a tractable SDP for distributionally robust
individual and joint chance-constrained problems when the first-order and second-order moments as well as the support of the uncertain parameters were given. Cheng et al. [38] solved a distributionally robust version of a quadratic knapsack problem by assuming that the first and second moment information was given. The binary constraints in their problem were challenging and they provided a tractable approach by utilizing the SDP relaxation. Liu et al. [113] considered a reward-risk ratio optimization problem where the ambiguity set was constructed through the prior moment information. They used the well-known entropic risk measure to construct an approximation of the semi-infinite constraints and then solved the latter by using an implicit Dinkelbach method.

Example 3.1 below demonstrates a classical statistical problem about optimal probability bound. This example shows a tractable DRO reformulation when computing the Chebyshev bounds.

**Example 3.1. **Chebyshev bounds over $\mathcal{P}_{MS}$ [172, 175].

In probability theory, based on the limited moment information, the Chebyshev inequality gives an upper bound on the tail probability of a univariate random variable $\xi \in \mathbb{R}$ with

$$
P(\xi - \mu \geq \kappa \sigma) \leq \begin{cases} 
\frac{1}{\kappa^2}, & \text{if } \kappa > 1, \\
1, & \text{otherwise}, 
\end{cases} 
(17)
$$

where $\kappa > 0$ is a constant, $\mu$ and $\sigma$ are the mean and standard deviation of $\xi$ with respect to $P$ respectively. The Chebyshev inequality provides upper bounds on the probability of a multivariate random vector $\xi$ that falls outside a prescribed confidence region $\Xi$ if only a few low-order moments are known. The best upper bound can be given by a worst-case probability problem $\sup_{P \in \mathcal{P}_{MS}} P(\xi \notin \Xi)$. Assume $\Xi$ is an open polyhedron representable as a finite intersection of open half spaces,

$$
\Xi = \{ \xi \in \mathbb{R}^k | a_j^T \xi < b_j, \ j = 1, \ldots, m \}, 
(18)
$$

where $a_j \in \mathbb{R}^k$, $b_j \in \mathbb{R}$, $\forall j = 1, \ldots, m$. Then the worst-case problem over $\mathcal{P}_{MS}$ is equivalent to a tractable SDP problem via the Lagrangian dual approach.

$$
\sup_{P \in \mathcal{P}_{MS}} P(\xi \notin \Xi) = \max_{\sum_{j=1}^m \lambda_j} \sum_{j=1}^m \left( \begin{array}{c} Z_j \\ z_j^T \lambda_j \end{array} \right) \preceq \left( \begin{array}{c} \Sigma_0 \\ \mu_0^T \mu_0 \end{array} \right),
(19)
$$

Problem (19) is an SDP problem which can be solved in polynomial time using the interior point method [197]. It can also be an inner maximization problem of many DRO problems with the similar setting.

3.1.2 Bounded moment information. Sometimes, it is challenging to get accurate moment information. Hence, we consider the situation that the exact first and second moment information is unknown but bounded by a set.

3.1.2.1 Conic moment-based ambiguity sets. Taking into account the knowledge of the distributions support and the confidence region for its mean and its second-moment matrix, Delage and Ye [47] used two constraints parameterized by $\gamma_1 \geq 0$
and $\gamma_2 \geq 1$ to represent the ambiguity set denoted as $\mathcal{P}_{MDY}$.

\[
\mathcal{P}_{MDY} = \left\{ P \in \mathcal{M}(\Xi, \mathcal{F}) \mid \begin{array}{c}
E_P[\xi - \mu_0]^T \Sigma_0^{-1} E[\xi - \mu_0] \leq \gamma_1, \\
E_P[(\xi - \mu_0)(\xi - \mu_0)^T] \leq \gamma_2 \Sigma_0
\end{array} \right\},
\] (20)

where $\mu_0$ and $\Sigma_0$ are given by using empirical estimation. Let $\Sigma_0 = I$, $\gamma_1 = 0$, and $\gamma_2 = \infty$, then $\mathcal{P}_{MDY}$ imposes exact mean $\mu_0$ as support constraints. If $\gamma_1 = 0$, and $\gamma_2 = \infty$, $\mathcal{P}_{MDY}$ relates closely to $\mathcal{P}_{MS}$. The worst case under $\mathcal{P}_{MDY}$ is hard to give a closed form as in Scarf's [145].

One can easily find the matrix form of $\Psi$ in $\mathcal{P}_{MG}$ to generate $\mathcal{P}_{MDY}$.

\[
E_P[\xi - \mu_0]^T \Sigma_0^{-1} E[\xi - \mu_0] \leq \gamma_1 \iff E_P\left[ \frac{\Sigma_0}{\gamma_1} (\xi - \mu_0)^T (\xi - \mu_0) \right] \geq 0.
\] (21)

The worst case in model (12) over $\mathcal{P}_{MDY}$ can be solved by the following model. Readers may refer to Lin's review work [110] for more details.

\[
\begin{align*}
\max_{P \in \mathcal{P}_{MDY}} & \int_\Xi f(x, \xi)dP(\xi) \\
s.t. & \int_\Xi dP(\xi) = 1 \\
& \int_\Xi (\xi - \mu_0)(\xi - \mu_0)^T dP(\xi) \leq \gamma_2 \Sigma_0 \\
& \int_\Xi (\xi - \mu_0)^T (\xi - \mu_0) \geq \gamma_1 \\
& dP(\xi) \geq 0.
\end{align*}
\] (22)

Delage and Ye [47] also proposed a dual approach to solve (22). With some mild assumptions on the convexity of the support set and the objective function, the dual problem (23) can be solved in polynomial time.

\[
\begin{align*}
\min_{r, t, Q, q} & \quad r + t \\
s.t. & \quad f(x, \xi) - r - \xi^T Q \xi - \xi^T q \leq 0, \forall \xi \in \Xi \\
& \quad (\gamma_2 \Sigma_0 + \mu_0 \mu_0^T) \cdot Q + \mu_0^T q + \sqrt{\gamma_1} \left| |\Sigma_0^{1/2}(q + 2Q\mu_0)| \right| \leq 0
\end{align*}
\] (23)

Furthermore, Delage and Ye [47] derived a confidence region for the mean and the covariance when solving the data-driven version of the same problem when $(\mu_0, \Sigma_0, \gamma_1, \gamma_2)$ is not given directly but estimated from data. Data-driven problems usually have plenty of historical data that can be used to estimate a confidence region of the true distribution. Let $\{\xi_i\}_{i=1}^N$ be the data generated independently and randomly according to an unknown distribution. A common assumption is that the true value of moment lies in the neighborhood of the empirical value. A common approach to give the empirical value is as follows [47].

\[
\begin{align*}
\hat{\mu} &= \frac{1}{N} \sum_{k=1}^N \xi_k, \\
\hat{\Sigma} &= \frac{1}{N} \sum_{k=1}^N (\xi_k - \hat{\mu})(\xi_k - \hat{\mu})^T, \\
\hat{\gamma}_1 &= \frac{1 - \alpha(\delta/4 - \beta(\delta/2))}{1 + \beta(\delta/2)}, \\
\hat{\gamma}_2 &= \frac{1 - \alpha(\delta/4 - \beta(\delta/2))}{1 + \beta(\delta/2)},
\end{align*}
\] (24)

where $\delta = 1 - \sqrt{1 - \delta}$. They proved that $\mathcal{P}_{MDY}$ is the $1 - \delta$ confidence region of the true distribution of $\xi$ when the number of samples, i.e. $N$, is large enough as follows.

\[
\begin{align*}
\hat{R} &= \sup_{\xi \in \Xi} \left\| \Sigma^{-1/2}(\xi_k - \hat{\mu}) \right\|_2, \\
N &> \max \left\{ \left( \frac{\hat{R}^2 + 2}{2} \right)^2 \left( 2 + \sqrt{2 \ln 4 / \delta} \right)^2, \left( \frac{8 + \sqrt{32 \ln 4 / \delta}}{\sqrt{\hat{R}^2 + \hat{R}}^4} \right) \right\}.
\end{align*}
\] (25)
Corollary 1. (Corollary 4 in [47]) Let \( \{\xi_k\}_{k=1}^N \) be a set of \( N \) samples generated independently and randomly by the distribution of \( \xi \). If \( N \) satisfies constraint (25) and \( \Xi \) is bounded, then with a probability greater than \( 1 - \delta \) over the choice of \( \{\xi_k\}_{k=1}^N \), the distribution of \( \xi \) lies in the \( \mathcal{P}_{MDY} \) with \( (\mu_0, \Sigma_0, \gamma_1, \gamma_2) \) set as in (24).

Delage and Ye’s arguments [47] were consolidated by So [164] with weaker moment conditions. Zhang et al. [200] considered the distributionally robust chance-constrained binary programs over the ambiguity set \( \mathcal{P}_{MDY} \) and equivalently reformulated the problem as a 0-1 SOCP problem. A more general second order cone (SOC) moment-based ambiguity set was studied by Bertsimas [19].

\[
\mathcal{P}_{SOC} = \left\{ P \in \mathcal{M}(\Xi, \mathcal{F}) \left| \begin{array}{c}
\mathbb{E}_P[\mathbf{G}\xi] = \mu, \\
\mathbb{E}_P[g_i(\xi)] \leq \sigma_i, i = 1, \ldots, r_2.
\end{array} \right. \right\},
\]

where parameters \( \mathbf{G} \in \mathbb{R}^{l_1 \times k} \), \( \mu \in \mathbb{R}^{l_1} \), and functions \( g_i : \mathbb{R}^k \to \mathbb{R} \). The support set \( \Xi \) is an SOC representable set and the epigraph of each \( g_i \), \( i = 1, \ldots, r_2 \),

\[
\text{epi } g_i = \left\{ (\xi, u) \in \mathbb{R}^{k \times l_2} \mid g_i(\xi) \leq u \right\}
\]

is an SOC representable set.

Ghosal and Wiesemann [74] studied the capacitated vehicle routing problem (CVRP) under a large class of moment-based ambiguity sets. They especially proposed an efficient generation schemes to solve the DRO CVRP over \( \mathcal{P}_{SOC} \).

3.1.2.2 Convex moment uncertainty sets. Ghaoni et al. [73] considered the partial information in which the mean and covariance matrix belong to the given convex sets, respectively. Their work was inspired by Bertsimas [13]’s result that uses SDP to find the bounds for probabilities under partial information.

\[
\mathcal{P}_{MGh} = \{ P \in \mathcal{M}(\Xi, \mathcal{F}) \mid P(\xi \in \Xi) = 1, (\mathbb{E}[\xi], \mathbb{E}[(\xi - \mu)(\xi - \mu)^T]) \in \mathcal{U} \},
\]

where \( \mathcal{U} \) is a convex set, for example, a box of lower and upper bounds on the components of the mean and the covariance matrix, or a convex hull of several known possible pairs. The tractability of the DRO over \( \mathcal{P}_{MGh} \) depends on the structure of this bounded convex set \( \mathcal{U} \). Denote the mean \( \mu = \mathbb{E}[\xi] \), and the covariance matrix \( \Sigma = \mathbb{E}[(\xi - \mu)(\xi - \mu)^T] \).

- Polyphonic uncertainty.
  - \( \mathcal{U}_P \) is the convex hull of the given vertices \( (\mu_1, \Sigma_1), \ldots, (\mu_l, \Sigma_l) \),

\[
\mathcal{U} = \text{Co}\{(\mu_1, \Sigma_1), \ldots, (\mu_l, \Sigma_l)\}.
\]

Another interesting case is when the mean and covariance matrix are subject to independent polytopic uncertainty.

\[
\mathcal{U}_{P_I} = \text{Co}\{\mu_1, \ldots, \mu_l\} \times \text{Co}\{\Sigma_1, \ldots, \Sigma_l\}.
\]

- Component-wise bounds.

\[
\mathcal{U}_B = \{ \mu \in \mathbb{R}^k, \Sigma \in S^{k \times k}_+ \mid \mu_- \leq \mu \leq \mu_+, \Sigma_- \preceq \Sigma \preceq \Sigma_+ \}.
\]

Goh and Sim [77] studied the linear optimization problem that allows for expectations of recourse variables in the constraint specifications. They specifically discussed three different categories when constructing the ambiguity distribution set. Assuming the mean having an uncertain support, when the support of the random variable, the covariance matrix, and the direction deviation were known, they gave the corresponding worst-case bound of the minimax stochastic model.
We show an example of DRO in portfolio optimization problem that considers different convex ambiguity sets.

**Example 3.2.** *The Value-at-Risk (VaR) over admissible portfolios [73].*

Consider a classical portfolio problem where the risk is the minimized objective and requires a lower bound of the total return.

\[
\min_{\mathbf{x} \in X} \quad \mathbf{x}^T \Sigma_0 \mathbf{x} \\
\text{s.t.} \quad \mu_0^T \mathbf{x} \geq \epsilon,
\]

(32)

where \(\mu_0, \Sigma_0\) are the mean and variance of the return vector, and \(\epsilon\) is a pre-defined lower bound on the mean return.

The Value-at-Risk (VaR) framework instead looks at the probability of losses. It can be implemented by the following problem,

\[
\min_{\mathbf{x} \in X} V(\mathbf{x}) = \kappa(\epsilon) \sqrt{\mathbf{x}^T \Sigma_0 \mathbf{x} - \mu_0^T \mathbf{x}},
\]

(33)

where \(\kappa(\epsilon)\) is a risk factor depending on the prior assumptions on the distribution of returns. Assume the distribution of the return vector \(\mathbf{x}\) is unknown and locates on an ambiguity set \(\mathcal{P}_{M^2}\), then the worst-case VaR problem becomes

\[
\min_{\mathbf{x} \in X} V_{M^2}(\mathbf{x}) = \min_{\mathbf{x} \in X} \sup_{(\mu, \Sigma) \in \mathbb{U}} \kappa(\epsilon) \sqrt{\mathbf{x}^T \Sigma \mathbf{x} - \mu^T \mathbf{x}}.
\]

(34)

- \(\mathbb{U}_P = \mathbb{U}_{P^I}\).

\[
V_{M^2-P^I}(\mathbf{x}) = \kappa(\epsilon) \sqrt{\max_{\Sigma \in \mathcal{C}(\Sigma_i, \nu_i)} \mathbf{x}^T \Sigma \mathbf{x} - \min_{\mu \in \mathcal{C}(\mu_i, \nu_i)} \mu^T \mathbf{x}}.
\]

(35)

Then the worst-case VaR problem (33) can be reformulated as

\[
\min_{\mathbf{x} \in X} \quad s - t \\
\text{s.t.} \quad \|\Sigma_{1/2}\|_2 \leq s, \quad 1 \leq i \leq l \\
\quad \|\mathbf{x}\|_2 \leq \mu + 1 \leq i \leq l.
\]

(36)

This problem is an SOCP problem which can be efficiently solved in polynomial time.

- \(\mathbb{U}_P = \mathbb{U}_B\). The inner problem \(V_{M^2-B}(\mathbf{x})\) can be reformulated as the following SDP problem, which can also be efficiently solved with existing methods.

\[
V_{M^2-B}(\mathbf{x}) = \max_{\mathbf{x}} -\mathbf{\xi}^T \mathbf{x} \\
\text{s.t.} \quad \mu \leq \mu \leq \mu, \\
\quad \Sigma \leq \Sigma \leq \Sigma \\
\quad \left(\begin{array}{cc}
\mathbf{\xi}^T & \kappa(\epsilon)2
\end{array}\right) \geq 0.
\]

(37)

3.1.2.3 *General moment constraints.* Shapiro and Ahmed [157] analyzed the ambiguity set \(\mathcal{P}\) with bounded measures. For any two probability measures \(P_1, P_2 \in \mathcal{M}(\Xi, \mathcal{F})\), \(P_1 \geq P_2\) if \(P_1(A) \geq P_2(A), \forall A \in \mathcal{F}\). Let

\[
\mathcal{P}_{MC} = \left\{ P \in \mathcal{M}(\Xi, \mathcal{F}) \left| \begin{array}{c}
P \preceq P \preceq \overline{P} \\
\mathbb{E}_P[g_i(\mathbf{\xi})] = b_i, \quad i = 1, \ldots, r \\
\mathbb{E}_P[g_i(\mathbf{\xi})] \leq b_i, \quad i = r + 1, \ldots, q.
\end{array}\right. \right\},
\]

(38)

where \(P, \overline{P}\) are two given nonnegative measures in the space of all real measures on \((\Xi, \mathcal{F})\). Note that \(P\) is nonnegative if \(P(A) \geq 0, \forall A \in \mathcal{F}\). And \(b_i \in \mathbb{R}, g_i : \Xi \to \mathbb{R},\)
i = 1, \ldots, q are real valued measurable functions. Then the worst case in (12) over \( \mathcal{P}_{MC} \) can be modeled as the following formulation
\[
\max_{P \succeq P \succeq \mathcal{P}} \int_{\mathcal{X}} f(x, \xi)dP(\xi)
\text{s.t.}
\int_{\mathcal{X}} dP(\xi) = 1
\int_{\mathcal{X}} g_i(\xi)dP(\xi) = b_i, \ i = 1, \ldots, r
\int_{\mathcal{X}} g_i(\xi)dP(\xi) \leq b_i, \ i = r + 1, \ldots, q.
\] (39)

Note that when the measure constraint is replaced by \( P \succeq 0 \), then the above problem (39) becomes the classical problem of moments [104, 162, 172, 175]. Shapiro and Ahmed [157] showed that the problem \( \min_{x \in \mathcal{X}} \max_{P \in \mathcal{P}_{MC}} \mathbb{E}_P[f(x, \xi)] \) can be reformulated into an SO problem under mild assumptions:

(A1) The functions \( g_i(\cdot) \) are \( \overline{\mathcal{P}} \)–integrable, i.e., \( \int_{\mathcal{X}} |g_i(\xi)|d\overline{\mathcal{P}}(\xi) < \infty \), \( \forall i \);

(A2) There exists a \( P^* \), \( P \succeq P^* \succeq \overline{\mathcal{P}} \) such that \( \int_{\mathcal{X}} g_i(\xi)dP^*(\xi) = b_i \), \( \forall i \);

(A3) The problem (39) has finite optimal value and there exists a feasible solution for sufficiently small perturbations of all constraints.

(A1)-(A3) imply that we can find a probability measure \( Q \) such that \( P = P^* + Q \) for any feasible \( P \). The problem (39) is equivalent to the following problem (40) and its dual problem (41) without duality gap. Let \( Q = \overline{P} - P^* \) and \( \tilde{Q} = \overline{P} - P^* \), then we have
\[
\max_{Q \leq Q \leq \overline{Q}} \int_{\mathcal{X}} f(x, \xi)dP^*(\xi) + \int_{\mathcal{X}} f(x, \xi)dQ(\xi)
\text{s.t.}
\int_{\mathcal{X}} dQ(\xi) = 0
\int_{\mathcal{X}} g_i(\xi)dQ(\xi) = 0, \ i = 1, \ldots, r
\int_{\mathcal{X}} g_i(\xi)dQ(\xi) \leq 0, \ i = r + 1, \ldots, q.
\] (40)

and its Lagrangian dual problem can be computed as
\[
\min_{\lambda \in \mathbb{R}^{r+1}} \int_{\mathcal{X}} f(x, \xi)dP^*(\xi) + \int_{\mathcal{X}} [L_\lambda(x, \xi), \xi]_+d\overline{Q}(\xi) - \int_{\mathcal{X}} [-L_\lambda(x, \xi)]_+d\overline{Q}(\xi)
\text{s.t.}
\lambda_i \geq 0, \ i = r + 1, \ldots, q.
\] (41)

where \( L_\lambda(x, \xi) = f(x, \xi) - \lambda_0 - \sum_{i=1}^q \lambda_i g_i(\xi) \) and \( [a]_+ = \max\{a, 0\} \). Therefore, the DRO problem under an ambiguity set with the lower and upper bounds of the probability measures can be equivalently transformed into an SO problem. SO problems have mature algorithms and off-the-shelf software to use.

3.1.3 Nested moment-based ambiguity sets. Wiesemann et al. [186] proposed a standardized ambiguity set that contains possible distributions whose mean values reside on an affine manifold and confidence sets are conic representable. This type of ambiguity sets have been cited as nested moment-based ambiguity sets. The regularity conditions of nested moment-based ambiguity sets can ensure the tractability of these DRO problems. A nested moment-based ambiguity set \( \mathcal{P}_{MN} \) has the following standard form.
\[
\mathcal{P}_{MN} = \left\{ P \in \mathcal{M}(\Xi \times \Omega, \mathcal{F} \times \mathcal{G}) \left| \begin{array}{c}
\mathbb{E}_P[A\xi + Bu] = \mathbf{b}
\mathbb{P}'((\xi, u) \in \mathcal{C}_i) \in [p^i, p^i], \ \forall i \in J,
\end{array} \right. \right\},
\] (42)

where \( P \) is a joint distribution of the random vector \( \xi \in \Xi \subseteq \mathbb{R}^k \) appearing in the objective function, and some auxiliary random vector \( u \in \Omega \subseteq \mathbb{R}^d \). \( \mathcal{F} \) and \( \mathcal{G} \) are the \( \sigma \)–algebras associated with \( \Xi \) and \( \Omega \), respectively. Let \( A \in \mathbb{R}^{m \times k}, B \in \mathbb{R}^{m \times d}, \mathbf{b} \in \mathbb{R}^M, p_j, p_i \in [0, 1], p_i \leq p_j, \forall i \in J, \) and \( J = \{1, \ldots, I\} \). For any \( i \in J \), let the confidence set \( \mathcal{C}_i \) be
\[
\mathcal{C}_i = \{(\xi, u) \in \Xi \times \Omega \mid C_i\xi + D_iu \leq K_i d_i\},
\] (43)
where $C_i \in \mathbb{R}^{L_i \times k}$, $D_i \in \mathbb{R}^{L_i \times d}$, $d_i \in \mathbb{R}^{L_i}$, and $X_i$ is a proper cone. $\Xi$ and $\Omega$ can be simply chosen as $\mathbb{R}^k$ and $\mathbb{R}^d$, respectively. Notice that the expectation equality will be void if $M = 0$, and there will be no auxiliary random vector if $d = 0$. The following two regularity conditions and another nesting condition [186] completely define the nested ambiguity set $P_{MN}$:

(R1) The confidence set $C_i$ is bounded and has probability one, that is, $P_i = P_i = 1$.

(R2) $\exists P \in P_{MN}, P((\xi, u) \in C_i) \leq P_i, P_i \leq P_i, \forall i \in J$.

(N) For $i, i' \in J, i \neq i'$, we have either $C_i \subseteq C_i'$, $C_i' \subseteq C_i$, or $C_i \cap C_i' = \emptyset$.

Condition (R1) ensures $\Xi \times \Omega \subseteq C_i$ and condition (R2) guarantees the existence of a distribution $P$ that satisfies the inequalities in (42) for a nondegenerate probability interval. These two regularization conditions ensures a tractable reformulation of the corresponding DRO problem via strong duality [90, 152]. Condition (N) is the nesting condition which stipulates a strict partial order on the confidence sets. $C_i \subseteq C_i'$ denotes that $C_i$ is compactly embedded in $C_i'$, i.e., the closure of $C_i$ is a compact subset of the interior of $C_i'$. Wiesemann et al. [186] listed efficient ways to check this relationship in some special cases. The nesting condition is important when writing the tractable reformulations of DRO problems with nested moment-based ambiguity sets.

**Theorem 3.1.** (Theorem 2 in [186]) Verifying whether the ambiguity set $P_{MN}$ is empty is strongly $NP$-hard, even if the specification of $P_{MN}$ does not involve any expectation conditions (i.e., $M = 0$) and there are only two confidence sets $C_1, C_2$ that satisfy $C_1 \subseteq C_2$ but $C_1 \not\subseteq C_2$.

Theorem 3.1 implies that if the condition (N) is violated, then DRO problems will be strongly $NP$-hard in the form $\sup_{P \in P_{MN}} E_P[g(x, \xi)] \leq \omega$, for example, the distributionally robust uncertainty quantification problems and chance-constrained programming problems.

Assume the constraint function can be written as $g_l(x, \xi) = \max_{i \in L} g_l(x, \xi)$, where $L = \{1, \ldots, L\}$, and the auxiliary functions $g_l(x, \xi) : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ have the form

$$g_l(x, \xi) = (S_i^T \xi + s_i)^T x + (t_i^T \xi + t_i),$$

where $S_i \in \mathbb{R}^{n \times k}$, $s_i \in \mathbb{R}^n$, $t_i \in \mathbb{R}^k$, and $t_i \in \mathbb{R}, \forall i$. Then, $g$ is convex and piecewise affine in the decision variable $x$ and the random vector $\xi$. With the above assumptions of $g$ and conditions (R1), (R2) and (N), the chance-constrained problem over $P_{MN}$ can be reformulated as a tractable problem as follows.

$$\sup_{P \in P_{MN}} E_P[g(x, \xi)] \leq \omega \iff \begin{cases} \textbf{b}^T \beta + \sum_{i \in J}(P_i \kappa_i + P_i \lambda_i) \leq \omega \\
 \sum_{i \in A(j)} A_i^T \phi_i + S_i^T x + t_i \leq \sum_{i \in A(j)} (\kappa_i - \lambda_i), \forall i \in J, l \in L \\
 C_i^T \phi_i + A_i^T \beta = S_i^T x + t_i, \forall i \in J, l \in L \\
 D_i^T \phi_i + B_i^T \beta = 0, \forall i \in J, l \in L, \end{cases}$$

where $\beta \in \mathbb{R}^M$, $\kappa, \lambda \in \mathbb{R}^I$, and $X_i^*$ is the dual cone of $X_i$, $\phi_i \in \mathbb{X}_i^*$, $i \in J, l \in L$. Readers may refer to [186] for a step-by-step reformulation procedure. Moreover, the tractable reformulation (45) can have more favorable forms with different design of the confidence sets.
- $\mathcal{C}_i$ are linear inequalities, (45) is referred to an LP;
- $\mathcal{C}_i$ are conic quadratic inequalities, (45) is referred to an SOCP;
- $\mathcal{C}_i$ are semidefinite inequalities, (45) is referred to an SDP.

The nested moment-based ambiguity set is a canonical ambiguity set that may cover many moment-based ambiguity sets in current DRO problems. Hanasusanto et al. [84] presented a unifying framework for the modeling and solving of distributionally robust uncertainty quantification problems and robust chance constrained programs. Theorem 3.1 provides them with a theoretical support. Wiesemann et al. then proposed a lifting technique [186] to encode higher-order moments information. The lifted nested moment-based ambiguity sets cover several ambiguity sets from the literature as special cases. Inspired by ideas from machine learning, Shang and You [151] designed a novel approach to construct nested moment-based ambiguity sets based on principal component analysis (PCA) and first-order deviation functions. They reformulated mixed-linear programming problems for industrial-scale process network planning and batch process scheduling problems that effectively leverage uncertainty and yield more profits.

3.2 Structural ambiguity sets. The moment-based ambiguity set $\mathcal{P}$, consisting of all distributions sharing known mean and variance, may contain the distributions that are not realistic in many applications. Consequently, it renders the moment constraints overly pessimistic [173]. Recent work has shown that this pessimism can be partially mitigated by restricting the ambiguity set $\mathcal{P}$ with some structural requirements. In this section, three popular types of structural ambiguity sets are discussed, which request the symmetry, unimodality and monotonicity of the distributions, respectively.

The structural ambiguity can be represented as

$$\mathcal{P}_{St} = \mathcal{P}_M \cap \mathcal{P}_S,$$

where $\mathcal{P}_M$ is often a moment-based ambiguity set and $\mathcal{P}_S$ is a distribution family with desired structural properties. And let $\mathcal{P}_{SS}, \mathcal{P}_{SU},$ and $\mathcal{P}_{SM}$ be the structural sets with symmetry, unimodality, and monotonicity respectively.

Assume $\mathcal{P}_S$ is a convex ambiguity set. A distribution $P_E \in \mathcal{P}_S$ is said to be an extremal distribution of $\mathcal{P}_S$ if it can not be expressed as a strict convex combination of any two distinct distributions in $\mathcal{P}_S$ [173]. Denote $\mathcal{P}_{SE}$ as the family of extremal distributions of $\mathcal{P}_S$. Popescu [133] and Hanasusanto et al. [84] described the structural properties via the Choquet representation, whereby every distribution $P \in \mathcal{P}_S$ can be written as a convex combination of extremal distributions of $P$.

The Choquet representation of a convex ambiguity set $\mathcal{P}_S$ will reduce the worst-case expectation problem over $\mathcal{P}_S$ to a related worst-case expectation problem over the standard simplex $\mathcal{P}_{NE}$, where $N_E = |\mathcal{P}_{SE}|$. Please refer to the book of Choquet theory for more details [131]. We will introduce three main structural properties including symmetry, unimodality, and monotonicity.

**Symmetry** [84]. Let $\mathcal{P}_{SS}$ be the set of all point symmetric distributions with center $p_c$. Thus, $P \in \mathcal{P}_{SS}$ if and only if $P(B) = P(2p_c - B)$ for all Borel set $B \in \mathcal{B}(\mathbb{R}^k)$. The extremal distributions of $\mathcal{P}_{SS}$ are

$$P_E(\xi) = \frac{1}{2} \delta_\xi + \frac{1}{2} \delta_{2p_c - \xi}, \; \xi \in \Xi,$$

where $\delta_\xi$ and $\delta_{2p_c - \xi}$ denote the Dirac distributions that place all probability mass on the point $\xi$ and $2p_c - \xi$ respectively.
Unimodality. Unimodality is one of the underlying structures of distributions. It can describe skewness and kurtosis. Assume the continuous distribution $P$ has a centre. Generally speaking, $P$ is unimodal if its probability density function is decreasing in the distance from the centre [173].

**Definition 3.2.** ($\alpha-$Unimodal distributions $[48, 173]$.) For any fixed $\alpha \in \mathbb{R}_+$, a distribution $P \in \mathcal{P}_{SU}$ is $\alpha-$unimodal with centre 0 if $t^\alpha P(B/t)$ is non-decreasing in $t \in (0, \infty)$ for every Borel set $B \in \mathcal{B}$. The corresponding radial $\alpha$-unimodal distributions are denoted as $\delta^\alpha_{[0,y]}$, $y \in \mathbb{R}^k$ such that $\delta^\alpha_{[0,y]}([0,ty]) = t^\alpha \forall t \in [0,1]$. And $\delta^\alpha_{[0,y]}$ are the extremal distributions of the $\alpha$-unimodal set $\mathcal{P}_{SU}$ (Corollary 1 in [172]). The unimodal property can then be described by a unique Choquet representation by the extremal distributions (Theorem 3 in [172]).

Monotonicity. Monotonicity is closely related to unimodality. While unimodality defines a non-decreasing function $t^\alpha P(B/t)$, the monotonicity describes the smoothness of the function by the number of continuous derivatives.

**Definition 3.3.** ($\gamma-$monotone distributions [173]$.) For any fixed $1 \leq \gamma \in \mathbb{N}$, a distribution $P \in \mathcal{P}_{SM}$ is called $\gamma$-monotone with centre 0 if $t^{\gamma+k-1}P(B/t)$ is $\gamma$-monotone in $t \in (0, \infty)$ for every Borel set $B \in \mathcal{M}$, that is, if $t^{\gamma+k-1}P(B/t)$ is $\gamma$ times differentiable and $(-i)^j(t^{\gamma+k-1}P(B/t))^{(j)}(t) \geq 0$, $\forall t > 0, i \in \{0, \ldots, \gamma\}$.

Now we provide an example to show how the unimodality affects the DRO results.

**Example 3.3. Gauss bounds via DRO over an unimodal ambiguity set [172].** We have shown how the Chebyshev bound can be obtained by the DRO approach over the exact moment-based ambiguity set $\mathcal{P}_{MS}$ in Example 3.1. The Gauss bound improves the Chebyshev bound by a factor of $\frac{4}{9}$ when $P$ is known to be unimodal, that is,

$\mathcal{P}((\xi - \mu) \geq \kappa \sigma) \leq \begin{cases} 
\frac{4}{9\pi^2}, & \text{if } \kappa \geq \frac{2}{\sqrt{3}}, \\
1 - \frac{2}{\sqrt{3}}, & \text{otherwise.}
\end{cases}$

(48)

The Gauss inequality can reduce the conservativeness of the bound than the Chebyshev inequality. Similarly, we consider the multivariate case in this example. Adding the unimodal requirement into the ambiguity set, we have $\mathcal{P}_{SU} = \mathcal{P}_{MS} \cap \mathcal{P}_{SU}$. Thus we add constraints of $\alpha$-unimodality derived by the Choquet representation into the model of Chebyshev bound (19). Then the worst case problem over $\mathcal{P}_{SU}$ is equivalent to a tractable SDP via the Lagrangian dual approach.

$$
\sup_{P \in \mathcal{P}_{SU}} P(\xi \notin \Xi) = \max \sum_{j=1}^m (\lambda_j - t_{j,0})
$$

s.t. $\left\|
\begin{bmatrix}
\mathbf{b}_j \lambda_j \\
t_{j,i} \lambda_j - \mathbf{a}^T \mathbf{z}_j \\
t_{j,i+1} \lambda_j - \mathbf{a}^T \mathbf{z}_j \\
t_{j,i+1} \lambda_j - \mathbf{a}^T \mathbf{z}_j \\
t_{j,i} - t_{j,i+1} \lambda_j
\end{bmatrix}
\right\| \leq \mathbf{a}^T \mathbf{z}_j + t_{j,i} \mathbf{a}_j \lambda_j, \quad j = 1, \ldots, m$

$$
\begin{bmatrix}
\mathbf{z}_j \\
\lambda_j
\end{bmatrix} \geq 0, \mathbf{a}^T \mathbf{z}_j \geq 0, t_{j,i} \geq 0, \quad j = 1, \ldots, m
$$

$\mathbf{z}_j \in \mathbb{R}^k, \mathbf{z}_j \in \mathbb{S}^k, \lambda_j \in \mathbb{R}, t_{j,i} \in \mathbb{R}^{l+1} \quad j = 1, \ldots, m$, $\mathbf{a}_j \in \mathbb{R}^{k+l}, \mathbf{a}_j \in \mathbb{S}^{k+l}, \lambda_j \in \mathbb{R}$.
where \( l = \lceil \log_2 k \rceil \), \( E = \{ i \in \{0, \ldots, l - 1 \} : \lceil k/2^i \rceil \text{ is even} \} \) and \( O = \{ i \in \{0, \ldots, l - 1 \} : \lceil k/2^i \rceil \text{ is odd} \} \). The problem (49) is an SDP problem. Detailed procedure of the dual problem deduction can be found in [172]. This model can be a inner maximization problem of many DRO problems with the similar setting.

Shapiro and Kleywegt [159] declared that under mild regularity conditions, the worst case over \( \mathcal{P}_{SU} \) (unimodal ambiguity sets) is obtained on its average distribution. They consider examples of the news vendor problem, the problem of moments, and problems involving unimodal distributions. Shapiro [153] considered a set containing unimodal distributions that satisfy some given support constraints. Under some conditions on \( f(x, \xi) \), they characterize the worst distribution as being the uniform distribution.

Grani et al. [84] studied the DRO approach in uncertainty quantification and chance constrained programming when the ambiguity set was constructed according to generalized moment bounds and structural properties such as symmetry, unimodality, and independence patterns. Explicit conic reformulations for tractable problem are given in their work.

Based on the ambiguity sets constructed from the first and second moments of its distribution, Parys et al. [172] used the DRO approach to study the Gauss inequalities in probability theory. The unimodal structural ambiguity set enable them to obtain a less pessimistic Gauss-type bound via reformulated SDP model solving. Parys [173] showed that the worst-case DRO problems with Unimodal or monotone structural properties can both be reformulated as SDP problems.

Kocyigit et al. [99] explored the distributional auction problem in which multiple bidders have their private value of each indivisible good and the agents collect the ensemble of all bidders value as a random vector following an unknown distribution. The unknown distribution is under a structural independence ambiguity set. They proved that the Vickrey auction is the unique optimal mechanism when the support set of the ambiguity set is a hypercube. Moreover, they also proved that the new proposed highest-bidder lotteries mechanism is a 2-approximation optimal mechanism over the Markov ambiguity set.

3.3 Distance-based ambiguity sets. Distance-based ambiguity set \( \mathcal{P}_D \) is a distribution family in which each distribution is close to a nominal distribution \( \hat{P} \) within distance \( \epsilon_D \). Three elements would be required to construct such an ambiguity set: the nominal distribution, the way to measure the distance between two distributions, and the size of this set.

We first discuss the nominal distribution \( \hat{P} \). In some cases, \( \hat{P} \) can be obtained by some prior knowledge or by expert experience. Under a data-driven setting, a set with \( N \) independent observations \( \hat{\Xi}^N = \{ \hat{\xi}^i \mid i = 1, \ldots, N \} \subseteq \Xi \) is often available. Hence, a sample-based empirical distribution \( \hat{P}^N \) can be adopted as \( \hat{P} \). A widely used discrete empirical probability distribution \( \hat{P}^N \) is defined below.

\[
\hat{P}^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi^i}, \tag{50}
\]

where \( \delta_{\xi^i} \) denotes the Dirac point mass at the \( i \)-th training sample \( \xi^i \).

Another option is the elliptical probability distribution introduced by Kuhn et al. [100].
Definition 3.4 (Elliptical probability distribution [100].) \( P = \mathcal{E}_g(\mu, \Sigma) \) is an elliptical probability distribution if it has a density function of the form \( f(\xi) = C \det(\Sigma)^{-\frac{1}{2}} g((\xi - \mu)^T \Sigma^{-1} (\xi - \mu)) \) with density generator \( g(u) \geq 0 \) for all \( u \leq 0 \), normalization constant \( C > 0 \), mean vector \( \mu \in \mathbb{R}^k \), and covariance matrix \( \Sigma \in S^k_{++} \).

For example, Gaussian distribution, logistic distribution and t-distribution are elliptical distributions. Researchers may find it convenient to set \( \hat{P}_N \) as an elliptical distribution with a structure-dependent density generator function \( g \),

\[ \hat{P}_N = \mathcal{E}_g(\hat{\mu}, \hat{\Sigma}), \quad (51) \]

where \( \hat{\mu} \) and \( \hat{\Sigma} \) are respectively the maximum likelihood estimators of the mean and the covariance matrix.

Notice that \( \hat{P}_N \) inevitably differs from the unknown true distribution \( P \) even it is constructed by using the most sophisticated statistical tools. The inherit estimation error of \( \hat{P}_N \) will be exposed in the solutions. It has been observed in the financial portfolio theory that the estimation errors are often amplified by the optimization process [41,119]. The optimizer’s curse (see Section 2.1) also states that even if the the distributional input parameters of a decision problem are unbiased, the optimization results tend to be optimistically biased. It is not wise to use an estimated distribution. In contrast, a ball centered by the nominal distribution is a better choice.

The authors in article [179] have discussed how to measure the distance between two distributions. Here, we will briefly introduce the probability metrics [120] and statistical distance that are used for the distance measurement of two distributions. We first give the definition of metrics as follows.

Definition 3.5. (Probability metric) A metric on the measure space \( \mathcal{M}(\Xi, \mathcal{F}) \) is a function \( D : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R} \) satisfying that, \( \forall P_1, P_2, P_3 \in \mathcal{M} \),

1. \( D(P_1, P_2) \geq 0 \);
2. \( D(P_1, P_2) = 0 \) if and only if \( P_1 = P_2 \);
3. \( D(P_1, P_2) = D(P_2, P_1) \);
4. \( D(P_1, P_2) \leq D(P_1, P_3) + D(P_2, P_3) \).

Statistical distances are not all metrics as some of them do not satisfy all of the properties. Particularly the statistical distance that satisfies 1 and 2 are known as divergences. Divergence also can measure the distance between two probability distributions, but it is weaker than a metric.

Definition 3.6. (Divergence) A divergence on the measure space \( \mathcal{M}(\Xi, \mathcal{F}) \) is a function \( D(\cdot | \cdot) : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R} \) satisfying that, \( \forall P_1, P_2 \in \mathcal{M} \),

1. \( D(P_1 | P_2) \geq 0 \);
2. \( D(P_1 | P_2) = 0 \) if and only if \( P_1 = P_2 \).

In the view of divergence, one should notice that \( D(P_1 | P_2) \) may not equal to \( D(P_2 | P_1) \), and the triangle inequality dose not hold.

Shapiro introduced how to construct an ambiguity set via distance-based approaches [155]. Let \( D \) be the well-defined statistical distance, then the ambiguity set \( \mathcal{P}_D \) can be defined as follows if the size \( \epsilon_D \in \mathbb{R}_+ \) and the nominal distribution \( \hat{P} \in \mathcal{M}(\Xi, \mathcal{F}) \) are given,

\[ \mathcal{P}_D = \{ P \in \mathcal{M}(\Xi, \mathcal{F}) \mid D(P | \hat{P}) \leq \epsilon_D \}. \quad (52) \]
We have discussed the methods to estimate $\hat{P}$. The radius $\epsilon_D$ determines the goodness of the ambiguity set $\mathcal{P}_D$. If $\epsilon_D = 0$, $\mathcal{P}_D$ is reduced to a singleton, that is, the nominal distribution $\hat{P}$, and thus DRO becomes SO. If $\epsilon_D \to \infty$, $\mathcal{P}_D$ consists of all possible probabilities, then DRO may reach to a pathological distribution. As discussed at the beginning of this chapter, the choice of $\epsilon_D$ should obey the following principles.

- $\epsilon_D$ should be large enough to contain the unknown true distribution with a high probability. In the meanwhile, $\epsilon_D$ should be chosen as small as possible to preclude shunned distributions.
- For data-driven problems, if historical data is adequate, $\epsilon_D$ is always chosen to be small. In contrast, if the data is insufficient or unreliable, $\epsilon_D$ is preferably to be large to avoid risk.

How to choose the best radius still remains an open problem. Esfahani and Kuhn [61] analyzed the impact of the Wasserstein radius $\epsilon_D$ on the out-of-sample performance in the context of a stylized portfolio selection problem and an uncertainty quantification problem.

In the following section, we will introduce the definition and properties of different metrics and divergence used in DRO, as well as their pros and cons. The tractability is discussed over each ambiguity set via the reformulated optimization model.

3.3.1 *Metric-based ambiguity sets.* Metric-based ambiguity sets contain all distributions that are close to a nominal or most likely distribution with respect to the prescribed probability metric. Various metrics have been explored in the literature, for instance, Wasserstein (or Kantorovich) distance, total variation metric, Hellinger metric, relative entropy distance, Prohorov metric and so on. The Wasserstein distance has recently become a popular one in machine learning. In the subsequent subsection, we will introduce the metric-based ambiguity sets defined by the Wasserstein distance, the total variation metric, and the Prohorov metric. The relative entropy distance is also known as the Kullback-Leibler (KL) divergence which is suitable to be discussed in the divergence-based ambiguity sets.

3.3.1.1 *Wasserstein-distance-based ambiguity sets.*

**Definition 3.7. (Wasserstein distance)** For any $r \geq 1$, let $\mathcal{M}(\Xi, \mathcal{F})$ be the set of all probability distributions $P$ supported on $\Xi$ satisfying $E_P[d(\xi_1, \xi_2)^r] = \int_\Xi d(\xi_1, \xi_2)^r P(d\xi) < \infty$. Let $P_1, P_2 \in \mathcal{M}(\Xi, \mathcal{F})$. The $r$-Wasserstein distance between $P_1$ and $P_2$ is defined as

$$D_W^r(P_1, P_2) = \left\{ \inf_{\pi \in \Pi(P_1, P_2)} \left( \int_{\Xi^2} \|\xi_1 - \xi_2\|^r \pi(d\xi_1, d\xi_2) \right) \right\}^{1/r}, \quad (53)$$

where $\| \cdot \|$ is a norm in $\mathbb{R}^k$, while $\Pi(P_1, P_2)$ denotes the set of all joint probability distributions of $\xi_1 \in \Xi$ and $\xi_2 \in \Xi$ with marginals $P_1$ and $P_2$, respectively.

The Wasserstein distance is a metric with properties of nonnegative, symmetric and subadditive. It vanishes only if $P_1 = P_2$ [176]. $D_W^r(P_1, P_2)$ is finite whenever both $P_1$ and $P_2$ have finite $r^{th}$-order moments. The Wasserstein-distance-based ambiguity set can be defined as follows:

$$\mathcal{P}^r_{DW} = \{ P \in \mathcal{M}(\Xi, \mathcal{F}) \mid D_W^r(P, \hat{P}) \leq \epsilon_D \}.$$

(54)
The Wasserstein distance between $P_1$ and $P_2$ is actually the minimum transportation cost for moving the probability mass from $P_1$ to $P_2$. The original problem can trace back to the Monge-Kantorovich problem [30, 97, 141, 176].

Wasserstein metric has been used to study the convergence of an empirical distribution $\hat{P}_N$ from the i.i.d. samples to the true distribution. Specifically, Barrio et al. [46] showed that under the Wasserstein metric the empirical distribution converges to the true distribution almost surely as the number of samples goes to infinity. A priori estimate of the probability that the unknown true distribution $P$ locates outside $\mathcal{P}_{DW}$ is given as follows,

**Theorem 3.8.** (Measure concentration [61, 64]) If the light-tailed distribution assumption holds, that is, $A = \mathbb{E}_P[\exp(||\xi||^a)] < \infty, a > 1$, we have

$$P_N\{D_W(P, \hat{P}_N) \geq \epsilon_D\} \leq \begin{cases} c_1 \exp(-c_2 N \epsilon_D^{\max\{k,2\}}), & \text{if } \epsilon_D \leq 1, \\ c_1 \exp(-c_2 N \epsilon_D^a), & \text{if } \epsilon_D > 1, \end{cases}$$

for all $N \geq 1$, $k \neq 2$, and $\epsilon_D > 0$, where $c_1, c_2$ are positive constants that only depend on $a$, $A$, and $k$.

Thus, given that $\mathcal{P}_{DW}$ contains $P$ with confidence $1 - \delta$ for some prescribed $\delta \in (0,1)$, the Wasserstein radius can be computed,

$$\epsilon_D(\delta) = \begin{cases} \left(\frac{\log(c_1/\delta)}{c_2^2}\right)^{1/\max\{k,2\}}, & \text{if } N \geq \frac{\log(c_1/\delta)}{c_2^2}, \\ \left(\frac{\log(c_1/\delta)}{c_2}\right)^{1/a}, & \text{if } N < \frac{\log(c_1/\delta)}{c_2}. \end{cases}$$

There are two main reasons that the Wasserstein metric is widely utilized in DRO.

- Wasserstein ambiguity sets $\mathcal{P}_{DW}$ can contain both discrete and continuous distributions no matter the nominal distribution is discrete or continuous. However, for example, the KL-divergence-based ambiguity sets (discussed in Section 3.3.2.1) fail to capture possible continuous distributions when the nominal distribution is discrete [61, 70, 79].

- Rigorous theoretical proofs [31, 64] have shown that the unknown true distribution $P$ is guaranteed to be located in $\mathcal{P}_{DW}$ with a high probability $1 - \delta$ if its radius is sublinear to $\log(1/\delta)/N$. Besides, the data-driven DRO over $\mathcal{P}_{DW}$ has promising out-of-sample performance (see Section 4.5).

Although the Wasserstein ambiguity sets offer robust out-of-sample performance guarantees for decision-makers [61], the Wasserstein ambiguity sets and the moment-based ambiguity sets are complementary to each other, each with distinct advantages. Gao and Kleywegt [71] showed that DRO with moment-based ambiguity set performs better than DRO with Wasserstein ambiguity set in high-correlation regimes; in contrast, DRO with Wasserstein ambiguity set performs better in medium- and low-correlation regimes.

Most of the DRO problems with the Wasserstein ambiguity sets have tractable reformulations. Wang et al. [183] studied a data-driven DRO shortest path problem, in which the distribution of the travel time is located on a Wasserstein ball. The robust counterpart can be reformulated as a 0-1 mixed convex program, which can be solved by commercial solvers. Lee et al. [106] derived a closed form solution and an explicit characterization of the worst-case distribution for the data-driven DRO newsvendor model. The model was studied with a $p$-Wasserstein ambiguity set ($p \in [1, \infty]$). Ran and Lejeune [91] studied a group of chance-constrained DRO problems with data-driven Wasserstein ambiguity sets. The mixed integer LP
reformulation and the SOCP reformulation were provided. Here we demonstrate an example of the DRO problem with a piecewise affine loss function and Wasserstein ambiguity set.

Example 3.4. DRO over Wasserstein-distance-based ambiguity set with piecewise affine loss functions.

Given a loss function \( l(\xi) = \max_{1 \leq i \leq m} a_i(\xi) \), where \( a_i(\xi) = a_i^T \xi + b_i \), \( 1 \leq i \leq m \) are affine functions. Notice that the loss function \( l \) is decision-independent. Suppose that the support set is a polytope, i.e., \( \Xi = \{ \xi \in \mathbb{R}^k \mid \mathbf{C} \xi \leq \mathbf{d} \} \), where \( \mathbf{C} \) is a matrix and \( \mathbf{d} \) is a vector of appropriate dimensions. Then the inner worst-case problem \( \sup_{P \in \mathcal{P}_{DW}} \mathbb{E}_P[l(\xi)] \) can be reformulated as

\[
\inf_{\lambda, s, \mathbf{z}_j} \quad \epsilon_D \lambda + \frac{1}{N} \sum_{j=1}^N s_j \\
\text{s.t.} \quad b_i + a_i^T \xi + \mathbf{z}_j^T (\mathbf{d} - \mathbf{C} \xi) \leq s_j, \quad \forall 1 \leq j \leq N, 1 \leq i \leq m \\
\| G^T \mathbf{z}_j - a_i \|_* \leq \lambda, \quad \forall 1 \leq j \leq N, 1 \leq i \leq m
\]

(57)

where \( \| \cdot \|_* \) is the dual norm. The above problem is an SOCP problem which can be solved in polynomial time. The derivation of the dual problem is neglected, but readers can refer to convex optimization books [140] for more details.

DRO approaches over the Wasserstein ambiguity sets have attracted attentions in many fields. Some examples include automatic control [40, 148, 194, 195], energy systems [52, 180, 181], statistics [25, 72, 114, 126, 127, 147], and especially, machine learning. The applications of Wasserstein DRO approaches in machine learning will be discussed in Section 5.1. Kuhn et al. [100] recently also provided a survey of Wasserstein DRO with its theory and applications in machine learning.

3.3.1.2 Total-variation-metric-based ambiguity sets. We first present the definition of total variation metric (also known as \( l_1 \) distance) below.

Definition 3.9. (Total variation metric [2]) Let \( P_1, P_2 \in \mathcal{M}(\Xi, \mathcal{F}) \). The total variation metric between \( P_1 \) and \( P_2 \) is defined as

\[
D_{TV}(P_1, P_2) = \sup_{h \in \mathcal{M}} \{ \mathbb{E}_{P_1}[h(\xi)] - \mathbb{E}_{P_2}[h(\xi)] \},
\]

(58)

Notice that \( D_{TV} \) is equivalent to the Wasserstein distance (53) with the following norm.

\[
\| \xi_1 - \xi_2 \| = \begin{cases} 0, & \text{if } \xi_1 = \xi_2, \\ 1, & \text{if } \xi_1 \neq \xi_2. \end{cases}
\]

(59)

If the measurable functions in \( \mathcal{M} \) are restricted to be uniformly Lipschitz continuous, that is,

\[
\mathcal{M} = \left\{ h : \mathbb{R}^k \to \mathbb{R} \mid \sup_{\xi \in \Xi} |h(\xi)| \leq 1, L_1(h) \leq 1 \right\},
\]

(60)

where \( L_1(h) = \inf \{ L : |h(\xi') - h(\xi'')| \leq L |\xi' - \xi''|, \forall \xi', \xi'' \in \Xi \} \), then \( D_{TV}(P_1, P_2) \) is known as bounded Lipschitz metric [130]. Then the metric-based ambiguity set based on total variation metric can be formulated as

\[
\mathcal{P}_{D_{TV}} = \{ P \in \mathcal{M}(\Xi, \mathcal{F}) \mid D_{TV}(P, P^N) \leq \epsilon_D \}.
\]

(61)

We then check the tractability of the DRO based on the total-variation-metric-based ambiguity set. Considering the inner worst-case maximization problem (12)
over $\mathcal{P}_{DTV}$, i.e. sup$_{P \in \mathcal{P}_{DTV}} \{ \mathbb{E}_P[f(x, \xi)] \}$, we define an essential supremum [93] of a general cost function $f(x, \xi)$

$$\text{ess sup} f(x, \xi) = \inf \{ a \in \mathbb{R} \mid \mu(\{ \xi \in \Xi : f(x, \xi) > a \}) = 0 \}, \quad (62)$$

where $\mu(\cdot)$ is the Lebesgue measure. The ess sup$_{\xi \in \Xi} f(x, \xi)$ represents the almost surely worst-case cost $f(x, \xi)$ over set $\Xi$ with regard to the Lebesgue measure. Then we give the reformulation result [93,137,155] of sup$_{P \in \mathcal{P}_{DTV}} \{ \mathbb{E}_P[f(x, \xi)] \}$.

$$\sup_{P \in \mathcal{P}_{DTV}} \{ \mathbb{E}_P[f(x, \xi)] \} = \begin{cases} \epsilon_{D}\text{ess sup}_f(x, \xi), & \epsilon_D \geq 1, \\ \epsilon_{D}\text{ess sup}_f(x, \xi) + (1 - \epsilon_D)\text{CVaR}_{\hat{P}_D}^{\rho_N}[f(x, \xi)], & 0 \leq \epsilon_D \leq 1. \end{cases} \quad (63)$$

where CVaR$_{\hat{P}_D}^{\rho_N}[f(x, \xi)]$ represents the conditional value-at-risk of $f(x, \xi)$ with respect to $\hat{P}_D$ and confidence level $\epsilon_D$.

$$\text{CVaR}_{\hat{P}_D}^{\rho_N}[f(x, \xi)] = \min_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1 - \epsilon_D} \mathbb{E}_{\hat{P}_D}[f(x, \xi) - \eta]_+ \right\}. \quad (64)$$

Thus, we can see that the worst-case average cost over the total-variation-metric-based ambiguity set is a convex combination of the conditional value-at-risk and the worst-case cost under the nominal distribution. Then, the DRO problem over $\mathcal{P}_{DTV}$ can be reformulated as

$$\begin{cases} \min_{x \in X, \nu \in \mathbb{R}} \epsilon_{D}\sup_{\xi \in \Xi} f(x, \xi) + (1 - \epsilon_D)\nu + \mathbb{E}_{\hat{P}_D}[f(x, \xi) - \nu]_+, & 0 \leq \epsilon_D \leq 1, \\ \min_{x \in X, \nu \in \mathbb{R}} \sup_{\xi \in \Xi} f(x, \xi), & \epsilon_D \geq 1. \end{cases} \quad (65)$$

We can see that original DRO problems reduces to an SIP. The infinitely many constraints in this SIP is induced by $\xi \in \Xi$. Tractable reformulations can be obtained if $\Xi$ has special structures, for example, $\Xi$ is a finite set. Please refer to Section 4.1 for more efficient algorithms.

In a fundamental work [169], the convergence analysis of DRO problems with the total variation metric-based ambiguity was provided. Jiang and Guan [93] developed a risk-averse two-stage stochastic program with the total variation metric-based ambiguity set. They simplified the DRO problems to the convex combinations of a CVaR and an essential supremum and verified the convergence of the simplified model.

### 3.3.1.3 Prohorov-metric-based ambiguity sets.

**Definition 3.10. (Prohorov metric)** Let $P_1, P_2 \in \mathcal{M}(\Xi, \mathcal{F})$. The Prohorov metric between $P_1$ and $P_2$ is defined as

$$D_{Pr}(P_1, P_2) = \inf \{ \beta > 0 \mid P_1(B) \leq P_2(B^\beta) + \beta, P_2(B) \leq P_1(B^\beta) + \beta, \forall B \in \mathcal{F} \}, \quad (66)$$

where $B^\beta = \{ s \in \Xi \mid \inf_{s' \in B} ||s - s'|| \leq \beta \}$.

The Prohorov metric is essential as it metrizes weak convergence in probability theory [2,75]. The Prohorov metric-based ambiguity set is characterized by the following:

$$\mathcal{P}_{DPr} = \{ P \in \mathcal{M}(\Xi, \mathcal{F}) \mid D_{Pr}(P, \hat{P}) \leq \epsilon_D \}. \quad (67)$$
Lemma 3.11. Relationships between the Prohorov metric and other distances [60]. Given any $P_1, P_2 \in \mathcal{M}(\Xi, \mathcal{F})$,

- Prohorov and Wasserstein metrics: $D_{Pr}(P_1, P_2)^2 \leq D_W^1(P_1, P_2)$;
- Prohorov and Total variation metrics: $D_{Pr}(P_1, P_2) \leq D_{TV}(P_1, P_2)$

This lemma implies that, by adjusting the radius, the Prohorov-metric-based ambiguity set can provide a conservative approximation of the Wasserstein-distance-based ambiguity set and the total-variation-metric-based ambiguity set.

The chance-constrained problem is a classical problem of uncertain optimization. The following example illustrates how it can be solved efficiently using the DRO approach with a Prohorov ambiguity set.

Example 3.5. **The distributionally robust chance-constrained problem.** Given a deterministic optimization problem $\min_{x \in \mathcal{X}} \{ c^T x \mid x \in \mathcal{X}, g(x) \leq 0 \}$. Consider its ambiguous chance-constrained problem in DRO format over the Prohorov-metric-based ambiguity set.

$$
\min_{x \in \mathcal{X}} c^T x \quad \text{s.t.} \quad \sup_{P \in \mathcal{P}_{DRO}} P(g(x, \xi) > 0) \leq \delta. \tag{68}
$$

Denote the feasible set of the problem (68) as

$$\mathcal{X}_D(\delta) = \left\{ x \in \mathcal{X} \mid \sup_{P \in \mathcal{P}_{DRO}} P(g(x, \xi) > 0) \leq \delta \right\}. \tag{69}
$$

It has been proved that, with a high probability, the set $\mathcal{X}_D(\delta)$ can be approximated by the following simpler set under certain conditions [60]:

$$\left\{ x \in \mathcal{X} \mid g(x, \xi) \leq 0, \forall \xi \in \Xi \text{ s.t. } ||\xi - \xi_0|| \leq \epsilon_D, i = 1, \ldots, N \right\}, \tag{70}
$$

where $\xi_0$ are $N$ i.i.d. samples according to the nominal probability measure $\hat{P}^N$.

When $g(x, \xi)$ is an affine function, problem (68) can be reformulated as a tractable convex optimization problem.

$$\min_{x \in \mathcal{X}} c^T x \quad \text{s.t.} \quad g(x, \xi_0) \leq -\epsilon_D, \quad i = 1, \ldots, N \nonumber$$

$$|x_j| \leq t^1_j, \quad i = 1, \ldots, N, \quad j = 1, \ldots, n \nonumber$$

$$1 \leq t^1_j, \quad i = 1, \ldots, N \nonumber$$

$$||t||_* \leq y_i, \quad i = 1, \ldots, N \nonumber$$

$$y \in \mathbb{R}^N, t^i \in \mathbb{R}^{n+1}, \quad i = 1, \ldots, N \nonumber$$

$$x \in \mathcal{X}, \nonumber$$

where $|| \cdot ||_*$ denotes the dual norm of $|| \cdot ||$. For $l_1$ or $l_\infty$ norms, (71) reduces to an LP problem; for $l_q$, $1 < q \leq \infty$, (71) becomes an SOCP problem.

3.3.2. $\phi$-divergence-based ambiguity sets. Divergence plays an important role in measuring the distance between two distributions. $\phi$-divergence, also referred to as Phi-divergence or f-divergence in the literature, is the main divergence used in DRO. It adopts a $\phi$-function to define the divergence. Different implements of the $\phi$-function give different divergences.

Definition 3.12. ($\phi$-divergence [4, 92]) Let $\mathcal{M}(\Xi, \mathcal{F})$ be a probability space, and function $D_{\phi}: \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$. For any two probability distributions $P_1$ and $P_2$ in $\mathcal{M}$, the $\phi$-divergence of $P_1$ from $P_2$ is defined as

$$D_{\phi}(P_1 || P_2) = \mathbb{E}_{P_2} \left[ \phi \left( \frac{dP_1}{dP_2} \right) \right] = \int_{\Xi} \phi \left( \frac{dP_1}{dP_2} \right) dP_2, \tag{72}$$
where $\phi: \mathbb{R} \to \mathbb{R}$ is a convex function such that

(C1) $\phi(1) = 0$. 

(C2) $\phi(0) := \begin{cases} \lim_{t \to -\infty} \frac{\phi(t)}{t}, & \text{if } t > 0, \\ 0, & \text{if } t = 0. \end{cases}$

(C3) $\phi(t) = +\infty$ for $t < 0$.

(C1) implies that when $P_1 = P_2$, $D_\phi(P_1 || P_2) = 0$. Even though $\phi$-divergences can quantify distances between distributions, they are not metrics in general. Most $\phi$-divergences do not satisfy the triangle inequality, and many are not symmetric in the sense that $D_\phi(P_1 || P_2) \neq D_\phi(P_2 || P_1)$. One exception is the variation distance, which is equivalent to the $l_1$-distance between vectors.

Tables 1 and 2 show the common choices of $\phi$-functions and the corresponding adjoint functions, $\phi$-divergences and DRO counterparts. Note that $\phi^* = \phi$ is the adjoint function of $\phi(t)$. We have $D_\phi(P_1 || P_2) = D_{\phi^*}(P_2 || P_1)$. Table 1 describes their relationships and properties. When it comes to the conjugate dual problems, the conjugate function is important to know. $\phi^* : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is the conjugate function of $\phi(t)$ with $\phi^*(s) = \sup_{t \geq 0} \{st - \phi(t)\}$. $\phi^*(s)$ is a convex function. When $\phi(\cdot)$ is a proper closed convex function, we have $\phi^{**} = \phi$ and there is no duality gap [140]. Conjugate dual problems help a lot when we conduct the reformulation. Table 2 shows that the conjugate dual problem over the above $\phi$-divergence-based ambiguity sets are all convex problems that can be solved in polynomial time via varieties of existing algorithms and commercial software.

| Divergence          | $\phi(t)$ | $\hat{\phi}(t)$ | $\phi(t), t \geq 0$ | $D_\phi(P_1 || P_2)$ |
|---------------------|-----------|------------------|---------------------|----------------------|
| Kullback-Leibler    | $\phi_{KL}$ | $\hat{\phi}_B$ | $t \log t - t + 1$ | $\int_{\Xi} \log \left( \frac{dP_1}{dP_2} \right) dP_1$ |
| Burg entropy        | $\phi_B$  | $\phi_{KL}$     | $- t \log t + t - 1$ | $\int_{\Xi} \log \left( \frac{dP_1}{dP_2} \right) dP_2$ |
| J-divergence        | $\phi_J$  | $\phi_J$        | $(t - 1) \log t$    | $\int_{\Xi} \log \left( \frac{dP_1}{dP_2} \right) (dP_1 - dP_2)$ |
| $\chi^2$-divergence | $\phi_{\chi^2}$ | $\phi_{M\chi^2}$ | $\frac{1}{2} (t - 1)^2$ | $\int_{\Xi} \left( \frac{dP_1}{dP_2} \right)^2 dP_2$ |
| Modified $\chi^2$-divergence | $\phi_{M\chi^2}$ | $\phi_{\chi^2}$ | $(t - 1)^2$ | $\int_{\Xi} \left( \frac{dP_1}{dP_2} \right)^2 dP_2$ |
| Variation distance  | $\phi_V$  | $\phi_V$        | $|t - 1|$            | $\int_{\Xi} |dP_1 - dP_2|_2$ |
| Hellinger distance  | $\phi_H$  | $\phi_H$        | $(\sqrt{t} - 1)^2$ | $\int_{\Xi} \left( \sqrt{dP_1} - \sqrt{dP_2} \right)^2$ |

Table 1. Examples of $\phi$-functions, their adjoints and $\phi$-divergence [4].

Then, the $\phi$-divergence-based ambiguity set with radius $\epsilon_D$ can be defined as

$P_\phi = \{ P \in \mathcal{M}(\Xi, \mathcal{F}) \mid D_\phi(P || \bar{P}) \leq \epsilon_D \}$. \hspace{1cm} (73)

Given any two distributions $P_1$ and $P_2$, the following lemma compares the magnitude of each $\phi$-divergence in Table 1.

**Lemma 3.13. (Lemma 1 in [92])** The bounding relations among $\phi$-divergences can be described as follows:

1. Variation distance and Hellinger distance:

$D_{\phi-H}(P_1 || P_2) \leq D_{\phi-V}(P_1 || P_2) \leq 2D_{\phi-H}(P_1 || P_2)$. 

2. Hellinger distance and KL-divergence:

$D_{\phi-H}(P_1 || P_2) \leq \sqrt{D_{\phi-KL}(P_1 || P_2)}$. 


3. KL-divergence and J-divergence:
\[ D_{\phi-KL}(P_1||P_2) \leq D_{\phi-J}(P_1||P_2). \]

4. J-divergence and modified \(\chi^2\)-divergence:
\[ D_{\phi-J}(P_1||P_2) \leq D_{\phi-M\chi^2}(P_1||P_2). \]

Thus, we can derive the following inclusion relations among different \(\phi\)-divergences:
\[ \mathcal{P}_{\phi-M\chi^2}(\epsilon_D) \subseteq \mathcal{P}_{\phi-J}(\epsilon_D) \subseteq \mathcal{P}_{\phi-KL}(\epsilon_D) \subseteq \mathcal{P}_{\phi-H}(\sqrt{\epsilon_D}) \subseteq \mathcal{P}_{\phi-V}(2\epsilon_D^{1/4}). \]  

Similar to Lemma 3.11, the formula (74) allows us to construct approximated ambiguity sets by using one \(\phi\)-divergence for another.

DRO with \(\phi\)-divergence-based ambiguity sets has been widely used in classical uncertain optimization problems. Bayraksan and Love [4] presented distributionally robust two-stage models, especially for risk-averse optimization problems, by using \(\phi\)-divergence-based ambiguity sets. They also offered concluding marks for a classification of \(\phi\)-divergence. This classification results in four types of \(\phi\)-divergences that depend on the data properties and decision-making rules. Jiang and Guan [92] studied the classical stochastic chance-constrained problems in DRO setting with \(\phi\)-divergence-based ambiguity sets. In addition, over \(\phi\)-divergence-based ambiguity sets, Ben-Tal et al. [5] analyzed the distributionally robust linear optimization problems; Namkoong and Duchi [121] applied stochastic gradient method to the distributionally robust empirical risk minimization problems. Recently, Duchi and Namkoong [54] developed a DRO framework based on \(\phi\)-divergence-based ambiguity sets to minimize the total loss in general machine learning models. Great progress in \(\phi\)-divergence-based DRO indicates the following pros and cons.

- **Pros of \(\phi\)-divergence-based ambiguity sets in DRO.**
  1. \(\phi\)-divergence is a common measure tool in statistics. For example, it can be used to conduct goodness-of-fit tests [129]. This essential connection enables \(\phi\)-divergence-based ambiguity sets to capture the uncertainty in distributions.
  2. The convexity of \(\phi\)-functions allows its \(\phi\)-divergence to preserve convexity, resulting in computationally tractable models.

- **Cons of \(\phi\)-divergence-based ambiguity sets in DRO.**

| Divergence            | \(\phi^*(s)\) | DRO Counterpart |
|-----------------------|----------------|-----------------|
| Kullback-Leibler      | \(e^s - 1\)   | Convex program  |
| Burg entropy          | \(- \log(1-s), s < 1\) | Convex program  |
| J-divergence          | No closed form | Convex program  |
| \(\chi^2\)-divergence | \(2 - 2\sqrt{1-s}, s < 1\) | SOCP           |
| Modified \(\chi^2\)-divergence | \(\begin{cases} -1, & s < -2 \\ s + \frac{2}{1}, & s \geq -2 \end{cases}\) | SOCP           |
| Variation distance    | \(\begin{cases} -1, & s \leq -1 \\ s, & -1 \leq s \leq 1 \end{cases}\) | LP             |
| Hellinger distance    | \(\frac{s}{s-1}, s < 1\) | SOCP           |

**Table 2.** Examples of \(\phi\)-divergence, their conjugates and DRO counterparts [4,138].
1. The ambiguity set $\mathcal{P}_\phi$ may not be rich enough to contain closely relevant distributions. For example, if the nominal distribution $\hat{P}$ is chosen as a discrete distribution, $\mathcal{P}_\phi - \text{KL}$ fails to be constructed in a continuous space (see Section 3.3.2.1).

2. Moreover, $\phi$-divergence may fail to comprise the true distribution in some cases. Gao and Kleywegt [70] provided an image processing example to illustrate it. Their numerical results show that the KL-divergence-based ambiguity set constructed by artificial data does not include the true distribution which generates the artificial data, and the corresponding DRO model obtained a wrong image result.

In the following subsections, we introduce two favorite $\phi$-divergence-based ambiguity sets, Kullback-Leibler-divergence-based ambiguity sets and modified-$\chi^2$-divergence-based ambiguity sets in the DRO literature.

### 3.3.2.1 Kullback-Leibler(KL)-divergence-based ambiguity sets

KL-divergence is also known as the relative entropy, which is popularly used in the literature of information theory [44,101]. We first give its definition.

**Definition 3.14.** (Kullback-Leibler divergence) Given two probability distributions $P_1$ and $P_2$, $P_1 \preceq P_2$, the Kullback-Leibler divergence of $P_1$ from $P_2$ is

$$D_{\phi - \text{KL}}(P_1 || P_2) = \mathbb{E}_{P_1} \left[ \log \frac{dP_1}{dP_2} \right].$$

(75)

Then the KL-divergence-based ambiguity set is defined as follows.

$$\mathcal{P}_{\phi - \text{KL}} = \{ P \in \mathcal{M}(\Xi, \mathcal{F}) \mid D_{\phi - \text{KL}}(P || \hat{P}) \leq \epsilon_D \}. \quad (76)$$

The definition of KL-divergence in format of likelihood ratio provides an easy way to derive an explicit model. Denote $L(\xi) = P(\xi)/\hat{P}(\xi)$ as a likelihood ratio (also referred to as Radon-Nikodym derivative [87]). It is easy to check that $\mathbb{E}_{\hat{P}}[L(\xi)] = 1$. Thus let $L$ be the set of likelihood ratios in the form of $L$ with the former characteristic. Then we have $D_{\phi - \text{KL}}(P || \hat{P}^N) = \mathbb{E}_{\hat{P}^N}[L(\xi) \log L(\xi)]$ by the change-of-measure technique. Similarly, applying the change-of-measure technique to the objective function, we have $\mathbb{E}_{P}[f(\mathbf{x}, \xi)] = \mathbb{E}_{\hat{P}^N}[f(\mathbf{x}, \xi)L(\xi)]$. Hence, the worst-case inner problem $\sup_{P \in \mathcal{P}_{\phi - \text{KL}}} \mathbb{E}_{P}[f(\mathbf{x}, \xi)]$ in (12) can be reformulated as the following optimization model:

$$\sup_{L \in \mathbb{L}} \mathbb{E}_{\hat{P}^N}[f(\mathbf{x}, \xi)L(\xi)]$$

s.t. $\mathbb{E}_{\hat{P}^N}[L(\xi) \log L(\xi)] \leq \epsilon_D$, \quad (77)

Therefore, an optimization problem with an unknown $P$ becomes a standard SO problem in the form of (2) on $\mathbb{L}$ with a known $\hat{P}^N$.

The KL-divergence-based ambiguity set has attracted much attention from classical uncertain optimization problems. It was first used in portfolio optimization to depict the distributional uncertainty of asset returns. The KL-divergence-based ambiguity set was constructed as a ball centered in Gaussian nominal distribution [73]. Calafiore [34] also studied a distributionally robust portfolio selection problem in which its KL-divergence-based ambiguity of the return distribution is constructed around a discrete nominal distribution. Chen et al. [37] applied the DRO model over KL-divergence-based ambiguity set to the unit commitment problem and Faury et al. [63] illustrated how the KL DRO can serve as a principle tool for the counterfactual risk minimization problem.
Hu and Hong [87] considered the complexity of the new KL-divergence-based DRO approach for general problem with format (12). They showed that this type of DRO problems often have the same complexity as the nominal stochastic optimization problems. Namkoong and Duchi [121] explored efficient algorithms to solve a KL-divergence-based distributionally robust empirical risk minimization problem. Redesigning data structures empower their algorithms to reach the same computational complexity as the stochastic gradient descent methods. The distributionally robust chance-constrained problems over KL-divergence-based ambiguity sets were proved to be equivalent to the classical problem under the rescaled nominal distribution [92].

Recently, Parys et al. [171] gave a strong conclusion that is reflected in the article entitled “From data to decisions: distributionally robust optimization is optimal”. They proved that, by solving a distributionally robust optimization problem over $P_{\phi-KL}$, the best data-driven decision can be obtained with guarantees of the best out-of-sample performance. We summarize the pros and cons of the KL-divergence-based ambiguity sets in DRO as follows.

- **Pros of KL-divergence-based ambiguity sets in DRO.**
  1. KL-divergence is a widely accepted measure of distances between distributions.
  2. The worst-case expectation of a random performance under KL-divergence may be derived explicitly.

- **Cons of KL-divergence-based ambiguity sets in DRO.**
  1. There may not be any practical guidelines in determining the size of the KL-divergence-based ambiguity set.
  2. KL-divergence has difficulty in handling random functions that are heavily right tailed under the nominal distribution.
  3. KL-divergence-based ambiguity sets typically fail to represent the confidence sets of an unknown distribution $P$ [61].

### 3.3.2.2 Modified-χ²-divergence-based ambiguity sets.

The summaries of $\phi$–divergence in Table 1 and Table 2 show that the $\chi^2$-divergence and the modified $\chi^2$-divergence are adjoints. They are both related to the famous $\chi^2$ statistical test, which plays an important role in statistics. Moreover, the modified $\chi^2$-divergence can produce the smallest ambiguity sets compared to others according to (74). Therefore, we introduce the ambiguity problem based on the modified $\chi^2$-divergence in this section.

**Definition 3.15.** (Modified $\chi^2$-divergence.) Given two probability distributions $P_1$ and $P_2 \in \mathcal{M}(\Xi, \mathcal{F})$, the modified $\chi^2$-divergence between $P_1$ and $P_2$ is

$$D_{\phi-M\chi^2}(P_1||P_2) = \int_\Xi \frac{(dP_1 - dP_2)^2}{dP_2}. \tag{78}$$

Then, the modified-χ²-divergence-based ambiguity set can be defined as

$$\mathcal{P}_{\phi-M\chi^2} = \{ P \in \mathcal{M}(\Xi, \mathcal{F}) \mid D_{\phi-M\chi^2}(P||\hat{P}) \leq \epsilon_D \}. \tag{79}$$

As for this type of ambiguity sets, we give an example of the classical two-stage DRO problems to illustrate how to reformulate the DRO with $\mathcal{P}_{\phi-M\chi^2}$.

**Example 3.6.** A discrete distributionally dynamic stochastic problem under an ambiguity set based on $\mathcal{P}_{\phi-M\chi^2}$. 

Given a general distributionally robust linear two-stage problem over $\mathcal{P}_{\phi-M\chi^2}$, its first-stage subproblem is

$$\min_{x \in \mathbb{R}^n} \ c^T x + \max_{P \in \mathcal{P}_{\phi-M\chi^2}} \mathbb{E}_P[(f(x, \xi))]$$

s.t. $\ Ax = b,$ $\ x \geq 0,$

(80)

where $A \in \mathbb{R}^{m \times n}, \ b \in \mathbb{R}^m, \ c \in \mathbb{R}^n$.

We assume that, given a dataset $\Xi_N \subseteq \mathbb{R}^k$ of size $N$ for the uncertain parameter $\xi \in \mathbb{R}^k$, the nominal distribution $\hat{P}_N$ assigns equal probability $\frac{1}{k}$ to each component, that is, $\hat{P}_N = (\frac{1}{k}, \ldots, \frac{1}{k})^T$. Then, given a radius as $\epsilon^2/N > 0$, a discrete ambiguity set $\mathcal{P}_{\phi-M\chi^2}$ is as follows according to (79).

$$\mathcal{P}_{\phi-M\chi^2} = \{ p \in \mathbb{R}^k \ | \ \sum_{i=1}^k p_i = 1, p_i \geq 0, \forall i, \sum_{i=1}^k \frac{(p_i - 1/k)^2}{1/k} \leq \frac{\epsilon^2}{N} \}. \quad (81)$$

The inner worst-case maximization subproblem of (80) can be reformulated as

$$\max_{P \in \mathcal{P}_{\phi-M\chi^2}} \mathbb{E}_P[(f(x, \xi))] = \max_{p \in \mathbb{R}^k} \ \sum_{i=1}^k p_i f(x, \xi_i)$$

s.t. $\ \sum_{i=1}^k p_i = 1$ 

$$\sum_{i=1}^k \frac{(p_i - 1/k)^2}{1/k} \leq \frac{\epsilon^2}{N} \quad (82)$$

The original problem is then equivalent to

$$\min_{x \in \mathbb{R}^n, p \in \mathbb{R}^k, t \in \mathbb{R}} \ c^T x + t$$

s.t. $\ Ax = b,$ $\ \sum_{i=1}^k p_i f(x, \xi_i) \leq t,$

$$\ \sum_{i=1}^k p_i = 1,$$ 

$$\sum_{i=1}^k \frac{(p_i - 1/k)^2}{1/k} \leq \frac{\epsilon^2}{N}$$ 

$$p \geq 0,$$ 

$$x \geq 0.$$ 

(83)

The feasible region of (83) is conic representable. In other words, its constraints only involve linear and quadratic (in)equalities. When the second-stage objective function $f(x, \xi)$ is convex, (83) is a tractable optimization problem. For example, if $f(x, \xi)$ is SOC representable, (83) is a quadratic programming problem.

Klabjan et al. [98] solved a single-item multi-period periodic review stochastic lot-sizing problem where the demand is random over a $\chi^2$-divergence-based ambiguity set. The DRO model converges to the SO model with the true distribution under mild conditions. Philpott et al. [132] studied the stochastic dual dynamic programming problem in which the objective function in each stage has an uncertain parameter over a modified-$\chi^2$-based ambiguity set. The correctness and the almost sure convergence were verified in the paper. Levy et al. [107] proposed algorithms to solve distributionally robust optimization of convex losses with CVaR over $\mathcal{P}_{\phi-\chi^2}$. They are the first to give guarantees that the number of gradient evaluations is independent with the training set size and the number of parameters. The new contributions make their methods suitable for large-scale applications. Ben-Tal et al. [5] focused on robust linear optimization problems with uncertainty regions defined by $\phi$-divergences including $\chi^2$-divergence. They proposed a strong
The dual problem of the inner maximization problem is formulated as
\[ \exists \text{efficiently solvable with linear programming.} \]
Duchi and Namkoong [53] proposed a new approach to automatically balance the two errors via DRO over \( \chi^2 \)-divergence-based ambiguity sets and Owen’s empirical likelihood [128]. Theoretical guarantees and empirical evidence in the paper show a certificated out-of-sample performance.

### 3.4 Other ambiguity sets
Various ambiguity sets have been studied in the literature in addition to the popular ones we have discussed above. Over a box support, the Hoeffding ambiguity sets collect all component-wise independent distributions [10, 17, 35]. The Bernstein ambiguity sets add the marginal moments bound on the Hoeffding ambiguity sets [125]. The ambiguity sets consisting of all distributions that pass prescribed statistical tests are cited as the goodness-of-fit bounds on the Hoeffding ambiguity sets [12].

#### 3.4.1 Frchet ambiguity sets
Frchet ambiguity set is a class of joint distributions with fixed marginal distributions. We first introduce some definitions to help understand the Frchet class.

Given a set \( S \), let \( W = \{W_1, \ldots, W_R\} \subseteq 2^S \) be a cover of \( S \), i.e. \( \cup_{r \in R} W_r = S \), where \( R = \{1, \ldots, R\} \). For \( r \neq q \), \( W_r \nsubseteq W_q \). Let \( \text{proj}_{W_r}(P) \) denote the marginal distribution of the subvector \( \xi_r \) formed from the components in the subset \( W_r \). Then the general Frchet class of distributions can be defined as follows.

**Definition 3.16.** The Frchet class [144] of distributions \( \mathcal{P}_F \) for a cover \( W = \{W_1, \ldots, W_R\} \) is defined as the set of all possible joint distributions of the random vector \( \xi \) with the given multivariate marginal distributions \( \{P_r\}_{r \in R} \) of the subvectors \( \{\xi_r\}_{r \in R} \) as projections:

\[
\mathcal{P}_F = \{ P \in \mathcal{M}(\Xi, \mathcal{F}) | \text{proj}_{W_r}(P) = P_r, \forall r \in R \}.
\] (84)

For a Frchet class of discrete distributions with overlapping marginals, Doan et al. [50] showed that the distributionally robust portfolio optimization problem is efficiently solvable with linear programming. The overlapping cover means that \( \exists r \neq q \), \( W_r \cap W_q \neq \emptyset \). For an overlapping cover, the pairwise consistency of the multivariate marginals is coincide with the existence of a joint distribution, that is \( \text{proj}_{W_r \cap W_q}(P_r) = \text{proj}_{W_r \cap W_q}(P_q), \forall r \neq q \). For a given distributional robust portfolio problem with CVaR measure (See equation (64) for its measure definition), i.e.,

\[
\min_{x \in X} \max_{P \in \mathcal{P}_F} \text{CVaR}_{\beta}^P(\xi^T x). \tag{85}
\]

The dual problem of the inner maximization problem is formulated as

\[
\min_{f_r(\cdot)} \beta + \frac{1}{1-\alpha} \sum_{r \in R} \mathbb{E} \left[ \left( \xi_r^T (\omega_r \circ x_r) - f_r(\xi_{W_r}) + \sum_{t > r; \sigma_t = r} f_r(\xi_{W_t}) - \frac{\beta}{R} \right)_+ \right], \tag{86}
\]

where \( \omega \) is a vector that expresses the separation with respect to the cover \( S \), that is, \( \omega_i = \sum_{r \in R} I_{i \in W_r}, \forall i \in S \); \( \circ \) is the Hadamard (entry-wise) product operator; and the decision variables \( f_r(\cdot) \) are measures, i.e. \( f_r(\xi_r) = \mathcal{P}_r(\xi_r = \xi_r) \). Doan et
al. [50] showed that under mild conditions, the model (86) can be reformulated as a LP problem.

Given a finite set of possible distributions and marginal distributions, Dupaˇ covˇa [56] showed that the DRO in linear form could be transferred to an ordinary LP in some particular cases. Dupaˇ covˇa [57] then enhanced the result into more general stochastic models.

3.4.2 Bayesian ambiguity sets. The Bayesian type approach is using a priori probability measure on the set $\mathcal{P}$ to help reduce the problem to the standard expected value formulation. For instance, let $\mathcal{P}$ be a finite set, i.e., $\mathcal{P} = \{P_1, \ldots, P_m\}$. The priori probability of the possible distributions is given as $\{p_1, \ldots, p_m\}$ with $\sum_{i=1}^{m} p_i = 1, p_i \in [0, 1], i = 1, \ldots, m$. Then an average distribution over the set $\mathcal{P}$ is $\bar{P} = \sum_{i=1}^{m} p_i P_i$. Recall the classical DRO problem (12) $\inf_{x \in X} \sup_{P \in \mathcal{P}} E_P[f(x, \xi)]$. It becomes an ordinary SO problem with the fixed distribution $\bar{P}$, $\inf_{x \in X} E_{\bar{P}}[f(x, \xi)]$. (87)

Therefore, for a given finite set of possible distributions $\mathcal{P}_I = \{P_1, \ldots, P_I, |I| < \infty\}$, it can be replaced by its convex hull when the priori probability is provided as $\{p_1, \ldots, p_I\}$ with $\sum_{i=1}^{I} p_i = 1, p_i \in [0, 1], i = 1, \ldots, I$. Denote this convex hull as the so-called Bayesian ambiguity set $\mathcal{P}_B$,

$$
\mathcal{P}_B = \left\{ P \in \mathcal{M}(\Xi, \mathcal{F}) \mid P = \sum_{i=1}^{I} p_i P_i, P_i \in \mathcal{P}_I, \sum_{i=1}^{I} p_i = 1, p_i \in [0, 1], |I| < \infty \right\}.
$$

Gupta [80] proposed a Bayesian framework for DRO under a verified asymptotically Bayesian optimal set. They also proved that their Bayesian DRO solutions converge to the stochastic optimization problem’s solutions when the true distribution is known. Numerical results in their work support their claim that the near-optimal Bayesian DRO models can significantly outperform existing models for the same application. Please refer to the paper [80] for more information about the ambiguity set construction and asymptotic analysis.

Wiesemann et al. [185] and Wang et al. [184] used the Bayesian method to specify a set of parameterized distributions that make the observed data achieve a certain level of likelihood.

4 Solution methods and performance guarantees. Distributionally robust optimization models attract much attention in uncertain problems recently. One of the most charming advantages is the tractability. Many DRO problems can be solved in polynomial time, even though its corresponding stochastic problem cannot be solved efficiently due to the expansive computation cost of multivariate integral [61]. In this section, we summarize the tractable solution methods used in the DRO literature as well as their performance analysis.

Consider the classical DRO model (12), it can be reformulated as the following SIP model,

$$
\inf_{x \in X, t} \begin{array}{c}
t \\
\text{s.t.} \\
 t \geq \int_{\Xi} f(x, \xi) dP, \forall P \in \mathcal{P}.
\end{array}
$$

(89)

Since the SIP problems are not often tractable, researchers always look for efficient near-optimal algorithms to solve the problems. We refer to review papers and books of SIP for more information [76, 86, 154]. Hettich et al. [85] also provided
useful numerical methods including exchange methods, discretization methods, and reduction methods to solve the SIP problems.

Moreover, the tractability of DRO has been exhibited with regard to the structure of ambiguity sets in Section 3. This section will introduce four main types of solution methods to solve DRO problems. The first type focuses on the numerical methods for solving the reformulated SIP problems. In the second type, the DRO problems could be reformulated as finite convex optimization problems which have well-developed polynomial-time solvable algorithms. Another type is to reduce the DRO problem to an SO problem by seeking the extremal distributions for the inner maximization problem. In the last type, we discuss some special DRO problems for which the direct efficient algorithms have been developed. And finally we review the convergence and out-of-sample analysis in the DRO literature.

4.1 Numerical methods on SIP. The main difficulty in solving an SIP problem is associated with the infinitely many constraints. The most straightforward method is the discretization method which seeks a discrete construction of $\mathcal{P}^M = \{P_i\}_{i=1}^M$ with a large $M$. Then the SIP problem becomes a solvable problem with finite constraints. However, the discrete set $\mathcal{P}^M$ is often hard to build up. If $\mathcal{P}^M$ is small, it may be not representable enough. A large $\mathcal{P}^M$ generally yields a better solution but the computation cost may be too high.

Another common method is the cutting surface method. It starts with finding a local optimal solution on a subset of $\mathcal{P}$. A constraint will be added as a cut until the current solution becomes globally optimal.

Readers may find useful numerical methods for solving convex SIP problem in literature, for instance, penalty methods [43, 111, 196], primal methods [182], dual methods [86], and smooth approximation and projection methods [193]. In this section, we mainly introduce the cutting surface method in the context of DRO.

- Cutting surface method.

We first introduce the basic idea to solve a DRO problem by solving its reformulated SIP problem (89) via cutting surface method. The cutting surface method is also called the exchange method in some SIP literature.

Suppose we can find a finite subset $\mathcal{P}^M = \{P_1, \ldots, P_m\} \subseteq \mathcal{P}$, then the optimal solution $(\hat{x}^m, t^m)$ is computed for the following problem with finite constraints:

$$\inf_{x \in \mathcal{X}, t} \quad t \quad \text{s.t.} \quad t \geq \int_{\Xi} f(x, \xi) dP_i, \quad i = 1, \ldots, m. \quad (90)$$

Then compute $P_{m+1} = \arg \min_{P \in \mathcal{P}} \{t^m - \int_{\Xi} f(\hat{x}^m, \xi) dP\}$. If $t^m - \int_{\Xi} f(\hat{x}^m, \xi) dP_{m+1} \geq 0$, then $(\hat{x}^m, t^m)$ is the optimal solution to (89); otherwise, let $\mathcal{P}^{M+1} = \mathcal{P}^M \cup \{P_{m+1}\}$ for the next iteration. The computation bottleneck of the above procedures exists in solving (90) and finding the minimizer $P_{m+1}$. Finding the exact minimizer $P_{m+1}$ is often difficult. People then turn to find a near-optimal solution by solving a relaxed tractable problem. For a small $\rho \in [0, 1]$, $x^*_\rho \in \mathcal{X}$ is $\rho$-optimal if $\int_{\Xi} f(x^*_\rho, \xi) dP - t \leq \rho, \forall P \in \mathcal{P}$ and the corresponding objective value $t^*_\rho \leq t^*$.

Cutting surface methods have been well-developed in the literature. Mehrotra and Papp [117] proposed a cutting surface method to solve a DRO problem with a moment-based ambiguity set whose lower and upper bounds of the first moment are given. The $\rho$-optimality and linear convergence rate of the algorithm are analyzed. Lee and Mehrotra [105] studied a Wassertein DRO approach to construct support vector machines. They applied the cutting surface method proposed in [117] to
solve the reformulated SIP problem. The numerical results on both simulated data and real data show the advantages of DRO models over the traditional SVMs in terms of the Area Under Curve measures. Luo and Mehrotra [114] studied the Wasserstein DRO problems in the context of logistic regression models and used the cutting surface method to solve the SIP problems induced by the dual of the inner maximization problem. Their algorithm is guaranteed to terminate in finite number of iterations and to return a $\rho$-optimal solution. Rahimian et al. [137] considered a convex DRO model with the total-variation-distance-based ambiguity set. They proposed a cutting plane method called Primal Decomposition to solve the problems. The cutting plane method exploits the polyhedral structure of the ambiguity set $\mathcal{P}_{DTV}$ which might have computational benefits over other decomposition methods.

4.2 Convex reformulation based on duality theory. Optimization problems under uncertainty, including SO, RO, and DRO problems, are $NP$-hard in general. DRO outperforms RO and SO in tractability because of the structural properties of the ambiguity sets which yield convex reformulations based on the duality theory [140]. Some examples of reformulations have been presented in Section 3. In this section, we give a general summary of different kinds of reformulated convex problems under different conditions, such as the structures of the objective function $f(x, \xi)$, the feasible set $\mathcal{X}$, the support set $\Xi$, and the ambiguity sets $\mathcal{P}$.

4.2.1 Finite convex programming reformulation. An equivalent convex reformulation generally requires a proper, convex and bounded objective function and a convex and closed feasible region [140] for continuous optimization problems. Existed DRO problems follow the principle. For discrete cases, it depends on the specific problems. Esfahani and Kuhn [61] studied a group of DRO problems over Wasserstein balls that can be reformulated as finite convex problems. The objective function can be expressed as the point-wise maximum of finitely many concave functions in the uncertain parameter, i.e., $l(x, \xi) : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$, $l(x, \xi) = \max_{1 \leq i \leq m} l_i(x, \xi)$, where the negative constituent functions $-l_i(x, \xi)$ are proper, convex, bounded and lower semicontinuous. The support set $\Xi$ is convex and closed. Computing the conjugates of $l_i$ over Wasserstein ambiguity sets gives a finite dual convex problem. They applied their DRO models on mean-risk portfolio optimization and uncertainty quantification problems. The simulation results show the impact of the Wasserstein radius and the sample size on the out-of-sample performance.

Chen et al. [37] applied the reformulation skills and stochastic programming technologies to a distributionally robust unit commitment problem over a KL-divergence-based ambiguity set. The deterministic unit commitment problem is a mixed-integer LP problem. The corresponding DRO model could be tantamount to a mixed-integer convex programming problem whose objective function is convex of form $\log \sum e^{x_i}$. They proposed a two-level decomposition method based on the generalized Bender’s decomposition to solve a large-scale problem.

4.2.2 Second order cone programming reformulation. Some DRO problems have SOCP reformulations more than the general finite convex reformulations. This type of problems commonly have SOC representable sets and functions [1]. For example, the SOC-moment-based ambiguity sets (26) is a SOC representable set, and linear functions, convex quadratic functions and norm functions are SOC representable functions.
Mehrotra and Zhang [118] gave SOCP reformulations of distributionally robust least squares problems with a finite discrete conic moment-based ambiguity sets. Both the objective function and the ambiguity set are SOC representable. Nataraajan et al. [122] proposed an SOCP reformulation for a class of robust expected utility models over the moment-based ambiguity sets with known mean and covariance matrix. The objective utility function is a piecewise-linear concave function. The feasible region $X$ is SOC representable. Moment-based ambiguity sets with known mean and covariance matrix in their work are also SOC representable. Bertsimas et al. [19] studied a two-stage adaptive linear DRO problem over a moment-based SOC ambiguity set. If the ambiguity set has expectation constraints over variances, then the dual problem can be reformulated as an SOCP problem.

We have discussed in Section 3.3.2 that the dual of DRO problems with certain divergence-based ambiguity sets can be reformulated into SOCP problems. For example, the $\chi^2$-divergence and the modified $\chi^2$-divergence function are convex quadratic functions and the Hellinger-divergence function is a square root of homogeneous convex quadratic function, which are all SOC representable. The corresponding DRO counterparts are SOCPs.

### 4.2.3 Semidefinite programming reformulation.

Most convex reformulations of DRO problems are SDP reformulations. The construction of most ambiguity sets show the semidefinite representable property.

Ghaoui et al. [73] used DRO to study the data sensitive by mean-variance or Value-at-Risk (VaR) approaches on portfolio optimization problems. They considered the bounded moment-based ambiguity sets and showed that the problems can be reformulated as SDP problems. Delage and Ye [47] studied a DRO problem with a piecewise linear and concave utility objective function in the whole Euclidean space or an ellipsoidal set. The ambiguity set is a bounded moment-based ambiguity set under first and second moment information. An SDP reformulation is given for the worst-case inner problem. Based on Delage and Ye’s model, Li and Kwon [109] added a penalty term on the objective function with respect to the discrepancy between a pair of first two moments and the set of moments specified in the confidence region. The semidefinite representable objective functions, support sets and ambiguity sets enables the authors to get a SDP reformulation.

Bertsimas et al. [13] studied the distributionally robust two-stage linear optimization problems over moment-based ambiguity sets with exact first two moments. The DRO problems then can be tightly reformulated into SDP problems. Hanasusanto and Kuhn [82] explored the same problems in [13] over a 2-Wasserstein-distance-based ambiguity set. If the problem has complete recourse, a conic copositive problem is obtained; if the recourse is sufficiently expensive, the problem still can be approximated arbitrarily closely by a sequence of copositive programs.

The Chebyshev inequality is important in classical probability theory. It provides an upper bound on the tail probability of a univariate random variable based on limited moment information. The best upper bound can be obtained by a inner maximization problem of DRO, that is, \( \sup_{P \in \mathcal{P}} P(\xi \notin \Xi) \), with a moment-based ambiguity set $\mathcal{P}$. Bertsimas and Popescu [16] gave a hierarchy of increasingly accurate bounds when the moments can be expressed as expectations of polynomials. The conic duality enables them to solve a tractable SDP problem to obtain the bounds. Popescu [133] considered adding the structural information, such as symmetry, unimodality and convexity, to the moment information in Bertsimas and Popescu’s work [16] for a generalized problem. The problem can also be reformulated
as a SDP to be solved. Parys et al. [172] used an unimodal ambiguity set to mitigate the conservatism of the generalized Chebyshev bounds. The unimodality can help exclude the pathological discrete distributions from the general moment-based ambiguity set. The new problem can also keep the exactness and computational tractability of the SDP reformulation of the corresponding worst-case probability problem. Napat et al. [143] studied the Chebyshev bounds for sums, maxima, and minima of nonnegative random variables when given only information about first two moments. They could reformulate the problem as an SDP when the exact covariance matrix is known.

Wagner [177] considered a binary linear DRO problem where the parameters of the linear inequality constraints are within a $k$-th order moment-based ambiguity set. The DRO problem is equivalent to a deterministic integer LP problem when given first-order information, an integer SOCP problem when given second-order moment information, and a $\rho$-optimal SDP problem when given $m \geq 3$ order moment information. Natarajan and Teo [123] also studied a binary linear DRO problem over a Markov ambiguity set. They proposed a new SDP relaxation of the distributionally robust multi-item news vendor problem. Another contribution of their work is that they proposed a method to reduce the size of the SDP problem when the feasible region of the decision variable includes finitely many points.

Cheng et al. [38] studied a distributionally robust quadratic knapsack problem. They assumed that the first moment information is given and the second information is within a given bound. The knapsack problem with binary constraints can be relaxed by an SDP problem, which achieved a better lower bound than the classic quadratic knapsack problem.

Although the SDP problem is perfectly tractable in theory, it is computationally expensive when the size of the problem increases. Cheng et al. [39] took advantage of the variable selection given by principal component analysis (PCA) to reduce the size of the SDP reformulations of the moment-based DRO problems. The PCA approximation achieves a near-optimal solution within 1% gap when using only half amounts of the variables.

4.2.4 Linear programming reformulation. Plenty of DRO problems are equivalent to or tightly bounded by LP problems in real applications such as revenue management problems, scheduling problems, and dispatch problems. The models have a particular structure with linear objective functions or amounts of linear constraints.

Goh and Sim [77] investigated a two-stage DRO model whose objective has a linear structure. The linear structure can be used to express piece-wise linear utility functions and CVaR constraints. This two-stage DRO model can be reformulated as a minimax LP problem. Doan et al. [50] studied the classic portfolio optimization model in a DRO manner over a Fréchet class of discrete distributions (introduced in Section 3.4.1). The DRO problem can be tightly bounded by an LP problem.

Wiesemann et al. [186] proposed a framework for modeling and solving convex DRO problems, in which the constraint functions are piece-wise affine with respect to both decision variables and random parameters. They showed that the reformulated problem is polynomial-time solvable over a nested moment-based ambiguity set. When linear inequalities describe the given confidence set of the random vector, the linear reformulation is obtained.

Shehadeh et al. [160] investigated a distributionally robust outpatient colonoscopy scheduling problem over a moment-based ambiguity set by first-order moment information that finds an optimal appointment sequence to minimize the worst-case
expected total waiting time. The distributionally robust scheduling problem can be reformulated as a mixed-integer LP problem.

Zhou et al. [202] proposed a data-driven distributionally robust chance-constrained real-time dispatch problem over a Wasserstein ambiguity set. They used linear relaxation techniques to relax the quadratic constraints to seek the tractability. A linear reformulation was provided finally.

4.3 Extremal distributions. Recall the classic DRO model (12)

\[
\inf_{x \in X} \sup_{P \in \mathcal{P}} \{E_P[f(x, \xi)]\},
\]

we have discussed the SIP equivalence and the convex dual reformulation of the inner problem. In some cases, the extremal distribution \(P^*\) can be obtained under mild conditions on the ambiguity set.

\[
P^* = \arg \sup_{P \in \mathcal{P}} \{E_P[f(x, \xi)]\}. \tag{91}
\]

The DRO problem (12) reduces to an SO problem under \(P^*\), i.e., \(\inf_{x \in X} E_{P^*}[f(x, \xi)]\). Researchers have shown a class of problems that can explicitly obtain an extremal distribution.

Section 3.4.2 displays a class of Bayesian ambiguity sets that can induce extremal distributions into a weighted distribution. The prerequisites of this type of ambiguity sets include a finite set \(\mathcal{P}\) and a priori probability distribution for each elements in \(\mathcal{P}\).

Given exact moment-based ambiguity sets with first two moments information, Scarf [145] and Jitka [57] derived a two-point extremal distribution for the inner worst-case problem. While previous work [57] examined the expected value of distribution information (EVDI), Yue et al. [199] extended it to computing the maximum EVDI. The extremal distribution is also a two-point distribution in the extended work.

Shapiro and Kleywegt [159] showed that a weighted distribution can be obtained for the maximization inner problem under certain regularity assumptions. This extremal distribution depends on the existence of a priori distribution over a finite ambiguity set \(\mathcal{P}\). They then extended their work with a general moment-based ambiguity set [157] bounded by finite real measures.

The extremal distributions of DRO problems over Wasserstein ambiguity sets was explored by Gao and Kleywegt [70]. They claimed the necessary and sufficient conditions for the existence of a worst-case distribution.

4.4 Approximation algorithm. The cone duality of moment-based problems and the conjugate dual of divergence functions provide good properties to construct tractable dual problems for most DRO problems. Some good research works provide tractable algorithms by using new methodologies in closed fields such as SO, machine learning, etc. We introduce some exciting and new methods to solve DRO problems in this section.

4.4.1 Stochastic gradient method. Consider minimizing an average of functions \(\min_x \frac{1}{n} \sum_{i=1}^{n} f_i(x)\). It is a common setting in machine learning problems, where this average of functions is often equivalent to a loss function and each \(f_i(x)\) is
associated to the loss of the data point \( x_i \). The full gradient decent step is given by

\[
x^{(t)} = x^{(t-1)} - \alpha_t \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^{(t-1)}), \quad t = 1, \ldots, T,
\]

where \( x^{(t)} \) is the solution in \( t \)-th iteration and \( \alpha_t \) is step size. Assume the gradient exists and let \( \nabla f_i \) denote the gradient of \( f_i \). The stochastic gradient method (SGD) uses a subset of all samples to approximate the full gradient. More formally, stochastic gradient repeats

\[
x(t) = x(t-1) - \alpha_t \nabla f_{i_t}(x(t-1)), \quad t = 1, \ldots, T,
\]

where \( i_t \in \{1, \ldots, n\} \) is a randomly chosen index at iteration \( t \). The indices \( i_t \) are usually chosen without replacement until we complete one full cycle through the entire data set.

Ghosh et al. [69] proposed an SGD algorithm to efficiently solve the DRO problems. Given a dataset with size \( n \), a size sequence \( \{n_t\}_{t=1}^{T} \) is generated. At each iteration \( t = 1, \ldots, T \), do

1. Sample \( n_t \) data points with indices uniformly from \( \{1, \ldots, n\} \) without replacement;
2. Solve the inner maximization problem to obtain an optimal solution for a discrete distribution \( P_t \);
3. Use \( P_t \) to update the gradient and then update the variable \( x^{t+1} \) by a descent step.

They applied gradient descent to the outer minimization problem by estimating the gradient of the inner maximization based on a sample average approximation (SAA). Growing the support size in each iteration ensures the convergence and balances the stochastic error and computation cost. Their work also states the benefit in generalization by using DRO learning approach.

Namkoong and Duchi [121] applied SGD to the DRO problem

\[
\min_{x \in X} \sup_{p \in \mathcal{P}_\phi} \frac{1}{n} \sum_{i=1}^{n} p_l_i(x),
\]

where \( p = (p_1, \ldots, p_n)^T \) and \( \mathcal{P}_\phi \) is a \( \phi \)-divergence-based ambiguity set centered in the uniform distribution \( \mathbb{1}/n \).

\[
\mathcal{P}_\phi = \{ p \in \mathbb{R}^n \mid p^T \mathbb{1} = 1, p \geq 0, D_\phi(p\|\mathbb{1}/n) \leq \frac{\epsilon}{n} \}. \tag{95}
\]

The functions \( l_i : X \to \mathbb{R}_+ \) are convex and subdifferentiable and \( X \) is compact. The dual of the inner maximization problem can be formulated as

\[
\inf_{\lambda \geq 0, \eta \in \mathbb{R}} \sum_{i=1}^{n} \lambda \phi^*(\frac{l_i(x) - \eta}{\lambda}) + \frac{\epsilon}{n} \lambda + \eta, \tag{96}
\]

where \( \phi^* \) is the conjugate of the convex divergence function \( \phi \). The dual problem is jointly convex in \( (x, \lambda, \eta) \), thus the classical SGD generally fails because the variance and subgradients of the objective explode as \( \lambda \to 0 \). Therefore, Namkoong and Duchi [121] solved the problem (94) as a mirror descent two-player stochastic games [124]. The similar game strategy was also studied in [7, 42, 124, 150].
4.4.2 Branch and cut method. Branch-and-cut is an efficient method to solve integer programming problems. It has been well developed under deterministic settings. The branch-and-cut approach uses two techniques including branch-and-bound and the cutting planes. The algorithm starts with a relaxed problem to give an initial upper bound and an initial solution. If the solution is feasible, then stop; otherwise, new sub-problems are generated by branching. Cutting planes will be added by solving a separation problem to find validate cuts. Then fathoming and pruning will be handled then. The basic procedure can be described as follows. At each iteration, do the following:

1. Solve a relaxed problem. If the optimal solution is infeasible, prune.
2. Solve a separation problem to find validate cuts. If validate cuts are found, add them and go to (1).
3. Compare the bounds to prune meaningless nodes;
4. Create two or more new problems to do the branching procedure.

Wolsey’s book [187] provided mature discussion of the basic branch-and-cut algorithms. They have been applied to mixed integer DRO problems such as network flow problems, vehicle routing problems (VRP) and so on.

Wagner [177] investigated a 0-1 linear DRO problem where the parameters in the linear constraints are random vectors with partial moments information. The binary linear DRO problem is equivalent to a deterministic integer LP problem with only restrictions of first moment information, and to an SOCP problem with second moment information. Both cases in their work were solved exactly by using a branch-and-bound algorithm. Dinh et al. [49] studied a chance-constrained VRP with stochastic demands and proposed a branch-and-cut algorithm. The branch-and-cut method is considered as the state of the art in solving the deterministic VRP. They were the first to extended the method to the VRP with stochastic demands. Ghosal and Wiesemann [74] studied a distributionally robust chance-constrained VRP based on Dinh’s work.

\[
\min_{\text{R} \in \mathcal{B}(V_C, m)} c(R) \quad \text{s.t.} \quad P(R_i \in \mathbb{R}(\xi)) \geq 1 - \eta, \quad \forall i, \forall \mathbf{P} \in \mathcal{P}_{BB},
\]

(97)

where \(\mathcal{B}(V_C, m)\) is the set of all partitions of the customer set \(V_C\) into \(m\) mutually disjoint and collectively exhaustive (ordered) routes \({\mathbf{R}_1, \ldots, \mathbf{R}_m}\). The notation \(R_i \in \mathbb{R}(\xi)\) expresses the capacity constraint for the \(i\)-th vehicle, and capacity \(\xi\) is a random vector following an unknown distribution \(P \in \mathcal{P}_{BB}\). The ambiguity set \(\mathcal{P}_{BB}\) is a special case of the SOC ambiguity sets \(\mathcal{P}_{SOC}\) (26) as follows.

\[
\mathcal{P}_{BB} = \left\{ P \in \mathcal{M}(\Xi, \mathcal{F}) \middle| \begin{array}{l}
P(\xi \in \Xi = [\underline{\xi}, \overline{\xi}]) = 1, \\
\mathbb{E}_P[\xi] = \mu, \\
\mathbb{E}_P[g(\xi)] \leq \sigma
\end{array} \right\},
\]

(98)

where \(g : \mathbb{R}^k \rightarrow \mathbb{R}^r\) is a dispersion measure of demand variations with \(g = (g_1, \ldots, g_r)\). Ghosal and Wiesemann’s work over \(\mathcal{P}_{BB}\) covers a large class of moment-based ambiguity sets. They provided an efficient cut generating method for this generalized SOC moment-based ambiguity set. Numerical results showed their distributionally robust CVRP beat the deterministic CVRP both in scaling properties and computational time.

4.5 Performance guarantee and convergence analysis. As mentioned in Section 2, the ambiguity sets help DRO to overcome the difficulty of finding the true distribution or estimated distribution in SO and reduce the conservativeness in lack
of distribution information in RO. An important part is to evaluate the performance of a DRO problem. During the ambiguity construction process, once its type is determined, the parameters design becomes decisive. For instance, when choosing moment-based ambiguity sets, the moments information needs to be obtained; and when selecting distance-based ambiguity sets, the nominal distribution and the radius need to be given. In some cases, these information may be provided by an expert or industrial standard parameter values. However, in most DRO problems, data-driven setting is more valuable. For a given dataset \( \Xi^N \) with \( N \) sample points, it is used to estimate the moment information \( \bar{\mu}, \bar{\Sigma} \) by (24) or to estimate the nominal distribution \( \hat{P} \) by (50) or (51) and radius \( \epsilon_D \) by (56). Therefore, for a data-driven problem, the out-of-sample performance or the convergence of the model are important.

Let \( J^* \) be the optimal value under the unknown true distribution \( P \), that is,
\[
J^* = \inf_{x \in \mathcal{X}} \{ \mathbb{E}_P[f(x, \xi)] \},
\]
and similarly, \( x^* \) is the optimal solution,
\[
x^* = \arg \inf_{x \in \mathcal{X}} \{ \mathbb{E}_P[f(x, \xi)] \}.
\]
Denote \( \hat{x}_N \in \mathcal{X} \) as a feasible solution of (100) based on the training dataset \( \hat{\Xi}_N \). The out-of-sample performance of \( \hat{x}_N \) is defined as \( \mathbb{E}_{\hat{P}}[f(\hat{x}_N, \xi)] \) under a new sample \( \xi \). Since the true distribution \( P \) is unknown, the exact out-of-sample performance is impossible to compute. An alternative way is to construct a bound to evaluate the performance [61],
\[
P^N \left\{ \hat{\Xi}_N : \mathbb{E}_{\hat{P}}[f(\hat{x}_N, \xi)] \leq \hat{J}^N \right\} \geq 1 - \eta,
\]
where \( P^N \) is a distribution constructed from any \( N \) random samples, and \( \hat{J}^N \) is an upper bound that may related to the training dataset. \( \eta \) is a confidence level of the out-of-sample of \( \hat{x}_N \) below the upper bound \( \hat{J}^N \). Then we can evaluate the out-of-sample performance of the DRO model (12) as follows. Let
\[
\hat{J}^N = \inf_{x \in \mathcal{X}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[f(x, \xi)],
\]
and
\[
\hat{x}_N = \arg \inf_{x \in \mathcal{X}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[f(x, \xi)].
\]
If \( \hat{x}_N \) and \( \hat{J}^N \) satisfy (101), then we can say the optimal value of the DRO problem provides a \( 1 - \eta \) confidence bound of the out-of-sample performance of its decision value.

The size of dataset is finite so that researchers seek a bound of the performance guarantee. However, it is important to provide the convergence analysis, i.e., \( \hat{x}_N \rightarrow x^* \), \( \hat{J}_N \rightarrow J^* \), as \( N \rightarrow \infty \), which refers to the perfect information.

Note that the above performance evaluation is for exact algorithms. The near-optimal algorithms yield gaps to the exact optimal values and it is crucial to provide the near-optimality analysis.

Researchers [32,33,170] did some stability analysis for some DRO problems with finite discrete probability distributions. Riis and Andersen [139] then extended the stability analysis to continuous probability distributions. A bunch of works addresses the convergence of distributionally robust optimization problems, where the unknown distribution \( P \) is estimated by Monte Carlo sampling [184,185,191].
Dupacova [58] showed the $\rho-$convergence of the optimal value for the DRO problems over the moment-based ambiguity sets. Sun and Xu [169] provided a unified framework for the convergence of some DRO problems. They focused on the case in which the ambiguity set is approximated by a sequence of ambiguity sets constructed through samples or other means. They investigated convergence of the ambiguity sets under total variation metric as sample size increases. In addition, they conducted the convergence analysis work on the moment-based ambiguity sets which were adopted in [47] and [164].

Gupta [80] proposed a Bayesian framework for DRO under a verified asymptotically Bayesian optimal set. Gupta also proved that the solutions of their Bayesian DRO converge to solutions to the full-information SO problems when the true distribution is known.

Esfahani and Kuhn [61] proved that for a carefully chosen size of the Wasserstein-distance-based ambiguity set (see Section 3.3.1.1) in the order of $N^{1/k}$, the data-driven optimal solution can have certified confidence level out-of-sample performance. They provided the first out-of-sample performance guarantees for Wasserstein DRO. Besides, they gave a finite-sample non-asymptotic guarantee for the data-driven optimal solution. However, the bound has a problem with the curse of dimensionality since the radius of the Wasserstein ball shrinks too slow, even for moderate dimensional problems.

To overcome the curse of dimensionality, Blanchet et al. [26, 28, 29] finished a series of work by an approach inspired from the empirical likelihood. They sought the smallest radius of the Wasserstein ball such that the ball contains at least one distribution for which the optimal solution is also optimal to the true distribution, with high probability. An asymptotic analysis gives a $1/\sqrt{N}$ order of the radius selection. However, this bound only works in the asymptotic sense. Applying the Wasserstein DRO in linear regression/classification models, Shafieezadeh-Abadeh et al. [146] provided a $1/\sqrt{N}$ order radius selection in finite-sample performance guarantee. Chen et al. [36] achieved a generalization bounds in the context of DRO with norm regularization. In addition to the out-of-sample loss (prediction bias), they also gave the performance guarantees on the discrepancy between the estimated and the true regression coefficients (estimation bias). Gao [69] provided the first finite-sample guarantee for generic Wasserstein DRO problems and showed that the non-asymptotic $1/\sqrt{N}$ rate holds for general loss functions.

5 Applications. In this section, we bring some applications of DRO. We put emphasis on the machine learning and operations research fields.

5.1 DRO in machine learning. Machine learning is a research field of analyzing and drawing useful information from the data. The uncertainty exists in the process of analyzing data and the behavior of a system [11]. In this section, we bring a brief review of DRO applied to machine learning. We first introduce the supervised learning models in DRO and then summarize other types of DRO related machine learning models.

5.1.1 Supervised learning. Supervised learning is one of the most important parts in machine learning. Assume that the input and output pair $(y, z) \in \mathbb{R} \times \mathbb{R}^k$ follows the distribution $P$, then most of the supervised learning problems in the Euclidean space can be generalized as the following optimization problem:

$$\min_{\alpha \in \mathbb{R}^k} \mathbb{E}_P[\ell(y, z; \alpha)],$$

(104)
where $\alpha$ is the parameters to be learned and $\ell$ is the loss function provided by the given learning machine. The distribution $P$ is usually unknown and even hard to assume a certain one. Hence, given $N$ sample points $\{(y^{(i)}, z^{(i)})\}_{i=1}^N$, the problem (104) can be approximated by solving the following problem in a data-driven version:

$$
\min_{\alpha \in \mathbb{R}^k} \frac{1}{N} \sum_{i=1}^N \ell(y^{(i)}, z^{(i)}; \alpha). \tag{105}
$$

Problem (105) is solved with finite observations, thus it may be sensitive to the perturbation in data points, which may cause overfitting problems [146]. To avoid overfitting, regularization terms are often added, which strongly improve the accuracy of the learning models. Consider the following $\ell_p$ norm regression problem in matrix form

$$
\min_{\alpha \in \mathbb{R}^k} \|y - Z^T \alpha\|_p, \tag{106}
$$

where $p \in [1, +\infty]$, $y = (y^{(1)}, \ldots, y^{(N)})^T \in \mathbb{R}^N$ and $Z = (z^{(1)}, \ldots, z^{(N)})^T \in \mathbb{R}^{k \times N}$. Literature [12,59,108,190] shows that the regularized regression is equivalent to the a worst-case RO problem by involving perturbation of $Z$.

Theorem 5.1 (Corollary 1 in [12]). For any $p, q \in [1, +\infty]$, 

$$
\min_{\alpha \in \mathbb{R}^k} \max_{\Delta \in \mathcal{U}_{(p,q)}} \|y - (Z + \Delta)^T \alpha\|_p = \min_{\alpha \in \mathbb{R}^k} \|y - Z^T \alpha\|_p + \lambda \|\alpha\|_q,
$$

where the uncertainty set $\mathcal{U}_{(p,q)}^\lambda$ is defined by

$$
\mathcal{U}_{(p,q)}^\lambda \triangleq \left\{ A \in \mathbb{R}^{k \times N} \left| \max_{\beta \in \mathbb{R}^k} \frac{\|A^T \beta\|_q}{\|\beta\|_p} \leq \lambda \right. \right\}. \tag{107}
$$

Notice that when $p = 2, q = 2$, it reveals the case of ridge regression. When $p = 2, q = 1$, it reveals the case of Lasso. This theorem theoretically supports the robustness that the regularization terms bring to the original regression models.

The worst-case RO optimization in the theorem is also referred as the robustification of regression in literature. The connection between this robustification and the regularized regression models is established in [6,190] and is generalized in [12]. It is also investigated on other supervised learning models, such as the support vector machine (SVM) [189], neural networks and kernel methods [146].

The uncertainty factors have been addressed through rigorous and comprehensive procedures which yields various formulations in RO problems. However, recall that our original goal is to solve the optimization problem (104) and the uncertainty of the SO problem is governed by the probability distribution $P$. It is different from the RO problems which ignore the distributional information and solve the worst-case over all of the possible realizations of the uncertainty. Filling the gap between RO and SO, the DRO formulation (108) considers the uncertainty by minimizing the worst-case expectation over the ambiguity set of distributions.

$$
\min_{\alpha \in \mathbb{R}^k} \sup_{P \in \mathcal{P}} \mathbb{E}_P \ell(y, Z; \alpha), \tag{108}
$$

where $\mathcal{P}$ is the ambiguity set, which contains the unknown distribution $P$ of the given data set with high confidence [147].

A majority of literature in machine learning focuses on designing the ambiguity set with Wasserstein distance. We have discussed in Section 3.3.1.1 that Wasserstein
DRO is highly related to data-driven RO and the RO has exhibited certain regularization effects. The distributionally robust least square problem has been studied with Wasserstein-distance-based ambiguity set (hereinafter called the “Wasserstein ambiguity set”) as early as in [118]. Besides, it has been rigorously shown that how the regularized machine learning models, such as regularized $\ell_1$ regression [36], regularized logistic regression [28,147], square-root Lasso [28] and SVM [28] are related to the DRO formulations with ambiguity sets utilizing Wasserstein distance. The DRO formulations reveal the commonly used regularized learning models when the Wasserstein ambiguity sets are set properly [28, 36]. For example, adopting $\mathcal{P} = \mathcal{P}_{DW}^\infty$ as defined in (53), the DRO formulation reveals the square-root Lasso model [28].

$$\min_{\alpha \in \mathbb{R}^k} \left[ \frac{1}{\sqrt{N}} \|y - Z^T \alpha\|_2 + \lambda \|\alpha\|_1 \right]^2 = \min_{\alpha \in \mathbb{R}^k} \max_{D_{\infty}^\infty, (p, \hat{p}_N) \in \lambda^2} \mathbb{E}_p \left[ (y - Z^T \alpha)^2 \right], \quad (109)$$

where $\lambda > 0$ is the given parameter, and it determines the radius of the Wasserstein ambiguity set. Both of the theoretical proofs and the numerical experiments of asymptotic performance were conducted in the above literature, which indicated that the DRO approaches yield guaranteed out-of-sample performances. Besides, an inference methodology were introduced in [28] to select the size of the ambiguity set.

We consider the ambiguity sets that are characterized in moment conditions in the following. A distributionally robust minimax probability machine for binary classification was proposed in [88, 102] and was developed in [103]. The idea is to minimizes the worst case misclassification error within the ambiguity set with prior moment information. It shares the similar idea of DRO, even though not in the DRO formulation. The model was later extended for regression [168] and feature selection [21]. A more general investigation to the distributionally robust learning models with moment ambiguity sets was provided in [62]. A general bound of the worst-case error was given under specific conditions.

Moreover, Duchi and Namkoong [54] studied the regression models from the perspective of DRO by utilizing the $\phi$-divergence ambiguity set. Specifically, the authors conducted massive numerical experiments to verify the theoretical results of lower bounds, asymptotics, tail performances, etc.

5.1.2 Other types of machine learning models. The robustness of DRO impacts other fields of machine learning. In recent years, the adversarial robust learning [78] has attracted much attention in data security. Adversarial data points are intentionally designed as barely-perceivable illusions for learning machines to make mistakes. Robustifications and regularizations [115,149] have been utilized to deal with the adversarial examples in training the models. Sinha et al. [161] studied the robustness of the Wasserstein DRO in the training process with adversarial examples. The study provides a certificate of robustness for any fixed size of the Wasserstein ambiguity set.

In addition, Blanchet and Kang [27] derived a DRO procedure for semi-supervised learning. The ambiguity is constructed based on the optimal transport metric, of which the Wasserstein metric is a special case. Frogner et al. [66] also proposed their work in distributionally robust learning with unlabeled data. By removing the distributions whose marginals do not resemble the unlabeled data from the ambiguity set, the decision set becomes smaller. The stochastic gradient descent method was adopted in [27,66] to implement the models.
5.2 DRO in operations research. We review the DRO applications in operations research in this section. Operations research is an interdiscipline that helps make better decisions with advanced analytical methods from mathematical modeling, statistical analysis, industrial engineering, operations management, and so on. As a powerful modeling tool, DRO has been widely applied to the research field of operations research. A detailed review of portfolio optimization problems are presented in this section. Applications in other topics are listed including Newsvendor problems, resource control problems, supply chain management problems, transportation problems, scheduling problems and so on.

5.2.1 Portfolio optimization. Portfolio optimization has been widely studied in the research field of RO. Recently, the DRO methodologies have achieved much success in this topic. As we introduced in Section 3, various types of distribution ambiguity sets have been adopted in the distributionally portfolio optimization problems. In Section 5.2.1.1, we discuss the DRO in portfolio optimization in terms of the type of distribution ambiguity. Then we bring a brief review of the solution approaches in the literature. One portfolio optimization model has been introduced in Example 3.2 over the convex moment-based ambiguity set. Thus We neglect the formulation review here but the problem setting in the following sections.

5.2.1.1 Various distributional ambiguity. Moment-based ambiguity sets have been adopted frequently. Back to 2003, El Ghaoui et al. [73] studied a portfolio optimization problem by minimizing the VaR over portfolios (see Example 3.2). Popescu [134] maximized the worst-case expected utility of the portfolios. The uncertainty in both works is based on the given mean and covariance. [47, 122] considered a piecewise linear concave utility of portfolios. Natarajan et al. [122] analyzed the different uncertainty occasions based on given mean, variance and support information. Delage and Ye [47] focused on the uncertainty set as an ellipsoidal set of mean vectors and a conic set of covariance matrices. [109] incorporated a penalty function on the moment discrepancy in order to avoid an overly-conservative worst-case performance. Rujeerapaiboon et al. [142] designed a distributionally robust fixed-mix-strategy approach with a moment-based ambiguity set that offers a similar performance as the classic growth-optimal portfolio model does. A stock portfolio pricing problem was studied by Van Parys et al. [173] by considering structural information with the moment-based distribution ambiguity. The mean-CVaR portfolio selection problem was studied in [96, 112]. [3] also adopted the moment-based ambiguity set in their studies.

There are distributionally robust portfolio selection models that adopt ambiguity sets based on the $\phi$-divergence. Calafiore et al. [34] solved the mean-variance and mean-absolute deviation models with ambiguity sets measured by the KL-divergence. The Wasserstein ambiguity is also adopted in the portfolio selection models. Blanchet et al. [24] studied the distributionally robust mean-variance portfolio selection model with a Wasserstein ambiguity set, while Du et al. [51] focused on the data-driven mean-CVaR portfolio selection problem.

In addition, there are other approaches to depict the distribution ambiguity and risk measurements that were investigated in literature. Wolzabal [188] introduced a framework for robustifying convex, law invariant risk measures, and applied to portfolio optimization problems. Bertsimas et al. [15] proposed a robust SAA approach by combining the idea of data-driven DRO and the traditional SAA. The distributional ambiguity was constructed by the hypothesis testing of goodness-of-fit. Motivated by asymptotic analysis in a Bayesian setting, Gupta [80] proposed a
near-optimal Bayesian ambiguity set for DRO, and applied to portfolio allocation problems.

5.2.1.2 Approaches to solve the problem. Most of the literature reviewed in Section 5.2.1.1 provided solution methods for their models. Closed forms solutions were obtained in [112, 188]. Besides, convex or conic reformulations were performed in [3, 15, 50, 96, 142, 173], which made the problems computationally tractable. In addition, a line-search approach was adopted to improve the computational efficiency in [3]. Doan et al. [50] provided an efficient linear reformulation for their model with a specific Fréchet class of discrete distributions with overlapping marginals.

5.2.2 Applications in other operations research topics. In addition to portfolio optimization problems, the idea of DRO has been applied to many real-world applications.

Newsvendor problems. In 1957, Scarf [145] studied the min-max stochastic newsvendor problem under exact moment ambiguity set. Shapiro and Kleywegt [159] showed that the worst-case problem generates a distribution, which can be regarded as a weighted distribution. With certain mild regularity conditions, the DRO problem can be equivalently reformulated as an SO problem with respect to the weighted distribution. They adopted the newsvendor problem as an example in their study. Yue et al. [199] studied the distribution-free newsvendor problem by assuming the known mean and variance information. Zhu et al. [203] proposed a stochastic robust model for the newsvendor problem. They considered the distribution of the random demand with a known mean value and one of the following: the variance or the support. Grani et al. [83] studied a risk-averse multi-dimensional newsvendor problem for a special class of products whose demands are highly correlated. A Markov ambiguity set is considered in their DRO formulation, which was shown to be NP-hard. In addition, they provided a conservative and computationally tractable approximation, which was verified to be accurate numerically. Lee et al. [106] revisited the newsvendor models by considering the risk-neutral and the risk-averse cases in DRO with Wasserstein ambiguity sets. The closed-form solutions were provided and numerical experiments were conducted to verify the out-of-sample performance of their approach.

Resource control problems. Jitka [57] studied the resources allocation problem for the water resource system in DRO, by considering a moment-based ambiguity set. Zymler et al. [205] investigated a DRO problem by approximating the joint chance constraints with the worst-case CVaR constraints. Their model was evaluated with a dynamic water reservoir control problem in their numerical study.

Supply chain management problems. Klabjan et al. [98] investigated the inventory management problem where the ambiguity set of the demand is based on histogram and the $\chi^2$-distance. Bertsimas et al. studied a facility location problem in their work [13]. They proposed an SDP model for a class of two-stage stochastic linear optimization problems with risk aversion. The ambiguity set of distributions for the second stage is based on known first and second moments.

Transportation problems. The core of intelligent transportation system considers the uncertain routing services. It has attracted attention to use DRO solving shortest path problems and vehicle routing problems (VRP) by considering the uncertainty. Zhang et al. [201] presented a distributionally robust shortest path problem based on partial information of travel times, including the support set, mean, variance and correlation matrix. They reformulated the DRO problem into
an SDP problem to give tight bounds. The numerical experiments validated the effectiveness of the DRO formulation. Wang et al. [183] proposed a data-driven distributionally robust shortest path model where the distribution of the travel time is located on a Wasserstein ball. Its robust counterpart could be reformulated as a 0-1 mixed convex program, which is easy to be solved by commercial solvers. Ghosal and Wiesemann [74] studied the capacity-constrained VRP problem in DRO with a chance-constrained ambiguity set. An efficient branch-and-cut algorithm was also provided to solve the problem under certain conditions. The author also conducted numerical experiments by comparing their model with deterministic capacity-constrained VRP problem.

Scheduling problems. There are applications of DRO on hospital planning problems. Shehadeh et al. [160] formulated the outpatient colonoscopy scheduling problem as a min-max problem considering the mean-support ambiguity set. By adopting the LP and MILP reformulation, they conducted numerical experiments and analysis on their models and showed the out-of-sample performances. Yang [194] studied a distributionally robust Markov Decision Process (MDP) problem with Wasserstein ambiguity set. An example in human-centered air conditioning problem was provided in the study.

Other problems. DRO is also utilized in the field of energy systems. Wang et al. [180, 181] studied risk-based DRO optimal power flow problems with Wasserstein ambiguity sets. Duan et al. [52] considered the chance-constrained AC power flow model with Wasserstein ambiguity set. Zhou et al. [202] proposed a data-driven distributionally robust chance-constrained real-time dispatch model based on Wasserstein ambiguity set. Koçyiğit et al. [99] studied a mechanism design problem in online auction. The problem can be formulated to help the seller achieve the goal by maximizing the worst-case expected revenue across the ambiguity set. Three different types of ambiguity sets were considered in their study.

6 Future research Directions. The above sections survey the known landscapes of the theory and applications of DRO. This section proposes some possible future research directions of DRO.

Tractable algorithms. DRO is attractive to lots of research fields due to its tractability. Although many particular DRO problems have tractable reformulations or efficient algorithms, there is still some space to improve for more general and complicated problems. For instance, most Wasserstein DRO problems focus on the Wasserstein ambiguity sets of order one. However, the more general DRO models with Wasserstein ambiguity sets of order ∞, which has been widely used in adversarial robust learning, need more tractable algorithms to be developed [69]. Moreover, most DRO problems in the literature are constructed with decision independent ambiguity sets, which means that the decision variables do not affect the construction of the ambiguity sets. Luo and Mehrotra [114] demonstrated that the dual reformulation of decision dependent DRO problems are often non-convex. Further tractable algorithms need to be explored for this type of DRO problems.

Ambiguity set construction. The principle and procedures to construct an ambiguity set are clearly described in Section 3. Nevertheless, how to construct the perfect ambiguity set is not given systematically. 1) Choice of statistic distance. The integral probability metrics and $\phi$-divergences discussed in Section 3.3 are two main approaches to measure the distance between two distributions. Birrell et al. [23] proposed a new divergence that interpolates between $\phi$-divergences and probability
metrics to range two distributions. The interpolated divergence inherits the strict concavity from $\phi$-divergences and the ability to handle not absolute continuous distributions from metrics. Applications in image generation via generative adversarial networks show the new interpolated divergence’s performance advantages over pure distances. Researchers may construct new distance-based ambiguity sets based on an interpolated divergence in the context of DRO problems to seek a better model. 2) Automatic selection of the ambiguity sets. Most DRO problems in literature define parameters in the ambiguity construction manually. Esfahani and Kuhn [61] proposed a selection rule for the radius of Wasserstein balls. Inspired by the machine learning idea, Shang and You [151] proposed a novel approach to construct ambiguity sets based on principal component analysis and first-order deviation functions. Most DRO problems cannot assure their ambiguity set is the most fitted one. Researchers may have interests to develop a generalized work to derive a perfect ambiguity set, at least for a specific class of DRO problem.

Convergence analysis. In Section 4.5, we introduce some papers that analyze the convergence of the DRO problems, including moment-based DRO and distance-based DRO. However, lots of works verified the convergence by numerical experiments. A much more compelling theoretical analysis is needed.

Benchmark DRO problems. There are some excellent benchmark works about DRO. For instance, Goh and Sim [77] designed a general framework for modeling and solving linear DRO problems with flexible piecewise linear decision rules. Wiesemann et al. [186] proposed a unifying work for convex DRO problems under a highly expressive standardized ambiguity. Bertsimas et al. [19] developed a modular and tractable framework for solving an adaptive distributionally robust linear optimization problem. There are still open topics waiting to be proposed in a high-level framework.

Apply DRO approaches to the existing problems. 1) The DRO approach to classical theoretical problems. For instance, we have seen many research works [16, 17, 133, 172] to derive optimal probability bounds (e.g. Chebyshev bounds and Gaussian bounds) in statistics. With natural connections to statistics and optimization, the DRO approach may provide a new path to other classical problems in these fields. 2) The DRO approach to classical practical problems.

This paper has shown many examples that DRO models outperform other models under uncertainty in machine learning, operations research, and other fields. Nevertheless, there are amounts of undeveloped problems to be explored via the DRO approach. One potential research topic is the DRO approach on general uncertain optimization problems. For instance, DRO approach has shown its effectiveness in general continuous multistage stochastic programming problems [156]. Also, regarding to the recent progress on integer multistage stochastic programming problems induced by stochastic dual dynamic approach [204], DRO approach may be promising to be explored for the respective integer problems. In addition, plenty of realistic problems may obtain satisfaction from DRO approach. A possible real-world problem is the machine maintenance problem with uncertain machine life time. While the traditional method relies on estimating a distribution of the remaining life time [198], DRO approach may avoid the cost of the estimation error.
Acknowledgement  The authors are grateful to the editors and referees for their valuable comments and suggestions, which have significantly improved the quality of this paper.

REFERENCES

[1] F. Alizadeh and D. Goldfarb, Second-order cone programming, Mathematical Programming, 95 (2003), 3–51.
[2] K. B. Athreya and S. N. Lahiri, Measure Theory and Probability Theory, Springer Science & Business Media, 2006.
[3] I. E. Bardakci and C. M. Lagoa, Distributionally robust portfolio optimization, in 2019 IEEE 58th Conference on Decision and Control (CDC), IEEE, (2019), 1526–1531.
[4] G. Bayraksan and D. K. Love, Data-driven stochastic programming using phi-divergences, in The Operations Research Revolution, INFORMS, (2015), 1–19.
[5] A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg and G. Rennen, Robust solutions of optimization problems affected by uncertain probabilities, Management Science, 59 (2013), 341–357.
[6] A. Ben-Tal, L. El Ghaoui and A. Nemirovski, Robust Optimization, Princeton University Press, 2009.
[7] A. Ben-Tal, E. Hazan, T. Koren and S. Mannor, Oracle-based robust optimization via online learning, Operations Research, 63 (2015), 628–638.
[8] A. Ben-Tal and A. Nemirovski, Robust convex optimization, Mathematics of Operations Research, 23 (1998), 769–805.
[9] A. Ben-Tal and A. Nemirovski, Robust solutions of uncertain linear programs, Operations Research Letters, 25 (1999), 1–13.
[10] A. Ben-Tal and A. Nemirovski, Robust solutions of linear programming problems contaminated with uncertain data, Mathematical Programming, 88 (2000), 411–424.
[11] D. Bertsimas, D. B. Brown and C. Caramanis, Theory and applications of robust optimization, SIAM Review, 53 (2011), 464–501.
[12] D. Bertsimas and M. S. Copenhaver, Characterization of the equivalence of robustification and regularization in linear and matrix regression, European Journal of Operational Research, 270 (2018), 931–942.
[13] D. Bertsimas, X. V. Doan, K. Natarajan and C.-P. Teo, Models for minimax stochastic linear optimization problems with risk aversion, Mathematics of Operations Research, 35 (2010), 580–602.
[14] D. Bertsimas, V. Gupta and N. Kallus, Data-driven robust optimization, Mathematical Programming, 167 (2018), 235–292.
[15] D. Bertsimas, V. Gupta and N. Kallus, Robust sample average approximation, Mathematical Programming, 171 (2018), 217–282.
[16] D. Bertsimas and I. Popescu, Optimal inequalities in probability theory: A convex optimization approach, SIAM Journal on Optimization, 15 (2005), 780–804.
[17] D. Bertsimas and M. Sim, The price of robustness, Operations Research, 52 (2004), 35–53.
[18] D. Bertsimas and M. Sim, Tractable approximations to robust conic optimization problems, Mathematical Programming, 107 (2006), 5–36.
[19] D. Bertsimas, M. Sim and M. Zhang, Adaptive distributionally robust optimization, Management Science, 65 (2019), 604–618.
[20] H.-G. Beyer and B. Sendhoff, Robust optimization—a comprehensive survey, Computer Methods in Applied Mechanics and Engineering, 196 (2007), 3190–3218.
[21] C. Bhattacharyya, Second order cone programming formulations for feature selection, Journal of Machine Learning Research, 5 (2004), 1417–1433.
[22] J. R. Birge and F. Louveaux, Introduction to Stochastic Programming, 2nd edition, Springer Publishing Company, Incorporated, 2011.
[23] J. Birrell, P. Dupuis, M. A. Katsoulakis, Y. Pantazis and L. Rey-Bellet, $f$-divergences: Interpolating between $f$-divergences and integral probability metrics, arXiv:2011.05953.
[24] J. Blanchet, L. Chen and X. Y. Zhou, Distributionally robust mean-variance portfolio selection with wasserstein distances, arXiv:1802.04885.
[25] J. Blanchet, P. W. Glynn, J. Yan and Z. Zhou, Multivariate distributionally robust convex regression under absolute error loss, arXiv:1905.12231.
[26] J. Blanchet and Y. Kang, Sample out-of-sample inference based on wasserstein distance, arXiv:1605.01340.
[27] J. Blanchet and Y. Kang, Semi-supervised learning based on distributionally robust optimization, Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, 5 (2020), 1–33.
[28] J. Blanchet, Y. Kang and K. Murthy, Robust wasserstein profile inference and applications to machine learning, Journal of Applied Probability, 56 (2019), 830–857.
[29] J. Blanchet, K. Murthy and N. Si, Confidence regions in wasserstein distributionally robust estimation, reprint arXiv:1906.01614.
[30] V. I. Bogachev and A. V. Kolesnikov, The monge-kantorovich problem: achievements, connections, and perspectives, Russian Mathematical Surveys, 67 (2012), 785–890.
[31] F. Bolley, A. Guillin and C. Villani, Quantitative concentration inequalities for empirical measures on non-compact spaces, Probability Theory and Related Fields, 137 (2007), 541–593.
[32] M. Breton and S. El Hachem, Algorithms for the solution of stochastic dynamic minimax problems, Computational Optimization and Applications, 4 (1995), 317–345.
[33] M. Breton and S. El Hachem, A scenario aggregation algorithm for the solution of stochastic dynamic minimax problems, Stochastics and Stochastic Reports, 53 (1995), 305–322.
[34] G. C. Calafiore, Ambiguous risk measures and optimal robust portfolios, SIAM Journal on Optimization, 18 (2007), 853–877.
[35] G. C. Calafiore and L. El Ghaoui, On distributionally robust chance-constrained linear programs, Journal of Optimization Theory and Applications, 130 (2006), 1–22.
[36] R. Chen and I. C. Paschalidis, A robust learning approach for regression models based on distributionally robust optimization, Journal of Machine Learning Research, 19 (2018), 1–48.
[37] Y. Chen, Q. Guo, H. Sun, Z. Li, W. Wu and Z. Li, A distributionally robust optimization model for unit commitment based on kullback–leibler divergence, IEEE Transactions on Power Systems, 33 (2018), 5147–5160.
[38] J. Cheng, E. Delage and A. Lässer, Distributionally robust stochastic knapsack problem, SIAM Journal on Optimization, 24 (2014), 1485–1506.
[39] 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 Journal on Optimization, 28 (2018), 1817–1841.
[40] A. Cherukuri and J. Cortés, Cooperative data-driven distributionally robust optimization, IEEE Transactions on Automatic Control, 65 (2019), 4400–4407.
[41] V. K. Chopra and W. T. Ziemba, The effect of errors in means, variances, and covariances on optimal portfolio choice, in Handbook of the Fundamentals of Financial Decision Making: Part I, World Scientific, (2013), 365–373.
[42] K. L. Clarkson, E. Hazan and D. P. Woodruff, Sublinear optimization for machine learning, Journal of the ACM (JACM), 59 (2012), 1–49.
[43] A. R. Conn and N. I. Gould, An exact penalty function for semi-infinite programming, Mathematical Programming, 37 (1987), 19–40.
[44] T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd edition, Wiley Online Library, Hoboken, 2012.
[45] G. B. Dantzig, Linear programming under uncertainty, Management Science, 50 (2004), 1764–1769.
[46] E. del Barrio, E. Giné and C. Matrán, Central limit theorems for the wasserstein distance between the empirical and the true distributions, Annals of Probability, 27 (1999), 1009–1071.
[47] E. Delage and Y. Ye, Distributionally robust optimization under moment uncertainty with application to data-driven problems, Operations Research, 58 (2010), 595–612.
[48] S. Dharmadhikari and K. Joag-Dev, Unimodality, Convexity, and Applications, Elsevier, 1988.
[49] T. Dinh, R. Fukasawa and J. Luedtke, Exact algorithms for the chance-constrained vehicle routing problem, Mathematical Programming, 172 (2018), 105–138.
[50] X. V. Doan, X. Li and K. Natarajan, Robustness to dependency in portfolio optimization using overlapping marginals, Operations Research, 63 (2015), 1468–1488.
[51] N. Du, Y. Liu and Y. Liu, A new data-driven distributionally robust portfolio optimization method based on wasserstein ambiguity set, IEEE Access, 9 (2021), 3174–3194.
[52] C. Duan, W. Fang, L. Jiang, L. Yao and J. Liu, Distributionally robust chance-constrained approximate ac-opf with wasserstein metric, IEEE Transactions on Power Systems, 33 (2018), 4924–4936.
[53] J. C. Duchi and H. Namkoong, Variance-based regularization with convex objectives, Journal of Machine Learning Research, 20 (2019), 1–68.
[54] J. C. Duchi and H. Namkoong, Learning models with uniform performance via distributionally robust optimization, arXiv:1810.08750.
[55] J. Dupaˇcová, Stochastic programming: Minimax approach, Encyclopedia Optimization, 5 (2001), 327–330.
[56] J. Dupaˇcová, On minimax solutions of stochastic linear programming problems, Časopis Pro Pˇestov´an´ı Matematiky, 091 (1966), 423–430.
[57] J. Dupaˇcová, The minimax approach to stochastic programming and an illustrative application, Stochastics, 20 (1987), 73–88.
[58] J. Dupaˇcová, Uncertainties in minimax stochastic programs, Optimization, 60 (2011), 1235–1250.
[59] L. El Ghaoui and H. Lebret, Robust solutions to least-squares problems with uncertain data, SIAM Journal on Matrix Analysis and Applications, 18 (1997), 1035–1064.
[60] E. Erdogan and G. Iyengar, Ambiguous chance constrained problems and robust optimization, Mathematical Programming, 107 (2006), 37–61.
[61] P. M. Esfahani and D. Kuhn, Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations, Mathematical Programming, 171 (2018), 115–166.
[62] F. Farnia and D. Tse, A minimax approach to supervised learning, Advances in Neural Information Processing Systems, 29 (2016), 4240–4248.
[63] L. Faury, U. Tanielian, E. Dohmatob, E. Smirnova and F. Vasile, Distributionally robust counterfactual risk minimization, Proceedings of the AAAI Conference on Artificial Intelligence, 34 (2020), 3850–3857.
[64] N. Fournier and A. Guillin, On the rate of convergence in wasserstein distance of the empirical measure, Probability Theory and Related Fields, 162 (2015), 707–738.
[65] D. Fouskakis and D. Draper, Stochastic optimization: a review, International Statistical Review, 70 (2002), 315–349.
[66] C. Frogner, S. Claicì, E. Chien and J. Solomon, Incorporating unlabeled data into distributionally robust learning, Journal of Machine Learning Research, 22 (2021), 1–46.
[67] V. Gabrel, C. Murat and A. Thiele, Recent advances in robust optimization: An overview, European Journal of Operational Research, 235 (2014), 471–483.
[68] G. Gallego and I. Moon, The distribution free newsboy problem: review and extensions, Journal of the Operational Research Society, 44 (1993), 825–834.
[69] R. Gao, Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality, arXiv:2009.04382.
[70] R. Gao and A. J. Kleywegt, Distributionally robust stochastic optimization with wasserstein distance, arXiv:1604.02199.
[71] R. Gao and A. J. Kleywegt, Distributionally robust stochastic optimization with dependence structure, arXiv:1701.04200.
[72] R. Gao, L. Xie, Y. Xie and H. Xu, Robust hypothesis testing using wasserstein uncertainty sets, in NeurIPS, (2018), 7913–7923.
[73] L. E. Ghaoui, M. Oks and F. Oustry, Worst-case value-at-risk and robust portfolio optimization: A conic programming approach, Operations Research, 51 (2003), 543–556.
[74] S. Ghosal and W. Wiesemann, The distributionally robust chance-constrained vehicle routing problem, Operations Research, 68 (2020), 716–732.
[75] A. L. Gibbs and F. E. Su, On choosing and bounding probability metrics, International Statistical Review, 70 (2002), 419–435.
[76] M. Á. Goberna and M. A. López, Semi-Infinite Programming: Recent Advances, Springer Science & Business Media, 2013.
[77] J. Goh and M. Sim, Distributionally robust optimization and its tractable approximations, Operations Research, 58 (2010), 902–917.
[78] I. J. Goodfellow, J. Shlens and C. Szegedy, Explaining and harnessing adversarial examples, arXiv:1412.6572.
[79] S. Guo, H. Xu and L. Zhang, Probability approximation schemes for stochastic programs with distributionally robust second-order dominance constraints, Optimization Methods and Software, 32 (2017), 770–789.

[80] V. Gupta, Near-optimal bayesian ambiguity sets for distributionally robust optimization, Management Science, 65 (2019), 4242–4260.

[81] S. A. Gustafson, Semi-infinite programming: Methods for linear problems, in Encyclopedia of Optimization, Springer, (2009), 3424–3429.

[82] G. A. Hanasusanto and D. Kuhn, Conic programming reformulations of two-stage distributionally robust linear programs over wasserstein balls, Operations Research, 66 (2018), 849–869.

[83] G. A. Hanasusanto, D. Kuhn, S. W. Wallace and S. Zymler, Distributionally robust multi-item newsvendor problems with multimodal demand distributions, Mathematical Programming, 152 (2015), 1–32.

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

[85] R. Hettich, A. Kaplan and R. Tichatschke, Semi-infinite programming: Numerical methods, in Encyclopedia of Optimization, Springer, (2009), 3429–3434.

[86] R. Hettich and K. O. Kortanek, Semi-infinite programming: theory, methods, and applications, SIAM Review, 35 (1993), 380–429.

[87] Z. Hu and L. J. Hong, Kullback-leibler divergence constrained distributionally robust optimization, Available at Optimization Online.

[88] K. Huang, H. Yang, I. King, M. R. Lyu and L. Chan, The minimum error minimax probability machine, Journal of Machine Learning Research, 5 (2004), 1253–1286.

[89] G. Infanger, Planning under uncertainty solving large-scale stochastic linear programs, Technical report, Stanford University, 1992.

[90] K. Isii, On sharpness of tchebycheff-type inequalities, Annals of the Institute of Statistical Mathematics, 14 (1962), 185–197.

[91] R. Ji and M. A. Lejeune, Data-driven distributionally robust chance-constrained optimization with wasserstein metric, Journal of Global Optimization, (2020), 1–33.

[92] R. Jiang and Y. Guan, Data-driven distributionally robust appointment scheduling over wasserstein balls, arXiv:1907.03219.

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

[94] C. Koçyiğit, G. Iyengar, D. Kuhn and W. Wiesemann, Distributionally robust mechanism design, Management Science, 66 (2020), 159–189.

[95] D. Kuhn, P. M. Esfahani, V. A. Nguyen and S. Shafieezadeh-Abadeh, Wasserstein distributionally robust optimization: Theory and applications in machine learning, in Operations Research & Management Science in the Age of Analytics, INFORMS, (2019), 130–166.

[96] S. Kullback, Information Theory and Statistics, Courier Corporation, Mineola, 1997.
[107] D. Levy, Y. Carmon, J. C. Duchi and A. Sidford, Large-scale methods for distributionally robust optimization, *Advances in Neural Information Processing Systems*, 33.

[108] A. S. Lewis and C. J. Pang, Lipschitz behavior of the robust regularization, *SIAM Journal on Control and Optimization*, 48 (2010), 3080–3104.

[109] J. Y. Li and R. H. Kwon, Portfolio selection under model uncertainty: a penalized moment-based optimization approach, *Journal of Global Optimization*, 56 (2013), 131–164.

[110] G. D. Lin, Recent developments on the moment problem, *Journal of Statistical Distributions and Applications*, 4 (2017), 1–17.

[111] Q. Lin, R. Loxton, K. L. Teo, Y. H. Wu and C. Yu, A new exact penalty method for semi-infinite programming problems, *Journal of Computational and Applied Mathematics*, 261 (2014), 271–286.

[112] J. Liu, Z. Chen, A. Lisser and Z. Xu, Closed-form optimal portfolios of distributionally robust mean-cvar problems with unknown mean and variance, *Applied Mathematics & Optimization*, 79 (2019), 671–693.

[113] Y. Liu, R. Meskarian and H. Xu, Distributionally robust reward-risk ratio optimization with moment constraints, *SIAM Journal on Optimization*, 27 (2017), 957–985.

[114] F. Luo and S. Mehrotra, Decomposition algorithm for distributionally robust optimization using wasserstein metric with an application to a class of regression models, *European Journal of Operational Research*, 278 (2019), 20–35.

[115] C. Lyu, K. Huang and H.-N. Liang, A unified gradient regularization family for adversarial examples, in *2015 IEEE International Conference on Data Mining*, IEEE, (2015), 301–309.

[116] A. Majumdar, G. Hall and A. A. Ahmadi, Recent scalability improvements for semidefinite programming with applications in machine learning, control, and robotics, *Annual Review of Control, Robotics, and Autonomous Systems*, 3 (2020), 331–360.

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

[118] S. Mehrotra and H. Zhang, Models and algorithms for distributionally robust least squares problems, *Mathematical Programming*, 146 (2014), 123–141.

[119] R. O. Michaud, The markowitz optimization enigma: Is ‘optimized’ optimal? *Financial Analysts Journal*, 45 (1989), 31–42.

[120] H. Mostafaei and S. Kordnourie, Probability metrics and their applications, *Applied Mathematical Sciences*, 5 (2011), 181–192.

[121] H. Namkoong and J. C. Duchi, Stochastic gradient methods for distributionally robust optimization with f-divergences, in *NIPS*, 29 (2016), 2208–2216.

[122] K. Natarajan, M. Sim and J. Uichanco, Tractable robust expected utility and risk models for portfolio optimization, *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 20 (2010), 695–731.

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

[124] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, Robust stochastic approximation approach to stochastic programming, *SIAM Journal on Optimization*, 19 (2009), 1574–1609.

[125] A. Nemirovski and A. Shapiro, Convex approximations of chance constrained programs, *SIAM Journal on Optimization*, 17 (2007), 969–996.

[126] V. A. Nguyen, D. Kuhn and P. M. Esfahani, Distributionally robust inverse covariance estimation: The wasserstein shrinkage estimator, arXiv:1805.07194.

[127] V. A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn and P. M. Esfahani, Bridging bayesian and minimax mean square error estimation via wasserstein distributionally robust optimization, arXiv:1911.03539.

[128] A. B. Owen, *Empirical Likelihood*, CRC Press, 2001.

[129] L. Pardo, *Statistical Inference Based on Divergence Measures*, CRC Press, 2018.

[130] G. C. Pflug and A. Pichler, Approximations for probability distributions and stochastic optimization problems, in *Stochastic Optimization Methods in Finance and Energy*, Springer, 2011, 343–387.

[131] R. R. Phelps, *Lectures on Choquet’s Theorem*, Springer Science & Business Media, 2001.

[132] A. B. Philpott, V. L. de Matos and L. Kapelevich, Distributionally robust sadp, *Computational Management Science*, 15 (2018), 431–454.

[133] I. Popescu, A semidefinite programming approach to optimal-moment bounds for convex classes of distributions, *Mathematics of Operations Research*, 30 (2005), 632–657.
[134] I. Popescu, Robust mean-covariance solutions for stochastic optimization, *Operations Research*, 55 (2007), 98–112.
[135] K. Postek, D. den Hertog and B. Melenberg, Computationally tractable counterparts of distributionally robust constraints on risk measures, *SIAM Review*, 58 (2016), 603–650.
[136] M. A. Proschan and P. A. Shaw, *Essentials of Probability Theory for Statisticians*, CRC Press, 2018.
[137] H. Rahimian, G. Bayraksan and T. Homem-de Mello, Identifying effective scenarios in distributionally robust stochastic programs with total variation distance, *Mathematical Programming*, 173 (2019), 393–430.
[138] H. Rahimian and S. Mehrotra, Distributionally robust optimization: A review, arXiv:1908.05659.
[139] M. Riis and K. A. Andersen, Applying the minimax criterion in stochastic recourse programs, *European Journal of Operational Research*, 165 (2005), 569–584.
[140] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.
[141] Y. Rubner, C. Tomasi and L. J. Guibas, The earth mover’s distance as a metric for image retrieval, *International Journal of Computer Vision*, 40 (2000), 99–121.
[142] N. Rujerapaiboon, D. Kuhn and W. Wiesemann, Robust growth-optimal portfolios, *Management Science*, 62 (2016), 2090–2109.
[143] N. Rujerapaiboon, D. Kuhn and W. Wiesemann, Chebyshev inequalities for products of random variables, *Mathematics of Operations Research*, 43 (2018), 887–918.
[144] L. R¨uschendorf, Bounds for distributions with multivariate marginals, *Lecture Notes-Monograph Series*, (1991), 285–310.
[145] H. Scarf, A min-max solution of an inventory problem, *Studies in the Mathematical Theory of Inventory and Production*.
[146] S. Shafieezadeh-Abadeh, D. Kuhn and P. M. Esfahani, Regularization via mass transportation, *Journal of Machine Learning Research*, 20 (2019), 1–68.
[147] S. Shafieezadeh Abadeh, P. M. Mohajerin Esfahani and D. Kuhn, Distributionally robust logistic regression, *Advances in Neural Information Processing Systems*, 28 (2015), 1576–1584.
[148] S. Shafieezadeh-Abadeh, V. A. Nguyen, D. Kuhn and P. M. Esfahani, Wasserstein distributionally robust kalman filtering, in *Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018, 8474–8483.
[149] U. Shaham, Y. Yamada and S. Negahban, Understanding adversarial training: Increasing local stability of supervised models through robust optimization, *Neurocomputing*, 307 (2018), 195–204.
[150] S. Shalev-Shwartz and Y. Wexler, Minimizing the maximal loss: How and why, in *International Conference on Machine Learning*, PMLR, 2016, 793–801.
[151] C. Shang and F. You, Distributionally robust optimization for planning and scheduling under uncertainty, *Computers & Chemical Engineering*, 110 (2001), 135–165.
[152] A. Shapiro, On duality theory of conic linear programs, in *Semi-infinite Programming*, Springer, Boston, MA, 2001, 135–165.
[153] A. Shapiro, *Worst-case distribution analysis of stochastic programs*, *Mathematical Programming*, 107 (2006), 91–96.
[154] A. Shapiro, *Semi-infinite programming, duality, discretization and optimality conditions*, *Optimization*, 58 (2009), 133–161.
[155] A. Shapiro, Distributionally robust stochastic programming, *SIAM Journal on Optimization*, 27 (2017), 2258–2275.
[156] A. Shapiro, Tutorial on risk neutral, distributionally robust and risk averse multistage stochastic programming, *European Journal of Operational Research*, 288 (2020), 1–13.
[157] A. Shapiro and S. Ahmed, On a class of minimax stochastic programs, *SIAM Journal on Optimization*, 14 (2004), 1237–1249.
[158] A. Shapiro, D. Dentcheva and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory*, 2nd edition, SIAM, Philadelphia, PA, 2014.
[159] A. Shapiro and A. Kleywegt, Minimax analysis of stochastic problems, *Optimization Methods and Software*, 17 (2002), 523–542.
[160] K. S. Shehadeh, A. E. Cohn and R. Jiang, A distributionally robust optimization approach for outpatient colonoscopy scheduling, *European Journal of Operational Research*, 283 (2020), 549–561.
A. Sinha, H. Namkoong and J. Duchi, Certifying some distributional robustness with principled adversarial training, in International Conference on Learning Representations, 2018.

J. E. Smith, Generalized chebyshev inequalities: theory and applications in decision analysis, Operations Research, 43 (1995), 807–825.

J. E. Smith and R. L. Winkler, The optimizer’s curse: Skepticism and postdecision surprise in decision analysis, Management Science, 52 (2006), 311–322.

A. M.-C. So, Moment inequalities for sums of random matrices and their applications in optimization, Mathematical Programming, 130 (2011), 125–151.

A. L. Soyster, Convex programming with set-inclusive constraints and applications to inexact linear programming, Operations Research, 21 (1973), 1154–1157.

A. Spanos, Probability Theory and Statistical Inference: Econometric Modeling with Observational Data, Cambridge University Press, Cambridge, 1999.

G. Still, Optimization problems with infinitely many constraints, Buletinul tainnic al Universitatu Bi Bi Mare, Seria B, Fascicola matematic˘ a-informatic˘ a, 18 (2002), 343–354.

T. Strohmann and G. Z. Gradic, A formulation for minimax probability machine regression, in NIPS, Citeseer, 2002, 769–776.

H. Sun and H. Xu, Convergence analysis for distributionally robust optimization and equilibrium problems, Mathematics of Operations Research, 41 (2016), 377–401.

S. Takriti and S. Ahmed, Managing short-term electricity contracts under uncertainty: A minimax approach, 2002.

B. P. Van Parys, P. M. Esfahani and D. Kuhn, From data to decisions: Distributionally robust optimization is optimal, Management Science, (2020), preprint.

B. P. Van Parys, P. J. Goulart and D. Kuhn, Generalized gauss inequalities via semidefinite programming, Mathematical Programming, 156 (2016), 271–302.

B. P. Van Parys, P. J. Goulart and M. Morari, Distributionally robust expectation inequalities for structured distributions, Mathematical Programming, 173 (2019), 251–280.

L. Vandenberghe and S. Boyd, Semidefinite programming, SIAM Review, 38 (1996), 49–95.

L. Vandenberghe, S. Boyd and K. Comanor, Generalized chebyshev bounds via semidefinite programming, SIAM Review, 49 (2007), 52–64.

C. Villani, Optimal Transport: Old and New, vol. 338, Springer Science & Business Media, 2008.

M. R. Wagner, Stochastic 0–1 linear programming under limited distributional information, Operations Research Letters, 36 (2008), 150–156.

A. Wald, Statistical decision functions which minimize the maximum risk, Annals of Mathematics, 46 (1945), 265–280.

A. Wald, Statistical decision functions, in Breakthroughs in Statistics, Springer, 1992, 342–357.

C. Wang, R. Gao, F. Qiu, J. Wang and L. Xin, Risk-based distributionally robust optimal power flow with dynamic line rating, IEEE Transactions on Power Systems, 33 (2018), 6074–6086.

C. Wang, R. Gao, W. Wei, M. Shafie-khah, T. Bi and J. P. Catalao, Risk-based distributionally robust optimal gas-power flow with wasserstein distance, IEEE Transactions on Power Systems, 34 (2018), 2190–2204.

S. Wang and Y. Yuan, Feasible method for semi-infinite programs, SIAM Journal on Optimization, 25 (2015), 2537–2560.

Z. Wang, K. You, S. Song and Y. Zhang, Wasserstein distributionally robust shortest path problem, European Journal of Operational Research, 284 (2020), 31–43.

Z. Wang, P. W. Glynn and Y. Ye, Likelihood robust optimization for data-driven problems, Computational Management Science, 13 (2016), 241–261.
[190] H. Xu, C. Caramanis and S. Mannor, Robust regression and lasso, *IEEE Transactions on Information Theory*, 56 (2010), 3561–3574.

[191] H. Xu, C. Caramanis and S. Mannor, A distributional interpretation of robust optimization, *Mathematics of Operations Research*, 37 (2012), 95–110.

[192] H. Xu, Y. Liu and H. Sun, Distributionally robust optimization with matrix moment constraints: Lagrange duality and cutting plane methods, *Mathematical Programming*, 169 (2018), 489–529.

[193] M. Xu, S.-Y. Wu and J. Y. Jane, Solving semi-infinite programs by smoothing projected gradient method, *Computational Optimization and Applications*, 59 (2014), 591–616.

[194] I. Yang, A convex optimization approach to distributionally robust markov decision processes with wasserstein distance, *IEEE Control Systems Letters*, 1 (2017), 164–169.

[195] I. Yang, Wasserstein distributionally robust stochastic control: A data-driven approach, *IEEE Transactions on Automatic Control*, (2020), 1–8.

[196] X. Yang, Z. Chen and J. Zhou, Optimality conditions for semi-infinite and generalized semi-infinite programs via lower order exact penalty functions, *Journal of Optimization Theory and Applications*, 169 (2016), 984–1012.

[197] Y. Ye, *Interior Point Algorithms: Theory and Analysis*, John Wiley & Sons, 2011.

[198] M. Yildirim, X. A. Sun and N. Z. Gebraeel, Sensor-driven condition-based generator maintenance scheduling-part i: Maintenance problem, *IEEE Transactions on Power Systems*, 31 (2016), 4253–4262.

[199] J. Yue, B. Chen and M.-C. Wang, Expected value of distribution information for the newsvendor problem, *Operations Research*, 54 (2006), 1128–1136.

[200] Y. Zhang, R. Jiang and S. Shen, Ambiguous chance-constrained binary programs under mean-covariance information, *SIAM Journal on Optimization*, 28 (2018), 2922–2944.

[201] Y. Zhang, S. Song, Z.-J. M. Shen and C. Wu, Robust shortest path problem with distributional uncertainty, *IEEE transactions on intelligent transportation systems*, 19 (2017), 1080–1090.

[202] A. Zhou, M. Yang, M. Wang and Y. Zhang, A linear programming approximation of distributionally robust chance-constrained dispatch with wasserstein distance, *IEEE Transactions on Power Systems*, 35 (2020), 3366–3377.

[203] Z. Zhu, J. Zhang and Y. Ye, Newsvendor optimization with limited distribution information, *Optimization Methods and Software*, 28 (2013), 640–667.

[204] J. Zou, S. Ahmed and X. A. Sun, Stochastic dual dynamic integer programming, *Mathematical Programming*, 175 (2019), 461–502.

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

Received February 2021; revised May 2021, Final revision June 2021; Early access November 2021.

E-mail address: flin6@ncsu.edu
E-mail address: xfang8@ncsu.edu
E-mail address: gaozheming@ise.neu.edu.cn