Data-driven distributionally robust optimization: Intersecting ambiguity sets, performance analysis and tractability

Neda Tanoumand, Merve Bodur, Joe Naoum-Sawaya

[1]Department of Mechanical and Industrial Engineering,
University of Toronto, Toronto, Ontario M5S 3G8, Canada
{neda.tanoumand,bodur}@mail.utoronto.ca

[2]Ivey Business School, University of Western Ontario,
1255 Western Rd, London, ON N6G 0N1
jnaoum-sawaya@ivey.ca

Abstract We consider stochastic programs in which the probability distribution of uncertain parameters is unknown and partial information about it can only be captured from limited data. We use distributionally robust optimization (DRO) to model such problems. As opposed to the commonly used approach for DRO problems that suggests creating an ambiguity set by following a specific procedure, we propose to build it by adopting multiple procedures. Specifically, we design the ambiguity set by intersecting sets that are constructed by different discrepancy-based measures. The new ambiguity set excludes the probability distributions that reside in only one set, while preserving the common ones, which prevents the DRO problem from producing overly conservative solutions. We derive single-level convex programming reformulations for the resulting two-level DRO problems with various supports, namely, discrete known, univariate, and multivariate supports. We additionally design a three-level cutting-plane algorithm and a relaxation-based technique to tackle computationally challenging DRO problems with multivariate supports. To evaluate the quality of the solutions that our reformulations yield and the performance of these techniques, we conduct experiments on newsvendor and mean-risk portfolio allocation problems. Our results suggest that using multiple measures to build the ambiguity set decreases robustness and provides high-quality solutions, especially for small-size samples. Moreover, the outcomes illustrate that our proposed algorithm and the upper-bounding technique are indeed effective.

1 Introduction

Many decision problems in practice are solved under uncertainty when the exact values of the parameters are not available or difficult to obtain. Stochastic programming (SP) is a powerful modeling approach which enables the incorporation of uncertainty into a mathematical model [Ghasemi et al., 2020, Hong et al., 2015]. Given an underlying probability distribution of uncertain parameters, i.e. true data-generating distribution, SP often aims to optimize the expected value of the objective function. In such settings, the goal is to find a solution that minimizes/maximizes the expected value of a function, which is a function of decision variables and a random vector of parameters following the given distribution [Prékopa, 2013]. Computing the expected value requires the calculation of multiple integrals which can be computationally intractable and requires

approximation methods. One of the well-known techniques for approximating the SP problem is the sample average approximation (SAA), which uses a sample of random vectors drawn from the underlying probability distribution [Birge and Louveaux, 2011].

Given a sample of observations for a set of independent and identically distributed (IID) random variables, SAA approximates the empirical distribution by assigning equal probabilities to each data point in the sample. It has been shown in the literature that under mild conditions SAA is computationally tractable and provides asymptotic convergence, meaning that as the sample size tends to infinity, the optimal objective value and the optimal solution of the SAA problem converge to those of the SP problem [Kleywegt et al., 2002]. In such a setting, the true data-generating distribution is assumed to be known and sampling from it is possible. However, there are applications where only a small sample of data is available or there is limited information on the true distribution [Kapteyn et al., 2019]. Additionally, there are many applications such as healthcare related problems and emergency resource allocation problems where the events are rare and data is scarce. In such cases, SAA might be used in which case one approximates the underlying distribution by an empirical distribution which assigns equal probability for each observation. However, the optimal objective values provided by SAA have been shown to be highly optimistic and sample dependent, especially for small sample sizes. Moreover, considering a minimization problem, SAA provides only a valid lower bound, and its optimal solution has poor out-of-sample performance. Besides, the optimal objective value and the optimal solution have large variations [Bertsimas et al., 2018]. For the problems with scarce data, efficient data-driven techniques are required in order to use the available data effectively.

Distributionally robust optimization (DRO) is a data-driven modeling approach that assumes the availability of partial information on the underlying probability distribution, while SP assumes the full availability of the distribution. DRO aims at finding a solution that minimizes the worst-case expected cost with respect to a set of distributions which is called ambiguity set. This set contains a family of probability distributions that share properties similar to the true data-generating distribution. The set can be constructed based on the information that the available data implies about the underlying true distribution. In the DRO literature, various procedures are proposed for constructing ambiguity sets where these methods are designed based on different measures such as discrepancy metrics and moments of the probability distributions. Following a particular procedure, the ambiguity set contains all distributions satisfying measure-specific properties. By choosing the measure carefully, the resulting ambiguity set provides beneficial properties for the DRO problem, namely asymptotic convergence, performance guarantee, and tractability [Bertsimas et al., 2018].

To elaborate, under mild assumptions the optimal objective value and the optimal solution of the DRO problem converge to those of the SP problem as the sample sizes increase. Moreover, the optimal objective value of the DRO problem can provide a probabilistically guaranteed upper bound over its out-of-sample performance. In other words, considering a minimization problem, solving the DRO problem given a sample, yields an optimal objective value and a solution such that evaluation of the solution on another sample is bounded above by the optimal objective value with high probability. Additionally, for many problems of practical interest, such as two-stage stochastic programming problems and risk-averse problems, DRO is equivalent to a tractable single-level convex optimization problem that can be tackled by off-the-shelf solvers. However, the ambiguity set might contain distributions which lead to extremely conservative solution and objective value for the DRO problem especially for small sample sizes.

In order to prevent this robustness, we propose to build the ambiguity set as an intersection of multiple ambiguity sets. The intersection yields a *joint* set that may exclude the extreme distributions that belong to individual sets and result in robust outcomes to the DRO problem. The idea is to exclude any distributions

that reside in only individual sets and preserve the common ones. In this case, the joint set can be constructed by incorporating multiple measures and the DRO problem with the joint ambiguity set can inherit certain useful measure-specific properties of the DRO problems with individual sets. In this work, we examine the conditions under which the DRO problem with joint set can preserve those properties while still maintains tractability. We illustrate that under mild assumptions, this DRO problem provides at least as good in-sample performance as each of the individual problems.

More specifically, we construct a joint ambiguity set for the DRO problem using Wasserstein metric and Goodness-of-Fit (GoF) tests. For problems with discrete-known and univariate supports, we intersect the sets generated with various GoF tests. For problems with multivariate support, we intersect the Wasserstein ball and the confidence region of the linear-convex ordering test. We discuss the assumptions under which the DRO problems with individual and joint ambiguity sets possess performance guarantee and convergence. Moreover, for the resulting problems, we derive tractable single-level convex reformulations which are solvable by the state-of-the-art solvers. For computationally challenging problems in the multivariate setting, we also propose a three-level cutting-plane algorithm and a relaxation-based procedure to find lower and upper bounds on the optimal value of the DRO problem. We conduct computational experiments to compare the quality of the bounds provided by the DRO problem with joint ambiguity set and those from the DRO problems with individual ambiguity set. We illustrate the performance of our approach on newsvendor problem with random demand and mean-risk portfolio allocation problem with uncertain returns. Our experiments indicate that the DRO problems with joint ambiguity set in most of the cases can yield better solutions than the DRO problems with single-measure based ambiguity sets in terms of in-sample and out-of-sample performance, especially for small sample sizes.

Our contributions can be listed as follows:

- In modeling SP problems using a DRO framework, we propose to construct the ambiguity set as an intersection of multiple single-measure based ambiguity sets.
- We study the conditions under which the proposed DRO problem possesses useful properties.
- We derive single-level convex reformulations for two-level DRO problems with joint ambiguity sets for problems with discrete known, univariate, and multivariate supports.
- We propose lower and upper bounding techniques for computationally demanding problems in the multivariate setting and show the efficacy and efficiency of the techniques.
- We evaluate the quality of the solutions provided by the DRO problem with a joint ambiguity set in terms of their in-sample and out-of-sample performances using the two well-known problems from the literature, namely, newsvendor problem with random demand and mean-risk portfolio optimization problem with random returns.

The rest of the paper is organized as follows. In Section 2, we review the literature and discuss the proposed techniques therein for constructing ambiguity sets. In Section 3, we provide a brief background and mathematical formulation of the problems as well as the measures that we use later in this work. In Section 4, we elaborate on our idea of constructing a joint ambiguity set and provide the conditions under which the DRO problem can achieve beneficial properties. In Sections 5, 6, and 7, we propose tractable reformulations for DRO problems with joint ambiguity sets under discrete known, univariate, and multivariate supports, respectively. In Section 7, we propose bounding techniques for problems with multivariate support. In Section

8, we illustrate the quality of the bounds provided by the DRO problems over newsvendor problem and mean-risk portfolio optimization problem. In Section 9, we explain the advantages of using our approach over existing approaches in the literature. Additionally, we provide a summary table on the various problem settings that we study and discuss the tractability of the reformulations we derive. Next, in Section 10, we conclude the paper with a summary of our findings and discuss future work.

2 Literature Review

In this section, we review the relevant literature regarding constructing an ambiguity set for DRO problems. In the literature, ambiguity sets can be categorized into two main groups, namely, moment-based and discrepancy-based ambiguity sets. While the moment-based ambiguity sets contain all distributions whose moments satisfy specific properties, discrepancy-based ambiguity sets contain all distributions in a certain distance with respect to a specific discrepancy measure from a reference distribution which can be obtained from available data.

Moment-based ambiguity sets are usually constructed based on the first and second moment information of a reference distribution. Scarf [1957] presents one of the first studies in the DRO context and studies a distributionally robust newsvendor problem with random demand where the ambiguity set is constructed based on the known mean and variance of the demand. Extending the work, Gallego and Moon [1993] assume that mean and covariance matrix are unknown and reside in a polytopic and interval uncertainty sets, while Lotfi and Zenios [2018] consider them to reside in an ellipsoidal set. Rujeerapaiboon et al. [2018] study a DRO problem where they assume that the mean is known, but the covariance matrix is unknown or bounded above. Delage and Ye [2010] assume that the moments of a random variable are unknown and construct an ambiguity set using a data-driven procedure to build a confidence region for them. They consider a minimization problem and prove that the optimal objective value of the DRO problem with the proposed ambiguity set provides a probabilistic upper bound on the out-of-sample performance of the optimal solution of the DRO problem.

Shapiro and Ahmed [2004] propose a framework based on generalized inequalities which can be used for modeling the support of a random vector, defining bounds on the probability measure, and defining bounds on the function of the random variable. Particular cases of the framework appears in some other works such as [Mehrotra and Papp, 2014, Perakis and Roels, 2008]. Using the same idea, Royset and Wets [2017] propose a decision-dependent ambiguity set which is constructed by imposing bounds on the decision-dependent cumulative distribution of a random vector and analyze the convergence results of the corresponding DRO problem. While the studies mentioned so far construct an ambiguity set based on the available information on the joint distribution of a random vector, there are studies that assume the availability of additional information on marginal probability distributions as well [Chen et al., 2022a, Dhara et al., 2021, Doan et al., 2015].

Moment-based ambiguity sets are studied in the context of risk-averse decision making. Liu et al. [2017] study reward-risk ratio model where the ambiguity set is constructed by restricting mean and covariance matrix entries to an interval. Natarajan et al. [2014] consider CVaR minimization problem with ambiguity set created based on the information on marginal distributions. Beside risk-averse decision making, moment-based ambiguity sets are incorporated into problems with individual and joint chance-constraints. Hanasusanto et al. [2017], Xie et al. [2022], and Xie and Ahmed [2018] study chance-constrained problems where the ambiguity set is constructed by imposing generalized inequalities on the first/second moment of a random

vector. For information on various moment-based ambiguity sets and their mathematical foundations, we refer the interested reader to Rahimian and Mehrotra [2019].

Alternatively, using a discrepancy measure the ambiguity set can be restricted to the probability distributions with small dissimilarity to the nominal distribution. Optimal transport discrepancy measures such as Wasserstein metric are very well-known ones in this category which have gained significant popularity in recent years. The metrics compute the minimum cost associated with transporting a mass from a probability distribution to another one with respect to a discrepancy measure such as p-Wasserstein metric. Modeling a DRO problem with optimal transport metrics can be seen in many operations research and machine learning studies such as [Gao and Kleywegt, 2022, Chen et al., 2022b, Blanchet et al., 2022, Luo and Mehrotra, 2019, Lee and Raginsky, 2018, Esfahani and Kuhn, 2018, Mehrotra and Zhang, 2014].

It has been illustrated by Esfahani and Kuhn [2018] that the ambiguity set can be constructed using the Wasserstein distance measure. In this case, the modern measure concentration result guarantees the existence of the true data-generating probability distribution in the Wasserstein ball with high confidence [Bolley et al., 2007]. While the authors use 1-Wasserstein metric, they show that under mild assumptions the optimal objective value of the DRO problem with the Wasserstein ambiguity set provides a probabilistic guarantee on the out-of-sample performance of its optimal solution, also enjoys asymptotic convergence and tractability for numerous cost functions. In a similar work, Gao and Kleywegt [2022] study a DRO problem with an ambiguity set constructed by p-Wasserstein metric where p-norm is utilized as a distance measure. The authors identify necessary and sufficient conditions for the existence of a worst-case distribution and show that the distribution has a certain structure. Based on the structure of the distribution they argue that any data-driven DRO can be approximated by an appropriate robust optimization problem.

In an alternative way, GoF tests can be considered as discrepancy measures and used to construct the ambiguity sets. Bertsimas et al. [2018] propose constructing ambiguity sets as a confidence region of GoF tests which contain all probability distributions that pass the corresponding hypothesis test. The framework is called Robust SAA and aims at leveraging useful properties provided by SAA while possessing a guarantee on the quality of its performance. The authors show that under mild assumptions the DRO problem with GoF test-based ambiguity set can be reformulated as a tractable convex optimization problem, also provides asymptotic convergence and performance guarantees. Postek et al. [2016] study DRO problems with risk constraints where GoF tests are used for constructing ambiguity sets. The authors propose tractable reformulations for these problems.

There are other studies in the literature that consider different discrepancy measures such as ϕ-divergence [Blanchet et al., 2022, Lam, 2019, Jiang and Guan, 2016, Wang et al., 2016, Bayraksan and Love, 2015, Ben-Tal et al., 2013, Yanıkoğlu and den Hertog, 2013], total variation distance [Rahimian et al., 2019a,b, Shapiro, 2017], Prohorov metric [Gibbs and Su, 2002, Erdoğan and Iyengar, 2006], and l_p-norm [Jiang and Guan, 2018, Huang et al., 2017].

In the reviewed articles, the authors consider a single measure for constructing the ambiguity set whereas in this work we propose to use multiple of them. We note that similar idea is used in Bertsimas et al. [2018] where the authors construct the ambiguity set using a moment-based measure and GoF tests in order to ensure a finite optimal solution to the DRO problem. However, their goal of using the intersection idea is different from our work, where we propose the construction of a joint ambiguity set to improve the quality of the solution by removing highly conservative distributions from the set. Moreover, the authors present the idea for a special problem setting, namely for problems with univariate and unbounded support, whereas we propose the idea for the general case and illustrate its use in detail for various problem settings. In the next

section, we provide a background and mathematical basis for the measures that we consider in this study.

3 Preliminaries

In this paper, we consider the stochastic programming problem

$$z^{\texttt{SP}} = \min_{x \in X} \mathbb{E}_{\mathbf{F}}[c(x, \xi)] \tag{1}$$

where the goal is to find $x \in X \subseteq \mathbb{R}^n$ that minimizes the expected value of a cost function, $c(x, \xi)$, which is a function of the decision variables x and the random vector ξ following the joint probability distribution $\mathbf{F}$ that is defined on a support Ξ.

In the cases where $\mathbf{F}$ is known and sampling from it is possible, SAA approximates problem (1) by

$$z^{\texttt{SAA}} = \min_{x \in X} \frac{1}{N} \sum_{j=1}^{N} c(x, \xi^j) \tag{2}$$

where $\xi^1, \ldots, \xi^N$ is an IID sample of random vectors drawn from $\mathbf{F}$. On the other hand, DRO is a modeling approach aiming at finding a solution that minimizes the worst-case expected cost

$$z^{\texttt{DRO}} = \min_{x \in X} \max_{F \in \mathcal{F}} \mathbb{E}_F[c(x, \xi)] \tag{3}$$

where $\mathcal{F}$ is an ambiguity set that contains probability distributions of interest. The set $\mathcal{F}$ plays a crucial role in this context as the DRO problem can possess desirable properties based on the measure used to build the set. Useful properties for the DRO problem, namely performance guarantee, convergence, and tractability are introduced in Bertsimas et al. [2018], which we overview in what follows for the sake of completeness.

Performance guarantee: Let $x^{\texttt{DRO}}$ denote an optimal solution of problem (3). The out-of-sample performance of $x^{\texttt{DRO}}$ is represented by $\mathbb{E}_{\mathbf{F}}[c(x^{\texttt{DRO}}, \xi)]$ which is the expected cost of $x^{\texttt{DRO}}$ under the true data-generating distribution $\mathbf{F}$. Since $x^{\texttt{DRO}}$ is a feasible solution to the SP problem (1), its out-of-sample performance is an upper bound on the optimal objective value of the SP problem, $z^{\texttt{SP}}$. In settings where $\mathbf{F}$ is not available, obtaining $z^{\texttt{SP}}$ and $\mathbb{E}_{\mathbf{F}}[c(x^{\texttt{DRO}}, \xi)]$ is impossible. However, $z^{\texttt{DRO}}$ can provide a probabilistic upper bound on them under certain conditions.

An ambiguity set is called at significance level α for a given $\alpha \in [0, 1]$, if it contains the true data-generating distribution $\mathbf{F}$ with probability $1 - \alpha$. Let $\mathcal{F}(\alpha)$ denote an ambiguity set at significance level α. The optimal objective value of the DRO problem (3) with ambiguity set $\mathcal{F}(\alpha)$ provides a probabilistic upper bound on the out-of-sample performance of its optimal solution. The performance guarantee of the DRO problem can be formalized as

$$\mathbb{P}(\mathbb{E}_{\mathbf{F}}[c(x^{\texttt{DRO}}, \xi)] \leq z^{\texttt{DRO}}) \geq 1 - \alpha, \tag{4}$$

meaning that the probability that $z^{\texttt{DRO}}$ is an upper bound on $\mathbb{E}_{\mathbf{F}}[c(x^{\texttt{DRO}}, \xi)]$ is at least $1 - \alpha$. This probability is called *reliability* in some references [Esfahani and Kuhn, 2018].

Asymptotic convergence: Let $\mathcal{F}^N(\alpha)$ denote an ambiguity set at significance level α which is constructed based on an available sample of observations $\xi^1, \ldots, \xi^N$. Let $x^{\texttt{DRO}}$ and $x^{\texttt{SP}}$ denote an optimal solution of the DRO problem and the SP problem, respectively. Under the following condition on $\mathcal{F}^N(\alpha)$ along with measure specific assumptions on the cost function $c(x, \xi)$, feasible region X, and support Ξ, the DRO problem achieves

the convergence of $z^{\texttt{DRO}} \to z^{\texttt{SP}}$ and $x^{\texttt{DRO}} \to x^{\texttt{SP}}$ as $N \to \infty$:

$$\mathbb{P}(F_N \nrightarrow \mathbf{F} \Longrightarrow F_N \notin \mathcal{F}^N) = 1. \tag{5}$$

This condition means that every sequence F_N that does not converge weakly to the true distribution should not be included in the ambiguity set $\mathcal{F}^N$ almost surely, as N tends to infinity (see Definition 3 of [Bertsimas et al., 2018]).

Tractability: Structure of the cost function and ambiguity set can provide theoretical and practical tractability to the DRO problem. By theoretical tractability, we mean that there exists a polynomial-time algorithm for solving problem (3), while by practical tractability we mean that the problem can be reformulated as a single-level convex optimization problem and in turn can be solved by state-of-the-art solvers. Consider the problem of finding a worst-case expected value over an ambiguity set $\mathcal{F}$ for a given solution x

$$\mathcal{C}(x, \mathcal{F}) = \max_{F \in \mathcal{F}} \mathbb{E}[c(x, \xi)]. \tag{6}$$

For many ambiguity sets, it has been shown in the literature that the problem can be reformulated as a single-level convex optimization problem for various cost functions of practical interest. By taking the dual of $\mathcal{C}(x, \mathcal{F})$ and transforming it into a minimization problem, it can be merged with $\min_{x \in X}(\cdot)$ so that the two-level DRO problem (3) can be written as a single-level problem. For instance, problem (3) with a Wasserstein-based ambiguity set reduces to a linear optimization problem if the 1-norm is used in the definition of the Wasserstein metric and the cost function is the maximum or minimum of affine functions. Additionally, the DRO problem with a certain GoF test based ambiguity set can be reformulated as a single-level convex optimization problem which can be solved by polynomial-time algorithms.

In this study, we mainly focus on building an ambiguity set using discrepancy-based measures, specifically, GoF tests and the Wasserstein metric. In what follows we briefly provide mathematical bases for creating individual sets using these measures. In all upcoming sections, we use the notation $[m]$ to denote the set $\{1, \ldots, m\}$.

3.1 GoF test-based ambiguity sets

The ambiguity set of a DRO problem can be the set of all probability distributions that pass a particular GoF test. A GoF test evaluates whether or not a given sample of observations follows a hypothetical distribution. The null hypothesis is that the provided sample is drawn from the hypothetical distribution. A test is said to be at *significance level* $\alpha \in [0, 1]$, if the probability of incorrectly rejecting the null hypothesis is at most α. In GoF tests, a test-specific statistic is calculated based on the given sample and a given hypothetical distribution. If the value of the test statistic is strictly larger than the threshold implied by the significance level α, then the null hypothesis is rejected. The set of all probability distributions that pass the test is called the *confidence region* of a test. Therefore, the ambiguity set of a DRO problem can be constructed as the confidence region of a GoF test.

Let $\xi^1, \ldots, \xi^N$ be an IID sample and F be a hypothetical distribution. Let $\texttt{T}$ specify a GoF test and $S_{\texttt{T}}^N(F, \xi^1, \ldots, \xi^N)$ denote a test-specific statistic that is calculated based on the given sample and the hypothetical distribution. An ambiguity set that is equivalent to the confidence region of a GoF test at

significance level α can be constructed as

$$\mathcal{F}_{\mathtt{T}}^{\mathtt{GoF}}(\alpha) := \{F \in \mathcal{P}(\Xi) : S_{\mathtt{T}}^{N}(F, \xi^1, \ldots, \xi^N) \leq \mathcal{Q}_{\mathtt{T}}(\alpha)\} \tag{7}$$

where $\Xi \subseteq \mathbb{R}^m$ denotes the support of the random vector, $\mathcal{P}(\Xi)$ is the set of probability distributions over Ξ, and $\mathcal{Q}_{\mathtt{T}}(\alpha)$ is a test-specific threshold at significance level of α. By construction, the ambiguity set $\mathcal{F}_{\mathtt{T}}^{\mathtt{GoF}}(\alpha)$ contains the true data-generating distribution with probability of at least $1-\alpha$ which provides a guarantee on the performance of the DRO problem with $\mathcal{F}_{\mathtt{T}}^{\mathtt{GoF}}(\alpha)$ as its ambiguity set. If $c(x,\xi)$ is continuous in x over all $\xi \in \Xi$, feasible set X is close and bounded, and support Ξ is bounded, then problem (3) with ambiguity set $\mathcal{F}_{\mathtt{T}}^{\mathtt{GoF}}(\alpha)$ enjoys asymptotic convergence and performance guarantee while it can be reformulated as a single-level convex optimization problem. For more detailed information regarding the assumptions, interested readers are referred to [Bertsimas et al., 2018].

3.2 Wasserstein metric-based ambiguity sets

One of the most widely used probability metrics on the space of probability distributions is the Wasserstein metric. It is utilized for measuring the distance between two probability distributions and interpreted as the optimal mass transportation plan from one distribution to another one. Let $\mathcal{P}'(\Xi)$ denote the set of all probability distributions F on the support Ξ where $\mathbb{E}_F\big[||\xi||\big] = \int_{\Xi} ||\xi||\ F(\xi) \leq \infty$. Wasserstein distance $d^{\mathtt{Wass}} : \mathcal{P}'(\Xi) \times \mathcal{P}'(\Xi) \to \mathbb{R}_+$ between two probability distributions is defined as

$$d^{\mathtt{Wass}}(F_1, F_2) := \inf_{\Pi} \int_{\Xi \times \Xi} ||\xi_1 - \xi_2||_p \ \Pi(\xi_1, \xi_2) \tag{8}$$

where the decision variable Π is a joint probability distribution of ξ_1 and ξ_2 with marginal distributions F_1 and F_2, respectively. Additionally, $||\cdot||_p$ for $p \geq 1$ represents the p-norm on $\mathbb{R}^m$ resulting in a generalized p-Wasserstein metric. For simplicity, in this paper, we consider the 1-Wasserstein metric ($p = 1$) which is a commonly use metric in the literature.

Esfahani and Kuhn [2018] propose to construct the ambiguity set of a DRO problem using the Wasserstein metric. Given an IID sample $\xi^1, \ldots, \xi^N$ and the corresponding empirical distribution $\hat{F}_N$, the idea is to construct a Wasserstein ball of radius ϵ_N^α centered at the empirical distribution such that

$$\mathcal{F}^{\mathtt{Wass}}(\epsilon_N^\alpha) := \{F \in \mathcal{P}'(\Xi) : d^{\mathtt{Wass}}(F, \hat{F}_N) \leq \epsilon_N^\alpha\} \tag{9}$$

This ambiguity set $\mathcal{F}^{\mathtt{Wass}}(\epsilon_N^\alpha)$ contains all probability distributions that are within a certain distance from the empirical distribution. Under a light tail assumption on the true data-generating distribution, the ambiguity set provides a performance guarantee for the DRO problem with $\mathcal{F}^{\mathtt{Wass}}(\epsilon_N^\alpha)$ ambiguity set. This assumption means that the tail of the true distribution $\mathbf{F}$ should decay at an exponential rate, and the modern measure concentration theorem [Fournier and Guillin, 2015] suggests that $\mathcal{F}^{\mathtt{Wass}}(\epsilon_N^\alpha)$ contains $\mathbf{F}$ with probability of at least $1-\alpha$ if ϵ_N^α is a sublinearly growing function of $\log(1/\alpha)/N$. Based on the measure concentration theorem, the performance guarantee of DRO problem with $\mathcal{F}^{\mathtt{Wass}}(\epsilon_N^\alpha)$ ambiguity set follows. If $c(x,\xi)$ is upper semicontinuous in ξ and there exists a constant $L \geq 0$ where $|c(x,\xi)| \leq L(1 + ||\xi||)$ for all $x \in X$, the optimal objective value of the DRO problem converges to that of the SP problem almost surely as the sample size tends to infinity. Moreover, if the assumptions hold and $c(x,\xi)$ is lower semicontinuous in x for all $\xi \in \Xi$, the optimal solution of the DRO problem converges to that of SP problem almost surely as the sample size

tends to infinity. Furthermore, assuming $c(x,\xi) := \max_{k\in[K]} c_k(x,\xi)$ where $-c_k(x,\xi)$ are proper, convex, and lower semicontinuous in ξ for all $x \in X$ and $k \in [K]$, also Ξ is closed and convex, the worst-case expected value problem $\max_{F\in\mathcal{F}^{\texttt{Wass}}(\epsilon_N^\alpha)} \mathbb{E}_F[c(x,\xi)]$ can be reduced to a finite convex problem and, consequently, the DRO problem can be reformulated as a single-level convex problem and can be solved by off-the-shelf solvers. For more detailed information regarding the assumptions, an interested reading is referred to [Esfahani and Kuhn, 2018].

4 Joint ambiguity set

While a common approach suggested in the literature for DRO problems is to build an ambiguity set using a single measure, in this work we propose to construct the ambiguity set based on a combination of various measures. In this case, the ambiguity set can be regarded as a joint region of multiple ambiguity sets and can potentially inherit the desirable properties of the individual sets while cut off overly conservative distributions of each set and provide better results for the DRO problem in terms of optimal solution and objective value. Specifically, we use intersection as an operator to construct a joint ambiguity set which combines multiple sets, of which is each created by a single measure that guarantees the existence of the true distribution in the set at a certain confidence level. Let $\mathcal{F}_i(\alpha_i)$ denote an ambiguity set at significance level α_i for $i \in [m]$ where $\sum_{i=1}^m \alpha_i < 1$. We define

$$\mathcal{F}(\alpha) = \mathcal{F}_1(\alpha_1) \cap \mathcal{F}_2(\alpha_2) \cap \ldots \cap \mathcal{F}_{m-1}(\alpha_{m-1}) \cap \mathcal{F}_m(\alpha_m) \tag{10}$$

as the joint ambiguity set at significance level $\alpha \in [0,1]$ where $\alpha = \alpha_1 + \ldots + \alpha_m$ and by construction the joint ambiguity set $\mathcal{F}(\alpha)$ contains the true distribution $\mathbf{F}$ with a probability of at least $1-\alpha$. A DRO problem with each of the ambiguity sets $\mathcal{F}_i(\alpha_i)$ should satisfy measure specific assumptions in order to possess desirable properties that are discussed in Section 3. Based on the measure, each property requires a particular set of assumptions to hold. In what follows, we discuss assumptions required for a DRO problem with a joint ambiguity set to enjoy each of the properties.

In terms of performance guarantee, a DRO problem with ambiguity set $\mathcal{F}_i(\alpha_i)$ can provide a performance guarantee of type (4) under a set of assumptions which we denote as $\mathcal{A}_i^{\texttt{Valid}}$. The assumptions are on the cost function $c(x,\xi)$, feasible region X, structure of the ambiguity set $\mathcal{F}_i(\alpha_i)$, and support Ξ, that ensure the existence of the finite optimal objective value and optimal solution for the DRO problem. The list of assumptions required for the DRO problem with ambiguity set $\mathcal{F}_i(\alpha_i)$ can be found in related references. If the joint ambiguity set is constructed as in (10), set of assumptions $\mathcal{A}_{\texttt{J}}^{\texttt{Valid}}$, under which the DRO problem (3) with $\mathcal{F}(\alpha)$ provides a performance guarantee can be constructed as the union of $\mathcal{A}_i^{\texttt{Valid}}$ sets for $i \in [m]$.

Under assumptions $\mathcal{A}_{\texttt{J}}^{\texttt{Valid}}$, the optimal objective value of the DRO problem with the joint ambiguity set is a lower bound on the optimal value of the DRO problems with individual sets $\mathcal{F}_i(\alpha_i)$, under the assumptions $\mathcal{A}_i^{\texttt{Valid}}$ for $i \in [m]$. The reason is that the joint ambiguity set is a subset of individual sets which creates a smaller feasible region for problem (6). In this case, problem (3) with $\mathcal{F}(\alpha)$ ambiguity set can provide a tighter bound on the out-of-sample performance of its optimal solution compared to the bounds provided by DRO problems with individual ambiguity sets on the out-of-sample performance of their optimal solutions.

In terms of asymptotic convergence, a DRO problem with $\mathcal{F}_i(\alpha_i)$ ambiguity set which satisfies condition (5) possesses convergence under a set of measure-specific assumptions that is denoted by $\mathcal{A}_i^{\texttt{Conv}}$. The set may contain assumptions on the cost function $c(x,\xi)$, feasible set X, and support Ξ. Let $\mathcal{A}_{\texttt{J}}^{\texttt{Conv}}$ denote the set

of assumptions that should hold so that the DRO problem with the joint ambiguity set has asymptotic convergence. The joint ambiguity set (10), by construction, satisfies the condition (5) and in order to guarantee the convergence for the DRO problem with the joint ambiguity set, it is sufficient for $\mathcal{A}_{\mathtt{J}}^{\mathtt{Conv}}$ to contain all assumptions included in a single set $\mathcal{A}_i^{\mathtt{Conv}}$ as long as all $\mathcal{F}_j(\alpha_j)$ for $j \in [m]$ and $j \neq i$ contain the empirical distribution. However, in order to get the most benefit out of the joint ambiguity set in terms of convergence, the set $\mathcal{A}_{\mathtt{J}}^{\mathtt{Conv}}$ can be defined as the union of $\mathcal{A}_i^{\mathtt{Conv}}$ sets for $i \in [m]$. Therefore, the convergence property of the DRO problem with joint ambiguity set can be interpreted based on the convergence of the DRO problem with individual ambiguity sets.

In terms of tractability, let $\mathcal{A}_i^{\mathtt{Trac}}$ denote the set of all assumptions under which a DRO problem with an ambiguity set $\mathcal{F}_i(\alpha_i)$ can be reformulated as a single-level convex optimization problem. A DRO problem with joint ambiguity set (10) possesses tractable reformulation under assumptions $\mathcal{A}_{\mathtt{J}}^{\mathtt{Trac}}$ which is the union of $\mathcal{A}_i^{\mathtt{Trac}}$ for $i \in [m]$. It is important to note that although the set of assumptions for the DRO problem with joint ambiguity set can be created based on the assumptions of the individual problems, the derivation of a single-level formulation for the DRO problem with joint ambiguity set is not straightforward. In fact, a single-level reformulation of problem (3) with joint ambiguity set is not achievable by directly combining the reformulations of the DRO problems with individual ambiguity sets. In order to obtain the single-level reformulation, the feasible set of problem (6) should be replaced by the joint ambiguity set and the dual of the problem should be taken so that the resulting minimization problem can be merged with $\min_{x \in X}(\cdot)$ and yield a single-level optimization problem.

Various measures have been proposed in the literature for constructing an ambiguity set and all of them can be utilized to construct the joint ambiguity set as long as their underlying assumptions to obtain the desirable properties do not contradict with each other. In this paper, we study the intersection of multiple ambiguity sets in problems with various support structures, namely, problems with discrete known support, univariate support, and multivariate support. Choosing appropriate measures is crucial as the measures should be compatible with the problem setting. Therefore, we consider the intersection of multiple sets each created by an individual GoF test for problems with discrete and univariate supports, and the intersection of Wasserstein ball with a set created by a particular GoF test for the problems with multivariate support.

5 Problems with discrete known support

Consider ξ as a random variable with known discrete support $\Xi = \{\hat{\xi}^1, \ldots, \hat{\xi}^n\}$. In this case, two well-known statistical tests, namely Pearson's χ^2 test and G-test can be utilized for constructing the ambiguity set of a DRO problem. The test statistics for the tests are

$$X_N = \left(\sum_{j=1}^{n} \frac{(F(\hat{\xi}^j) - \hat{F}_N(\hat{\xi}^j))^2}{F(\hat{\xi}^j)} \right)^{1/2}$$

$$G_N = \left(2 \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j) \log \left(\frac{\hat{F}_N(\hat{\xi}^j)}{F(\hat{\xi}^j)} \right) \right)^{1/2}$$

where $F(\hat{\xi}^j)$ and $\hat{F}_N(\hat{\xi}^j)$ denote the hypothetical and empirical probabilities of observing $\hat{\xi}^j$ for $j \in [n]$, respectively. Let $\mathfrak{Q}_{\chi^2}(\alpha_1)$ and $\mathfrak{Q}_{\mathtt{G}}(\alpha_2)$ denote the thresholds of the χ^2-test and the G-test at significance

levels of α_1 and α_2, respectively. Also, let $\mathcal{F}_{\chi^2}(\alpha_1)$ and $\mathcal{F}_{\mathtt{G}}(\alpha_2)$ represent the individual confidence regions created by χ^2-test and G-test, respectively. A joint ambiguity set for a DRO problem using the tests can be constructed as

$$\mathcal{F}(\alpha) = \mathcal{F}_{\chi^2}(\alpha_1) \cap \mathcal{F}_{\mathtt{G}}(\alpha_2) = \{F \in \mathcal{P}(\Xi) : \ X_N \leq \mathcal{Q}_{\chi^2}(\alpha_1), \ G_N \leq \mathcal{Q}_{\mathtt{G}}(\alpha_2)\} \tag{11}$$

where $\alpha = \alpha_1 + \alpha_2$ is the significance level of the joint ambiguity set and $\alpha, \alpha_1, \alpha_2 \in [0, 1]$. Using the properties of two sets, the joint ambiguity set provides a performance guarantee with a probability of at least $1 - \alpha$. The DRO problem with joint ambiguity set can possess asymptotic convergence, which can be built upon the convergence property of the individual DRO problems with $\mathcal{F}_{\chi^2}(\alpha_1)$ and $\mathcal{F}_{\mathtt{G}}(\alpha_2)$ ambiguity sets.

For the χ^2-test and the G-test, there exist sets of assumptions, denoted by $\mathcal{A}_{\chi^2}^{\mathtt{Trac}}$ and $\mathcal{A}_{\mathtt{G}}^{\mathtt{Trac}}$, respectively, under which the DRO problems with ambiguity sets as a confidence region of the tests have tractability. Under these assumptions, problem (3) with $\mathcal{F}_{\chi^2}(\alpha_1)$ and $\mathcal{F}_G(\alpha_2)$ ambiguity sets can be reformulated as a single-level convex optimization problem that can be solved by off-the-shelf solvers. In this setting, the assumptions contained in both sets are the same [Bertsimas et al., 2018]. Under the same set of assumptions, the DRO problem with the joint ambiguity set have tractability meaning that we have $\mathcal{A}_{\mathtt{J}}^{\mathtt{Trac}} = \mathcal{A}_{\chi^2}^{\mathtt{Trac}} = \mathcal{A}_{\mathtt{G}}^{\mathtt{Trac}}$. In the following theorem, we provide an equivalent reformulation for the worst-case expected value problem.

Theorem 1. *Let $\mathcal{F}(\alpha)$ represent a joint ambiguity set defined as in (11). Under assumptions of $\mathcal{A}_{\mathtt{J}}^{\mathtt{Trac}}$, problem (6) can be reformulated as*

$$\mathcal{C}(x, \mathcal{F}(\alpha)) = \min_{r,s,s',t,t',\ell,y,\gamma} \ r + \mathcal{Q}^2_{\chi^2}(\alpha_1)\ s + \ \frac{1}{2}\mathcal{Q}^2_G(\alpha_2)\ s' - \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)(t_j + t'_j) \tag{12a}$$

$$\begin{aligned}
s.t.\ & \ell_j - \gamma_j \leq s && j \in [n] && \text{(12b)}\\
& 2s + t_j \leq y_j && j \in [n] && \text{(12c)}\\
& \gamma_j - r \leq s' && j \in [n] && \text{(12d)}\\
& \ell_j \geq c(x, \hat{\xi}^j) && j \in [n] && \text{(12e)}\\
& (y_j, \ell_j - \gamma_j, 2s - \ell_j + \gamma_j) \in C_{\mathtt{SOC}} && j \in [n] && \text{(12f)}\\
& (t'_j, s', s' - \gamma_j + r) \in C_{\mathtt{XC}} && j \in [n] && \text{(12g)}\\
& r \in \mathbb{R}, s, s' \in \mathbb{R}_+, \gamma, t, t', \ell, y \in \mathbb{R}^n && && \text{(12h)}
\end{aligned}$$

where

$$C_{\mathtt{SOC}} = \{(a, b, c) \in \mathbb{R}^3 : \sqrt{a^2 + b^2} \leq c\}$$

is a second-order cone, and

$$C_{\mathtt{XC}} = \mathrm{cl}\left(\{(a, b, c) : be^{a/b} \leq c, b > 0\}\right)$$

is an exponential cone, and $\mathrm{cl}(\cdot)$ *denotes the closure operator.*

Proof. To transform the two-level min-max DRO problem to a single-level convex optimization problem, we need to take the dual of the inner maximization problem and obtain a minimization problem. To begin with, we add all of the constraints contained in the definition of the joint ambiguity set (11) to the maximization problem (6). Specifically, we add the constraints which make the decision variable F a probability distribution

and to reside in the sets created by χ^2-test and G-test. We use definition of ϕ-divergence of the tests in order to formulate the related constraints.

In order to obtain a minimization problem we take the following steps. First, we use Fenchel duality in order to get the dual of the maximization problem. Next, we omit the maximization operator using the definitions of convex conjugates of the ϕ-divergences. Finally, we further simplify the resulted formulation with replacing the convex conjugates with their alternative definitions and add the required constraints to the model. For mathematical formulations and more details on the steps of the proof please see Appendix A.1. □

6 Problems with univariate distributions

Let ξ be a continuous univariate random variable that belongs to a bounded support Ξ. There are well-known GoF tests that can be used for constructing ambiguity sets, namely Kolmogorov–Smirnov test, Kuiper test, Cramér-von Mises test, Watson test, and Anderson-Darling test [D'Agostino, 2017, Wang et al., 2016, Nwaigwe et al., 2022]. Considering test $\mathtt{T}$ be one of the mentioned GoF tests with a test statistic $S_{\mathtt{T}}^N$ and a threshold $\mathcal{Q}_{\mathtt{T}}(\alpha)$ at significance level α, the individual ambiguity set of test $\mathtt{T}$, $\mathcal{F}_{\mathtt{T}}(\alpha)$ can be constructed as in (7) [Bertsimas et al., 2018].

Given a *sorted* (ascending according to their values) sample of observations $\{\xi^1, \ldots, \xi^N\}$, let ζ denote the hypothetical cumulative distribution that the sample can be drawn from. It has been shown in the literature that the inequality $S_{\mathtt{T}}^N(\zeta, \xi^1, \ldots, \xi^N) \leq \mathcal{Q}_{\mathtt{T}}(\alpha)$ has a conic representation in the form of $A_{S_{\mathtt{T}}^N}\zeta - b_{S_{\mathtt{T}}^N,\alpha} \in K_{S_{\mathtt{T}}^N}$ where cone $K_{S_{\mathtt{T}}^N}$, matrix $A_{S_{\mathtt{T}}^N}$, and the vector $b_{S_{\mathtt{T}}^N,\alpha}$ are defined based on the test statistic of the test $\mathtt{T}$. The cones associated with the mentioned tests are canonical cones meaning that they can be written as a Cartesian product of $\mathbb{R}^k$, $\{0\}$, $\mathbb{R}^k_+$, C_{SOC}, and semidefinite cone where k denotes an appropriate dimension. Therefore, a DRO problem with the ambiguity set as a confidence region of a single GoF test is an optimization problem over a cone which is tractable and can be solved using state-of-the-art solvers (see Theorem 10 of [Bertsimas et al., 2018]).

In this setting, we construct a joint ambiguity set as an intersection of multiple GoF tests. Let $\mathtt{T}_1, \ldots, \mathtt{T}_m$ be m tests at significance levels of $\alpha_1, \ldots, \alpha_m$, respectively, also, let $\mathcal{F}_{\mathtt{T}_1}(\alpha_1), \ldots, \mathcal{F}_{\mathtt{T}_m}(\alpha_m)$ be their individual confidence regions. The joint ambiguity set is

$$\mathcal{F}(\alpha) = \mathcal{F}_{\mathtt{T}_1}(\alpha_1) \cap \ldots \cap \mathcal{F}_{\mathtt{T}_m}(\alpha_m) = \{\zeta \in \mathcal{D}(\Xi) : A_{S_{\mathtt{T}_j}^N}\zeta - b_{S_{\mathtt{T}_j}^N,\alpha_j} \in K_{S_{\mathtt{T}_j}^N},\ j \in [m]\} \tag{13}$$

where $\mathcal{D}(\Xi)$ denotes the set of cumulative distributions over the support Ξ and $\alpha = \sum_{j=1}^m \alpha_j$ is the significance level of the joint ambiguity set. Using the properties of multiple sets, the joint ambiguity set provides a performance guarantee with a probability of at least $1 - \alpha$. The DRO problem with the joint ambiguity set can possess asymptotic convergence, which can be built upon the convergence property of the DRO problems with the $\mathcal{F}_{\mathtt{T}_j}(\alpha_j)$ ambiguity sets for $j \in [m]$.

For the tests $\mathtt{T}_j$, there exist sets of assumptions denoted by $\mathcal{A}_{\mathtt{T}_j}^{\mathtt{Trac}}$ for $j \in [m]$ under which the DRO problem with ambiguity sets as a confidence region of the tests have tractability. In this problem setting, considering the mentioned tests, the assumptions contained in all sets are the same [Bertsimas et al., 2018]. Under the assumptions, problem (3) with $\mathcal{F}_{\mathtt{T}_j}(\alpha_j)$ ambiguity sets for $j \in [m]$ can be reformulated as a single-level convex cone programming problem. The DRO problem with the joint ambiguity set (13) has tractability under the same set of assumptions, meaning that $\mathcal{A}_{\mathtt{J}}^{\mathtt{Trac}} = \mathcal{A}_{\mathtt{T}_1}^{\mathtt{Trac}} = \ldots = \mathcal{A}_{\mathtt{T}_m}^{\mathtt{Trac}}$. In the following theorem, we provide an equivalent reformulation for worst-case expected value problem.

Let $b_{\mathtt{T}_j} = b_{S^N_{\mathtt{T}_j},\alpha_j}$, $A_{\mathtt{T}_j} = A_{S^N_{\mathtt{T}_j}}$, and $K_{\mathtt{T}_j} = K_{S^N_{\mathtt{T}_j}}$ be the information associated with the test $\mathtt{T}_j$ for $j \in [m]$, and let $(\cdot)_i$ represent the i-th row of a given matrix. For cone $K \subseteq \mathbb{R}^k$ of suitable dimension k, let K^* denote its dual cone, i.e., $K^* = \{y \in \mathbb{R}^k : y^\top z \geq 0,\ \forall z \in K\}$.

Theorem 2. *Given a sorted sample of $\{\xi^1, \ldots, \xi^N\}$ and an ambiguity set defined as in (13), under assumptions of $\mathcal{A}_J^{\mathcal{Trac}}$, problem (6) can be reformulated as*

$$\mathcal{C}(x, \mathcal{F}(\alpha)) = \min_{r,\ell} \sum_{j=1}^{m} b_{T_j}^\top r_j + \ell_{N+1} \tag{14a}$$

$$s.t.\ \ell \in \mathbb{R}^{N+1},\ -r_j \in K^*_{T_j},\ j \in [m] \tag{14b}$$

$$\left(\sum_{j=1}^{m} A_{T_j}^\top r_j\right)_i = \ell_i - \ell_{i+1} \qquad i \in [N] \tag{14c}$$

$$\ell_i \geq \sup_{\xi \in (\xi^{i-1}, \xi^i]} c(x, \xi) \qquad i \in [N+1] \tag{14d}$$

where ξ^0 and ξ^{N+1} are lower bound and upper bound on the support, respectively.

Proof. Similar to the problems with discrete known support, in this setting we need to take the dual of the inner maximization problem in the two-level min-max DRO problem to convert it to a minimization problem and merge the resulting problem with the outer minimization problem. As opposed to the discrete known support problems where we build the ambiguity set for the probability distributions, in univariate setting, we construct the ambiguity set for cumulative probability distributions. Therefore, the maximization problem has constraints which assure that the decision variable F is a cumulative probability distribution and reside in the regions defined by the considered GoF tests. As we discussed earlier, the regions are canonical cones, so, the resulted maximization problem is a conic programming problem and conic duality results can be used in the reformulations.

In order to obtain a single-level convex programming problem, we use the final model proposed in Theorem 11 of [Bertsimas et al., 2018] and the conic duality theorem proposed in Section 1.4.5.1 of [Ben-Tal and Nemirovski, 2019] (see Appendix A.2 for more details). In this setting, we use multiple tests to build the ambiguity set; therefore, we have multiple conic constraints as opposed to a single constraint which is considered in [Bertsimas et al., 2018]. Our final reformulation (14) follows from the duality theorem mentioned in [Ben-Tal and Nemirovski, 2019] for $m > 2$ which allows us to incorporate the information of the multiple conic constraint in the forms of summations in the objective function (14a) and constraints (14c).

□

For the case of univariate distributions with unbounded support, a novel test has been proposed in [Bertsimas et al., 2018] where the ambiguity set is constructed by restricting the confidence region of an individual GoF test to contain only the distributions that have particular moment-based conditions. The same idea can be applied for constructing joint ambiguity set for the problems with unbounded support. In this case, multiple GoF tests can be intersected similar to the bounded case and restrict the joint region by the same moment based conditions that is proposed by the authors.

7 Problems with multivariate distributions

In this section, we discuss the performance guarantee and tractability of the DRO problem with joint ambiguity set for problems with multivariate distributions. We derive a tractable reformulation for the problem and propose bounding techniques for these computationally challenging problems.

7.1 Performance Guarantee and Tractability

Let ξ denote a random vector of dimension d that follows a multivariate distribution with support Ξ. In this setting, we consider an intersection of the ambiguity sets constructed by Wasserstein distance measure and a GoF test based on linear-convex ordering (LCX) of random vectors.

LCX-based GoF test constitutes of two test statistics namely R_N and C_N. Given a sample of random vector observations $\{\xi^1, \ldots, \xi^N\}$ and their corresponding empirical distribution $\hat{F}_N$, the test statistics are as follows:

$$R_N = \mathbb{E}_{\hat{F}_N}[||\xi||_2^2] - \mathbb{E}_F[||\xi||_2^2] \tag{15}$$

$$C_N = \sup_{|a_1|+\ldots,+|a_d|+|b|\leq 1} \Big(\mathbb{E}_F[\max\{a^\top \xi - b, 0\}] - \mathbb{E}_{\hat{F}_N}[\max\{a^\top \xi - b, 0\}]\Big) \tag{16}$$

where F is a hypothetical distribution and $||\xi||_2^2 = \sum_{j=1}^d \xi_j^2$. Let $\mathfrak{Q}_{\mathtt{R}_N}(\alpha_1)$ and $\mathfrak{Q}_{\mathtt{C}_N}(\alpha_2)$ denote two thresholds at significance levels of α_1 and α_2, respectively. An ambiguity set can be constructed based on the LCX ordering as follows:

$$\mathcal{F}_{\mathtt{LCX}}(\alpha) := \{F \in \mathcal{P}(\Xi) : \ R_N \leq \mathfrak{Q}_{\mathtt{R}_N}(\alpha_1), \ C_N \leq \mathfrak{Q}_{\mathtt{C}_N}(\alpha_2)\} \tag{17}$$

where $\alpha = \alpha_1 + \alpha_2$. The joint ambiguity set in the multivariate case can be considered as an intersection of the confidence region of LCX-based GoF test and Wasserstein ball centered at empirical distribution. Intersecting the ambiguity set (17) with Wasserstein ball (9) yields the following joint ambiguity set:

$$\mathcal{F}(\alpha, \beta) = \mathcal{F}_{\mathtt{LCX}}(\alpha) \cap \mathcal{F}^{\mathtt{Wass}}(\epsilon_N^\beta) = \{F \in \mathcal{P}(\Xi) : \ R_N \leq \mathfrak{Q}_{\mathtt{R}_N}(\alpha_1), \ C_N \leq \mathfrak{Q}_{\mathtt{C}_N}(\alpha_2), \ d^{\mathtt{Wass}}(F, \hat{F}_N) \leq \epsilon_N^\beta\} \tag{18}$$

where $\alpha = \alpha_1 + \alpha_2$. The joint ambiguity set is at significance level of $\alpha + \beta$ and provides a performance guarantee on the out-of-sample performance of the optimal solution of the corresponding DRO problem (18). The DRO problem can possess asymptotic convergence which can be built upon the convergence property of the DRO problems with ambiguity sets of $\mathcal{F}_{\mathtt{LCX}}(\alpha)$ and Wasserstein ball. Under the assumptions of at least one of the measures, the DRO problem with the joint ambiguity set enjoys asymptotic convergence. However, satisfying all the assumptions of both measures might improve the convergence rate.

Under the sets of assumptions $\mathcal{A}_{\mathtt{L}}^{\mathtt{Trac}}$ and $\mathcal{A}_{\mathtt{w}}^{\mathtt{Trac}}$, the DRO problems with ambiguity sets as a confidence region of the LCX-based test and the Wasserstein ball have tractability [Bertsimas et al., 2018, Esfahani and Kuhn, 2018]. Under the union of the set of assumptions, $\mathcal{A}_{\mathtt{J}}^{\mathtt{Trac}} = \mathcal{A}_{\mathtt{L}}^{\mathtt{Trac}} \cup \mathcal{A}_{\mathtt{w}}^{\mathtt{Trac}}$, the DRO problem with the joint ambiguity set have tractability. In the following theorem, we propose a tractable reformulation for problem (6) with ambiguity set (18) for the problems with cost functions which can be written as the maximum of concave functions in ξ.

Theorem 3. *Given a sample of random vector observations $\{\xi^1, \ldots, \xi^N\}$, let $c(x, \xi) = \max_{k\in[K]} c_k(x, \xi)$ denote the cost function and $\mathcal{F}(\alpha, \beta)$ represent a joint ambiguity set defined as in (18). Under the assumptions*

of $\mathcal{A}_J^{Trac}$, problem (6) can be reformulated as follows:

$$\mathcal{C}(x,\mathcal{F}) = \inf_{\substack{z,r,f,w,w',\\ y,y',e,\eta,g}} \quad \eta\epsilon_N^\beta + \frac{1}{N}\sum_{i=1}^N r_i + \sum_{\gamma\in\mathcal{G}} f_\gamma\ Q_{\mathcal{C}_N}(\alpha_1) + \sum_{\gamma\in\mathcal{G}}\sum_{i=1}^N e_{\gamma i} \tag{19a}$$

$$s.t. \quad \sum_{j=1}^d y_{\gamma j} + y'_\gamma \le f_\gamma \qquad \gamma\in\mathcal{G} \tag{19b}$$

$$\frac{1}{N}(\sum_{j=1}^d w_{\gamma j}\ \xi_j^i - w'_\gamma) \le e_{\gamma i} \qquad \gamma\in\mathcal{G},\ i\in[N] \tag{19c}$$

$$w_{\gamma j} - y_{\gamma j} \le 0 \qquad \gamma\in\mathcal{G},\ j\in[d] \tag{19d}$$

$$-w_{\gamma j} - y_{\gamma j} \le 0 \qquad \gamma\in\mathcal{G},\ j\in[d] \tag{19e}$$

$$w'_\gamma - y'_\gamma \le 0 \qquad \gamma\in\mathcal{G} \tag{19f}$$

$$-w'_\gamma - y'_\gamma \le 0 \qquad \gamma\in\mathcal{G} \tag{19g}$$

$$\langle z_{ik}, \xi^i\rangle - r_i + \sum_{\gamma\in\mathcal{G}} \gamma_{ik}\ w'_\gamma - g_k \le 0 \qquad k\in[K],\ i\in[N] \tag{19h}$$

$$g_k \le c_{k^*}(x,\ \sum_{\gamma\in\mathcal{G}} \gamma_{ik}\ w_\gamma + z_{ik}) \qquad k\in[K],\ i\in[N] \tag{19i}$$

$$\|z_{ik}\|_* \le \eta \qquad k\in[K],\ i\in[N] \tag{19j}$$

$$\eta\in\mathbb{R}_+,\ f\in\mathbb{R}_+^{|\mathcal{G}|},\ e\in\mathbb{R}_+^{|\mathcal{G}|\times N}, \tag{19k}$$

$$y\in\mathbb{R}_+^{|\mathcal{G}|\times d},\ y'\in\mathbb{R}_+^{|\mathcal{G}|},\ z\in\mathbb{R}^{N\times K\times d}, \tag{19l}$$

$$r\in\mathbb{R}^N, w\in\mathbb{R}^{|\mathcal{G}|\times d},\ w'\in\mathbb{R}^{|\mathcal{G}|}, g\in\mathbb{R}^K. \tag{19m}$$

where $c_{k^*}(x,\Gamma) := \inf_{\xi\in\Xi}\Gamma^\top\xi - c_k(x,\xi)$ is the concave conjugate of c_k, $\langle\cdot,\cdot\rangle$ is the inner product operator, $\|z_{ik}\|_* := \sum_{\|\xi\|\le 1}\langle z,\xi\rangle$ denotes the dual norm of $\|\cdot\|$, and $\mathcal{G} := \{0,1\}^{N\times K}\setminus\{(0,\ldots,0)\}$.

Proof. Here, we provide the main steps which are required in order to obtain a tractable reformulation. For a more detailed proof, see Appendix A.3. For the problems with multivariate distributions, problem (6) given the joint ambiguity set (18) can be written as

$$\mathcal{C}(x,\mathcal{F}) = \sup_{F,\Pi} \int_\Xi c(x,\xi)F(d\xi) \tag{20a}$$

$$\text{s.t.} \int_\Xi \max\{a^\top\xi - b, 0\}F(d\xi) \le \int_\Xi \max\{a^\top\xi' - b, 0\}\hat{F}_N(d\xi') + \mathcal{Q}_{\mathcal{C}_N}(\alpha_2) \quad \forall (a,b)\in S \tag{20b}$$

$$\int_\Xi ||\xi||_2^2 F(d\xi) \ge \mathcal{Q}_{\mathcal{R}_N}(\alpha_1) \tag{20c}$$

$$\int_{\Xi^2} ||\xi - \xi'||\Pi(d\xi, d\xi') \le \epsilon_N^\beta \tag{20d}$$

where Π is a joint distribution of ξ and ξ' with marginal distributions F and $\hat{F}_N$, respectively, and $S = \{(a,b)\in\mathbb{R}^d\times\mathbb{R}\mid \|a\|_1 + |b|\le 1\}$. Based on the law of total probability, F and Π can be replaced by their equivalent formulations in terms of marginal and conditional distributions. More specifically, $\Pi = \frac{1}{N}\sum_{i=1}^N \delta_i \times F_i$ where δ_i represents the Dirac distribution that dedicates the unit mass to ξ^i. Let S' denote the set of non-negative

measures over S, and let λ, θ, and η represent the dual variables associated with constraints (20b), (20c), and (20d), respectively. Using notions of Fenchel duality, the dual of problem (20) is

$$\inf_{\lambda\in S',\theta,\eta\in\mathbb{R}_+} \Big\langle \lambda, \frac{1}{N}\sum_{i=1}^{N} \max\{a^\top \xi^i - b, 0\} + \mathcal{Q}_{\mathtt{C}_N}(\alpha_2)\Big\rangle_S - \theta \mathcal{Q}_{\mathtt{R}_N}(\alpha_1) + \eta\epsilon_N^\beta + \\ \frac{1}{N}\sum_{i=1}^{N} \sup_{\xi\in\Xi}\Big(c(x,\xi) - \Big\langle \lambda, \max\{a^\top\xi - b, 0\}\Big\rangle_S + \theta||\xi||_2^2 - \eta||\xi - \xi^i||\Big). \tag{21}$$

where $\langle\cdot\,,\cdot\rangle_S$ is the inner product operator defined over space S. By defining a new decision variable s, problem (21) is equivalent to

$$\inf_{\lambda\in S',\theta,\eta\in\mathbb{R}_+,s\in\mathbb{R}^N} \Big\langle \lambda, \frac{1}{N}\sum_{i=1}^{N} \max\{a^\top \xi^i - b, 0\} + \mathcal{Q}_{\mathtt{C}_N}(\alpha_2)\Big\rangle_S - \theta \mathcal{Q}_{\mathtt{R}_N}(\alpha_1) + \eta\epsilon_N^\beta + \frac{1}{N}\sum_{i=1}^{N} s_i \tag{22a}$$

$$\text{s.t. } s_i \geq \sup_{\xi\in\Xi}\Big(c(x,\xi) - \Big\langle \lambda, \max\{a^\top\xi - b, 0\}\Big\rangle_S + \theta||\xi||_2^2 - \eta||\xi - \xi^i||\Big) \quad i\in[N]. \tag{22b}$$

Looking closely at constraints (22b), based on the assumptions in $\mathcal{A}_{\mathtt{L}}^{\mathtt{Trac}}$ and by the same reasoning mentioned in [Bertsimas et al., 2018], the only feasible value for θ is zero. Therefore, the above formulation can be simplified by replacing θ by zero. Next, using the definition of the cost function and the dual norm, the constraints (22b) can be reformulated as follows.

$$s_i \geq \sup_{\xi\in\Xi}\Big(c_k(x,\xi) - \Big\langle \lambda, \max\{a^\top\xi - b, 0\}\Big\rangle_S - \max_{||z_{ik}||_*\leq\eta}\langle z_{ik}, \xi - \xi^i\rangle\Big) \quad k\in[K], i\in[N]. \tag{23a}$$

Therefore, problem (22) is equivalent to

$$\inf_{\lambda\in S',\eta\in\mathbb{R}_+,s\in\mathbb{R}^N} \Big\langle \lambda, \frac{1}{N}\sum_{i=1}^{N} \max\{a^\top \xi^i - b, 0\} + \mathcal{Q}_{\mathtt{C}_N}(\alpha_2)\Big\rangle_S + \eta\epsilon_N^\beta + \frac{1}{N}\sum_{i=1}^{N} s_i \tag{24a}$$

$$\text{s.t. } s_i \geq \min_{||z_{ik}||_*\leq\eta}\sup_{\xi\in\Xi}\Big(c_k(x,\xi) - \Big\langle \lambda, \max\{a^\top\xi - b, 0\}\Big\rangle_S - \langle z_{ik}, \xi - \xi^i\rangle\Big) \\ k\in[K], i\in[N]. \tag{24b}$$

where the minimization in constraint (24b) can be eliminated and a constraint $||z_{ik}||_* \leq \eta$ can be added to obtain the following model:

$$\inf_{\lambda\in S',\eta\in\mathbb{R}_+,s\in\mathbb{R}^N} \Big\langle \lambda, \frac{1}{N}\sum_{i=1}^{N} \max\{a^\top \xi^i - b, 0\} + \mathcal{Q}_{\mathtt{C}_N}(\alpha_2)\Big\rangle_S + \eta\epsilon_N^\beta + \frac{1}{N}\sum_{i=1}^{N} s_i \tag{25a}$$

$$\text{s.t. } s_i \geq \sup_{\xi\in\Xi}\Big(c_k(x,\xi) - \Big\langle \lambda, \max\{a^\top\xi - b, 0\}\Big\rangle_S - \langle z_{ik}, \xi - \xi^i\rangle\Big) \quad k\in[K], i\in[N] \tag{25b}$$

$$||z_{ik}||_* \leq \eta \qquad k\in[K], i\in[N]. \tag{25c}$$

Next, we take the dual of the supremum problem on the right-hand side of the constraints (25b) and convert it to an infimum problem so that the infimum can be dropped and the generated constraints can be added to the problem (25). For a given i and k, we take the dual of the supremum problem in the constraint.

To simplify the dual problem, we use the definition of convex conjugate of the cost function along with defining a new decision variables r_{ik}. By eliminating the infimum and adding the constraints of the dual problem to the problem (25) we obtain

$$\inf_{\lambda,\eta,s,H,r,z} \left\langle \lambda, \frac{1}{N}\sum_{i=1}^{N} \max\{a^\top \xi^i - b, 0\} + \mathcal{Q}_{\mathfrak{c}_N}(\alpha_2) \right\rangle_S + \eta\epsilon_N^\beta + \frac{1}{N}\sum_{i=1}^{N} s_i \tag{26a}$$

$$\text{s.t. } \lambda \in S', \eta \in \mathbb{R}_+, s \in \mathbb{R}^N, H_{ik} \in S', r, z \in \mathbb{R}^{N\times K} \tag{26b}$$

$$-s_i \leq \Big(c_{k*}(x, r_{ik}) - \langle z_{ik}, \xi^i\rangle - \langle H_{ik}, b\rangle_S\Big) \qquad k \in [K], i \in [N] \tag{26c}$$

$$\lambda - H_{ik} \in S' \qquad k \in [K], i \in [N] \tag{26d}$$

$$r_{ik} = z_{ik} + \langle H_{ik}, a\rangle_S \qquad k \in [K], i \in [N] \tag{26e}$$

$$||z_{ik}||_* \leq \eta \qquad k \in [K], i \in [N] \tag{26f}$$

Now, one can write the above problem as $\inf_{\eta,z} \inf_{\lambda,s,H,r}(\cdot)$ and take the dual of inner infimum problem using the dual variables p_{ik}, $\psi_{ik}(a,b)$, and q_{ik} associated with constraints (26c),(26d), and (26e), respectively. Notice that the decision variables in [Bertsimas et al., 2018] have index k only, but, in our reformulations since we also incorporate the decision variables associated with the Wasserstein ball, the decision variables have indices i and k, which means that the size of the problem is multiplied by the sample size N. After taking the dual, the inner infimum problem is equivalent to

$$\sup_{q\in\mathbb{R}^{K\times N\times d}, p\in\mathbb{R}_+^{K\times N}} \eta\epsilon_N^\beta + \sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik}\langle z_{ik}, \xi^i\rangle - \langle z_{ik}, q_{ik}\rangle + p_{ik} c\left(x, \frac{q_{ik}}{p_{ik}}\right) \tag{27a}$$

$$\text{s.t. } \sum_{k=1}^{K} p_{ik} = 1 \qquad i \in [N] \tag{27b}$$

$$\inf_{(a,b)\in S} \frac{1}{N}\sum_{i=1}^{N} \max\{a^\top \xi^i - b, 0\} + \mathcal{Q}_{\mathfrak{c}_N}(\alpha_2) - \sum_{i=1}^{N}\sum_{k=1}^{K} \max\{a^\top q_{ik} - p_{ik} b, 0\} \geq 0. \tag{27c}$$

To simplify constraint (27c), we define the parameter $\gamma \in \mathcal{G}$ where $\mathcal{G} = \{0,1\}^{K\times N}\backslash\{(0,\ldots,0)\}$. This parameter enumerates all possibilities of $\max\{a^\top q_{ik} - p_{ik} b, 0\}$. Therefore, constraint (27c) can be reformulated as

$$\mathcal{Q}_{\mathfrak{c}_N}(\alpha_2) \geq \sup_{a\in\mathbb{R}^d, b\in\mathbb{R}} \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}(a^\top q_{ik} - p_{ik} b) - \frac{1}{N}\sum_{i=1}^{N} \max\{a^\top \xi^i - b, 0\} \quad \gamma \in \mathcal{G} \tag{28a}$$

$$\text{s.t. } \sum_{j=1}^{d} |a_j| + |b| \leq 1. \tag{28b}$$

Problem (28) can be linearized and the linear dual of the problem can be taken, which results in an infimum problem. Constraint (27c) can be replaced by the infimum problem along with its constraints, and the infimum operator can be removed from the constraint. A dual can be taken from the modified version of the supremum problem (27), which results in an infimum problem that can be merged with $\inf_{\eta,z}(\cdot)$ and yield a single-level convex optimization problem (19). This reformulation can be further simplified, and the variable z can be omitted from the reformulation. For more details on the formulations, interested reader is referred

to Appendix A.3.

□

In the next section, we propose techniques for lower and upper bounding the reformulation (19) when it is computationally challenging.

7.2 Bounds for problems with multivariate distributions

The size of problem (19) is highly dependent on the size of the set $\mathcal{G}$ which grows exponentially in K and N. For many problems of practical interest, K may be 1 or 2. Therefore, dependence on K may not cause a lot of computational burden when solving the problem. However, growing the size of the set at an exponential rate in N makes the problem computationally expensive. In the next sections, we propose procedures to obtain upper and lower bounds for the problem (19) with larger samples.

7.2.1 A cutting-plane algorithm

We propose a three-level cutting-plane algorithm that can be applied for solving problem (19). This algorithm is indeed an exact approach but we will propose to use it to derive a lower bound in practice. We use an intermediate step in the proof of Theorem 3 to decompose the problem. Consider problem (27) merged with $\inf_{x,z,\eta}(\cdot)$ which yields

$$\inf_{x,z,\eta} \sup_{q,p} \eta\epsilon_N + \sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik}\langle z_{ik}, \xi^i\rangle - \langle z_{ik}, q_{ik}\rangle + p_{ik}c\left(x, \frac{q_{ik}}{p_{ik}}\right) \tag{29a}$$

$$\text{s.t. } q \in \mathbb{R}^{K\times N\times d}, p \in \mathbb{R}_+^{K\times N}, x \in X, z \in \mathbb{R}^{K\times N}, \eta \in \mathbb{R}_+ \tag{29b}$$

$$\sum_{k=1}^{K} p_{ik} = 1 \qquad i \in [N] \tag{29c}$$

$$\mathcal{Q}_{C_N}(\alpha_2) \geq \sup_{a\in\mathbb{R}^d, b\in\mathbb{R}} \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}(a^\top q_{ik} - p_{ik}b) - \frac{1}{N}\sum_{i=1}^{N}\max\{a^\top\xi^i - b, 0\} \quad \gamma \in \mathcal{G} \tag{29d}$$

$$\text{s.t. } \sum_{j=1}^{d}|a_j| + |b| \leq 1. \tag{29e}$$

Note that problem (19) is equivalent to problem (29) which can be decomposed into three problems: the grand master problem (GMP), the master problem (MP), and the subproblem (SP). Let $\mathcal{S}$ denote pairs of (p^*, q^*) obtained from MP. Given the set $\mathcal{S}$, the formulation of GMP is

$$\text{GMP}(\mathcal{S}) = \inf_{x,z,\tau} \eta\epsilon_N + \tau \tag{30a}$$

$$\text{s.t. } x \in X, \tau \in \mathbb{R}, z \in \mathbb{R}^{K\times N}, \eta \in \mathbb{R}_+ \tag{30b}$$

$$\tau \geq \sum_{i=1}^{N}\sum_{k=1}^{K} p^*_{ik}\langle z_{ik}, \xi^i\rangle - \langle z_{ik}, q^*_{ik}\rangle + p^*_{ik}c\left(x, \frac{q^*_{ik}}{p^*_{ik}}\right) \qquad (p^*, q^*) \in \mathcal{S} \tag{30c}$$

GMP yields an optimal solution $(\hat{x}, \hat{z}, \hat{\tau})$ that is passed on to MP. Let $\mathcal{R}$ and $\mathcal{G}'$ denote pairs of $(\hat{a}, \hat{b})$ and γ

yielded by SP. Then, MP can be formulated as

$$\mathrm{MP}(\mathcal{G}', \mathcal{R}, \hat{x}, \hat{z}) = \sup_{q,p} \sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik}\langle \hat{z}_{ik}, \xi^i\rangle - \langle \hat{z}_{ik}, q_{ik}\rangle + p_{ik} c\left(\hat{x}, \frac{q_{ik}}{p_{ik}}\right) \tag{31a}$$

$$\text{s.t. } q \in \mathbb{R}^{K\times N\times d}, p \in \mathbb{R}_+^{K\times N} \tag{31b}$$

$$\sum_{k=1}^{K} p_{ik} = 1 \qquad i \in [N] \tag{31c}$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K} \hat{\gamma}_{ik}(\hat{a}^\top q_{ik} - p_{ik}\hat{b}) \le \mathcal{Q}_{C_N}(\alpha_2) + \frac{1}{N}\sum_{i=1}^{N} \max\{\hat{a}^\top \xi^i - \hat{b}, 0\} \quad \hat{\gamma} \in \mathcal{G}', (\hat{a}, \hat{b}) \in \mathcal{R}. \tag{31d}$$

For the cost functions $c(x, \xi)$ which are affine in ξ, MP is a linear program. Given a candidate solution of MP, $(\hat{p}, \hat{q})$, SP can be formulated as

$$\mathrm{SP}(\hat{p}, \hat{q}) = \sup_{a\in\mathbb{R}^d, b\in\mathbb{R}} \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}(a^\top \hat{q}_{ik} - \hat{p}_{ik} b) - \frac{1}{N}\sum_{i=1}^{N} \max\{a^\top \xi^i - b, 0\} \tag{32a}$$

$$\text{s.t. } \sum_{j=1}^{d} |a_j| + |b| \le 1 \tag{32b}$$

$\mathrm{SP}(\hat{p}, \hat{q})$ can be linearized and solved by off-the-shelf solvers. In an intermediate iteration of the algorithm, GMP provides a candidate solution $(\hat{x}, \hat{z}, \hat{\tau})$. Given the solution, MP is solved, and its candidate solution $(\hat{p}, \hat{q})$ is passed to SP to find the optimal solution of SP, $(\hat{\gamma}, \hat{a}, \hat{b})$. By plugging the candidate solutions of MP and SP into the constraint (31d), we check whether there is a violated cut

$$\sum_{i=1}^{N}\sum_{k=1}^{K} \hat{\gamma}_{ik}(\hat{a}^\top \hat{q}_{ik} - \hat{p}_{ik}\hat{b}) > \mathcal{Q}_{\mathtt{C}_N}(\alpha_2) + \frac{1}{N}\sum_{i=1}^{N} \max\{\hat{a}^\top \xi^i - \hat{b}, 0\}.$$

If there is such a cut, $\hat{\gamma}$ and $(\hat{a}, \hat{b})$ are added to $\mathcal{G}'$ and $\mathcal{R}$, respectively, also a constraint of the form (31d) is added to MP. The loop between MP and SP continues until there is no violated cut for MP. At this stage, the optimal solution of MP is denoted by (p^*, q^*), which along with $(\hat{x}, \hat{z}, \hat{\tau})$ are used to detect the existence of a violated cut for GMP. By plugging in the solution into constraint (30c), we are looking for a violated cut of the form

$$\hat{\tau} < \sum_{i=1}^{N}\sum_{k=1}^{K} p^*_{ik}\langle \hat{z}_{ik}, \xi^i\rangle - \langle \hat{z}_{ik}, q^*_{ik}\rangle + p^*_{ik} c\left(\hat{x}, \frac{q^*_{ik}}{p^*_{ik}}\right).$$

If there is such a violation, (p^*, q^*) is added to $\mathcal{S}$ and a constraint of the form (30c) is added to GMP. The algorithm continues until there is no violated cut for GMP. We summarize the procedure of the three-level cutting-plane algorithm for solving problem (3) in multivariate setting in Algorithm 1.

Decomposing the problem into three smaller problems helps us to solve it for larger sample sizes. Furthermore, early termination of the algorithm yields a valid lower bound on the optimal objective value of problem (3). However, our computational experiments show that the convergence rate of the algorithm is very slow and the quality of the lower bound provided by the algorithm is not satisfactory. In the next section, we introduce a relaxation-based upper bounding technique which is proved to be computationally effective

Algorithm 1 Three-level Cutting-plane Algorithm for Solving (29)

Initialization: $\mathcal{S} = \mathcal{R} = \mathcal{G}' = \emptyset$
loop
  Solve GMP($\mathcal{S}$) to find $(\hat{x}, \hat{z}, \hat{\tau})$.
  loop
    Solve MP($\mathcal{G}', \mathcal{R}, \hat{x}, \hat{z}$) to find $(\hat{p}, \hat{q})$.
    Solve SP($\hat{p}, \hat{q}$) to find $(\hat{\gamma}, \hat{a}, \hat{b})$.
    if $\sum_{i=1}^{N}\sum_{k=1}^{K} \hat{\gamma}_{ik}(\hat{a}^\top \hat{q}_{ik} - \hat{p}_{ik}\hat{b}) > \mathcal{Q}_{\mathcal{C}_N}(\alpha_2) + \frac{1}{N}\sum_{i=1}^{N} \max\{\hat{a}^\top \xi^i - \hat{b}, 0\}$ **then**
      Add cut (31d) at $(\hat{\gamma}, \hat{a}, \hat{b})$.
    else
      Stop.
    end if
  end loop
  Set $p^* \leftarrow \hat{p}$ and $q^* \leftarrow \hat{q}$
  if $\hat{\tau} < \sum_{i=1}^{N}\sum_{k=1}^{K} p^*_{ik}\langle \hat{z}_{ik}, \xi^i\rangle - \langle \hat{z}_{ik}, q^*_{ik}\rangle + p^*_{ik} c\left(\hat{x}, \frac{q^*_{ik}}{p^*_{ik}}\right)$ **then**
    Add cut (30c) at (p^*, q^*)
  else
    Stop.
  end if
end loop

and can provide high-quality upper bounds. We design a procedure which runs the three-level cutting-plane algorithm to provide inputs for the relaxation-based techniques, and our experiments show that the procedure is indeed effective. We discuss the details of the procedure and illustrate its performance in Section 8.3.

7.2.2 Upper bounding: Relaxation of the supremum problem

In order to obtain an upper bound on the problem (3) in a multivariate setting, we consider a relaxation of an intermediate step of the proof of Theorem 3. The relaxation can be obtained by replacing set $\mathcal{G}$ by set $\mathcal{G}'$ where $\mathcal{G}' \subseteq \mathcal{G}$ in constraints (28a). The following proposition, highlights the fact that problem (19) with $\mathcal{G}'$ set provides a valid upper bound on problem (19) with $\mathcal{G}$ set.

Proposition 1. *Let $\mathcal{C}'(x, \mathcal{F})$ denote the optimal objective value of problem (19) given $\mathcal{G}' \subseteq \mathcal{G}$. Under assumptions of $\mathcal{A}_J^{\texttt{Valid}}$, $\mathcal{C}'(x, \mathcal{F})$ is a valid upper bound for $\mathcal{C}(x, \mathcal{F})$.*

Proof. Problem (27) is equivalent to the following problem after replacing constraint (27c) by the reformulation of problem (28).

$$\sup_{\substack{q \in \mathbb{R}^{K\times N\times d},\\ p \in \mathbb{R}_+^{K\times N}}} \eta\epsilon_N^\beta + \sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik}\langle z_{ik}, \xi^i\rangle - \langle z_{ik}, q_{ik}\rangle + p_{ik} c\left(x, \frac{q_{ik}}{p_{ik}}\right) \tag{33a}$$

$$\text{s.t.} \sum_{k=1}^{K} p_{ik} = 1 \qquad i \in [N] \tag{33b}$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}(a^\top q_{ik}-p_{ik}b)-\frac{1}{N}\sum_{i=1}^{N}\max\{a^\top\xi^i-b,0\}\leq \mathcal{Q}_{\mathfrak{C}_N}(\alpha_2)$$

$$\gamma\in\mathcal{G},(a,b)\in S \qquad (33c)$$

Replacing set $\mathcal{G}$ in constraints (33c) with $\mathcal{G}'$ yields a relaxation of the supremum problem which consequently provides an upper bound on problem (19). □

The set $\mathcal{G}'$ can be a random subset of $\mathcal{G}$ or can be created based on the information obtained from solving problem (19) with $\mathcal{G}$ for small sample sizes. Our computational studies revealed that even small sized $\mathcal{G}'$ sets can provide good results while requiring less computational effort.

8 Numerical Results

We conduct numerical experiments on a newsvendor problem and a mean-risk portfolio allocation problem similar to Bertsimas et al. [2018] and Esfahani and Kuhn [2018]. The experiments validate the theoretical results on the convergence and performance guarantee of DRO problems with proposed joint ambiguity sets. All experiments are conducted on Niagara GNU-parallel [Tange, 2018], and all the models are implemented in Python 2.7.16 and solved via GUROBI 9.0.0 [Gurobi Optimization, LLC, 2022]. Moreover, we evaluate the effectiveness of our proposed bounding techniques by experimenting on a mean-risk portfolio allocation problem where these set of experiments are conducted on a Mac computer with 3 GHz Intel Core i5 CPU and 16 GB memory.

8.1 Single-item newsvendor problem

Consider a newsvendor problem where the goal is to decide on the ordering quantity, $x\geq 0$, to minimize a cost function while considering the uncertainty in demand. Let ξ denote the uncertain future demand following distribution $\mathbf{F}$. Let $g>0$ and $h>0$ represent the lost sale for unmet demand and the holding cost of excess inventory, respectively. In this problem, we assume a continuous demand with $\xi\in[\underline{\xi},\bar{\xi}]$ where $\bar{\xi}<\infty$. The cost function associated with this problem is $c(x,\xi)=\max\{g(\xi-x),h(x-\xi)\}$. The DRO problem for newsvendor problem can be written as

$$\bar{z}=\min_{x\geq 0}\ \max_{F\in\mathcal{F}}\ \mathbb{E}_F[\max\{g(\xi-x),h(x-\xi)\}]. \qquad (34)$$

To construct the uncertainty set $\mathcal{F}$, we use the Kolmogorov-Smirnov (KS), Kuiper, and Cramér-von Mises (CvM) tests. We consider pairwise intersection of the confidence regions and intersection of all three regions. Given a sorted sample of $\{\xi^1,\ldots,\xi^N\}$, we solve problem (14) where the constraint (14d) is replaced by

$$\ell_i\geq g(\xi^i-x)\quad i\in[N+1] \qquad (35a)$$

$$\ell_i\geq h(x-\xi^{i-1})\quad i\in[N+1] \qquad (35b)$$

where $\xi^0=\underline{\xi}$ and $\xi^{N+1}=\bar{\xi}$. The definition of dual cones K^*, matrices A, and vector $b(\alpha)$, mentioned in Section 6, are provided in [Bertsimas et al., 2018] for various GoF tests. For the sake of self-containment,

we mention the information regarding KS, CvM, and Kuiper tests here. Let $\mathcal{Q}_{\texttt{KS}}(\alpha)$, $\mathcal{Q}_{\texttt{CvM}}(\alpha)$, and $\mathcal{Q}_{\texttt{Kuiper}}(\alpha)$ denote the critical values of the KS, CvM, and Kuiper tests at the significance level of α, respectively, then

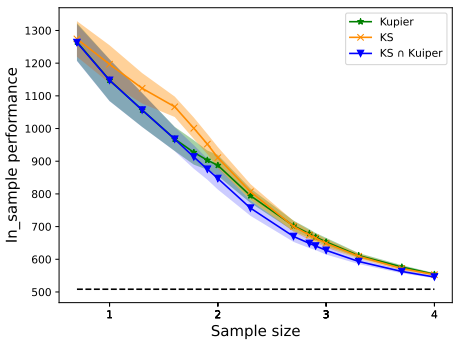
$$b_{\texttt{KS},\alpha} = \begin{bmatrix} \frac{1}{N} - \mathcal{Q}_{\texttt{KS}}(\alpha) \\ \vdots \\ \frac{N}{N} - \mathcal{Q}_{\texttt{KS}}(\alpha) \\ -\frac{0}{N} - \mathcal{Q}_{\texttt{KS}}(\alpha) \\ \vdots \\ -\frac{N-1}{N} - \mathcal{Q}_{\texttt{KS}}(\alpha) \end{bmatrix}, A_{\texttt{KS}} = \begin{bmatrix} [I_N] \\ [-I_N] \end{bmatrix}, K^*_{\texttt{KS}} = \mathbb{R}^{2N}_+,$$

$$b_{\texttt{CvM},\alpha} = \begin{bmatrix} \sqrt{N(\mathcal{Q}_{\texttt{CvM}}(\alpha))^2 - \frac{1}{12N}} \\ \frac{1}{2N} \\ \frac{3}{2N} \\ \vdots \\ \frac{2N-1}{2N} \end{bmatrix}, A_{\texttt{CvM}} = \begin{bmatrix} 0 \dots 0 \\ [I_N] \end{bmatrix}, K^*_{\texttt{CvM}} = C^{N+1}_{\texttt{SOC}},$$

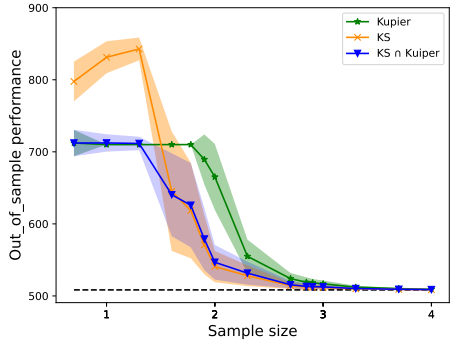
$$b_{\texttt{Kuiper},\alpha} = \begin{bmatrix} \frac{1}{N} - \frac{\mathcal{Q}_{\texttt{Kuiper}}(\alpha)}{2} \\ \vdots \\ \frac{N}{N} - \frac{\mathcal{Q}_{\texttt{Kuiper}}(\alpha)}{2} \\ -\frac{0}{N} - \frac{\mathcal{Q}_{\texttt{Kuiper}}(\alpha)}{2} \\ \vdots \\ -\frac{N-1}{N} - \frac{\mathcal{Q}_{\texttt{Kuiper}}(\alpha)}{2} \end{bmatrix}, A_{\texttt{Kuiper}} = \begin{bmatrix} [I_N] \\ [-I_N] \end{bmatrix}, K^*_{\texttt{Kuiper}} = \{(r, r') \in \mathbb{R}^{2N}_+ : \sum_{i=1}^{N} r_i = \sum_{i=1}^{N} r'_i\}.$$

Problem (14) is a linear optimization problem when we use the information corresponding to KS and Kuiper tests, and it is a second order cone programming problem with the information of CvM test. In our experiments, we consider $g = 19$, $h = 1$, and Normal distribution with $\mu = 200$ and $\sigma^2 = 70$ as the underlying unknown distribution with support $\Xi = [50, 400]$. We run our experiments using samples of 15 different sizes ranging from 5 to 10^4 and we run 100 independent for each sample size. Also, we calculate the out-of-sample performances of the obtained solutions by running SAA using a sample of size 10^5. In the experiments, all tests are at the $\alpha = 20\%$ significance level. The critical value associated with KS test is calculated using Python built-in function (*ksone*) and critical values of CvM and Kuiper are computed by bootstrap algorithm (see the details in Appendix B). In these experiments, all models are solved to optimality.

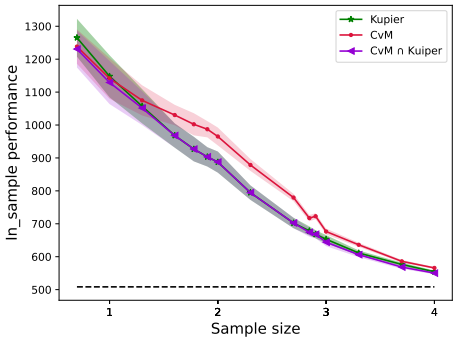
Figure 1 illustrates the performances associated with the pairwise intersections of the tests and the combination of all of them. The figures on the left present the in-sample performances of the different tests and their intersection with respect to the sample size, which are represented on a logarithmic base 10 scale. The figures on the right show the out-of-sample performances of the tests and their intersections with respect to the sample size, which are again represented in the same logarithmic scale. The solid lines in the figures are the average of the outcomes over 100 independent runs whereas the shaded areas around the lines are their corresponding variations. The dashed lines in the figures show the optimal objective value of the problem (34).



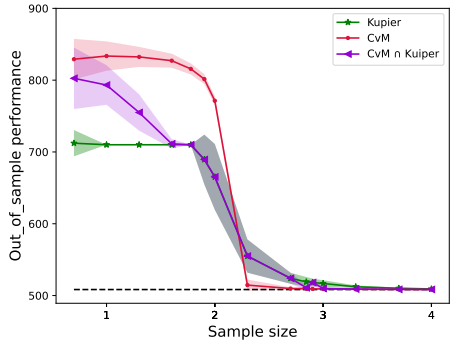
(a) In-sample results using Kuiper and KS



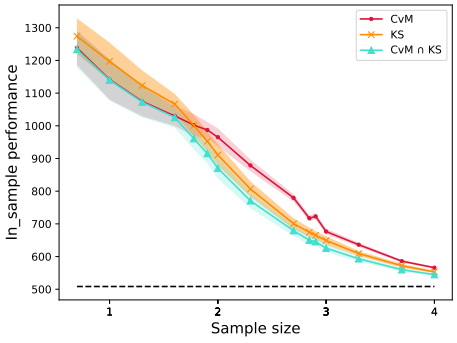
(b) Out-of-sample results using Kuiper and KS



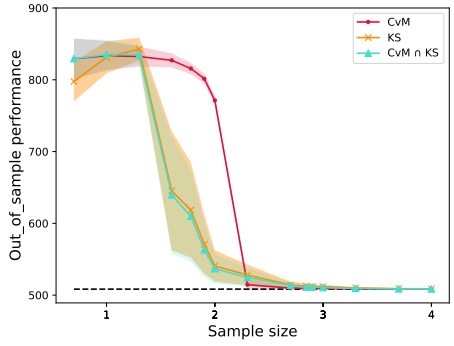
(c) In-sample results using Kuiper and CvM



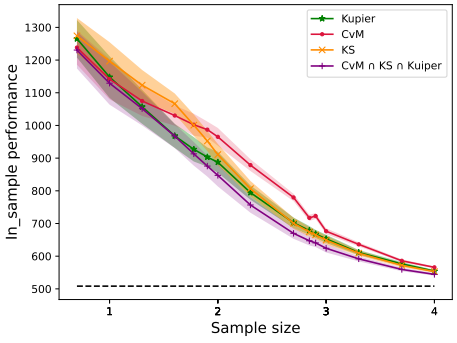
(d) Out-of-sample results using Kuiper and CvM



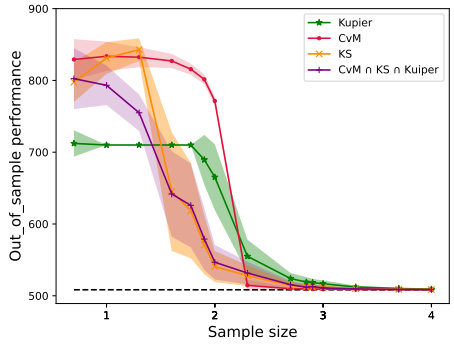
(e) In-sample results using CvM and KS



(f) Out-of-sample results using CvM and KS



(g) In-sample results using and Kuiper, CvM, and KS



(h) Out-of-sample results using Kuiper, CvM, and KS

Figure 1 Comparison of in-sample and out-of-sample performances of DRO problem with ambiguity sets created by Kuiper, KS, CvM tests, and their intersections.

In terms of in-sample performances, the figures illustrate that the results of the DRO problem with joint ambiguity sets is at least as good as the performance of the problems with individual ambiguity sets. In Figures 1(a), 1(e), and 1(g) the performance of the problems with joint ambiguity sets is equal to the best performance among the individual tests when the sample size is small, and the performance improves as the sample size grows larger. Additionally, Figure 1(c) shows that, for each sample size, the performance of the problem with the joint ambiguity set is equal to the performance of the test that produces the best results.

In terms of out-of-sample performances, Figure 1(b) illustrates that the performance of the DRO problem with joint ambiguity set is the same as the performance of the problem with individual ambiguity set, which produces the best outcome for each sample sizes. Figures 1(d), 1(f), and 1(h) show that the DRO problem with the joint ambiguity set is not the best performing problem in small sample sizes but as the sample size increases, its performance is as good as the performance of the problem with individual sets.

Additionally, in these experiments, we observe that the in-sample performances are always upper bounds on the out-of-sample performances of the optimal solutions of the DRO problems with joint and individual ambiguity sets, meaning that we achieve a reliability of 1.

8.2 Mean-risk portfolio optimization

Consider a portfolio optimization problem with multiple assets and random returns where the goal is to find the best portfolio allocation weights by optimizing over a weighted sum of the expectation and conditional value-at-risk (CVaR) of a loss function. Consider a portfolio problem with d assets where $\xi = (\xi_1, \ldots, \xi_d)$ denotes the random return vector associated with the assets following an unknown distribution $\mathbf{F}$ with support $\mathbb{R}^d$. The optimization problem will decide on $x = (x_1, \ldots, x_d)$, portfolio allocation vector representing the fraction of total budget that will be allocated to each asset which belongs to the set $X = \{x \in \mathbb{R}^d_+ \mid \langle x, 1\rangle \ = 1\}$. Using the mentioned notations, the mean-risk portfolio optimization problem can be formulated as

$$z^{\texttt{SP}} = \inf_{x \in X} \ \left\{ \mathbb{E}_{\mathbf{F}}[-\langle x, \xi\rangle] + \rho \ \mathrm{CVaR}^{\beta}_{\mathbf{F}}[-\langle x, \xi\rangle] \right\} \tag{36}$$

where $\rho \geq 0$ illustrates the decision maker's risk aversion and $\beta \in (0, 1]$ is the confidence level of CVaR. The corresponding DRO problem is equivalent to

$$z^{\texttt{DRO}} = \inf_{x \in X, \tau \in \mathbb{R}} \ \sup_{F \in \mathcal{F}} \ \mathbb{E}_F \left[\max_{k \in [K]} \ a_k \langle x, \xi\rangle \ + \ b_k \ \tau \right] \tag{37}$$

where $K = 2$, $a = [-1, \ -1 - \frac{\rho}{\beta}]$, $b = [\rho, \ \rho(1 - \frac{1}{\beta})]$ [Esfahani and Kuhn, 2018]. To construct the ambiguity set, $\mathcal{F}$, we consider the region created by the LCX test intersected with the Wasserstein ball. Therefore, we solve the following optimization model:

$$\inf_{\substack{r, f, w, w', \\ y, y', e, \eta, x, \tau}} \quad \eta\epsilon + \frac{1}{N} \sum_{i=1}^{N} r_i + \sum_{\gamma \in \mathcal{G}} f_\gamma \ Q_{\mathtt{c}_N}(\alpha_1) + \sum_{\gamma \in \mathcal{G}} \sum_{i=1}^{N} e_{\gamma i} \tag{38a}$$

$$\text{s.t.} \sum_{j=1}^{d} y_{\gamma j} + y'_\gamma \leq f_\gamma \qquad \gamma \in \mathcal{G} \tag{38b}$$

$$\frac{1}{N} (\sum_{j=1}^{d} w_{\gamma j} \ \xi^i_j - w'_\gamma) \leq e_{\gamma i} \qquad \gamma \in \mathcal{G}, \ i \in [N] \tag{38c}$$

$$w_{\gamma j} - y_{\gamma j} \leq 0 \qquad \gamma \in \mathcal{G},\ j \in [d] \tag{38d}$$

$$-w_{\gamma j} - y_{\gamma j} \leq 0 \qquad \gamma \in \mathcal{G},\ j \in [d] \tag{38e}$$

$$w'_{\gamma} - y'_{\gamma} \leq 0 \qquad \gamma \in \mathcal{G} \tag{38f}$$

$$-w'_{\gamma} - y'_{\gamma} \leq 0 \qquad \gamma \in \mathcal{G} \tag{38g}$$

$$\langle a_k x - \sum_{\gamma \in \mathcal{G}} \gamma_{ik}\ w_\gamma,\ \xi^i \rangle - r_i + \sum_{\gamma \in \mathcal{G}} \gamma_{ik}\ w'_\gamma + b_k \tau \leq 0 \qquad k \in [K],\ i \in [N] \tag{38h}$$

$$\| a_k x - \sum_{\gamma \in \mathcal{G}} \gamma_{ik}\ w_\gamma \|_* \leq \eta \qquad k \in [K],\ i \in [N] \tag{38i}$$

$$\sum_{j=1}^{d} x_j = 1 \tag{38j}$$

$$\begin{aligned} &\eta \in \mathbb{R}_+,\ f \in \mathbb{R}_+^{|\mathcal{G}|},\ e \in \mathbb{R}_+^{|\mathcal{G}| \times N}, \\ &y \in \mathbb{R}_+^{|\mathcal{G}| \times d},\ y' \in \mathbb{R}_+^{|\mathcal{G}|}, \\ &r \in \mathbb{R}^N, w \in \mathbb{R}^{|\mathcal{G}| \times d},\ w' \in \mathbb{R}^{|\mathcal{G}|} \\ &x \in \mathbb{R}_+^d,\ \tau \in \mathbb{R}. \end{aligned} \tag{38k}$$

We consider one-norm in Wasserstein formulation, so, the associated dual norm will be norm-infinity. As a result, constraint (38i) can be written as follows:

$$\| a_k x - \sum_{\gamma \in \mathcal{G}} \gamma_{ik}\ w_\gamma \|_* = \max_{j \leq d} |\ a_k x_j - \sum_{\gamma \in \mathcal{G}} \gamma_{ik}\ w_{\gamma j}| \leq \eta \qquad k \in [K],\ i \in [N]$$

$$-\eta \leq a_k x_j - \sum_{\gamma \in \mathcal{G}} \gamma_{ik}\ w_{\gamma j} \leq \eta \qquad j \in [d],\ k \in [K],\ i \in [N].$$

In our experiments, we consider a portfolio problem with $d = 10$ with the distributions similar to the ones mentioned by Esfahani and Kuhn [2018]. We assume $\xi_i = \psi + \varsigma_i$ where $\psi \sim N(0, 2\%)$ and $\varsigma_i \sim N(i \times 3\%, i \times 2.5\%)$ for each asset $i \in \{1, \ldots, d\}$. We also consider $\rho = 10$ and CVaR at confidence level $\beta = 20\%$. We focus on the quality of the upper bounds provided by the individual measures and their intersection. For this purpose, we implemented the relaxation-based technique described in Section 7.2.2. The method is computationally efficient and provides a valid upper bound on the in-sample performance of the DRO problem with a joint ambiguity set. In this procedure, the set $\mathcal{G}'$ can be chosen as any subset of $\mathcal{G}$.

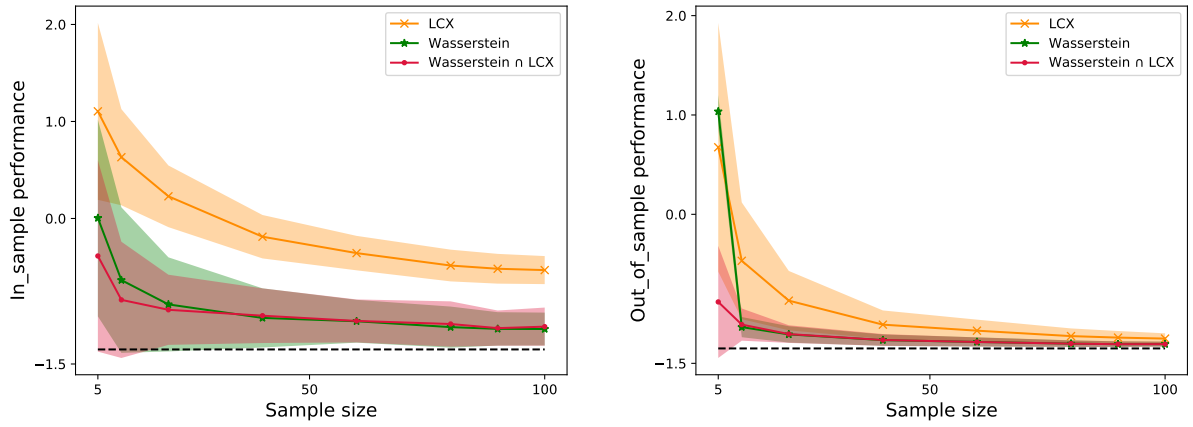
In our experiments, we set $\mathcal{G}' = \mathcal{G}$ when solving the portfolio problem (38) with small sample sizes. We analyze the obtained optimal solutions and observe that most of the variables with the γ subscript take zero values in the solutions. The γ's associated with the non-zero variables can be used to create $\mathcal{G}'$ for the problems with larger sample sizes. For these problems, we construct the set using 20 randomly generated γ's along with the ones associated with those non-zero decision variables with γ subscript. Our experiments reveal that this procedure of creating $\mathcal{G}'$ is effective as it results in high-quality upper bounds.

In these experiments, tuning the radius of the Wasserstein ball and critical value of LCX test is of foremost importance. If the radius is chosen to be significantly larger than the critical value of the LCX test, the ball would contain the confidence region of the test. In such a case, the joint ambiguity set would be the same as the confidence region of the LCX test, which would result in the same outcomes for the corresponding DRO problems. On the other hand, if the radius is chosen significantly smaller than the critical value of

the test, the ball would be the subset of the confidence region. In this case, the joint ambiguity set would be the same as the Wasserstein ball, and the outcomes of the corresponding DRO problems would be the same. Therefore, in these experiments we followed the same steps as proposed by Esfahani and Kuhn [2018] and Bertsimas et al. [2018] for choosing the radius of the Wasserstein ball and the critical value of the LCX test, respectively. The radius of the Wasserstein ball is selected from a discrete set proposed by Esfahani and Kuhn [2018], namely, from $E = \{\epsilon = b \cdot 10^c \ : \ b \in \{0, \ldots, 9\}, c \in \{-3, -2, -1\}\}$. In our experiments, for each run, we choose the best radius from the set in terms of its out-of-sample performance. Additionally, the threshold of the LCX test is calculated using the bootstrap procedure which is explained in Section 9.3 of Bertsimas et al. [2018] (see the details in Appendix B).

Figures 2 illustrate the results associated with the Wasserstein ball, the LCX test, and their intersection. While Figure 2(a) presents the in-sample performances of the individual measures and their intersection with respect to the sample size, Figure 2(b) shows their out-of-sample performances with respect to the sample size. The solid lines in the figures are the average of the outcomes over 200 independent runs whereas the shaded areas around the lines show are their corresponding variations. The dashed lines in the figures show the optimal objective value of the problem (37).

In terms of in-sample performances, Figure 2(a) illustrates that the results of the DRO problem with the joint ambiguity set are slightly better than those with Wasserstein and LCX for small sample sizes. As samples get larger, performance of the joint ambiguity set and Wasserstein ball become the same and both of them outperform LCX test. In terms of out-of-sample performances, Figure 2(b) illustrates that the performance of the DRO problem with joint ambiguity set is better than the Wasserstein ball and the LCX test for small sample sizes. However, as the sample size increases similar to in-sample results, the performances of the joint ambiguity set and Wasserstein ball become the same.



(a) In-sample performance analysis

(b) Out-of-sample performance analysis

Figure 2 Comparison of in-sample and out-of-sample performances of DRO problem with ambiguity sets created by LCX test, Wasserstein metric, and their intersection.

In terms of reliability, Figure 3 illustrates that the DRO problem with Wasserstein metric-based ambiguity set provides an optimal objective value which is always a valid upper bound on the out-of-sample performance of its optimal solution. On the other hand, it shows that the DRO problem with joint ambiguity set provides the same reliability as the problem with LCX test-based set. We, also, observe that for small sample sizes, the DRO problem provides a valid upper bound with probability of at least 0.7 and the reliability approaches

to 1 as the sample sizes increase. While in these experiments we expect the DRO problem with the joint ambiguity set to provide lower reliability than the DRO problem with individual sets, we observe that the reliability level is the same as the least reliable problem.

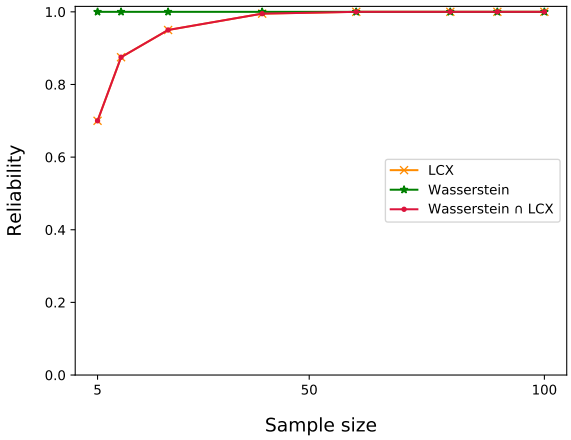


Figure 3 Comparison of the reliability of DRO problems with ambiguity sets created by LCX test, Wasserstein metric, and their intersection.

8.3 Evaluation of bounding techniques

In this section, we assess the performance of the proposed three-level cutting-plane algorithm and relaxation-based upper bounding technique and evaluate the quality of the provided bounds. We conduct experiments on the mean-risk portfolio optimization problem discussed in Section 8.2 where the joint ambiguity set of the DRO problem was constructed as an intersection of the sets created by Wasserstein metric and LCX test. In these experiments, we combine two bounding techniques where we run the three-level cutting-plane algorithm and at the end of each GMP iteration, we construct the $\mathcal{G}'$ set discussed in the relaxation-based technique using the γ's generated through MP-SP iterations. Using this procedure, at the end of each GMP iteration of the algorithm we obtain a lower bound and an upper bound on the optimal objective value of the DRO problem with joint ambiguity set.

In order to obtain valid bounds while keeping the computational burden manageable, we fix the radius of Wasserstein ball and the critical value of the LCX test to 0.7 and 0.05, respectively. Additionally, a lower bound of -10 is set on the optimal objective value provided by the three-level cutting-plane algorithm while initializing the algorithm. We run our experiments with GUROBI 9.0.3, on a Mac computer with 3 GHz Intel Core i5 CPU and 16 GB memory, also we set a time limit of 4 hours. For the experiments that cannot complete one GMP iteration within the time limit, we stop the procedure after one GMP iteration is done.

Results of our experiments are presented in Table 1 where the numbers are average values that are taken over 8 random samples of the same size. Note that the numbers in the parenthesis report the standard deviations. In what follows we explain the column labels used in the table.

- "N" is the sample size,
- "CuttingPlane LB" is the lower bound provided by the three-level cutting-plane algorithm,

- "Relaxation UB" is the upper bound provided by the relaxation-based technique,
- "First UB" is the upper bound provided by the relaxation-based technique at the end of the first iteration,
- "Wass Obj" is the optimal objective value of the DRO problems with Wasserstein metric-based ambiguity set,
- "% Improv in UB" is the percentage of the improvement on the upper bound yielded by the relaxation-based technique with respect to the DRO problem with Wasserstain metric-based ambiguity set,
- "Num γ in First Iter" is the number of γ's generated in the first iteration of the three-level cutting-plane algorithm,
- "Total Num γ" is the total number of γ's generated within the time limit,
- "Wass Time" is the time taken by the DRO problem with Wasserstein metric-based ambiguity set to be solved,
- "% First Iter Time" is the percentage of the total time that the first iteration of the three-level cutting-plane algorithm takes to be completed,
- "% MP Time" is the percentage of the total time that MP problem takes to be solved,
- "% SP Time" is the percentage of the total time that SP problem takes to be solved.

Table 1 Performance summary of our three-level cutting-plane algorithm and relaxation-based technique

N	CuttingPlane LB	Relaxation UB	First UB	Wass Obj	% Improv in UB	Num γ in First Iter	Total Num γ	Wass Time	% First Iter Time	% MP Time	% SP Time
10	-1.50 (0.37)	0.85 (0.24)	0.85 (0.24)	2.37 (0.14)	64.1	126.4	2^{20}	0.01	1.1	35.2	64.0
15	-7.55 (0.36)	0.95 (0.21)	0.95 (0.21)	2.36 (0.15)	59.7	251.9	2^{30}	0.01	18.7	5.5	94.2
20	-10.00 (0.00)	1.06 (0.24)	1.06 (0.24)	2.40 (0.14)	55.8	375.1	2^{40}	0.05	100.0	9.7	90.1
30	-10.00 (0.00)	2.33 (0.22)	2.33 (0.22)	2.46 (0.16)	5.3	279.5	2^{60}	0.02	100.0	5.5	94.3
50	-10.00 (0.00)	2.48 (0.05)	2.48 (0.05)	2.48 (0.05)	0.0	361.1	2^{100}	0.01	100.0	15.9	83.8

The results show that the quality of the lower bound provided by the three-level cutting-plane algorithm is not satisfactory, and as the sample size gets larger, there is no improvement on the lower bound. On the other hand, the relaxation-based technique provides high-quality upper bounds for all sample sizes. By comparing the upper bounds obtained at the end of the first and the last iterations, it can be inferred that the improvements in the quality of the upper bounds are negligible. In our experiments, we observed that the Wasserstein metric-based DRO problem outperforms the one with the LCX test-based set in terms of the quality of the bound and computational time. Therefore, herein we compare the quality of the upper bound obtained from relaxation-based technique at the end of the first iteration with the one provided by the DRO problem with Wasserstein metric-based ambiguity set. Note that the bounds provided by the relaxation-based technique are upper bound on the optimal objective value of the DRO problem with joint ambiguity set, which is a valid upper bound on the optimal objective value of the underlying stochastic programming problem. Looking into the percentage improvements, for small samples the bounds provided by

our technique are better than the ones yielded by the Wasserstein metric-based DRO problem up to 64%, and as the sample size increases the latter upper bounds converge to the former ones. The results suggest that with a small subset of γ's we can obtain high-quality upper bounds using the proposed relaxation-based technique. While using randomly generated γ's might not be beneficial, running one iteration of the three-level cutting-plane algorithm generates valuable ones. These outcomes indicate that only one iteration of the three-level cutting-plane algorithm suffices to produce high-quality upper bounds, which also highlights the quality of the γ's generated in the first iteration of the algorithm.

In term of computational time, the table illustrates that the DRO problem with Wasserstein metric-based ambiguity set can quickly reach to optimality, while the three-level cutting-plane algorithm is not computationally efficient. In our experiments, we look into different steps of the algorithm and analyze their time efficiency. While the time that is spent for solving GMP is negligible with respect to those of MP and SP, the most computationally challenging step of the algorithm is to solve SP. In terms of the number of iterations, for $N = 10$ ($N = 15$) GMP completes 7.8 (3) iterations on average which executes MP-SP loop 1472.9 (635.9) times on average to reach optimality. For $N \in [20, 30, 50]$, we terminate the algorithm after one full GMP iteration since the runs exceed the given time limit. Within the iteration, for $N = 20$ ($N = 30$, $N = 50$) the total number of MP-SP loop execution on average is 375.1 (279.5,361.1).

Our results reveal that early termination of the algorithm is beneficial in terms of upper bound as it is a valid bound for the optimal objective value of the underlying stochastic programming problem. However, this may not apply for the lower bound as it is not a valid bound on that optimal objective value. Based on our experiments, it is reasonable to run one iteration of the three-level cutting-plane algorithm and construct $\mathcal{G}'$ set based on generated γ's, which results in a high-quality upper bound.

9 Discussion

In this section, we summarize and discuss our findings on the DRO problems with joint ambiguity sets. In terms of performance guarantee, our proposed approach provides a valid lower bound on the optimal objective value of the DRO problems with a single-measure-based ambiguity set while producing a valid upper bound on the underlying SP problem. This means that the optimal objective value of the DRO problem with the joint ambiguity set is more representative of the SP problem and it is less robust than other approaches. Additionally, the optimal solution of the DRO problem with the joint ambiguity set is robust in the sense that it is obtained with respect to a worst-case distribution which belongs to a smaller ambiguity set and it is protected against small perturbations of the underlying probability distribution. Our experiments also show that while the joint ambiguity set improves the quality of the bound and the solution, the reliability of the joint set does not decline and indeed it is the same as the least reliable individual set that is incorporated in the construction of the joint set.

In terms of tractability, while we assume that the cost function can be written as the maximum of affine functions, we observe that the final reformulations of the DRO problem with the joint set include the same constraint structure as the formulations of the DRO problems with individual sets. The final reformulations have additional dual variables which couples the constraints coming from the formulations of the individual problems. In the case of problems with discrete known support, the formulation of the DRO problem with χ^2-test and G-test based ambiguity set include second-order cone (SOC) and exponential (Expo) cone, respectively, in addition to the linear (Linear) constraints. The final reformulation of the DRO problem with a joint set created by those tests includes constraints of the same type with additional dual

variables. In the case of problems with univariate support, the individual problems have conic and linear constraint and the same constraint structure appears in the final reformulation of the DRO problem with the joint set. Lastly, in the case of problems with multivariate supports, the DRO problems with individual sets only have linear constraints, and we note that the final reformulation of the DRO problem with the joint set includes only linear constraints. Table 2 illustrates the summary of our observations which suggests that intersecting several ambiguity set does not increase the complexity of the final reformulation by adding intractable constraints to the model. However, our numerical experiments highlight that the computational complexity of the DRO problems with the joint ambiguity sets is greater than the ones with individual sets due to the additional decision variables and constraints that are included in the final reformulations.

Table 2 Comparison of the constraint structure of the DRO problems

	Measures	Individual problems	Joint problems
Problems with discrete known support	χ^2 test	SOC + Linear	SOC + Expo +Linear
	G test	Expo + Linear	
Problems with univariate support	KS, Kuiper, CvM tests	Conic + Linear	Conic + Linear
Problems with multivariate support	LCX test, Wasserstein metric	Linear	Linear

10 Conclusions

In this study, we consider stochastic programs where the distribution of the uncertain parameters is unknown and partial information about it can only be captured from limited available data. We use the distributionally robust optimization framework for modeling such problems. We propose to construct the ambiguity set of DRO problems as a joint region that is an intersection of multiple regions each created by an individual measure from the literature. More specifically, we consider the joint region of the ambiguity sets created by discrepancy-based measures, namely, Wasserstein metric and Goodness-of-Fit tests. We look into the conditions under which the joint region can preserve useful properties of individual sets; in particular, performance guarantee, convergence, and tractability.

We derive tractable single-level convex reformulations for DRO problems with joint ambiguity set for three different problem settings. For computationally challenging problems, we additionally propose lower and upper bounding techniques, and illustrate the quality of the provided bounds. We conduct numerical experiments on two well-known problems from the literature, namely, the newsvendor and mean-risk portfolio allocation problems. Our results indicate that, for small sample sizes, the DRO problem with the joint ambiguity set has better in-sample and out-of-sample performances compared to the problems with individual ambiguity sets, which is in alignment with our theoretical results. The results also show that, as the samples get larger, the outcomes of the best performing DRO problem with individual ambiguity set converge to those of the DRO problem with joint ambiguity set.

We observe that the LCX test brings exponentially many decision variables and constraints to the resulting DRO problems. Therefore, as a future research direction, one can consider creating sets with other measures such as ϕ-divergences to intersect with the Wasserstein ball. Also, our experiments indicate that the three-level cutting-plane algorithm does not provide good-quality lower bounds and its convergence rate is low. Thus, more efficient lower bounding or general solution methods can be developed and their performances can be evaluated. Lastly, our proposed approach can be applied to various applications of practical interest where only limited data is available.

References

Peiman Ghasemi, Kaveh Khalili-Damghani, Ashkan Hafezalkotob, and Sadigh Raissi. Stochastic optimization model for distribution and evacuation planning (a case study of tehran earthquake). *Socio-Economic Planning Sciences*, 71:100745, 2020.

Xing Hong, Miguel A Lejeune, and Nilay Noyan. Stochastic network design for disaster preparedness. *IIE Transactions*, 47(4):329–357, 2015.

András Prékopa. *Stochastic programming*, volume 324. Springer Science & Business Media, 2013.

John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

Michael G Kapteyn, Karen E Willcox, and Andy B Philpott. Distributionally robust optimization for engineering design under uncertainty. *International Journal for Numerical Methods in Engineering*, 120(7): 835–859, 2019.

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

Herbert E Scarf. *A min-max solution of an inventory problem*. Rand Corporation Santa Monica, 1957.

Guillermo Gallego and Ilkyeong Moon. The distribution free newsboy problem: review and extensions. *Journal of the Operational Research Society*, 44(8):825–834, 1993.

Somayyeh Lotfi and Stavros A Zenios. Robust VaR and CVaR optimization under joint ambiguity in distributions, means, and covariances. *European Journal of Operational Research*, 269(2):556–576, 2018.

Napat Rujeerapaiboon, Daniel Kuhn, and Wolfram Wiesemann. Chebyshev inequalities for products of random variables. *Mathematics of Operations Research*, 43(3):887–918, 2018.

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

Alexander Shapiro and Shabbir Ahmed. On a class of minimax stochastic programs. *SIAM Journal on Optimization*, 14(4):1237–1249, 2004.

Sanjay Mehrotra and Dávid Papp. A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. *SIAM Journal on Optimization*, 24(4):1670–1697, 2014.

Georgia Perakis and Guillaume Roels. Regret in the newsvendor model with partial information. *Operations Research*, 56(1):188–203, 2008.

Johannes O Royset and Roger J-B Wets. Variational theory for optimization under stochastic ambiguity. *SIAM Journal on Optimization*, 27(2):1118–1149, 2017.

Louis Chen, Will Ma, Karthik Natarajan, David Simchi-Levi, and Zhenzhen Yan. Distributionally robust linear and discrete optimization with marginals. *Operations Research*, 2022a.

Anulekha Dhara, Bikramjit Das, and Karthik Natarajan. Worst-case expected shortfall with univariate and bivariate marginals. *INFORMS Journal on Computing*, 33(1):370–389, 2021.

Xuan V Doan, Xiaobo Li, and Karthik Natarajan. Robustness to dependency in portfolio optimization using overlapping marginals. *Operations Research*, 63(6):1468–1488, 2015.

Yongchao Liu, Rudabeh Meskarian, and Huifu Xu. Distributionally robust reward-risk ratio optimization with moment constraints. *SIAM Journal on Optimization*, 27(2):957–985, 2017.

Karthik Natarajan, Dongjian Shi, and Kim-Chuan Toh. A probabilistic model for minmax regret in combinatorial optimization. *Operations Research*, 62(1):160–181, 2014.

Grani A Hanasusanto, Vladimir Roitch, Daniel Kuhn, and Wolfram Wiesemann. Ambiguous joint chance constraints under mean and dispersion information. *Operations Research*, 65(3):751–767, 2017.

Weijun Xie, Shabbir Ahmed, and Ruiwei Jiang. Optimized Bonferroni approximations of distributionally robust joint chance constraints. *Mathematical Programming*, 191:79–112, 2022.

Weijun Xie and Shabbir Ahmed. On deterministic reformulations of distributionally robust joint chance constrained optimization problems. *SIAM Journal on Optimization*, 28(2):1151–1182, 2018.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 2022.

Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. Data-driven chance constrained programs over Wasserstein balls. *Operations Research*, 2022b.

Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research*, 47(2):1500–1529, 2022.

Fengqiao Luo and Sanjay 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(1):20–35, 2019.

Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. *Advances in Neural Information Processing Systems*, 31, 2018.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171 (1):115–166, 2018.

Sanjay Mehrotra and He Zhang. Models and algorithms for distributionally robust least squares problems. *Mathematical Programming*, 146(1):123–141, 2014.

Franccois Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3):541–593, 2007.

Krzysztof Postek, Dick den Hertog, and Bertrand Melenberg. Computationally tractable counterparts of distributionally robust constraints on risk measures. *SIAM Review*, 58(4):603–650, 2016.

Henry Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

Ruiwei Jiang and Yongpei Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, 158(1):291–327, 2016.

Zizhuo Wang, Peter W Glynn, and Yinyu Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261, 2016.

Güzin Bayraksan and David K Love. Data-driven stochastic programming using phi-divergences. In *The Operations Research Revolution*, pages 1–19. INFORMS, 2015.

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

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

Hamed Rahimian, Güzin Bayraksan, and Tito Homem-de Mello. Identifying effective scenarios in distributionally robust stochastic programs with total variation distance. *Mathematical Programming*, 173(1): 393–430, 2019a.

Hamed Rahimian, Güzin Bayraksan, and Tito Homem-de Mello. Controlling risk and demand ambiguity in newsvendor models. *European Journal of Operational Research*, 279(3):854–868, 2019b.

Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4): 2258–2275, 2017.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

Emre Erdoğan and Garud Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1):37–61, 2006.

Ruiwei Jiang and Yongpei Guan. Risk-averse two-stage stochastic program with distributional ambiguity. *Operations Research*, 66(5):1390–1405, 2018.

Jianqiu Huang, Kezhuo Zhou, and Yongpei Guan. A study of distributionally robust multistage stochastic optimization. *arXiv preprint arXiv:1708.07930*, 2017.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

RalphB D'Agostino. *Goodness-of-fit-techniques*. Routledge, 2017.

Chrysogonus Chinagorom Nwaigwe, Chukwudi Justin Ogbonna, and Emmanuel Uchechukwu Oliwe. Appropriate description of probability distribution of prostrate specific antigen (psa): An aid to early detection of prostrate cancer. *Asian J. Prob. Stat*, 20(4):39–50, 2022.

Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization–2019*. PhD thesis, School of Industrial & Systems Engineering, Georgia Institute of Technology, 2019.

Ole Tange. *GNU Parallel 2018*. Ole Tange, April 2018. https://doi.org/10.5281/zenodo.1146014.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2022. https://www.gurobi.com.

A Proof of Theorems

A.1 Proof of Theorem 1

Let $\phi_{\chi^2}(t) = (t-1)^2/t$ and $\phi_{\mathtt{G}}(t) = -\log(t) + t - 1$ be ϕ-divergence of χ^2-test and G-test, and t denote $\frac{F(\hat{\xi}^j)}{\hat{F}_N(\hat{\xi}^j)}$. We illustrate the equivalence of the test-related constraints to the ones defined based on ϕ-divergences as follows.

χ^2-test based constraint:

$$\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \phi_{\chi^2}\left(\frac{F(\hat{\xi}^j)}{\hat{F}_N(\hat{\xi}^j)}\right) \leq \mathcal{Q}^2_{\chi^2}(\alpha_1) \iff \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j) \frac{\left(\frac{F(\hat{\xi}^j)}{\hat{F}_N(\hat{\xi}^j)} - 1\right)^2}{\frac{F(\hat{\xi}^j)}{\hat{F}_N(\hat{\xi}^j)}} \leq \mathcal{Q}^2_{\chi^2}(\alpha_1) \iff$$

$$\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j) \frac{\left(F(\hat{\xi}^j) - \hat{F}_N(\hat{\xi}^j)\right)^2}{F(\hat{\xi}^j) \cdot \hat{F}_N(\hat{\xi}^j)} \leq \mathcal{Q}^2_{\chi^2}(\alpha_1) \iff \left(\sum_{j=1}^{n} \frac{(F(\hat{\xi}^j) - \hat{F}_N(\hat{\xi}^j))^2}{F(\hat{\xi}^j)}\right)^{1/2} \leq \mathcal{Q}_{\chi^2}(\alpha_1) \iff X_N \leq \mathcal{Q}_{\chi^2}(\alpha_1).$$

G-test based constraint:

$$\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \phi_{\mathtt{G}}\left(\frac{F(\hat{\xi}^j)}{\hat{F}_N(\hat{\xi}^j)}\right) \leq \frac{1}{2}\ \mathcal{Q}^2_{\mathtt{G}}(\alpha_2) \iff$$

$$\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \left(-\log\left(\frac{F(\hat{\xi}^j)}{\hat{F}_N(\hat{\xi}^j)}\right) + \frac{F(\hat{\xi}^j)}{\hat{F}_N(\hat{\xi}^j)} - 1\right) \leq \frac{1}{2}\ \mathcal{Q}^2_{\mathtt{G}}(\alpha_2) \iff$$

$$\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \log\left(\frac{\hat{F}_N(\hat{\xi}^j)}{F(\hat{\xi}^j)}\right) + F(\hat{\xi}^j) - \hat{F}_N(\hat{\xi}^j) \leq \frac{1}{2}\ \mathcal{Q}^2_{\mathtt{G}}(\alpha_2) \iff$$

$$\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \log\left(\frac{\hat{F}_N(\hat{\xi}^j)}{F(\hat{\xi}^j)}\right) + \sum_{j=1}^{n} F(\hat{\xi}^j) - \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j) \leq \frac{1}{2}\ \mathcal{Q}^2_{\mathtt{G}}(\alpha_2) \iff$$

$$\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \log\left(\frac{\hat{F}_N(\hat{\xi}^j)}{F(\hat{\xi}^j)}\right) + 1 - 1 \leq \frac{1}{2}\ \mathcal{Q}^2_{\mathtt{G}}(\alpha_2) \iff$$

$$\left(2\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \log\left(\frac{\hat{F}_N(\hat{\xi}^j)}{F(\hat{\xi}^j)}\right)\right)^{1/2} \leq\ \mathcal{Q}_{\mathtt{G}}(\alpha_2) \iff G_N \leq \mathcal{Q}_{\mathtt{G}}(\alpha_2).$$

As a result, the intersection of two test can be reformulated as follows:

$$\mathcal{C}(x,\mathcal{F}) = \max_{F\in\mathcal{F}} \sum_{j=1}^{n} F(\hat{\xi}^j)\ c(x,\hat{\xi}^j) \tag{39a}$$

$$\text{s.t.} \sum_{j=1}^{n} F(\hat{\xi}^j) = 1 \tag{39b}$$

$$\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \phi_{\chi^2}\left(\frac{F(\hat{\xi}^j)}{\hat{F}_N(\hat{\xi}^j)}\right) \le \mathcal{Q}^2_{\chi^2}(\alpha_1) \tag{39c}$$

$$\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \phi_{\mathtt{G}}\left(\frac{F(\hat{\xi}^j)}{\hat{F}_N(\hat{\xi}^j)}\right) \le \frac{1}{2}\ \mathcal{Q}^2_{\mathtt{G}}(\alpha_2) \tag{39d}$$

$$F(\hat{\xi}^j) \in \mathbb{R}_+ \qquad j\in[n] \tag{39e}$$

Using Fenchel duality $\mathcal{C}(x,\mathcal{F})$ is equal to:

$$\min_{\substack{r\in\mathbb{R},\\ s,s'\in\mathbb{R}_+}} \max_{F\in\mathbb{R}_+} \sum_{j=1}^{n} F(\hat{\xi}^j)\ c(x,\hat{\xi}^j) + r\left(1-\sum_{j=1}^{n} F(\hat{\xi}^j)\right) + s\left(\mathcal{Q}^2_{\chi^2}(\alpha_1) - \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \phi_{\chi^2}\left(\frac{F(\hat{\xi}^j)}{\hat{F}_N(\hat{\xi}^j)}\right)\right) \tag{40a}$$

$$+ s'\left(\frac{1}{2}\mathcal{Q}^2_{\mathtt{G}}(\alpha_2) - \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \phi_{\mathtt{G}}\left(\frac{F(\hat{\xi}^j)}{\hat{F}_N(\hat{\xi}^j)}\right)\right) \tag{40b}$$

Let $\rho_j = \rho'_j = \frac{F(\hat{\xi}^j)}{\hat{F}_N(\hat{\xi}^j)}$, we get:

$$\min_{r\in\mathbb{R},s,s'\in\mathbb{R}_+} \max_{\rho,\rho'\in\mathbb{R}^n_+} r + \mathcal{Q}^2_{\chi^2}(\alpha_1)\ s + \frac{1}{2}\mathcal{Q}^2_{\mathtt{G}}(\alpha_2)\ s' + \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ c(x,\hat{\xi}^j)\rho_j - r(\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \rho'_j)$$

$$- s(\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \phi_{\chi^2}(\rho_j)) - s'(\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \phi_{\mathtt{G}}(\rho'_j)) \tag{41a}$$

$$\text{s.t.}\ \ \rho = \rho' \tag{41b}$$

In order to make the reformulations easier, we multiple both sides of constraint (41b) with the empirical distribution, $\hat{F}_N$, then, we take the dual of the constraint by defining a new dual variable γ.

$$= \min_{r\in\mathbb{R},s,s'\in\mathbb{R}_+} \min_{\gamma\in\mathbb{R}^n} \max_{\rho,\rho'\in\mathbb{R}^n_+} r + \mathcal{Q}^2_{\chi^2}(\alpha_1)\ s + \frac{1}{2}\mathcal{Q}^2_{\mathtt{G}}(\alpha_2)\ s' + \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ c(x,\hat{\xi}^j)\rho_j - r(\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \rho'_j)$$

$$- s(\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \phi_{\chi^2}(\rho_j)) - s'(\sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\ \phi_{\mathtt{G}}(\rho'_j))\ + \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)\gamma_j(\rho'_j - \rho_j) \tag{42a}$$

$$= \min_{r\in\mathbb{R},s,s'\in\mathbb{R}_+,\gamma\in\mathbb{R}^n} \; r + \mathcal{Q}^2_{\chi^2}(\alpha_1)\; s + \; \frac{1}{2}\mathcal{Q}^2_{\mathtt{G}}(\alpha_2)\; s' + \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j) s \; \max_{\rho\in\mathbb{R}^n_+} \left(\frac{c(x,\hat{\xi}^j) - \gamma_j}{s} \rho_j - \; \phi_{\chi^2}(\rho_j) \right)$$
$$+ \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j) s' \; \max_{\rho'\in\mathbb{R}^n_+} \left(\frac{\gamma_j - r}{s'} \rho'_j - \; \phi_{\mathtt{G}}(\rho'_j) \right) \tag{43a}$$

$$= \min_{r\in\mathbb{R},s,s'\in\mathbb{R}_+,\gamma\in\mathbb{R}^n} \; r + \mathcal{Q}^2_{\chi^2}(\alpha_1)\; s + \; \frac{1}{2}\mathcal{Q}^2_{\mathtt{G}}(\alpha_2)\; s' + \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j) s \; \phi^*_{\chi^2} \left(\frac{c(x,\hat{\xi}^j) - \gamma_j}{s} \right)$$
$$+ \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j) s' \; \phi^*_{\mathtt{G}} \left(\frac{\gamma_j - r}{s'} \right) \tag{44a}$$

We can define decision variables, ℓ_j, to represent the value of $c(x,\hat{\xi}^j)$ and add constraints of the form $\ell_j \geq c(x,\hat{\xi}^j)$ to the problem due to the sense of the problem which is minimization.

$$= \min_{r,s,s',t,t',\ell,\gamma} \; r + \mathcal{Q}^2_{\chi^2}(\alpha_1)\; s + \; \frac{1}{2}\mathcal{Q}^2_{\mathtt{G}}(\alpha_2)\; s' - \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)(t_j + t'_j) \tag{45a}$$

$$\text{s.t. } t_j \leq -s \; \phi^*_{\chi^2} \left(\frac{\ell_j - \gamma_j}{s} \right) \qquad j \in [n] \tag{45b}$$
$$t'_j \leq -s' \phi^*_{\mathtt{G}} \left(\frac{\gamma_j - r}{s'} \right) \qquad j \in [n] \tag{45c}$$
$$\ell_j \geq c(x,\hat{\xi}^j) \qquad j \in [n] \tag{45d}$$
$$r \in \mathbb{R}, s, s' \in \mathbb{R}_+, \gamma, t, t', \ell \in \mathbb{R}^n \tag{45e}$$

Using the formulation of convex conjugates, one can further simplify the model as follows:

$$\phi^*_{\chi^2}(\tau) = \begin{cases} 2 - 2\sqrt{1-\tau} & \tau \leq 1 \\ \infty & \text{otherwise} \end{cases}, \quad \text{and} \quad \phi^*_{\mathtt{G}}(\tau) = \begin{cases} -\log(1-\tau) & \tau \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

$$= \min_{r,s,s',t,t',\ell,\gamma} \; r + \mathcal{Q}^2_{\chi^2}(\alpha_1)\; s + \; \frac{1}{2}\mathcal{Q}^2_{\mathtt{G}}(\alpha_2)\; s' - \sum_{j=1}^{n} \hat{F}_N(\hat{\xi}^j)(t_j + t'_j) \tag{46a}$$

$$\text{s.t. } \ell_j - \gamma_j \leq s \qquad j \in [n] \tag{46b}$$
$$2s + t_j \leq y_j \qquad j \in [n] \tag{46c}$$
$$y_j^2 + (\gamma_j - \ell_j)^2 \leq (2s - \ell_j + \gamma_j)^2 \qquad j \in [n] \tag{46d}$$
$$\gamma_j - r \leq s' \tag{46e}$$
$$s'(e^{t'_j/s'}) \leq s' - \gamma_j + r \qquad j \in [n] \tag{46f}$$
$$\ell_j \geq c(x,\hat{\xi}^j) \qquad j \in [n] \tag{46g}$$
$$r \in \mathbb{R}, s, s' \in \mathbb{R}_+, \gamma, t, t', \ell, y \in \mathbb{R}^n \tag{46h}$$

A.2 Proof of Theorem 2

In this section, for the sake of completeness, we provide the related theorems from the literature that we use in the reformulation of the problems with univariate support.

Theorem 11 of [Bertsimas et al., 2018]:
Under the assumptions of Theorem 1 in [Bertsimas et al., 2018], problem (6) is equivalent to the following problem:

$$\mathcal{C}(x, \mathcal{F}(\alpha)) = \min_{r,c} \; b_{\mathtt{T}}^{\top} r + c_{N+1} \tag{47a}$$

$$\text{s.t. } c \in \mathbb{R}^{N+1}, \; -r \in K_{\mathtt{T}}^{*} \tag{47b}$$

$$\left(A_{\mathtt{T}}^{\top} r\right)_i = c_i - c_{i+1} \qquad i \in [N] \tag{47c}$$

$$c_i \geq \sup_{\xi \in (\xi^{i-1}, \xi^{i}]} c(x, \xi) \qquad i \in [N+1] \tag{47d}$$

where $A_{\mathtt{T}}$, $b_{\mathtt{T}}$, and $K_{\mathtt{T}}^{*}$ are respectively matrix, vector, and dual cone associated with a specific test $\mathtt{T}$.

Section 1.4.5.1 of [Ben-Tal and Nemirovski, 2019]:
Given a primal problem (P):

$$(P) := \min_{x} \; c^{\top} x \tag{48a}$$

$$\text{s.t. } A_1 x - b_1 \geq 0 \tag{48b}$$

$$A_i x - b_i \in K_i \qquad 2 \leq i \leq m \tag{48c}$$

$$Rx = r \tag{48d}$$

where A_i, b_i for $i \in [m]$, R, r, and c are matrices and vectors of appropriate dimension, and K_i for $2 \leq i \leq m$ are regular cones in Euclidean spaces. As a result of conic duality, the dual of problem (P) is as follows:

$$D := \max_{z,y} \; r^{\top} z + b_1^{\top} y_1 + \sum_{i=2}^{m} \langle b_i, y_i \rangle \tag{49a}$$

$$\text{s.t. } y_1 \geq 0, \; y_i \in K_i^{*}, \; 2 \leq i \leq m \tag{49b}$$

$$R^{\top} z + \sum_{i=1}^{m} A_i^{\top} y_i = c \tag{49c}$$

where K_i^{*} for $2 \leq i \leq m$ are the dual cones.

A.3 Proof of Theorem 3

Given a sample of size N, $\{\xi^1, \ldots, \xi^N\}$, let denote d the dimension of the random vector ξ, and $S = \{(a, b) \in \mathbb{R}^d \times \mathbb{R} \mid |a_1| + \ldots, |a_d| + |b| \leq 1\}$. The supremum problem for problems with multivariate support can be

modeled as follows [Bertsimas et al., 2018]:

$$\mathcal{C}(x,\mathcal{F}) = \sup_{F,\Pi} \int_{\Xi} c(x,\xi)F(d\xi) \tag{50a}$$

$$\text{s.t.} \int_{\Xi} \max\{a^\top\xi - b, 0\}F(d\xi) \leq \int_{\Xi} \max\{a^\top\xi' - b, 0\}\hat{F}_N(d\xi') + \mathcal{Q}_{\mathtt{C}_N}(\alpha_2) \quad \forall (a,b) \in S \tag{50b}$$

$$\int_{\Xi} ||\xi||_2^2 F(d\xi) \geq \mathcal{Q}_{\mathtt{R}_N}(\alpha_1) \tag{50c}$$

$$\int_{\Xi^2} ||\xi - \xi'||\Pi(d\xi, d\xi') \leq \epsilon_N^\beta \tag{50d}$$

where Π is a joint distribution of ξ and ξ' with marginal distributions F and $\hat{F}_N$, respectively. Let $\otimes$ represent an operator which multiplies two probability distributions. Based on the law of total probability we have $\Pi = \hat{F}_N \otimes F$ where F_i are the conditional distributions of ξ when $\xi' = \xi^i$ for $i \in [N]$. Since we create the empirical distribution using a uniform distribution, we have $\Pi = \frac{1}{N}\sum_{i=1}^N \delta_{\xi^i} \otimes F_i$ where δ is the Dirac distribution and $\delta_{\xi^i} = 1$ if $\xi' = \xi^i$ and 0 otherwise [Esfahani and Kuhn, 2018].

$$\mathcal{C}(x,\mathcal{F}) = \sup_{F_i \in \mathcal{P}'(\Xi)} \frac{1}{N}\sum_{i=1}^N \int_{\Xi} c(x,\xi)F_i(d\xi) \tag{51a}$$

$$\text{s.t.} \ \frac{1}{N}\sum_{i=1}^N \int_{\Xi} \max\{a^\top\xi - b, 0\}F_i(d\xi) \leq \frac{1}{N}\sum_{i=1}^N \max\{a^\top\xi^i - b, 0\} + \mathcal{Q}_{\mathtt{C}_N}(\alpha_2) \quad \forall (a,b) \in S \tag{51b}$$

$$\frac{1}{N}\sum_{i=1}^N \int_{\Xi} ||\xi||_2^2 F_i(d\xi) \geq \mathcal{Q}_{\mathtt{R}_N}(\alpha_1) \tag{51c}$$

$$\frac{1}{N}\sum_{i=1}^N \int_{\Xi} ||\xi - \xi^i||F_i(d\xi) \leq \epsilon_N^\beta \tag{51d}$$

Let S' be the set of non-negative measures on S. Using Fenchel duality, taking the dual of the above model results in the following model:

$$\begin{aligned}\inf_{\lambda \in S', \theta, \eta \in \mathbb{R}_+} \sup_{F_i \in \mathcal{P}'(\Xi)} & \frac{1}{N}\sum_{i=1}^N \int_{\Xi} c(x,\xi)F_i(d\xi) \\ & + \Big\langle \lambda, \frac{1}{N}\sum_{i=1}^N \max\{a^\top\xi^i - b, 0\} + \mathcal{Q}_{\mathtt{C}_N}(\alpha_2) - \frac{1}{N}\sum_{i=1}^N \int_{\Xi} \max\{a^\top\xi - b, 0\}F_i(d\xi)\Big\rangle_S \\ & + \theta\Big(\frac{1}{N}\sum_{i=1}^N \int_{\Xi} ||\xi||_2^2 F_i(d\xi) - \mathcal{Q}_{\mathtt{R}_N}(\alpha_1)\Big) + \eta\Big(\epsilon_N^\beta - \frac{1}{N}\sum_{i=1}^N \int_{\Xi} ||\xi - \xi^i||F_i(d\xi)\Big)\end{aligned} \tag{52a}$$

$$\leq \inf_{\lambda\in S',\theta,\eta\in\mathbb{R}_+} \sup_{F_i\in\mathcal{P}'(\Xi)} \Big\langle \lambda, \frac{1}{N}\sum_{i=1}^{N} \max\{a^\top \xi^i - b, 0\} + \mathfrak{Q}_{\mathfrak{C}_N}(\alpha_2)\Big\rangle_S - \theta\mathfrak{Q}_{\mathfrak{R}_N}(\alpha_1) + \eta\epsilon_N^\beta +$$
$$\frac{1}{N}\sum_{i=1}^{N}\int_{\Xi}\Big(c(x,\xi) - \Big\langle \lambda, \max\{a^\top\xi - b, 0\}\Big\rangle_S + \theta||\xi||_2^2 - \eta||\xi - \xi^i||\Big)F_i(d\xi) \tag{53a}$$

since $\mathcal{P}'(\Xi)$ contains all the Dirac distributions supported on Ξ, the above problem is equivalent to the following one [Esfahani and Kuhn, 2018]:

$$\inf_{\lambda\in S',\theta,\eta\in\mathbb{R}_+} \Big\langle \lambda, \frac{1}{N}\sum_{i=1}^{N} \max\{a^\top \xi^i - b, 0\} + \mathfrak{Q}_{\mathfrak{C}_N}(\alpha_2)\Big\rangle_S - \theta\mathfrak{Q}_{\mathfrak{R}_N}(\alpha_1) + \eta\epsilon_N^\beta +$$
$$\frac{1}{N}\sum_{i=1}^{N}\sup_{\xi\in\Xi}\Big(c(x,\xi) - \Big\langle \lambda, \max\{a^\top\xi - b, 0\}\Big\rangle_S + \theta||\xi||_2^2 - \eta||\xi - \xi^i||\Big) \tag{54a}$$

We define epigraphical auxiliary variables s_i for $i \in [N]$ and reformulate the above problem as

$$= \inf_{\lambda\in S',\theta,\eta\in\mathbb{R}_+,s\in\mathbb{R}^N} \Big\langle \lambda, \frac{1}{N}\sum_{i=1}^{N} \max\{a^\top \xi^i - b, 0\} + \mathfrak{Q}_{\mathfrak{C}_N}(\alpha_2)\Big\rangle_S - \theta\mathfrak{Q}_{\mathfrak{R}_N}(\alpha_1) + \eta\epsilon_N^\beta + \frac{1}{N}\sum_{i=1}^{N} s_i \tag{55a}$$
$$\text{s.t. } s_i \geq \sup_{\xi\in\Xi}\Big(c(x,\xi) - \Big\langle \lambda, \max\{a^\top\xi - b, 0\}\Big\rangle_S + \theta||\xi||_2^2 - \eta||\xi - \xi^i||\Big) \quad i \in [N] \tag{55b}$$

Having a closer look on constraints (55b), one can obtain the following relation between the elements of the model:

$$\theta \leq \inf_{\xi\in\Xi} \frac{s_i + \Big\langle \lambda, \max\{a^\top\xi - b, 0\}\Big\rangle_S - c(x,\xi) + \eta||\xi - \xi^i||}{||\xi||_2^2}$$

With the same reasoning and assumptions provided in [Bertsimas et al., 2018], the only feasible solution for θ is zero. Using this information, the definition of cost function, and dual norm ($||x|| = \max_{||z||_*\leq 1} z^\top x$), we can reformulate constraints (55b) as follows:

$$s_i \geq \sup_{\xi\in\Xi}\Big(c_k(x,\xi) - \Big\langle \lambda, \max\{a^\top\xi - b, 0\}\Big\rangle_S - \max_{||z_{ik}||_*\leq\eta}\langle z_{ik}, \xi - \xi^i\rangle\Big) \quad k \in [K], i \in [N] \tag{56a}$$

We convert the maximization over z variable to a minimization using $-\max_z\langle z_{ik}, \xi - \xi^i\rangle = \min_z -\langle z_{ik}, \xi - \xi^i\rangle$. As a result, the minimization along with the supremum over ξ restricts the area for problem (56) and yields an upper bound on it.

$$\mathcal{C}(x,\mathcal{F}) \leq \inf_{\lambda\in S',\eta\in\mathbb{R}_+,s\in\mathbb{R}^N} \Big\langle \lambda, \frac{1}{N}\sum_{i=1}^{N} \max\{a^\top \xi^i - b, 0\} + \mathfrak{Q}_{\mathfrak{C}_N}(\alpha_2)\Big\rangle_S + \eta\epsilon_N^\beta + \frac{1}{N}\sum_{i=1}^{N} s_i \tag{57a}$$
$$\text{s.t. } s_i \geq \min_{||z_{ik}||_*\leq\eta}\sup_{\xi\in\Xi}\Big(c_k(x,\xi) - \Big\langle \lambda, \max\{a^\top\xi - b, 0\}\Big\rangle_S - \langle z_{ik}, \xi - \xi^i\rangle\Big)$$
$$k \in [K], i \in [N] \tag{57b}$$

Eliminating minimum operator from constraint (57b) and adding its constraint results in the following model:

$$\inf_{\lambda\in S',\eta\in\mathbb{R}_+,s\in\mathbb{R}^N} \Big\langle \lambda, \frac{1}{N}\sum_{i=1}^{N}\max\{a^\top\xi^i - b, 0\} + \mathfrak{Q}_{\mathfrak{c}_N}(\alpha_2)\Big\rangle_S + \eta\epsilon_N^\beta + \frac{1}{N}\sum_{i=1}^{N} s_i \tag{58a}$$

$$\text{s.t. } s_i \geq \sup_{\xi\in\Xi}\Big(c_k(x,\xi) - \Big\langle\lambda, \max\{a^\top\xi - b, 0\}\Big\rangle_S - \langle z_{ik}, \xi - \xi^i\rangle\Big) \quad k\in[K], i\in[N] \tag{58b}$$

$$||z_{ik}||_* \leq \eta \qquad k\in[K], i\in[N] \tag{58c}$$

For a given i and k, we can reformulate constraints (58b) by defining a new non-negative decision variable g.

$$-s_i \leq \inf_{\xi\in\Xi, g\in\mathbb{R}_+}\Big(-c_k(x,\xi) + \langle\lambda, g\rangle_S + \langle z_{ik}, \xi - \xi^i\rangle\Big) \tag{59a}$$

$$g \geq a^\top\xi - b \qquad (a,b)\in S \tag{59b}$$

Taking the dual of the above constraints results in the following constraint:

$$-s_i \leq \inf_{\xi\in\Xi, g\in\mathbb{R}_+}\sup_{H_{ik}\in S'}\Big(-c_k(x,\xi) + \langle\lambda, g\rangle_S + \langle z_{ik}, \xi - \xi^i\rangle + \Big\langle H_{ik}, a^\top\xi - b - g\Big\rangle_S\Big) \tag{60a}$$

By rearranging the terms and using the definition of convex conjugate we have:

$$-s_i \leq \sup_{H_{ik}\in S'}\Big(c_{k*}(x, z_{ik} + \langle H_{ik}, a\rangle_S) - \langle z_{ik}, \xi^i\rangle - \langle H_{ik}, b\rangle_S\Big) \tag{61a}$$

$$\text{s.t. } \lambda - H_{ik} \in S' \tag{61b}$$

Following the same steps in [Bertsimas et al., 2018], we define a new variable r_{ik}.

$$-s_i \leq \sup_{H_{ik}\in S', r_{ik}\in\mathbb{R}}\Big(c_{k*}(x, r_{ik}) - \langle z_{ik}, \xi^i\rangle - \langle H_{ik}, b\rangle_S\Big) \tag{62a}$$

$$\text{s.t. } \lambda - H_{ik} \in S' \tag{62b}$$

$$r_{ik} = z_{ik} + \langle H_{ik}, a\rangle_S \tag{62c}$$

We can now eliminate the supremum operator and add the corresponding constraints:

$$-s_i \leq \Big(c_{k*}(x, r_{ik}) - \langle z_{ik}, \xi^i\rangle - \langle H_{ik}, b\rangle_S\Big) \tag{63a}$$

$$\lambda - H_{ik} \in S' \tag{63b}$$

$$r_{ik} = z_{ik} + \langle H_{ik}, a\rangle_S \tag{63c}$$

Recall the reformulation (58), after following the above-mentioned steps, it is equivalent to the following problem:

$$\inf_{\lambda,\eta,s,H,r,z} \Big\langle \lambda, \frac{1}{N}\sum_{i=1}^{N}\max\{a^\top\xi^i - b, 0\} + \mathfrak{Q}_{\mathfrak{c}_N}(\alpha_2)\Big\rangle_S + \eta\epsilon_N^\beta + \frac{1}{N}\sum_{i=1}^{N} s_i \tag{64a}$$

$$\text{s.t. } \lambda\in S', \eta\in\mathbb{R}_+, s\in\mathbb{R}^N, H_{ik}\in S', r, z\in\mathbb{R}^{N\times K} \tag{64b}$$

$$-s_i \leq \Big(c_{k*}(x, r_{ik}) - \langle z_{ik}, \xi^i\rangle - \langle H_{ik}, b\rangle_S\Big) \qquad k\in[K], i\in[N] \tag{64c}$$

$$\lambda - H_{ik} \in S' \qquad k \in [K], i \in [N] \tag{64d}$$
$$r_{ik} = z_{ik} + \langle H_{ik}, a \rangle_S \qquad k \in [K], i \in [N] \tag{64e}$$
$$||z_{ik}||_* \leq \eta \qquad k \in [K], i \in [N] \tag{64f}$$

We write the above problem as $\inf_{\eta,z} \inf_{\lambda,s,H,r}(\cdot)$ and take the dual of inner infimum problem using the dual variables p_{ik}, $\psi_{ik}(a,b)$, and q_{ik} associated with constraints (64c),(64d), and (64e), respectively. Notice that in [Bertsimas et al., 2018] variables had only index k but now they have i and k, which means that the size of the problem is multiplied by sample size N. Following the same ideas provided in [Bertsimas et al., 2018], taking the dual of the above infimum problem results in the following supremum problem:

$$\sup_{\substack{q \in \mathbb{R}^{K \times N \times d}, \\ p \in \mathbb{R}_+^{K \times N}}} \eta \epsilon_N^\beta + \sum_{i=1}^N \sum_{k=1}^K p_{ik} \langle z_{ik}, \xi^i \rangle - \langle z_{ik}, q_{ik} \rangle + p_{ik} c\left(x, \frac{q_{ik}}{p_{ik}}\right) \tag{65a}$$
$$\text{s.t.} \sum_{k=1}^K p_{ik} = 1 \qquad i \in [N] \tag{65b}$$
$$\inf_{(a,b) \in S} \frac{1}{N} \sum_{i=1}^N \max\{a^\top \xi^i - b, 0\} + \mathcal{Q}_{\mathfrak{c}_N}(\alpha_2) - \sum_{i=1}^N \sum_{k=1}^K \max\{a^\top q_{ik} - p_{ik} b, 0\} \geq 0 \tag{65c}$$

We define parameter $\gamma \in \mathcal{G}$ where $\mathcal{G} = \{0,1\}^{K \times N} \backslash \{(0, \ldots, 0)\}$ in order to consider all possibilities of $\max\{a^\top q_{ik} - p_{ik} b, 0\}$. The constraint (65c) can be reformulated as follows:

$$\mathcal{Q}_{C_N}(\alpha_2) \geq \sup_{a \in \mathbb{R}^d, b \in \mathbb{R}} \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik}(a^\top q_{ik} - p_{ik} b) - \frac{1}{N} \sum_{i=1}^N \max\{a^\top \xi^i - b, 0\} \quad \gamma \in \mathcal{G} \tag{66a}$$
$$\text{s.t.} \sum_{j=1}^d |a_j| + |b| \leq 1 \tag{66b}$$

We linearize the supremum problem in the constraint and take its linear dual. As a result, the constraint is equivalent to the following set of constraints:

$$\mathcal{Q}_{\mathfrak{c}_N}(\alpha_2) \geq \rho_\gamma \qquad \gamma \in \mathcal{G} \tag{67a}$$
$$\frac{1}{N} \sum_{i=1}^N \mu_{\gamma,i} \xi^i + u_\gamma - v_\gamma = \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik} q_{ik} \qquad \gamma \in \mathcal{G} \tag{67b}$$
$$-\frac{1}{N} \sum_{i=1}^N \mu_{\gamma,i} + u'_\gamma - v'_\gamma = -\sum_{i=1}^N \sum_{k=1}^K \gamma_{ik} p_{ik} \qquad \gamma \in \mathcal{G} \tag{67c}$$
$$\mu_{\gamma,i} \leq 1 \qquad \gamma \in \mathcal{G}, i \in [N] \tag{67d}$$
$$-u_\gamma - v_\gamma + \rho_\gamma \geq 0 \qquad \gamma \in \mathcal{G} \tag{67e}$$
$$-u'_\gamma - v'_\gamma + \rho_\gamma \geq 0 \qquad \gamma \in \mathcal{G} \tag{67f}$$
$$\mu \in \mathbb{R}_+^{|\mathcal{G}| \times N}, \rho, u', v' \in \mathbb{R}_+^{|\mathcal{G}|}, u, v \in \mathbb{R}_+^{|\mathcal{G}| \times d} \tag{67g}$$

We can then replace constrain (65c) in (65) with constraints (67). After obtaining the supremum problem, we take its (LP) dual and obtain an infimum problem. This infimum can be merged with $\inf_{\eta,z}(\cdot)$ which we

mentioned after problem (64). The resulting infimum problem below can be solved by the state-of-the-art solvers.

$$\inf_{r,f,w,w',y,y',e,\eta,\tau} \quad \eta\epsilon_N^\beta + \frac{1}{N}\sum_{i=1}^N r_i + \sum_{\gamma\in\mathcal{G}} f_\gamma \ \mathcal{Q}_{\mathtt{C}_N}(\alpha_2) + \sum_{\gamma\in\mathcal{G}}\sum_{i=1}^N e_{\gamma i} \tag{68a}$$

$$\text{s.t.} \ \sum_{j=1}^d y_{\gamma j} + y'_\gamma \le f_\gamma \qquad \gamma\in\mathcal{G} \tag{68b}$$

$$\frac{1}{N}(\sum_{j=1}^d w_{\gamma j}\ \xi_j^i - w'_\gamma) \le e_{\gamma i} \qquad \gamma\in\mathcal{G},\ i\in[N] \tag{68c}$$

$$w_{\gamma j} - y_{\gamma j} \le 0 \qquad \gamma\in\mathcal{G},\ j\in[d] \tag{68d}$$

$$-w_{\gamma j} - y_{\gamma j} \le 0 \qquad \gamma\in\mathcal{G},\ j\in[d] \tag{68e}$$

$$w'_\gamma - y'_\gamma \le 0 \qquad \gamma\in\mathcal{G} \tag{68f}$$

$$-w'_\gamma - y'_\gamma \le 0 \qquad \gamma\in\mathcal{G} \tag{68g}$$

$$\langle z_{ik}, \xi^i\rangle - r_i + \sum_{\gamma\in\mathcal{G}} \gamma_{ik}\ w'_\gamma - g_k \le 0 \qquad k\in[K],\ i\in[N] \tag{68h}$$

$$g_k \le c_{k*}(x, \sum_{\gamma\in\mathcal{G}} \gamma_{ik} w_\gamma + z_{ik}) \qquad k\in[K],\ i\in[N] \tag{68i}$$

$$\|z_{ik}\|_* \le \eta \qquad k\in[K],\ i\in[N] \tag{68j}$$

$$\eta\in\mathbb{R}_+,\ f\in\mathbb{R}_+^{|\mathcal{G}|},\ e\in\mathbb{R}_+^{|\mathcal{G}|\times N}, \tag{68k}$$
$$y\in\mathbb{R}_+^{|\mathcal{G}|\times d},\ y'\in\mathbb{R}_+^{|\mathcal{G}|}, z\in\mathbb{R}^{K\times N\times d},$$
$$r\in\mathbb{R}^N, w\in\mathbb{R}^{|\mathcal{G}|\times d},\ w'\in\mathbb{R}^{|\mathcal{G}|}$$

This infimum problem can be merged with $\inf_{x\in X}(\cdot)$ and yield a single-level problem. This reformulation can be further simplified, and z variables can be omitted from the reformulation. The reformulation steps and final model are similar to the ones proposed by Bertsimas et al. [2018][1], however, our model has additional decision variables in order to incorporate information of the Wasserstein ball.

B Bootstrap Algorithm

In this section, we explain a bootstrap procedure that we follow in order to calculate the critical value of various GoF tests. This procedure is proposed by Bertsimas et al. [2018] and for the sake of completeness, we restate it here.

The procedure for the LCX test is as follows. Given an IID sample $\{\xi^1,\dots,\xi^N\}$, significance level α, and iteration count B, we first obtain the empirical distribution $\hat{F}_N$ from the given sample. Next, we start iterations $t\in[B]$, and in each of them we draw a sample $\{\xi^{t,1},\dots,\xi^{t,N}\}$ from the empirical distribution.

[1]We note that in Theorem 14 of Bertsimas et al. [2018], we encountered typos where w and w' are defined as non-negative variables; however, they should be free decision variables.

Using the sample in iteration t, we calculate

$$\mathcal{Q}^t_{\mathtt{C}_N} = \sup_{|a_1|+\ldots,|a_d|+|b|\leq 1} \left(\frac{1}{N}\sum_{i=1}^{N} \max\left\{a^\top \xi^i - b, 0\right\} - \frac{1}{N}\sum_{i=1}^{N} \max\left\{a^\top \xi^{t,i} - b, 0\right\} \right).$$

After calculating $\mathcal{Q}^t_{\mathtt{C}_N}$ for all iterations, we sort them in ascending order and record the $1-\alpha$ percentile of them. The recorded value is used as $\mathcal{Q}_{\mathtt{C}_N}(\alpha)$ in our experiments.

In order to calculate the thresholds for the KS, Kuiper, and CvM tests, we follow a similar procedure. Let $\hat{F}^i_N$ represent the empirical cumulative distribution of the random vector i. In this case, instead of having an IID sample and drawing samples from the empirical distribution, in each iteration t, we obtain a sample of $\hat{F}^i_N$'s from a uniform distribution on the support $[0,1]$. Next, we use the formulations

$$\mathcal{Q}^t_{\mathtt{KS}_N} = \max_{i\in[N]} \left\{ \max\left\{ \frac{i}{N} - \hat{F}^i_N, \hat{F}^i_N - \frac{i-1}{N} \right\} \right\},$$

$$\mathcal{Q}^t_{\mathtt{Kuiper}_N} = \max_{i\in[N]} \left(\hat{F}^i_N - \frac{i-1}{N} \right) + \max_{i\in[N]} \left(\frac{i}{N} - \hat{F}^i_N \right),$$

$$\mathcal{Q}^t_{\mathtt{CvM}_N} = \left(\frac{1}{12N^2} + \frac{1}{N}\sum_{i=1}^{N} \left(\frac{2i-1}{2N} - \hat{F}^i_N \right)^2 \right)^{1/2},$$

to calculate the values of $\mathcal{Q}^t_{\mathtt{T}_N}$ for test $\mathtt{T}$. After obtaining all values for all iterations $t \in [B]$, similar to the previous case, we sort them in ascending order and utilize the $1-\alpha$ percentiles of the calculated values as the thresholds of the corresponding tests.