A General Wasserstein Framework for Data-driven Distributionally Robust Optimization: Tractability and Applications

Jonathan Yu-Meng Li[†], Tiantian Mao[††]

[†] Telfer School of Management
University of Ottawa, Ottawa, ON, Canada

[††] Department of Statistics and Finance, School of Management
University of Science and Technology of China
Hefei, Anhui, China

July 19, 2022

Abstract

Data-driven distributionally robust optimization is a recently emerging paradigm aimed at finding a solution that is driven by sample data but is protected against sampling errors. An increasingly popular approach, known as Wasserstein distributionally robust optimization (DRO), achieves this by applying the Wasserstein metric to construct a ball centred at the empirical distribution and finding a solution that performs well against the most adversarial distribution from the ball. In this paper, we present a general framework for studying different choices of a Wasserstein metric and point out the limitation of the existing choices. In particular, while choosing a Wasserstein metric of a higher order is desirable from a data-driven perspective, given its less conservative nature, such a choice comes with a high price from a robustness perspective - it is no longer applicable to many heavy-tailed distributions of practical concern. We show that this seemingly inevitable trade-off can be resolved by our framework, where a new class of Wasserstein metrics, called coherent Wasserstein metrics, is introduced. Like Wasserstein DRO, distributionally robust optimization using the coherent Wasserstein metrics, termed generalized Wasserstein distributionally robust optimization (GW-DRO), has all the desirable performance guarantees: finite-sample guarantee, asymptotic consistency, and computational tractability. The worst-case expectation problem in GW-DRO is in general a nonconvex optimization problem, yet we provide new analysis to prove its tractability without relying on the common duality scheme. Our framework, as shown in this paper, offers a fruitful opportunity to design novel Wasserstein DRO models that can be applied in various contexts such as operations management, finance, and machine learning.

1 Introduction

Data-driven problems arise from many operations research and machine learning applications where a stochastic optimization problem needs to be solved using sample data drawn from a probability distribution of interest. The goal is to find a solution that performs well in out-of-sample tests against the distribution underlying the stochastic optimization problem. These problems are challenging to solve, because firstly the use of sample data to represent a distribution is prone to sampling errors, and secondly the underlying data-generating distribution in most real-life applications is fundamentally unknown. The increasing availability of large data sets in recent years has renewed the interest of exploring how to best exploit sample data to obtain solutions with favourable out-of-sample performances. One prominent idea is to find a solution that can perform well against distributions that are, in some sense, close to the empirical distribution constructed from the sample data. This, in principle, allows the solution to maximally leverage the information contained in sample data regarding the underlying data-generating distribution while at the same time ensuring the solution does not overly rely on sample data, i.e. avoids overfitting.

An increasingly popular framework to implement this idea is data-driven distributionally robust optimization (DD-DRO). It seeks a solution that performs the best with respect to the most adversarial distribution from a set of distributions, known as ambiguity set, that are close to the empirical distribution according to some predefined metric. The choice of a proper metric is crucial in DD-DRO. Ambiguity sets constructed from different metrics could contain distributions with distinctly different structural properties. A well-known example is ambiguity sets defined based on the Kullback-Leibler divergence metrics (Kullback and Leibler (1951)), which contain only discrete distributions whose support is limited to, i.e. a subset of, the support of the empirical distribution (Ben-Tal et al. (2013), Hu and Hong (2013), Shapiro (2017)). Solutions of DD-DRO that adopts such ambiguity sets may not generalize well to situations where the underlying distribution has a more general support structure, e.g. taking values other than the observed samples. Another metric that has now been more widely applied in DD-DRO is the Wasserstein metric (Kantorovich and Rubinshtein (1958)). Ambiguity sets defined based on the metric contain distributions with a more general distributional structure, including both discrete and continuous distributions. DD-DRO adopting such ambiguity sets, also known as Wasserstein distributionally robust optimization (Esfahani and Kuhn (2018)), is attractive in that its solution has potential to generalize well against various forms of distributions that may arise from practical applications. The Wasserstein metric consists of a family of metrics in different orders, namely the type-p Wasserstein metric, $p \in [1, \infty]$, each defined based on a transportation cost function with a different power order. The type-1 Wasserstein metric, extensively studied in Esfahani and Kuhn (2018), is so far the most popular choice in the applications of Wasserstein DRO. Kuhn et al. (2019) and Gao et al. (2022) provide a comprehensive study of Wasserstein DRO for the type-p Wasserstein metric of a higher order, $p > 1$. Several other works (Blanchet et al. (2019), Gao et al. (2020), Shafieezadeh-Abadeh et al. (2019), Sinha et al. (2020), Carlsson et al. (2018)) have explored its applications in machine learning and related data-driven problems. It is known that an ambiguity set defined based on the type-p Wasserstein metric, called type-p Wasserstein ball, would contain only distributions that have finite p^{th}-order moments (Villani (2008) p.95). With the same radius, a type-p Wasserstein ball is strictly smaller than a type-q Wasserstein ball, where $p > q$, and contains distributions that concentrate more heavily around the sample data. A Wasserstein ball of a higher order is thus less conservative or can be considered more data-driven. The type-∞ Wasserstein ball, in particular, is the most data-driven in that the distributions from the ball concentrate fully in a bounded neighbourhood of sample data.

One can see that when it comes to the choice of a Wasserstein metric, a metric with a lower order would appeal to those who seek an ambiguity set that offers protection against a more adversarial form of distributions, whereas a metric with a higher order would appeal to those who seek an ambiguity set that better exploits the sample data. In many real-life applications, however, a pursuit of both, i.e. exploiting well the information from the sample data yet without dismissing the possibility that the underlying distribution may take an extreme form, is of necessity. For instance, in the context of financial portfolio management, a portfolio needs to be optimized using much of the information from market data but without dismissing the possibility that the return distribution may have a heavy tail, i.e. non-negligible weights on rarely occurring events. We point out first in this paper that the Wasserstein metric as it stands, i.e. the family of type-p Wasserstein metrics, $p \in [1, \infty]$, cannot accommodate this simultaneous pursuit. This is because ambiguity sets constructed from a Wasserstein metric with a higher order, e.g. $p \geqslant 2$, inevitably exclude heavy-tailed distributions[1] (Birghila et al. (2022), de Haan and Ferreira (2006)). One may view this as a trade-off between the pursuit of robustness and data-drivenness when it comes to the choice of a Wasserstein metric. While this trade-off may appear inevitable, i.e. a less conservative choice would dismiss any heavy-tailed distribution, the primary goal of this paper is to present alternative

[1] In this paper, heavy-tailed distributions refer to distributions with finite mean but without finite variance.

families of Wasserstein metrics that could resolve, or at least lessen, this trade-off. We present a general framework that formalizes this simultaneous pursuit of robustness and data-drivenness in terms of the choice of a Wasserstein metric, and identify a large class of Wasserstein metrics, termed coherent Wasserstein metrics, that allows for exploring this pursuit.

The class of coherent Wasserstein metrics is motivated by the attempt to generalize the type-1 and the type-∞ Wasserstein metric from a new perspective. An observation can be made that the type-∞ Wasserstein metric can be viewed as a risk-averse counterpart of the type-1 Wasserstein metric, which replaces the expectation operator of the latter (see (5) in Section 2) with the worst-case risk measure, i.e. ess-sup, to summarize a transportation cost distribution (Kantorovich and Rubinshtein (1958), Rachev and Rüschendorf (1998), Villani (2008)). Coherent Wasserstein metrics generalize this observation by adopting a general class of risk measures, namely coherent risk measures (Artzner et al. (1999), Delbaen (2002)), to summarize a transportation cost distribution, which consists of the expectation and the worst-case risk measure as special cases. Coherent Wasserstein metrics can be interpreted also as general risk-averse formulations of the optimal transport problem arising from the classical definition of the Wasserstein metric (Villani (2008)). We show that coherent Wasserstein metrics provide a powerful means to reconcile the type-1 and the type-∞ Wasserstein metrics and offer the opportunity to identify new families of Wasserstein metrics motivated by the popularity of risk measures such as Conditional value-at-risk (CVaR) (Acerbi and Tasche (2002), Krätschmer et al. (2019), Embrechts et al. (2014)) and expectiles (Bellini et al. (2014), Bellini and Bernardino (2017), Gneiting (2011)). Like the Wasserstein metric, coherent Wasserstein metrics are theoretically sound in that they satisfy all necessary properties of a distance metric. We call the resulting DD-DRO formulation generalized Wasserstein distributionally robust optimization (GW-DRO) and show that GW-DRO generally satisfies the desirable conditions posed for DD-DRO, namely finite-sample guarantee, asymptotic consistency, and computational tractability (see Section 2.2 for detailed definitions).

From an optimization perspective, GW-DRO represents a new class of DRO problems that are distinctly different from existing DRO problems in two aspects. Firstly, the worst-case expectation problem embedded in GW-DRO has a nonlinear constraint in distribution, whereas existing worst-case expectation problems in DRO such as Wasserstein DRO generally have linear constraints in distributions, taking the form of moment constraints. Secondly, as shown in this paper, the worst-case expectation problem of GW-DRO is in general a nonconvex optimization problem in distribution, which to the best of our knowledge has not been studied in the DRO literature. These differences are significant, which render existing DRO analysis no longer applicable to studying the tractability of GW-DRO. In this paper, we take a different approach to studying the tractability of the worst-case expectation problems. Instead of relying on the common analysis starting from a dual problem formulation, we tackle directly the worst-case expectation problem from a primal perspective, and show how it can be reduced to a finite-dimensional optimization problem. Leveraging this finite-dimensional result, we then show how GW-DRO problems can be tractably solved as convex programs. Our approach of tackling the primal problem first and then deriving more tractable formulations is novel, which opens the door for solving a more general class of DRO problems. As a by-product, it provides an alternative, possibly simpler, way to derive the tractable formulation for Wasserstein DRO.

In addition to general tractability results, we provide also in-depth analysis of GW-DRO by focusing on two important instances of Wasserstein balls, defined by CVaR- and expectile-based Wasserstein metrics. We show that the worst-case expectation problems in these instances can be solved in closed-form when the loss function is convex Lipschitz continuous and the support set is unconstrained. These closed-form solutions are highly interpretable and structurally comparable to the solutions for the type-1 Wasserstein DRO. In particular, we show that the solutions for the case of expectile-based Wasserstein ball, CVaR-based Wasserstein ball, and type-1 Wasserstein ball are closely related in that they exhibit an "inclusion" relationship with the first being most general. This applies also to their respective worst-case distributions, with the first having the most flexible, or the richest, worst-case distribution structure. The closed-form solutions, when applied to contexts such as machine learning, could be interpreted also from a regularization perspective (c.f. Kuhn et al. (2019)). For instance, GW-DRO in these applications, when adopting CVaR-based Wasserstein balls, boils down to aggregating different regularized empirical minimization problems into a single minimization problem, and thus could be viewed as an ensemble of regularized models.

In the following, we summarize the key contributions of this paper.

1. We propose, in the spirit of data-driven distributionally robust optimization, a theoretically sound framework for studying different families of Wasserstein metrics. The framework sheds light on the potential limitation of the existing family of Wasserstein metrics and offers guidance to discover new families of Wasserstein metrics better suited for designing a richer, yet not only conservative, form of ambiguity sets.

2. We introduce a new family of coherent Wasserstein metrics and show that the corresponding distributionally robust optimization models, i.e. GW-DRO, enjoy all the desirable properties of a data-driven model for solving a stochastic program, namely finite-sample guarantee, asymptotic consistency, and tractability.

3. We provide a new systematic approach to studying the tractability of distributionally robust optimization problems without relying on the common duality scheme. It allows for tackling non-convex worst-case expectation problems naturally arising from GW-DRO and proving their tractability with the discovery of hidden convexity.

4. We present the application of GW-DRO to operations, finance, and machine learning problems. In particular, we show that in many of these problems a deep connection can be drawn between Wasserstein DRO and GW-DRO. Most notably, while the former is known to have a regularization interpretation in applications such as machine learning, the latter offers an even richer, and more novel, regularization interpretation.

2 Wasserstein data-driven distributionally robust optimization

As the basic setup, we denote a decision vector by $x \in \mathbb{R}^n$, a random vector of interest by $\xi \sim \mathbb{P}$, supported on a convex set $\Xi \subseteq \mathbb{R}^m$, i.e. $\mathbb{P}(\xi \in \Xi) = 1$, and a loss function $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ by $h(x, \xi)$, which depends on a made decision x and the realization of the random vector ξ. In many practical problems of interest, one seeks to find a decision x that minimizes the expected loss $\mathbb{E}^{\mathbb{P}}[h(x, \xi)]$, i.e. solving

$$J^\star := \inf_{x \in \mathbb{X}} \left\{ \mathbb{E}^{\mathbb{P}}[h(x,\xi)] = \int_{\Xi} h(x,\xi)\mathbb{P}(\mathrm{d}\xi) \right\}, \tag{1}$$

where $\mathbb{X}$ denotes a feasible set of solutions.

Data-driven optimization refers to finding a solution to the above problem when the distribution $\mathbb{P}$ can only be partially observed through a finite set of data $\widehat{\xi}_1, ..., \widehat{\xi}_N$ sampled independently from the distribution. One common data-driven method is to directly replace the distribution $\mathbb{P}$ with the empirical distribution $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\xi}_i}$ and solve instead the following optimization problem

$$\widehat{J}_{\mathrm{SAA}} := \inf_{x \in \mathbb{X}} \left\{ \mathbb{E}^{\widehat{\mathbb{P}}_N}[h(x,\xi)] = \frac{1}{N}\sum_{i=1}^{N} h(x, \widehat{\xi}_i) \right\}. \tag{2}$$

This method, also known as sample average approximation (SAA), is susceptible to sampling errors and suffers from the issue of the optimizer's curse (bias), i.e. disappointing out-of-sample performances (Kuhn et al. (2019)). As a remedy, data-driven distributionally robust optimization (DD-DRO) was proposed as a new data-driven method, which offers a solution that mitigates the adverse impact of sampling errors by solving the following minimax optimization problem

$$\widehat{J}_N := \inf_{x\in\mathbb{X}} \sup_{\mathbb{P}\in\mathbb{B}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[h(x,\xi)]. \tag{3}$$

The set $\mathbb{B}(\widehat{\mathbb{P}}_N)$, known as ambiguity set, is a set constructed based on the empirical distribution $\widehat{\mathbb{P}}_N$, which contains the unknown distribution $\mathbb{P}$ in (1) with high probability. A solution generated from (3) is robust against sampling errors in that it is guaranteed to perform the best with respect to the most adversarial distribution from the set $\mathbb{B}(\widehat{\mathbb{P}}_N)$. A natural construction of the set $\mathbb{B}(\widehat{\mathbb{P}}_N)$ takes the general form of

$$\mathbb{B}_{\varepsilon}^{d}\left(\widehat{\mathbb{P}}_N\right) := \left\{\mathbb{P}\in\mathcal{M}(\Xi) : d\left(\widehat{\mathbb{P}}_N,\mathbb{P}\right) \leqslant \varepsilon\right\}, \tag{4}$$

where $\mathcal{M}(\Xi)$ denotes the set of all distributions supported on Ξ, $d(\mathbb{P}_1,\mathbb{P}_2)$ stands for a probability metric that measures the distance between any two distributions $\mathbb{P}_1,\mathbb{P}_2\in\mathcal{M}(\Xi)$, and ε refers to the radius of the ball centred at the empirical distribution $\widehat{\mathbb{P}}_N$.

The quality of the solution generated from (3) depends critically on the structure of the ambiguity set $\mathbb{B}_{\varepsilon}^{d}(\widehat{\mathbb{P}}_N)$, which in turn depends on the choice of the probability metric d. Among several proposed probability metrics, the (type-1) Wasserstein metric (Kantorovich and Rubinshtein (1958))

$$d_{\mathrm{W}}\left(\mathbb{P}_1,\mathbb{P}_2\right) := \inf\left\{\mathbb{E}^{\Pi}\left[\|\xi_1-\xi_2\|\right] \,\middle|\, \begin{array}{c}\Pi \text{ is a joint distribution of } \xi_1 \text{ and } \xi_2\\ \text{with marginals } \mathbb{P}_1 \text{ and } \mathbb{P}_2, \text{ respectively}\end{array}\right\}, \tag{5}$$

has stood out as a popular choice, given its applicability to a large class of distributions, i.e. any distributions $\mathbb{P}_1,\mathbb{P}_2\in\mathcal{M}(\Xi)$ that have finite first moments. In particular, it allows for constructing a ball

$$\mathbb{B}_{\varepsilon}^{\mathrm{W}}\left(\widehat{\mathbb{P}}_N\right) := \left\{\mathbb{P}\in\mathcal{M}(\Xi) : d_{\mathrm{W}}\left(\widehat{\mathbb{P}}_N,\mathbb{P}\right) \leqslant \varepsilon\right\}$$

that contains a rich set of distributions.

The ball $\mathbb{B}_{\varepsilon}^{\mathrm{W}}(\widehat{\mathbb{P}}_N)$ is advantageous from a robustness perspective, i.e. containing various forms of distributions, but its flip side is less mentioned in the literature of Wasserstein DRO. Namely, it may contain overly-disperse distributions that differ too noticeably from the empirical distribution and thus be considered overly-conservative. As a useful contrast to highlight the limitation of the ball constructed from the type-1 Wasserstein metric, let us consider the following variant of Wasserstein metric, known as the type-∞ Wasserstein metric:

$$d_\infty\left(\mathbb{P}_1, \mathbb{P}_2\right) := \inf \left\{ \text{ess-sup}^\Pi \|\xi_1 - \xi_2\| \;\middle|\; \begin{array}{l} \Pi \text{ is a joint distribution of } \xi_1 \text{ and } \xi_2 \\ \text{with marginals } \mathbb{P}_1 \text{ and } \mathbb{P}_2, \text{ respectively} \end{array} \right\}. \tag{6}$$

Its induced ball

$$\mathbb{B}_\varepsilon^{\text{wc}}\left(\widehat{\mathbb{P}}_N\right) := \left\{\mathbb{P} \in \mathcal{M}(\Xi) : d_\infty\left(\widehat{\mathbb{P}}_N, \mathbb{P}\right) \leqslant \varepsilon\right\}$$

would contain only distributions that fully concentrate in a neighbourhood of samples $\widehat{\xi}_1, ..., \widehat{\xi}_N$, bounded by ε, and thus resemble to a greater extent the empirical distribution. The ball constructed from the type-∞ Wasserstein metric thus has the merit of data-drivenness. The price to pay to adopt the type-∞ Wasserstein metric is high, nonetheless, from a robustness perspective, as the metric is only applicable to distributions with bounded support.

One can see that the two Wasserstein metrics d_{W} and d_∞ essentially differ in how they summarize the distribution of $\|\xi_1 - \xi_2\|$. To formalize this point, we call a random variable X a transportation cost random variable from $\mathbb{P}_1$ to $\mathbb{P}_2$ if there exist $\xi_1 \sim \mathbb{P}_1,\ \xi_2 \sim \mathbb{P}_2$ such that $X \overset{\text{d}}{=} \|\xi_1 - \xi_2\|$. Let ρ denote a real-valued function that maps a random variable X to a real value. In the case of type-1 Wasserstein metric d_{W}, we have $\rho := \mathbb{E}$, whereas in the case of type-∞ Wasserstein metric we have $\rho :=$ ess-sup. The type-1 Wasserstein metric d_{W} could induce an overly-conservative ball $\mathbb{B}_\varepsilon^{\text{W}}(\widehat{\mathbb{P}}_N)$, because the expectation $\mathbb{E}$ is indistinguishable for deviations of X at different quantiles, whereas the type-∞ Wasserstein metric induces a ball $\mathbb{B}_\varepsilon^{\text{wc}}(\widehat{\mathbb{P}}_N)$ that can contain only distributions with bounded support, because esssup, as the worst-case risk measure, is the strongest tail measure.

Taking this perspective, we seek to identify in this paper a new class of Wasserstein metrics that can reconcile the type-1 and type-∞ Wasserstein metrics so that these metrics can be well justified from both robustness and data-drivenness perspective. We formalize this pursuit in the next section, where a new class of Wasserstein metrics, called coherent Wasserstein metrics, will be introduced.

2.1 Coherent Wasserstein metrics

We begin by defining $\{\rho_\alpha\}_{\alpha \in A}$ as a class of real-valued functions used to summarize the distribution of a transportation cost random variable, where A is an index set. The induced Wasserstein distance between two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as

$$d_{\rho_\alpha}(\mathbb{P}_1, \mathbb{P}_2) := \inf \left\{ \rho_\alpha^\Pi(\|\xi_1 - \xi_2\|) \;\middle|\; \begin{array}{l} \Pi \text{ is a joint distribution of } \xi_1 \text{ and } \xi_2 \\ \text{with marginals } \mathbb{P}_1 \text{ and } \mathbb{P}_2, \text{ respectively} \end{array} \right\}, \tag{7}$$

and a ball of radius ε centred at the empirical distribution $\widehat{\mathbb{P}}_N$ can be defined accordingly as

$$\mathbb{B}_{\varepsilon}^{\rho_\alpha}\left(\widehat{\mathbb{P}}_N\right) := \left\{\mathbb{P} \in \mathcal{M}(\Xi) : d_{\rho_\alpha}\left(\widehat{\mathbb{P}}_N, \mathbb{P}\right) \leqslant \varepsilon\right\}.$$

The novelty of our framework lies in taking a set perspective, i.e. $\alpha \in A$, to study properties that a whole family of Wasserstein metrics $\{d_{\rho_\alpha}\}_{\alpha \in A}$ should satisfy, rather than considering each metric separately. This perspective, which is largely missing in the literature of Wasserstein distributionally robust optimization is essential, we believe, when it comes to studying the choice of a Wasserstein metric. Built upon the observation made about the type-1 and type-∞ Wasserstein metrics d_{W} and d_∞, namely that the former is advantageous from a robustness perspective whereas the latter is advantageous from a data-driven perspective, we define the following two desirable properties for a family of metrics $\{d_{\rho_\alpha}\}_{\alpha \in A}$. These two properties capture the simultaneous pursuit of robustness and data-drivenness underlying the philosophy of data-driven distributionally robust optimization.

(i) **(Robustness)** A family of metrics $\{d_{\rho_\alpha}\}_{\alpha \in A}$ is said to have the property of robustness if for each $\alpha \in A$, ρ_α is well-defined (takes finite value) for any transportation cost random variable X that has finite first moment, i.e. L^1 random variables. Any distribution with finite mean is contained in a Wasserstein ball $\mathbb{B}_{\varepsilon}^{\rho_\alpha}(\widehat{\mathbb{P}}_N)$ for some $\varepsilon > 0$.

(ii) **(Data-drivenness)** A family of metrics $\{d_{\rho_\alpha}\}_{\alpha \in A}$ is said to have the property of data-drivenness if there exists a sequence of indices $\alpha_n \in A$, $n \in \mathbb{N}$, such that ρ_{α_n} converges to the worst-case risk measure ess-sup. The Wasserstein ball $\mathbb{B}_{\varepsilon}^{\rho_{\alpha_n}}(\widehat{\mathbb{P}}_N)$ converges to $\mathbb{B}_{\varepsilon}^{\mathrm{wc}}(\widehat{\mathbb{P}}_N)$, as $n \to \infty$.

These two properties together ensure that a family of metrics $\{d_{\rho_\alpha}\}_{\alpha \in A}$ is rich enough to, on the one hand, accommodate distributions with a more adversarial form, e.g. heavy-tailed distributions, like the type-1 Wasserstein metric, and on the other hand be used to approximate the functionality of the type-∞ Wasserstein metric. Clearly, the singleton $\{d_{\mathrm{W}}\}$ satisfies robustness but not data-drivenness, whereas the singleton $\{d_\infty\}$ satisfies data-drivenness but not robustness.

Definition 1. We call a family of metrics $\{d_{\rho_\alpha}\}_{\alpha \in A}$ data-driven distributionally robust Wasserstein **(DD-DRW)** metrics if they satisfy both the properties of robustness and data-drivenness.

When $\rho_p(X) = \mathbb{E}[X^p]^{1/p}$, $p \in [1, \infty)$, the induced distance is the Wasserstein metric of order p. It is clear that the family $\{d_{\rho_p}\}_{p \in [1,\infty)}$ is not **DD-DRW**, because it satisfies data-drivenness, i.e. $\rho_p(X)$ converges to ess-sup(X) as $p \to \infty$ but not robustness, i.e. the ambiguity set $\mathbb{B}_{\varepsilon}^{\rho_\alpha}(\widehat{\mathbb{P}}_N)$

fails to account for heavy-tailed distributions for some $p > 1$. This points out the potential limitation of applying the family of p^{th}-order Wasserstein metrics. Namely, the price that needs to be paid to construct a less conservative ambiguity set $\mathbb{B}_{\varepsilon}^{\rho_\alpha}(\widehat{\mathbb{P}}_N)$ is high from a distributionally robust perspective – one has to forgo any heavy-tailed distribution of practical interest.

It is natural to wonder if the limitation of the family $\{d_{\rho_p}\}_{p\in[1,\infty)}$ lies in its use of p^{th}-order power function. We show below that the limitation comes more fundamentally from the use of expected functionals to summarize the transportation cost random variable X.

Proposition 1. *The family of metrics $\{d_{\rho_\alpha}\}_{\alpha\in A}$, where $\rho_\alpha(X) = \ell_\alpha^{-1}(\mathbb{E}[\ell_\alpha(X)])$*[2]*, $\alpha \in A$ and ℓ_α is increasing convex function and $\ell_\alpha(0) = 0$, $\alpha \in A$, is not* **DD-DRW**.

We now introduce a new class of Wasserstein metrics, called coherent Wasserstein metrics, that generalize the type-1 and type-∞ Wasserstein metrics from a risk measure perspective.

Definition 2. (Coherent Wasserstein metrics) A metric $d_\rho(\cdot,\cdot) : \mathcal{M}^2 \to \mathbb{R}_+$ is called a coherent Wasserstein metric if it takes the form of

$$d_\rho(\mathbb{P}_1,\mathbb{P}_2) := \inf\left\{ \rho^{\Pi}(\|\xi_1 - \xi_2\|) \;\middle|\; \begin{array}{l} \Pi \text{ is a joint distribution of } \xi_1 \text{ and } \xi_2 \\ \text{with marginals } \mathbb{P}_1 \text{ and } \mathbb{P}_2, \text{ respectively} \end{array} \right\}, \tag{8}$$

where ρ is a law-invariant coherent risk measure, i.e. satisfying $\rho(0) = 0$ and the following properties:

(translation invariance) $\rho(X + c) = \rho(X) + c$ for any $c \geqslant 0$,

(monotonicity) $\rho(X_1) \geqslant \rho(X_2)$ for any $X_1 \geqslant X_2$,

(subadditvity) $\rho(X_1 + X_2) \leqslant \rho(X_1) + \rho(X_2)$,

(positive homogeneity) $\rho(cX) = c\rho(X)$ for any $c > 0$,

(law invariance) $\rho(X_1) = \rho(X_2)$ for any $X_1 \overset{\mathrm{d}}{=} X_2$.

The use of a law-invariant coherent risk measure ρ is motivated by its well-established properties in the literature of risk measures (Artzner et al. (1999), Kusuoka (2001)) and that it naturally includes the expectation $\mathbb{E}$ and the worst-case measure ess-sup, as limiting cases. Coherent Wasserstein metrics can be viewed as natural risk-averse formulations of the classical optimal transport problem (Villani (2008)). We show firstly that coherent Wasserstein metrics, like the classical Wasserstein metrics, are valid distance metrics.

Proposition 2. *Any coherent Wasserstein metric $d_\rho(\cdot,\cdot) : \mathcal{M}(\Xi) \times \mathcal{M}(\Xi) \to \mathbb{R}_+$ satisfies the following properties of a distance metric*

[2] For a non-decreasing function ℓ, its inverse function is defined as $\ell^{-1}(x) = \inf\{y : \ell(y) \geqslant x\}$.

(i) *(Identity of indiscernibles)* $d_\rho(\mathbb{P}_1, \mathbb{P}_2) = 0$ *if and only if* $\mathbb{P}_1 = \mathbb{P}_2$.

(ii) *(Symmetry)* $d_\rho(\mathbb{P}_1, \mathbb{P}_2) = d_\alpha(\mathbb{P}_2, \mathbb{P}_1)$ *for any* $\mathbb{P}_1, \mathbb{P}_2$.

(iii) *(Triangle inequality)* $d_\rho(\mathbb{P}_1, \mathbb{P}_2) + d_\rho(\mathbb{P}_2, \mathbb{P}_3) \geqslant d_\rho(\mathbb{P}_1, \mathbb{P}_3)$ *for any* $\mathbb{P}_1$, $\mathbb{P}_2$, $\mathbb{P}_3$.

(iv) *(Non-negativity)* $d_\rho(\mathbb{P}_1, \mathbb{P}_2) \geqslant 0$ *for any* $\mathbb{P}_1, \mathbb{P}_2$.

It turns out that coherent Wasserstein metrics offer the needed flexibility for building a family of metrics $\{d_{\rho_\alpha}\}_{\alpha \in A}$ satisfying the property of **DD-DRW**. We highlight below that a family of metrics $\{d_{\rho_\alpha}\}_{\alpha \in A}$ composed of coherent Wasserstein metrics naturally satisfies the property under very mild conditions. Recall that $\mathrm{VaR}_\alpha(X)$ is the Value-at-Risk of X at level α defined by

$$\mathrm{VaR}_\alpha(X) = F^{-1}(\alpha) = \inf\{x : F(x) \geqslant \alpha\}, \quad X \sim F,$$

and a function $g : [0,1] \to [0,1]$ is called a distortion function if it is increasing and satisfies $g(0) = 0, g(1) = 1$. We denote the left-derivative function of g by g'.

Proposition 3. *A family of metrics* $\{d_{\rho_\alpha}\}_{\alpha \in A}$*, where* d_{ρ_α} *is a coherent Wasserstein metric, satisfies* **DD-DRW** *if and only if for every* $\alpha \in A$, ρ_α *can be represented by*

$$\rho_\alpha(X) = \sup_{g \in \mathcal{H}_{\rho_\alpha}} \int_0^1 \mathrm{VaR}_\alpha(X) \mathrm{d}g(\alpha), \tag{9}$$

where $\mathcal{H}_{\rho_\alpha}$ *is a subset of convex distortion functions satisfying* $c_\alpha := \sup_{g \in \mathcal{H}_{\rho_\alpha}} \|g'\|_\infty < \infty$*, and* $\exists\, \alpha_n \in A$, $n \in \mathbb{N}$ *such that* $c_{\alpha_n} \to \infty$ *as* $n \to \infty$.

The representation (9) is known as the dual representation of a law-invariant coherent risk measure. The property of **DD-DRW**, particularly the **robustness** condition, boils down to requiring first the existence of such a representation. This is a very mild condition in that any lower-semicontinuous coherent risk measure is known to have a dual representation (see, e.g. Kusuoka (2001), Jouini et al. (2006) and Rüschendorf (2013)). This observation, more importantly, reveals that to build a family of metrics $\{d_{\rho_\alpha}\}_{\alpha \in A}$ satisfying **DD-DRW**, a generalization of Wasserstein metrics from a dual perspective, i.e. coherent Wasserstein metrics, is critical. This is in shape contrast to a generalization from an expected functional perspective, i.e. Proposition 1. The property of **DD-DRW** further requires that the set $\mathcal{H}_{\rho_\alpha}$ in (9) contains only Lipschitz continuous distortion functions with uniformly bounded Lipschitz constants, i.e. bounded by c_α, and that there exists a sequence of $\sup_{g \in \mathcal{H}_{\rho_\alpha}} \|g'\|_\infty$ in $\alpha \in A$ converges to the infinity. It is not hard to identify families of coherent risk measures satisfying these conditions. In particular, we identify the following two families of coherent Wasserstein metrics, defined through Conditional Value-at-Risk (CVaR) and expectiles, that satisfy **DD-DRW**. CVaR and expectiles are two most popular risk measures proposed as convex substitutes for the traditional risk measure, Value-at-Risk (VaR). In the remainder of this paper, we will pay particular attention to these two families of coherent Wasserstein metrics to demonstrate the practical value of our new framework.

CVaR-Wasserstein Metric

Take ρ as CVaR at level $\alpha \in [0,1)$, i.e.,

$$\rho(X) = \mathrm{CVaR}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(X)\mathrm{d}u, \quad \alpha \in [0,1).$$

We obtain the following metric

$$d_{\mathrm{CVaR}_\alpha}(\mathbb{P}_1,\mathbb{P}_2) := \inf\left\{\mathrm{CVaR}_\alpha^\Pi(\|\xi_1-\xi_2\|) : \Pi \in \Pi(\mathbb{P}_1,\mathbb{P}_2)\right\} \tag{10}$$

and the CVaR-Wasserstein ball

$$\mathbb{B}^\alpha_{(1),\varepsilon}\left(\widehat{\mathbb{P}}_N\right) := \left\{\mathbb{P} \in \mathcal{M}(\Xi) : d_{\mathrm{CVaR}_\alpha}\left(\widehat{\mathbb{P}}_N, \mathbb{P}\right) \leqslant \varepsilon\right\}, \ \alpha \in [0,1).$$

Expectile-Wasserstein Metric

Recall that the expectile $e_\alpha(X)$ of X at level $\alpha \in [0,1)$ is defined as the unique solution to

$$\alpha\mathbb{E}\left[(X-x)_+\right] = (1-\alpha)\mathbb{E}\left[(X-x)_-\right],$$

where $a_+ = \max\{a,0\}$ and $a_- = \max\{-a,0\}$. $e_\alpha(X)$ is coherent for any $\alpha \in [1/2,1)$, reduces to the mean $\mathbb{E}$ when $\alpha = 1/2$ and converges to the worst-case risk measure ess-sup as $\alpha \to 1$.

Taking ρ as expectile $e_\alpha(X)$ at level $\alpha \in [1/2,1)$, we have the following metric

$$d_{e_\alpha}(\mathbb{P}_1,\mathbb{P}_2) := \inf\left\{e_\alpha^\Pi(\|\xi_1-\xi_2\|) : \Pi \in \Pi(\mathbb{P}_1,\mathbb{P}_2)\right\} \tag{11}$$

and the expectile-Wasserstein ball

$$\mathbb{B}^\alpha_{(2),\varepsilon}\left(\widehat{\mathbb{P}}_N\right) := \left\{\mathbb{P} \in \mathcal{M}(\Xi) : d_{e_\alpha}\left(\widehat{\mathbb{P}}_N, \mathbb{P}\right) \leqslant \varepsilon\right\}, \ \alpha \in [1/2,1).$$

We close this section by providing a simple demonstration of how the family of CVaR-Wasserstein metrics allows for constructing Wasserstein balls that can, on the one hand, contain heavy-tailed distributions of practical interest and on the other hand converge to the type-∞ Wasserstein ball, as $\alpha \to 1$. It is worth noting that by adopting a family of coherent Wasserstein metrics $\{d_{\rho_\alpha}\}_{\alpha\in A}$, the property of **robustness** in fact implies that for any $\varepsilon > 0$, the ambiguity set $\mathbb{B}_\varepsilon^{\rho_\alpha}(\widehat{\mathbb{P}}_N)$ always contains a heavy-tailed distribution.

Example 1. A function $F_{\gamma,\beta}$ with $\gamma, \beta > 0$ is called a Pareto distribution if

$$F_{\gamma,\beta}(x) = 1 - \left(1 + \frac{x}{\beta}\right)^{-\gamma}, \quad x \geqslant 0.$$

Suppose that $\widehat{\mathbb{P}} = \delta_0$, i.e. a point mass at 0. Let us define the following two sets. The first is based on the p^{th}-order Wasserstein metric for some $p \geqslant 2$, whereas the second is based on the CVaR-Wasserstein metric for some $\alpha \in [0, 1)$

$$\mathbb{B}_1(p) = \{F_{\gamma,\beta} : d_{W_p}(F_{\gamma,\beta}, \widehat{\mathbb{P}}) \leqslant \varepsilon,\ \gamma, \beta > 0\}, \quad p \geqslant 2,$$

$$\mathbb{B}_2(\alpha) = \{F_{\gamma,\beta} : d_{\text{CVaR}_\alpha}(F_{\gamma,\beta}, \widehat{\mathbb{P}}) \leqslant \varepsilon,\ \gamma, \beta > 0\}, \quad \alpha \in [0, 1).$$

Figure 1 demonstrates Pareto distributions with different γ that are feasible to the CVaR-Wasserstein ball $\mathbb{B}_2(\alpha)$ for $\alpha = 0$ (the left figure) and for $\alpha = 0.99$ (the right figure). Note first that none of the heavy-tailed distributions in the figures are feasible to the type-p Wasserstein ball $\mathbb{B}_1(p)$, $p \geqslant 2$, since for any $p > \gamma$, $d_{W_p}(F_{\gamma,\beta}, \widehat{\mathbb{P}}) = \infty$, and thus, $F_{\gamma,\beta} \notin \mathbb{B}_1(p)$. In contrast, for any $\alpha \in [0, 1)$ and $\gamma > 1$, there always exists β such that $F_{\gamma,\beta} \in \mathbb{B}_2(\alpha)$.[3] Moreover, comparing the feasible Pareto distributions between the two figures, one can see that the Pareto distributions in the right figure (the case $\alpha = 0.99$) concentrate significantly around the sample point 0 while retaining "a bit of" heavy tail. This showcases how the family of CVaR-Wasserstein allows for the simultaneous pursuit of robustness and data-drivenness. In the case $\alpha = 0$ (the left figure), the CVaR-Wasserstein metric reduces to the type-1 Wasserstein metric and one can see from the figure that the feasible Pareto distributions disperse to the right noticeably away from the sample point,

[3] For $\alpha \in (0, 1)$, it holds that

$$d_{\text{CVaR}_\alpha}(F_{\gamma,\beta}, \widehat{\mathbb{P}}) = \frac{1}{1-\alpha}\int_\alpha^1 \left[\beta(1-u)^{-1/\gamma} - \beta\right] du = \frac{\gamma\beta}{\gamma-1}(1-\alpha)^{-1/\gamma} - \beta.$$

This implies

$$\mathbb{B}_2(\alpha) = \left\{F_{\gamma,\beta} : \frac{\gamma}{\gamma-1}(1-\alpha)^{-1/\gamma} - 1 \leqslant \frac{\varepsilon}{\beta},\ \gamma, \beta > 0\right\}.$$

Note that $\lim_{\beta\to 0} \varepsilon/\beta = \infty$ which implies for any $\gamma > 1$, there exists $\beta > 0$ small enough such that $\frac{\gamma}{\gamma-1}(1-\alpha)^{-1/\gamma} - 1 \leqslant \varepsilon/\beta$. So, for each $\gamma > 1$, there exists $\beta > 0$ such that $F_{\gamma,\beta} \in \mathbb{B}_2(\alpha)$.

which shows the conservative nature of the type-1 Wasserstein metric.

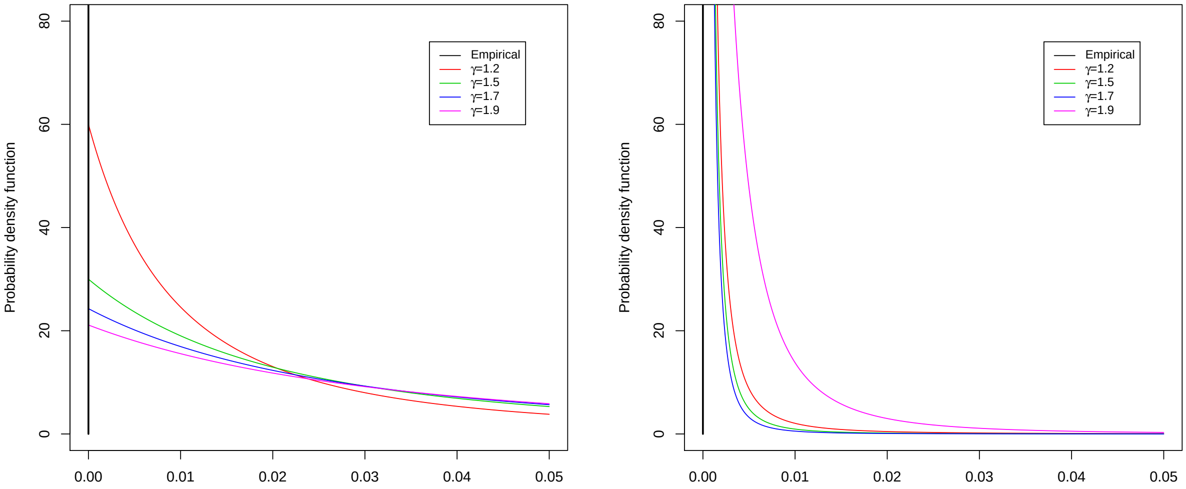


Figure 1: Feasible distributions in $\mathbb{B}_2(\alpha)$ with $\varepsilon = 0.1$ for $\alpha = 0$ (left) and $\alpha = 0.99$ (right).

2.2 Generalized Wasserstein distributionally robust optimization

We call the data-driven distributionally robust model (3) with an ambiguity set defined based on a coherent Wasserstein metric d_{ρ_α}, i.e. $\mathbb{B}(\widehat{\mathbb{P}}_N) := \mathbb{B}_\varepsilon^{\rho_\alpha}\left(\widehat{\mathbb{P}}_N\right)$, generalized Wasserstein distributionally robust optimization (GW-DRO) model. From this point on, we let $\widehat{J}_N$ and $\widehat{x}_N$ denote respectively the optimal value and the optimal solution to the GW-DRO model.

We will demonstrate throughout this paper that GW-DRO, like Wasserstein DRO (Esfahani and Kuhn (2018)), has all the desirable properties of a data-driven model for solving the stochastic program (1), namely finite-sample guarantee, asymptotic consistency, and tractability. The first refers to the guarantee that the out-of-sample performance of $\widehat{x}_N$ can be bounded, with some confidence level, by the optimal value $\widehat{J}_N$ when the ambiguity set $\mathbb{B}_\varepsilon^{\rho_\alpha}\left(\widehat{\mathbb{P}}_N\right)$ is properly calibrated, the second refers to assurance that $\widehat{J}_N$ and $\widehat{x}_N$ would converge respectively to the optimal value and solution to the nominal problem (1) as $N \to \infty$, and the third refers to the computational tractability of solving the minimax problem (3) for many loss functions $h(x, \xi)$ and sets $\mathbb{X}$.

We provide precise statements regarding the first two properties below. In particular, we highlight that GW-DRO enjoys these two properties under a rather mild condition on the risk measure ρ_α.

Proposition 4. *(Finite sample guarantee) Let ρ_α denote a risk measure with the representation* (9) *satisfying $g'(1) \leqslant c, \forall g \in \mathcal{H}_{\rho_\alpha}$ for some $c \in \mathbb{R}$, and $\mathbb{P}$ be a light-tailed distribution, i.e. satisfying $A := \mathbb{E}^{\mathbb{P}}\left[\exp\left(\|\xi\|^a\right)\right] = \int_\Xi \exp\left(\|\xi\|^a\right)\mathbb{P}(\mathrm{d}\xi) < \infty$, for some $a > 1$.*

Assume that $\widehat{J}_N$ and $\widehat{x}_N$ represent the optimal value and an optimizer of the distributionally robust program (3) *with an ambiguity set $\mathbb{B}(\widehat{\mathbb{P}}_N) = \mathbb{B}^{\rho_\alpha}_{\varepsilon_N(\eta)}\left(\widehat{\mathbb{P}}_N\right)$, $N \in \mathbb{N}$, for some $\eta \in (0,1)$, where*

$$\varepsilon_N(\eta) = \begin{cases} c\varepsilon_0^{1/m_2}, & \text{if } \varepsilon_0 \leqslant c, \\ c\varepsilon_0^{1/a}, & \text{if } \varepsilon_0 > c, \end{cases} \qquad \varepsilon_0 = \frac{\log\left(c_1/\eta\right)}{c_2 N},$$

for some constants c_1, c_2 only depending on a, A, m and $m_2 = \max\{m, 2\}$. Then, it holds the finite sample guarantee

$$\mathbb{P}^N\left\{\widehat{\Xi}_N : \mathbb{E}^{\mathbb{P}}\left[h\left(\widehat{x}_N, \xi\right)\right] \leqslant \widehat{J}_N\right\} \geqslant 1 - \eta. \tag{12}$$

Proposition 5. *(Asymptotic consistency) Under the condition of Proposition 4, let $\varepsilon_N = \left(\frac{k_N}{N}\right)^{1/m_2}$, $N \in \mathbb{N}$ where $k_N/N^\delta \to 0$ and $\log N/k_N \to 0$ as $N \to \infty$ for some $\delta < 1$, $m_2 = \max\{m, 2\}$, and assume that $\widehat{J}_N$ and $\widehat{x}_N$ represent the optimal value and an optimizer of the distributionally robust program* (3) *with an ambiguity set $\mathbb{B}(\widehat{\mathbb{P}}_N) = \mathbb{B}^{\rho_\alpha}_{\varepsilon_N}\left(\widehat{\mathbb{P}}_N\right)$, $N \in \mathbb{N}$.*

(i) If $h(x,\xi)$ is upper semicontinuous in ξ and there exists $L \geqslant 0$ with $|h(x, \xi)| \leqslant L(1 + \|\xi\|)$ for all $x \in \mathbb{X}$ and $\xi \in \Xi$, then $\mathbb{P}^\infty$-almost surely we have $\widehat{J}_N \downarrow J^\star$ as $N \to \infty$ where $J^\star$ is the optimal value of (1).

(ii) If the assumptions of assertion (i) hold, $\mathbb{X}$ is closed, and $h(x,\xi)$ is lower semicontinuous in x for every $\xi \in \Xi$, then any accumulation point of $\{\widehat{x}_N\}_{N\in\mathbb{N}}$ is $\mathbb{P}^\infty$-almost surely an optimal solution for (1).

These two guarantees are qualitatively identical to those of the Wasserstein DRO (c.f. Esfahani and Kuhn (2018)), except further parametrized by the exact specification of the risk measure ρ_α. Both guarantees require only that any distortion function g that may be invoked by (the dual representation of) the risk measure ρ_α has a bounded density at the worst-case value, i.e. the condition $g'(1) \leqslant c, \forall g \in \mathcal{H}_{\rho_\alpha}$. One can observe by Proposition 3 that, somewhat interestingly, any coherent Wasserstein metric d_{ρ_α} chosen from a **DD-DRW** family $\{d_{\rho_\alpha}\}_{\alpha\in A}$ would naturally meet this requirement. **DD-DRW** families of metrics are thus of rather convenient choices.

Lastly, with regard to the tractability of GW-DRO, we will demonstrate in Sections 3 and 4 that the minimax problem (3) can often be solved as finite-dimensional convex programs for many loss functions $h(x,\xi)$ arising from practical applications. In particular, we consider cases where the loss function $h(x, \xi)$ is either a general concave or convex function in ξ, and as motivating examples we present first in the next section a number of applications.

2.3 Illustrative examples

The first application is the classical two-stage planning problem with recourse decisions, which is particularly common in operations management contexts. The loss function $h(x, \xi)$ in this application could either be concave or convex in ξ, depending on the exact setting of the second-stage problem.

Example (i): Two-stage problems with recourse

Let x_0 denote the first stage, or "here-and-now", decision that needs to be made before the realization of a random vector $\xi \sim \mathbb{P}$. In the case where the distribution $\mathbb{P}$ is unknown, the following two-stage distributionally robust linear program naturally arises

$$\min_{x_0 \in \mathbb{X}} c^\top x_0 + \sup_{\mathbb{P} \in \mathbb{B}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[Q(x_0, \xi)],$$

where $Q(x_0, \xi)$ is the recourse function capturing the optimal value of the recourse problem. It can take either the formulation of

$$Q(x_0, \xi) = \min_{x_1} \left\{ x_1^\top \bar{Q} \xi \mid T x_0 + W x_1 \geqslant h \right\},$$

where the objective of the recourse problem is uncertain due to ξ, or the formulation of

$$Q(x_0, \xi) = \min_{x_1} \left\{ q^\top x_1 \mid T x_0 + W x_1 \geqslant h + H\xi \right\},$$

where the "right-hand-side" of the recourse problem is uncertain. Applications of these two cases can be found, for instance, in Bertsimas et al. (2010). Clearly, the first case corresponds to a loss function $h(x, \xi)$ in (1) that is concave in ξ, whereas the second case corresponds to a loss function $h(x, \xi)$ that is convex in ξ.

The next application is a problem of fundamental interest in finance.

Example (ii): Portfolio optimization

Let $\xi \sim \mathbb{P}$ denote a random vector of returns from n different financial assets. The problem of robust portfolio optimization is a widely studied topic (see, e.g. Delage and Ye (2010)), where a portfolio vector $x \in \mathbb{R}^n$ needs to be sought that maximizes the worst-case utility subject to investment constraints captured by $\mathbb{X}$

$$\max_{x \in \mathbb{X}} \inf_{\mathbb{P} \in \mathbb{B}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[u(\xi^\top x)].$$

The utility function $u : \mathbb{R} \to \mathbb{R}$ is non-decreasing and is often assumed to be concave so to capture the risk-aversion attitude of an investor. This corresponds to a loss function $h(x, \xi) := -u(\xi^\top x)$ in (1) that is convex in ξ.

The last application is motivated by the recent surge of interest in statistical learning.

Example (iii): Machine learning

In supervised learning, a random vector $\xi := (\xi^x, \xi^y) \sim \mathbb{P}$ represents an input-output pair and the goal is to seek a predictor (function) $f(\xi^x; \beta)$ parameterized by β that best maps a given input value ξ^x to a predicted output value. The issue of sampling errors, i.e. the uncertainty of $\mathbb{P}$, has motivated the recent study of the following distributionally robust statistical learning problem

$$\min_{\beta \in \mathcal{B}} \sup_{\mathbb{P} \in \mathbb{B}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\hat{\ell}(f(\xi^x; \beta), \xi^y)],$$

where $\hat{\ell}$ is a function capturing losses incurred from prediction errors. We make the common assumption of a linear predictor, i.e. $f(\xi^x; \beta) = \beta^\top \xi^x$ for some $\beta \in \mathcal{B}$. In the case of a regression problem, i.e. $\xi^y \in \mathbb{R}$, one can set

$$\hat{\ell}(f(\xi^x; \beta), \xi^y) := \ell(\beta^\top \xi^x - \xi^y), \quad \text{for some } \ell : \mathbb{R} \to \mathbb{R}_+$$

and $\|\xi_1 - \xi_2\| := \|(\xi_1^x, \xi_1^y) - (\xi_2^x, \xi_2^y)\|$ (in (7)), whereas in the case of a classification problem, i.e, $\xi^y \in \{-1, +1\}$, one can set

$$\hat{\ell}(f(\xi^x; \beta), \xi^y) := \ell(\xi^y \cdot \beta^\top \xi^x), \quad \text{for some non-increasing } \ell : \mathbb{R} \to \mathbb{R}_+,$$

and $\|\xi_1 - \xi_2\| := \|\xi_1^x - \xi_2^x\| + \mathbb{I}(\xi_1^y - \xi_2^y)$ (in (7)) where $\mathbb{I}(s) = 0$ if $s = 0$ and $\mathbb{I}(s) = \infty$ otherwise. The function ℓ chosen in most machine learning methods is a convex function (see e.g. Shafieezadeh-Abadeh et al. (2019)). The case of regression would thus correspond to a loss function $h(x, \xi)$ in (1) that is convex in ξ. In the case of classification, since the chosen norm $\|\xi_1 - \xi_2\|$ assumes the cost of perturbing an output is infinitely large, any distribution $\mathbb{P} \in \mathbb{B}(\widehat{\mathbb{P}}_N)$ would differ from the empirical distribution $\widehat{\mathbb{P}}_N$ only along the input space. We thus need only the observation that the loss function is convex in input variable ξ^x.

3 Solving GW-DRO with concave loss functions

To study the tractability of solving GW-DRO, we focus first on the inner maximization problem of (3) – the worst-case expectation problem. For notational convenience, we suppress the decision variable x in $\sup_{\mathbb{P}\in\mathbb{B}_{\rho,\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[h(x,\xi)]$ and write the problem as

$$\begin{aligned} \sup_{\mathbb{P}\in\mathcal{M}(\Xi)} \quad & \mathbb{E}^{\mathbb{P}}[\ell(\xi)] \\ \text{subject to} \quad & d_\rho\left(\widehat{\mathbb{P}}_N, \mathbb{P}\right) \leqslant \varepsilon. \end{aligned} \tag{13}$$

Throughout this section, we consider the case where the loss function ℓ is concave in ξ. Following the definition of coherent Wasserstein metrics, the above problem can be stated also in terms of the joint distributions Π

$$\begin{aligned} \sup_{\Pi} \quad & \mathbb{E}^{\Pi}[\ell(\xi)] \\ \text{subject to} \quad & \rho^{\Pi}(\|\widehat{\xi}-\xi\|) \leqslant \varepsilon, \\ & \Pi \in \mathcal{M}(\Xi^2) \text{ is a joint distribution of } \widehat{\xi} \text{ and } \xi \\ & \text{ with marginals } \widehat{\mathbb{P}}_N \text{ and } \mathbb{P}, \text{ respectively.} \end{aligned} \tag{14}$$

The above problem is distinctly different from and structurally more involved than the worst-case expectation problem from Wasserstein DRO in that it is a nonlinear optimization problem over the variable Π due to the nonlinearity of the function ρ^{Π}. An even more fundamental challenge of solving the problem (14) lies in the following observation.

Proposition 6. *The feasible set of* Π *in* (14) *may be nonconvex, and thus the worst-case expectation problem* (14) *is a nonconvex optimization problem in general.*

Proof. We show this by considering a representative class of metrics, CVaR-Wasserstein metrics. For $\rho = \mathrm{CVaR}_\alpha$ and $(1-\alpha) > 1/N$, let

$$\Pi_1 = \frac{1}{N}\sum_{i=1}^{N} \delta_{(\widehat{\xi}_i, \widehat{\xi}_i+\varepsilon e)} \quad \text{and} \quad \Pi_2 = \frac{1}{N}\sum_{i=2}^{N} \delta_{(\widehat{\xi}_i, \widehat{\xi}_i)} + \frac{1}{N}\delta_{\left(\widehat{\xi}_1, \widehat{\xi}_1+N\varepsilon e(1-\alpha)\right)},$$

where $e \in \mathbb{R}^m$ satisfies $\|e\| = 1$. One can verify that $\mathrm{CVaR}_\alpha^{\Pi_i}(\|\widehat{\xi}-\xi\|) = \varepsilon$, for $i = 1, 2$, that is, Π_1

and Π_2 are feasible solutions of the problem (14). Denote by $\Pi_\lambda = (1-\lambda)\Pi_1 + \lambda\Pi_2$ for $\lambda \in (0,1)$. We have

$$\mathrm{CVaR}_\alpha^{\Pi_\lambda}(\|\widehat{\xi} - \xi\|) = \varepsilon\frac{(1-\alpha)\lambda + [1-\alpha-\lambda/N]}{1-\alpha} = \varepsilon + \varepsilon\lambda\left(1 - \frac{1}{N(1-\alpha)}\right) > \varepsilon,$$

which means Π_λ is not a feasible solution of the problem (14). We therefore conclude that problem (14) is a nonconvex optimization problem in general. □

The problem (14) is thus an *infinite-dimensional nonconvex optimization problem* in general that is not amenable to convex analysis. The non-convexity of (14), as highlighted in the above proposition, naturally arises from coherent risk measures ρ^Π that are concave in distributions. These risk measures, such as CVaR and its extensions, could assign a higher risk value to a mixture distribution, i.e. a convex combination of two arbitrary distributions, reflecting the uncertainty resulting from the mixture. Because of the nonconvexity, the common strategy of analyzing the tractability of worst-case expectation problems, which starts from studying first their dual problems, is no longer applicable to (14). In this paper, we show how to bypass this difficulty by analyzing the worst-case expectation problem (14) first from a primal perspective, i.e. solving the problem (14) directly. Our analysis reveals that somewhat surprisingly, the problem (14) in its most general form can always be reduced to a structurally simple finite-dimensional convex optimization problem.

Theorem 1. *If ℓ is concave, then the worst-case expectation problem* (14) *is equivalent to*

$$\begin{aligned} &\sup_{y_1,\dots,y_N \in \mathbb{R}^m} \quad \frac{1}{N}\sum_{i=1}^N \ell(y_i) \\ &\text{subject to} \quad \rho^\Pi(\|\widehat{\xi} - \xi\|) \leqslant \varepsilon, \quad \Pi((\widehat{\xi}, \xi) = (\widehat{\xi}_i, y_i)) = \frac{1}{N}, \ y_i \in \Xi, \ i = 1,\dots,N. \end{aligned} \tag{15}$$

That is, the worst-case distribution to the problem (15) always takes the form of

$$\Pi = \frac{1}{N}\sum_{i=1}^N \delta_{(\widehat{\xi}_i, y_i^*)}, \quad \text{i.e.} \quad \mathbb{P} = \frac{1}{N}\sum_{i=1}^N \delta_{y_i^*},$$

where y_i^*, $i = 1,\dots,N$, is the optimal solution to the problem (15). The theorem is general in that it does not rely on any assumption of the function form of ρ, other than the general property of coherent risk measures. The feasible set of y_i, $i = 1,\dots,N$, is clearly a convex set, since for any two feasible solutions $y_i^{(1)}, y_i^{(2)} \in \Xi$, $i = 1,\dots,N$, and their convex combination $y_i^{(3)} = \lambda y_i^{(1)} + (1-\lambda)y_i^{(2)}$,

$i = 1, \ldots, N$, $\lambda \in [0,1]$, we have

$$\rho(\|\widehat{\xi} - \xi_{y^{(3)}}\|) \leqslant \rho(\lambda\|\widehat{\xi} - \xi_{y^{(1)}}\| + (1-\lambda)\|\widehat{\xi} - \xi_{y^{(2)}}\|) \leqslant \lambda\rho(\|\widehat{\xi} - \xi_{y^{(1)}}\|) + (1-\lambda)\rho(\|\widehat{\xi} - \xi_{y^{(2)}}\|) \leqslant \varepsilon,$$

where ξ_y denotes a random variable such that $(\widehat{\xi}, \xi_y)$ has the distribution $\Pi_y = \frac{1}{N}\sum_{i=1}^N \delta_{(\widehat{\xi}_i, y_i)}$. The theorem lays an important basis for studying the tractability of GW-DRO in general and can be extended further, as shown in the next section, to even more general class of loss functions ℓ.

Theorem 1 can be further exploited to identify the dual of the worst-case expectation problem (14). Namely, the convexity of the reduced problem (15) renders it now amenable to analysis using convex duality. We show in the following how the dual problem can be obtained also as a finite-dimensional convex minimization problem. This alternative formulation can be conveniently integrated with the outer minimization problem in (3) so to solve the overall problem of GW-DRO (3) as a single convex minimization problem.

Corollary 1. *If ℓ is concave, then the problem (14) is equivalent to*

$$\begin{aligned}
\inf_{\lambda, p, s, z_i, \nu_i} \quad & \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^N s_i \\
\text{subject to} \quad & [-\ell]^*(z_i - \nu_i) + \sigma_\Xi(\nu_i) - z_i^\top \widehat{\xi}_i \leqslant s_i, && i = 1, ..., N, && (16)\\
& \|z_i\|_* \leqslant p_i, && i = 1, ..., N, && (17)\\
& \sum_{i=1}^N p_i = \lambda, && && (18)\\
& \frac{p}{\lambda} \in A_\rho,
\end{aligned}$$

where $\lambda \in \mathbb{R}, p \in \mathbb{R}^N, s \in \mathbb{R}^N, z_i \in \mathbb{R}^m, v_i \in \mathbb{R}^m$, σ_Ξ is the support function of Ξ, A_ρ is a subset of a probability simplex, defined by

$$A_\rho = \left\{ y \in \mathbb{R}^N : \exists Z \sim \frac{1}{N}\sum_{i=1}^N \delta_{y_i},\ Z \in \mathcal{Z}_\rho \right\},$$

and $\mathcal{Z}_\rho$ denotes the risk envelope of ρ, i.e. $\mathcal{Z}_\rho = \{Z \geqslant 0 : \mathbb{E}[Z] = 1, \mathbb{E}[ZX] \leqslant \rho(X) \text{ for all } X\}$. It is defined that $\frac{p}{0} \notin A_\rho$ for any $p \neq \mathbf{0}$ and $\frac{\mathbf{0}}{0} \in A_\rho$.

We now demonstrate how Corollary 1 can be applied to solve GW-DRO problems when the ambiguity set $\mathbb{B}(\widehat{\mathbb{P}}_N)$ in (3) takes either the form of CVaR-Wasserstein ambiguity ball $\mathbb{B}^\alpha_{(1),\varepsilon}\left(\widehat{\mathbb{P}}_N\right)$ or expectile-Wasserstein ball $\mathbb{B}^\alpha_{(2),\varepsilon}\left(\widehat{\mathbb{P}}_N\right)$.

Example 2. (CVaR-Wasserstein) Note that CVaR_α can be represented as (Theorem 4.52 of Föllmer and Schied (2016))

$$\mathrm{CVaR}_\alpha(X) = \sup_{Z \in \mathcal{Z}_{\mathrm{CVaR}_\alpha}} \mathbb{E}[ZX] \quad \text{with} \quad \mathcal{Z}_{\mathrm{CVaR}_\alpha} = \left\{ Z \geqslant 0 : \mathbb{E}[Z] = 1, Z \leqslant \frac{1}{1-\alpha} \right\}.$$

Hence, we have $A_{\mathrm{CVaR}_\alpha} = \{y \in \mathbb{R}^N_+ : \sum_{i=1}^N y_i = 1, y_i \leqslant 1/(1-\alpha)\}$. Thus, following Corollary 1, the problem (3) with $\mathbb{B}(\widehat{\mathbb{P}}_N) = \mathbb{B}^\alpha_{(1),\varepsilon}\left(\widehat{\mathbb{P}}_N\right)$ can be solved by

$$\begin{aligned} &\inf_{\lambda,p,s,z_i,v_i} && \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^N s_i \\ &\text{subject to} && (16) - (18), \quad p_i \leqslant \frac{\lambda}{1-\alpha}, \quad i = 1, ..., N. \end{aligned}$$

Example 3. (Expectile-Wasserstein) Note that expectile e_α can be represented as (Proposition 8 of Bellini et al. (2014))

$$e_\alpha(X) = \sup_{Z \in \mathcal{Z}_{e_\alpha}} \mathbb{E}[XZ] \quad \text{with} \quad \mathcal{Z}_{e_\alpha} = \left\{ Z \geqslant 0 : \mathbb{E}[Z] = 1, \frac{\text{ess-sup} Z}{\text{ess-inf} Z} \leqslant \frac{\alpha}{1-\alpha} \right\}.$$

Hence, we have $A_{e_\alpha} = \{y \in \mathbb{R}^N_+ : \sum_{i=1}^N y_i = 1, \max y / \min y \leqslant \alpha/(1-\alpha)\}$. Thus, following Corollary 1, the problem (3) with $\mathbb{B}(\widehat{\mathbb{P}}_N) = \mathbb{B}^\alpha_{(2),\varepsilon}\left(\widehat{\mathbb{P}}_N\right)$ can be solved by

$$\begin{aligned} &\inf_{\lambda,p,s,z_i,v_i} && \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^N s_i, \\ &\text{subject to} && (16) - (18), \quad p_i \leqslant \frac{\alpha}{1-\alpha} p_j, \quad i, j = 1, ..., N. \end{aligned}$$

4 Solving GW-DRO with convex loss functions

Similar to the analysis presented in the previous section, we study the case of convex loss functions by investigating first how the worst-case expectation problem may be solved from a primal perspective. We show below that under the piecewise linear assumption, the worst-case expectation problem can also be reduced to a finite-dimensional optimization problem for any ambiguity set defined based on coherent Wasserstein metrics.

Theorem 2. *In the case where the loss function ℓ is convex piecewise linear, i.e. $\ell = \max_{k=1,\dots,K} \ell_k$, where ℓ_k, $k = 1, \dots, K$, are linear loss functions, the worst-case expectation problem (14) is equiv-*

alent to

$$\sup_{p_{ij},\xi_{ij}} \quad \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}\ell_j(\xi_{ij}) \tag{19}$$

$$\text{subject to} \quad \rho^{\Pi}(\|\widehat{\xi}-\xi\|) \leqslant \varepsilon, \tag{20}$$

$$\Pi((\widehat{\xi},\xi)=(\widehat{\xi}_i,\xi_{ij}))=p_{ij}\geqslant 0,\ \xi_{ij}\in\Xi,\ \forall\, i,j,$$

$$\sum_{j=1}^{K} p_{ij}=1,\quad i=1,\ldots,N.$$

The above result reveals the feasibility of solving the problem also as a finite-dimensional optimization problem for any ambiguity sets defined based on coherent Wasserstein metrics. The optimization problem (19), as it stands, differs from the finite-dimensional optimization problem (15) presented in the previous section in that it further requires determining the probability p_{ij} for each support ξ_{ij}, potentially rendering the problem (19) nonconvex. The tractability of the optimization problem (19) now depends more heavily on the exact specification of ρ and needs to be studied on a case-by-case basis. In the remainder of this section, we will focus on studying the worst-case expectation problem (14) in greater detail for CVaR-Wasserstien ambiguity sets $\mathbb{B}^{\alpha}_{(1),\varepsilon}\left(\widehat{\mathbb{P}}_N\right)$ and expectile-Wasserstein ambiguity sets $\mathbb{B}^{\alpha}_{(2),\varepsilon}\left(\widehat{\mathbb{P}}_N\right)$.

4.1 CVaR-Wasserstien ambiguity sets

By the well-known representation of CVaR (Rockafellar and Uryasev (2002)), $\mathrm{CVaR}_\alpha(X) = \inf_t\{t+\frac{1}{1-\alpha}\mathbb{E}[(X-t)_+]$, the problem (19) can be explicitly written as

$$\sup_{t,p_{ij},\xi_{ij}} \quad \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}\ell_j(\xi_{ij}) \tag{21}$$

$$\text{subject to} \quad t+\frac{1}{1-\alpha}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}(\|\xi_{ij}-\widehat{\xi}_i\|-t)_+ \leqslant \varepsilon,$$

$$\sum_{j=1}^{K} p_{ij}=1,\quad i=1,\ldots,N,\quad p_{ij}\geqslant 0,\ \xi_{ij}\in\Xi,\ \forall\, i,j.$$

The above problem is complicated by the need to handle the α-quantile variable t, a source of nonconvexity to the problem. Despite this nonconvexity, we show in this section that the above problem admits a more tractable reformulation in the case where the support set $\Xi=\mathbb{R}^m$. The reformulation not only enables us to demonstrate the tractability of the worst-case expectation problem $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)}\mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ for any Lipschitz continuous convex function ℓ, but also reveals a deep connection between the worst-case expectation problem $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ and the worst-case expectation problem formulated based on Wasserstein ambiguity sets, i.e. $\sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{W}}_{\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$.

Theorem 3. *In the case where the loss function ℓ is a convex function satisfying*

$$L := \max_{x\in\mathbb{R}^m} \|\partial\ell(x)\|_* < \infty$$

and $\Xi = \mathbb{R}^m$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, the worst-case expectation problem

$$\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)] \tag{22}$$

is equivalent to

$$\max\left\{ \sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{wc}}_{\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)],\ \sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)] \right\}, \tag{23}$$

where

$$\sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{wc}}_{\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)] = \frac{1}{N}\sum_{i=1}^{N} \max_{e:\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e), \tag{24}$$

and

$$\sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)] = \frac{1}{N}\sum_{i=1}^{N} \ell(\widehat{\xi}_i) + L(1-\alpha)\varepsilon. \tag{25}$$

The above result is striking for several reasons. First, it reveals via (23) an elegant connection between the worst-case expectation problem (22) and two other worst-case expectation problems, one formulated based on the worst-case ambiguity set $\mathbb{B}^{\mathrm{wc}}_{\varepsilon}\left(\widehat{\mathbb{P}}_N\right)$ and the other formulated based on the classical Wasserstein ambiguity set $\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)$ with the radius ε scaled by $1-\alpha$. These two latter problems can be solved respectively by a structurally simpler maximization problem (24) and in closed-form (25). Second, (23) is surprising because it implies that there is no loss of generality to reduce the set $\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)$ to the set

$$\overline{\mathbb{B}}^{\alpha}_{\varepsilon} := \mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N) \cup \mathbb{B}^{\mathrm{wc}}_{\varepsilon}(\widehat{\mathbb{P}}_N) \tag{26}$$

for solving the worst-case expectation problem. The latter is, however, a considerably smaller set of the former. To see this, note that the following set inclusion relationships follow straightforwardly the fact that $\mathbb{E}[\xi] \leqslant \mathrm{CVaR}_{\alpha}[\xi] \leqslant \frac{1}{(1-\alpha)}\mathbb{E}[\xi]$ for any nonnegative random variable ξ and $\alpha \in (0,1)$.

$$\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N) \subset \mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N) \subset \mathbb{B}^{\mathrm{W}}_{\varepsilon}(\widehat{\mathbb{P}}_N). \tag{27}$$

As α increases, the set $\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)$ converges towards $\widehat{\mathbb{P}}_N$ and thus is significantly smaller than $\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)$. The inclusion relationship $\mathbb{B}^{\mathrm{wc}}_{\varepsilon}(\widehat{\mathbb{P}}_N) \subset \mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)$ trivially holds, and the set $\mathbb{B}^{\mathrm{wc}}_{\varepsilon}(\widehat{\mathbb{P}}_N)$ is considerably smaller than $\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)$ in that $\mathbb{B}^{\mathrm{wc}}_{\varepsilon}(\widehat{\mathbb{P}}_N)$ contains only distributions whose support is uniformly bounded from the support of $\widehat{\mathbb{P}}_N$ by ε. Figure 2 demonstrates these inclusion relationships and highlights the considerable reduction from $\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)$ to $\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N) \cup \mathbb{B}^{\mathrm{wc}}_{\varepsilon}(\widehat{\mathbb{P}}_N)$.

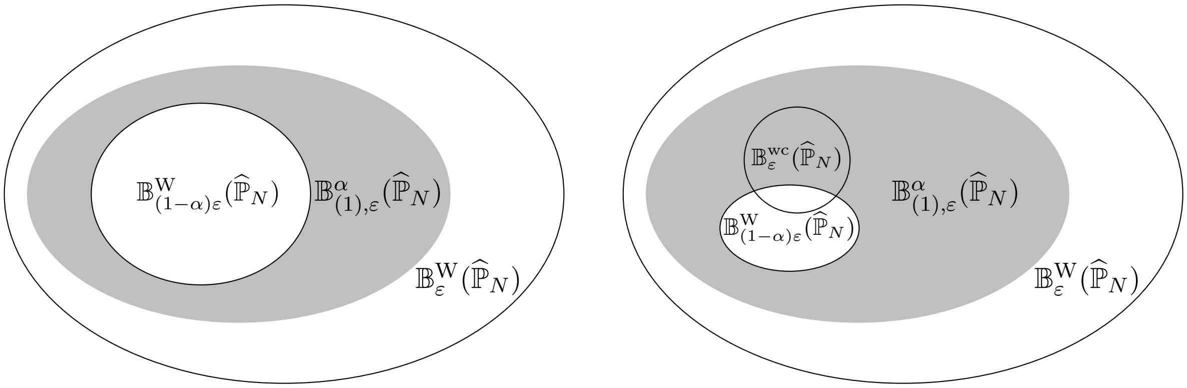

$\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)$
$\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)$
$\mathbb{B}^{\mathrm{W}}_{\varepsilon}(\widehat{\mathbb{P}}_N)$
$\mathbb{B}^{\mathrm{wc}}_{\varepsilon}(\widehat{\mathbb{P}}_N)$
$\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)$
$\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)$
$\mathbb{B}^{\mathrm{W}}_{\varepsilon}(\widehat{\mathbb{P}}_N)$

Figure 2: Relationships among the ambiguity sets

The figure and the set inclusion relationships (27) also demonstrate that while the worst-case expectation problem $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ can be bounded above and below respectively by $\sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{W}}_{\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ and $\sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$, neither of the two can be used as a reasonable proxy to $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$. The problem $\sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{W}}_{\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ is overly-conservative, i.e. less data-driven, whereas the problem $\sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ accounts for too little, or almost none, uncertainty, i.e. non-robust, as $\alpha \to 1$. Taking this perspective, we can further see how (23) sheds light on the underlying mechanism of the worst-case expectation problem $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ to offer data-driven evaluation of expected cost while maintaining some guaranteed level of robustness. Specifically, as $\alpha \to 1$, i.e. increasingly data-driven, the set $\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)$ in the reduced set $\overline{\mathbb{B}}^{\alpha}_{\varepsilon}$ (in (26)) would shrink towards $\widehat{\mathbb{P}}_N$, while at the same time the worst-case ambiguity set $\mathbb{B}^{\mathrm{wc}}_{\varepsilon}(\widehat{\mathbb{P}}_N)$ in $\overline{\mathbb{B}}^{\alpha}_{\varepsilon}$ guarantees the *minimum* level of robustness by taking into account any distribution with support maximally deviating by ε. The set $\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)$ can turn to be a set providing additional robustness when $\alpha \to 0$, in which case

$$\sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{wc}}_{\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)] < \sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]. \tag{28}$$

Third, this connection via (23) implies the following intriguing observation of the worst-case distributions for the worst-case expectation problem (22).

Corollary 2. *Under the condition of Theorem 3, if* $\max_{\|e\|\leqslant 1} \ell(\widehat{\xi}_i + \varepsilon e) > \ell(\widehat{\xi}_i)$ *for some* $\widehat{\xi}_i$*, then there always exists* $\alpha \in (0,1)$ *such that the worst-case expectation problem* (22) *is attainable, i.e. the worst-case distribution exists.*

This is in sharp contrast to the worst-case distributions for the worst-case expectation problem $\sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{W}}_{\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ formulated based on the Wasserstein ambiguity sets $\mathbb{B}^{\mathrm{W}}_{\varepsilon}(\widehat{\mathbb{P}}_N)$. Namely, as shown in Proposition 7 in the appendix, the worst-case distributions for the latter generally do not exist. The finding in Corollary 2 is difficult to identify directly from the property of the CVaR metric. It might be tempting to suppose that the worst-case distributions for (22) may not exist for any $\alpha \in [0,1)$, given that the CVaR metric is the conditional analog of the expectation for any $\alpha < 1$.

Finally, we demonstrate how Theorem 3 can be applied to solve problems such as portfolio optimization and machine learning mentioned in Section 2.3. Let us first make the following observation.

Corollary 3. *In the case where* $\ell(\xi) = f(x^\top \xi)$ *and* f *is a Lipschitz continuous convex function in* $\mathbb{R}$*, the worst-case expectation problem* (22) *can be solved by*

$$\min_{x\in\mathbb{X}} \max\left\{\frac{1}{N}\sum_{i=1}^{N} f(x^\top\widehat{\xi}_i) + \mathrm{Lip}(f)\|x\|_*(1-\alpha)\varepsilon, \frac{1}{N}\sum_{i=1}^{N}\max\left\{f_1(x^\top\widehat{\xi}_i - \varepsilon\|x\|_*),\ f_2(x^\top\widehat{\xi}_i + \varepsilon\|x\|_*)\right\}\right\},$$

where $f_1(t) = f(t)1_{\{t<t_0\}} + f(t_0)1_{\{t\geqslant t_0\}}$ *and* $f_2(t) = f(t_0)1_{\{t\leqslant t_0\}} + f(t)1_{\{t>t_0\}}$ *and* $t_0 \in [-\infty, \infty]$ *is such that* f *is decreasing on* $(-\infty, t_0)$ *and increasing on* (t_0, ∞)*.*

Example 4. (continue Example (ii), Section 2.3) Assuming the utility function u is Lipschitz continuous, we can apply Corollary 3 to the robust portfolio optimization problem

$$\min_{x\in\mathbb{X}} \sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[-u(\xi^\top x)]$$

by setting $f = -u$. We obtain the following convex program

$$\min_{x\in\mathbb{X}} \max\left\{\frac{1}{N}\sum_{i=1}^{N} -u(x^\top\widehat{\xi}_i) + \mathrm{Lip}(u)\|x\|_*(1-\alpha)\varepsilon, \frac{1}{N}\sum_{i=1}^{N} -u(x^\top\widehat{\xi}_i - \varepsilon\|x\|_*)\right\}.$$

Example 5. (continue Example (iii), Section 2.3) Let us consider first solving the distributionally robust regression problem

$$\min_{\beta\in\mathcal{B}} \sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\beta^\top\xi^x - \xi^y)],$$

where $\ell : \mathbb{R} \to \mathbb{R}_+$ is convex Lipschitz continuous. The function ℓ in regression is generally symmetric with respect to the origin and $\ell(0) = 0$. That is, $\ell(t) = \ell_1(t) + \ell_2(t)$, $\ell_1(t) = h(t_-)$ and $\ell_2(t) = h(t_+)$ for some non-decreasing convex function $h : \mathbb{R}_+ \to \mathbb{R}_+$, $h(0) = 0$. Applying Corollary 3, we arrive at the following convex program

$$\min_{\beta \in \mathcal{B}} \max \left\{ \begin{array}{l} \frac{1}{N}\sum_{i=1}^N \ell(\beta^\top \widehat{\xi}_i^x - \widehat{\xi}_i^y) + \mathrm{Lip}(\ell)\|(\beta,-1)\|_*(1-\alpha)\varepsilon, \\ \frac{1}{N}\sum_{i=1}^N \max\left\{\ell_1(\beta^\top \widehat{\xi}_i^x - \widehat{\xi}_i^y - \|(\beta,-1)\|_*\varepsilon), \ell_2(\beta^\top \widehat{\xi}_i^x - \widehat{\xi}_i^y + \|(\beta,-1)\|_*\varepsilon)\right\} \end{array} \right\}.$$

Next, consider solving the distributionally robust classification problem

$$\min_{\beta \in \mathcal{B}} \sup_{\mathbb{P} \in \mathbb{B}^\alpha_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi^y \cdot \beta^\top \xi^x)],$$

where $\ell : \mathbb{R} \to \mathbb{R}_+$ is non-increasing convex Lipschitz continuous. Applying Corollary 3, we arrive at the following convex program

$$\min_{\beta \in \mathcal{B}} \max \left\{ \frac{1}{N}\sum_{i=1}^N \ell(\widehat{\xi}_i^y \cdot \beta^\top \widehat{\xi}_i^x) + \mathrm{Lip}(\ell)\|\beta\|_*(1-\alpha)\varepsilon, \frac{1}{N}\sum_{i=1}^N \ell(\widehat{\xi}_i^y \cdot \beta^\top \widehat{\xi}_i^x - \|\beta\|_*\varepsilon) \right\}.$$

A list of Lipschitz continuous functions ℓ in regression and classification can be found for examples in Shafieezadeh-Abadeh et al. (2019).

A regularization perspective It is known that applying the type-1 Wasserstein DRO to statistical learning problems such as regression and classification is equivalent to solving a classical regularized empirical risk minimization problem (see e.g. Shafieezadeh-Abadeh et al. (2019).), i.e. (a):$\frac{1}{N}\sum_{i=1}^N f(x^\top \widehat{\xi}_i) + \mathrm{Lip}(f)\|x\|_*(1-\alpha)\varepsilon$ in Corollary 3, where the regularization term $\mathrm{Lip}(f)\|x\|_*(1-\alpha)\varepsilon$ controls the size of the decision variable x. One can see from Corollary 3 that applying GW-DRO to statistical learning problems essentially boils down to aggregating two different forms of regularized empirical problems, i.e. (a) and (b): $\frac{1}{N}\sum_{i=1}^N \max\left\{f_1(x^\top \widehat{\xi}_i - \varepsilon\|x\|_*),\ f_2(x^\top \widehat{\xi}_i + \varepsilon\|x\|_*)\right\}$, and GW-DRO determines which regularized problem, (a) or (b), to apply according to which one is more conservative, i.e. the one giving a larger value. Clearly, whether the regularized problem (a) or (b) is more conservative would depend on the exact value of the decision variable x, i.e. the choice of a regularized problem in GW-DRO is decision-dependent. One can observe that the value of (a) would tend to be larger than that of (b), when $\alpha \to 0$, since the regularization term $\mathrm{Lip}(f)\|x\|_*(1-\alpha)\varepsilon$ in (a) would become more dominating. In other words, GW-DRO would behave more similarly as the classical regularized problem as α decreases. This perspective that GW-DRO could serve as an ensemble of different regularized problems appears to be of high novelty. In particular, the idea of aggregating regularized problems by taking the pointwise maximum may offer a means to address the general challenge of regularization scheme selection.

4.2 Expectile-Wasserstein ambiguity sets

One may expect that the worst-case expectation problem $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)}\mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ defined over expectile-Wasserstein ambiguity sets is a more challenging, or less tractable, problem than the one studied in the previous section, i.e. $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)}\mathbb{E}^{\mathbb{P}}[\ell(\xi)]$, given the common belief that expectiles are more sophisticated forms of risk measures than CVaR. We show in this section that perhaps quite counter-intuitvely, the former is as tractable as, in fact generally more tractable than, the latter. In particular, in the case where $\rho = e_\alpha$, the finite-dimensional problem (19), as shown below, always admits a convex reformulation for any convex support set Ξ. The expectile-Wasserstein ambiguity sets $\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)$ could thus be a more appealing choice than the CVaR-Wasserstein ambiguity sets $\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)$ when the support set is constrained, i.e. $\Xi \neq \mathbb{R}^m$. We obtain the following two convex optimization reformulations, one in a maximization form and another in a minimization form.

Theorem 4. *In the case where the loss function ℓ is piecewise linear, taking the form of $\ell(x) = \max_{j=1,\ldots,K}\{a_j^\top x + b_j\}$, the worst-case expectation problem*

$$\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)] \tag{29}$$

is equivalent to the following convex maximization problem

$$\begin{aligned} \sup_{p_{ij}\geqslant 0, y_{ij}\in\mathbb{R}^m} \quad & \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\left(a_j^\top y_{ij} + (a_j^\top \widehat{\xi}_i + b_j)p_{ij}\right) \\ \text{subject to} \quad & \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}(\|y_{ij}\| - \varepsilon p_{ij})_+ + \frac{1-\alpha}{2\alpha-1}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\|y_{ij}\| \leqslant \frac{1-\alpha}{2\alpha-1}\varepsilon, \\ & \sum_{j=1}^{K} p_{ij} = 1, \quad i = 1,\ldots,N, \ \widehat{\xi}_i + \frac{y_{ij}}{p_{ij}} \in \Xi, \ \forall i,j, \end{aligned} \tag{30}$$

where y_{ij}/p_{ij} reads as ∞ if $y_{ij} \neq \mathbf{0}$, $p_{ij} = 0$, and $\mathbf{0}$ if $y_{ij} = \mathbf{0}$ and $p_{ij} = 0$. Moreover, the problem *is also equivalent to the following convex minimization problem*

$$\begin{aligned}
\inf_{\lambda, s, u_{ij}, v_{ij}} \quad & \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} s_i && && (31)\\
\text{subject to} \quad & b_j + \sigma_{\Xi}(u_{ij} + a_j) - u_{ij}^{\top}\xi_i + \|v_{ij}\|_* \varepsilon \leqslant s_i && \forall i, j,\\
& \|v_{ij}\|_* \leqslant \lambda && \forall i, j,\\
& \|u_{ij} + v_{ij}\|_* \leqslant \frac{1-\alpha}{2\alpha - 1}\lambda && \forall i, j,
\end{aligned}$$

where $\lambda \in \mathbb{R}, s \in \mathbb{R}^N, u_{ij} \in \mathbb{R}^m, v_{ij} \in \mathbb{R}^m$*, and* σ_Ξ *is the support function of* Ξ*.*

The formulation (30) is obtained from the problem (19) and can be used to compute the worst-case distributions, whereas the formulation (31) is the dual of (30) and can be used to solve the overall GW-DRO problem as a single minimization problem. The key observation from the above result, particularly the formulation (30), is that while expectiles are similar to CVaR in terms of general functional properties, i.e. both are coherent risk measures that are concave in distributions, the feasible sets of distributions induced from the two, i.e. the feasible set of (19), have different properties. Namely, the former is convex, whereas the latter is nonconvex. This demonstrates also why exploring different coherent risk measures in defining an ambiguity set can be useful. The above result may appear more limited than the result in the previous section in that the loss function ℓ is assumed to be piecewise linear. As another key finding, we show next that in the case where the support set Ξ is unconstrained, i.e. $\Xi = \mathbb{R}^m$, the worst-case expectation problem $\sup_{\mathbb{P} \in \mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ can also be solved more generally for any Lipschitz continuous convex function ℓ.

Theorem 5. *Under the condition of Theorem 3, the worst-case expectation problem*

$$\sup_{\mathbb{P} \in \mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)} \quad \mathbb{E}^{\mathbb{P}}[\ell(\xi)] \tag{32}$$

is equivalent to

$$\frac{1}{N}\sum_{i=1}^{N} \max\left\{\ell(\widehat{\xi}_i) + \beta L\varepsilon, \ \max_{\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e)\right\} \tag{33}$$

with $\beta = (1-\alpha)/\alpha$*.*

Besides shedding light on the tractability of solving the worst-case expectation problem (32) for a more general class of loss functions, the above result shows, rather unexpectedly, that the problem (32) also admits a structurally simple reformulation, i.e. (33), which is comparable to the reformulation (23) in the case of CVaR-Wasserstein ambiguity sets. Different from the reformulation (23), which requires comparing only the sample average of $\ell(\widehat{\xi}_i)+L(1-\alpha)\varepsilon$ and the sample average of $\max_{\|e\|=1}\ell(\widehat{\xi}_i+\varepsilon e)$, the reformulation (33) requires comparing first $\ell(\widehat{\xi}_i)+\beta L\varepsilon$ and $\max_{\|e\|=1}\ell(\widehat{\xi}_i+\varepsilon e)$ with respect to each sample $\widehat{\xi}_i$ before taking the sample average. Overall, the two formulations are very similar in terms of the total number of mathematical operations required for the computations. Some further insight about the reformulation (33) can be drawn from the structure of the worst-case distributions to the problem (32). Letting

$$I := \left\{ i = 1, \ldots, N : \ell(\widehat{\xi}_i) + \beta L\varepsilon > \max_{\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e) \right\},$$

one can verify that if $I \neq \emptyset$, then the discrete distributions

$$\mathbb{P}_n = \frac{1}{N}\sum_{i\notin I} \delta_{\widehat{\xi}_i+\varepsilon e_i} + \frac{1}{N}\sum_{i\in I}\left[\left(1-\frac{\varepsilon}{\|e_{ni}\|}\right)\delta_{\widehat{\xi}_i} + \frac{\varepsilon}{\|e_{ni}\|}\delta_{\widehat{\xi}_i+e_{ni}}\right], \tag{34}$$

where $e_i = \arg\max_{e:\|e\|=1}\ell(\widehat{\xi}_i+\varepsilon e)$, $i \notin I$, and $e_{ni} \in \mathbb{R}^m$ satisfies $\lim_{n\to\infty}\frac{\ell(\widehat{\xi}_i+e_{ni})-\ell(\widehat{\xi}_i)}{\|e_{ni}\|} = L$, $i \in I$, are feasible and asymptotically optimal to the problem (32) as $n \to \infty$. If $I = \emptyset$, then the worst-case distribution becomes

$$\mathbb{P}_\varepsilon = \frac{1}{N}\sum_{i=1}^{N}\delta_{\widehat{\xi}_i+\varepsilon e_i} \quad \text{with} \quad e_i = \arg\max_{\|e\|=1}\ell(\widehat{\xi}_i+\varepsilon e), \quad i = 1, \ldots, N.$$

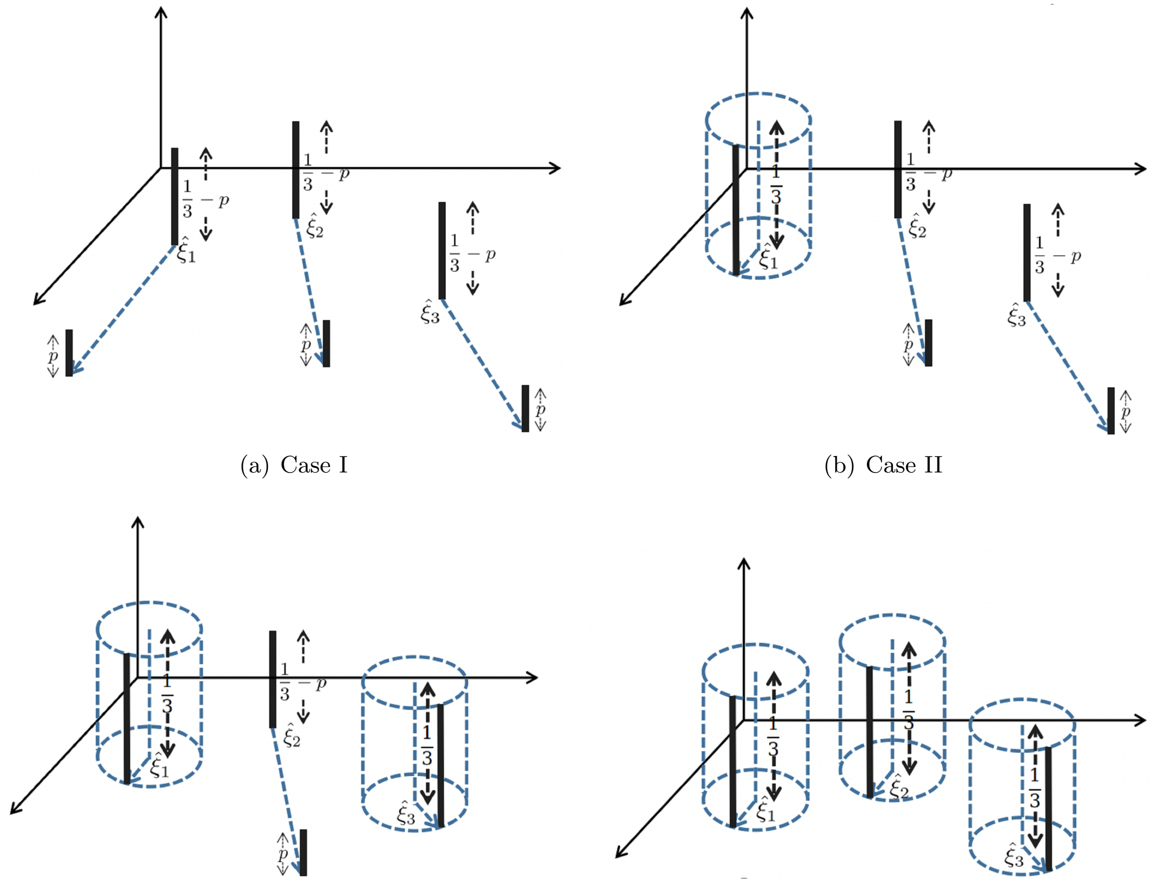
This second case, in particular, draws the connection between the problem $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ and the worst-case expectation problem $\sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{wc}}_{\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$, since $\mathbb{P}_\varepsilon$ is the worst-case distribution of the latter.

To best summarize and illustrate these distributions and their rich structure, we provide in Figure 1 an example based on three sample points. Each subfigure presents one representative structure of the worst-case distributions (or more precisely, distributions that are asymptotically optimal to the problem (32)). It is most useful to view these distributions by conditioning on each sample. In particular, the conditional distribution is either a two point mass distribution with an arbitrary small weight $p \to 0$ put on a point that is arbitrary far from a sample and the remaining weight $1-p$ put on the sample, or a distribution concentrated at a single point that is ε-away from the sample. The former is illustrated in the figure by placing no boundary (the cylinder) from a sample point, whereas the latter is illustrated by a cylinder with a fixed radius ε. These distributions are considerably richer than those of the Wasserstein worst-case expectation problem $\sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{W}}_{\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$

and the CVaR-Wasserstein worst-case expectation problem $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ in that

1. case I corresponds to the worst-case distributions of $\sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{W}}_{\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$, where conditional distributions with respect to all samples are not attainable,
2. case I and IV correspond to the worst-case distributions of $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$, where conditional distributions with respect to all samples are either all unattainable or all attainable,
3. and case I to IV correspond to the worst-case distributions of $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$, where the conditional distribution with respect to each sample is either attainable or unattainable.

In short, the structure of the worst-case distributions for the expectile-Wasserstein worst-case expectation problem $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ can flexibly vary with respect to each sample.



(a) Case I

(b) Case II

(c) Case III

(d) Case IV

Figure 3: Worst-case distributions for $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$

With this observation, one can also interpret the parameter $\alpha \in (1/2, 1)$ in the expectile-Wasserstein ambiguity sets $\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)$ as a parameter that fine tunes the number of sample points contaminated by only bounded perturbations. As $\alpha \to 1$, the number of such data points increases and the problem becomes increasingly data-driven and less conservative. The degree of conservativeness, reflected by the worst-case expected value, would more subtly depend on such a structural change of the worst-case distributions, as α varies. This can be seen by comparing the formulation (33) for the problem $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$ and the formulation (23) for the problem $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)]$. The former depends more nonlinearly on α in a convex fashion, whereas the latter is simply a two-piece linear function in α. Figure 4 illustrates this difference.

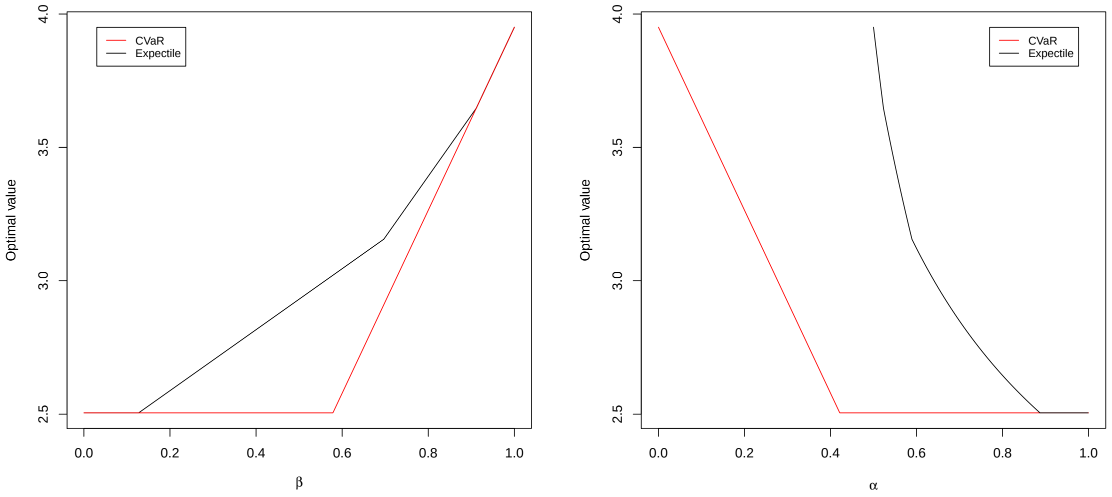


Figure 4: Optimal values of $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(1),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[f(x^\top\xi)]$ (denoted by CVaR) and $\sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[f(x^\top\xi)]$ (denoted by expectiles), where $f(t) = t^2 1_{\{\|t\|\leqslant 1\}} + (7|t|-6)1_{\{\|t\|>1\}}$ with $x = (1,2,-1)^\top$, $\varepsilon = 0.2$, and $\widehat{\mathbb{P}}_N := \frac{1}{3}\sum_{i=1}^{3}\delta_{\widehat{\xi}_i}$ with $\widehat{\xi}_1 = (0.2,-0.32,0.5)^\top$, $\widehat{\xi}_2 = (-0.2,-0.2,0.2)^\top$ and $\widehat{\xi}_3 = (0.3,-0.1,-0.1)^\top$. Left: The lines are with respect to $\beta = \beta_1 = 1-\alpha$ (CVaR) and $\beta = \beta_2 = (1-\alpha)/\alpha$ (Expectile). Right: The lines are with respect to α.

As done in the previous section, we identify below the conditions under which there always exists $\alpha < 1$ such that the worst-case distribution exists. It is worth noting that the condition is stronger than the condition identified in Corollary 2, but the two coincide when $\min_{x\in\mathbb{R}^m}\|\partial\ell(x)\| > 0$, i.e. the loss function ℓ does not contain any constant piece.

Corollary 4. *Under the condition of Theorem 3, if* $\max_{\|e\|\leqslant 1}\ell(\widehat{\xi}_i+\varepsilon e) > \ell(\widehat{\xi}_i)$ *for all* $i=1,\ldots,N$, *then there always exists* $\alpha\in(1/2,1)$ *such that the worst-case expectation problem (29) is attainable, i.e. the worst-case distribution exists.*

Finally, it is clear that Theorem 5 can be applied, just as Theorem 3, to solve problems in Section 2.3 as convex programs.

Corollary 5. *In the case where $\ell(\xi) = f(x^\top \xi)$ and f is a Lipschitz continuous convex function in $\mathbb{R}$, the worst-case expectation problem (32) can be solved by*

$$\min_{x\in\mathbb{X}} \frac{1}{N}\sum_{i=1}^{N} \max\left\{ f(x^\top \xi_i) + \mathrm{Lip}(f)\|x\|_*\beta\varepsilon,\ f_1(x^\top\xi_i - \varepsilon\|x\|_*),\ f_2(x^\top\xi_i + \varepsilon\|x\|_*)\right\}, \tag{35}$$

where $f_1(t) = f(t)1_{\{t<t_0\}} + f(t_0)1_{\{t\geqslant t_0\}}$ and $f_2(t) = f(t_0)1_{\{t\leqslant t_0\}} + f(t)1_{\{t>t_0\}}$ and $t_0 \in [-\infty,\infty]$ is such that f is decreasing on $(-\infty, t_0)$ and increasing on (t_0,∞).

Example 6. (continue Example (ii), Section 2.3) Applying Corollary 5 to the robust portfolio optimization problem

$$\min_{x\in\mathbb{X}} \sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[-u(\xi^\top x)],$$

where u is Lipschitz continuous, we obtain the following convex program

$$\min_{x\in\mathbb{X}} \frac{1}{N}\sum_{i=1}^{N} \max\left\{-u(x^\top\xi_i) + \mathrm{Lip}(u)\|x\|_*(1-\alpha)\varepsilon, -u(x^\top\xi_i - \varepsilon\|x\|_*)\right\}.$$

Example 7. (continue Example (iii), Section 2.3) Applying Corollary 5 to the distributionally robust regression problem

$$\min_{\beta\in\mathcal{B}} \sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\beta^\top\xi^x - \xi^y)]$$

and following Example 5, we obtain

$$\min_{\beta\in\mathcal{B}} \frac{1}{N}\sum_{i=1}^{N} \max\left\{ \begin{array}{l} \ell(\beta^\top\xi_i^x - \xi_i^y) + \mathrm{Lip}(\ell)\|(\beta,-1)\|_*(1-\alpha)\varepsilon, \\ \ell_1(\beta^\top\xi_i^x - \xi_i^y - \|(\beta,-1)\|_*\varepsilon), \ell_2(\beta^\top\xi_i^x - \xi_i^y + \|(\beta,-1)\|_*\varepsilon) \end{array}\right\},$$

whereas in the case of distributionally robust classification problem, i.e.

$$\min_{\beta\in\mathcal{B}} \sup_{\mathbb{P}\in\mathbb{B}^{\alpha}_{(2),\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi^y\cdot\beta^\top\xi^x)],$$

we obtain

$$\min_{\beta\in\mathcal{B}} \frac{1}{N}\sum_{i=1}^{N} \max\left\{\ell(\xi_i^y\cdot\beta^\top\xi_i^x) + \mathrm{Lip}(\ell)\|\beta\|_*(1-\alpha)\varepsilon,\ \ell(\xi_i^y\cdot\beta^\top\xi_i^x - \|\beta\|_*\varepsilon)\right\}.$$

5 Conclusion

In this paper, we propose a general framework for identifying families of Wasserstein metrics suited for data-driven distributionally robust optimization. We show that our framework offers a fruitful opportunity to design novel Wasserstein DRO models that can be theoretically sound and practically well-motivated. Necessary analysis is provided in this paper to facilitate tractable reformulations of the Wasserstein DRO models that adopt coherent Wasserstein metrics. We demonstrate the application of our framework using ambiguity sets constructed from CVaR- and expectile-Wasserstein metrics and provide an in-depth discussion of the connection between the new Wasserstein DRO models and the type-1 Wasserstein DRO model. The former generalizes the latter in an intuitive way, having a more enriched structure of optimization and worst-case distributions. In addition to several applications covered in this paper, the framework established in this paper shall provide an important basis for exploring further potential of Wasserstein DRO in a broader set of applications.

6 Appendix

6.1 Proofs of Section 2

Proof of Proposition 1. We show the result by proving the following two statements.

(i) For any $\alpha \in A$, if ρ_α takes finite value on L^1, then ℓ_α is a Lipschitz function.

(ii) If ℓ_α is a Lipschitz function for any $\alpha \in A$, then there does not exist a subsequence of ρ_α, $\alpha \in A$, converging to the worst-case risk measure ess-sup.

Obviously, with statements (i) and (ii), the **robustness** and **data-drivenness** of **DD-DRW** could not be satisfied simultaneously. We next show (i) and (ii).

To see (i), we show that ℓ_α is a Lipschitz function by contradiction. Suppose not, i.e., ℓ_α is not Lipschitz. We will construct a random variable X in L^1 whose risk measure $\rho_\alpha(X)$ takes infinity value. By the increase and convexity of ℓ_α, we have ℓ'_α is non-decreasing and unbounded, and thus, $\lim_{x\to\infty} \ell'_\alpha(x) = \infty$, where ℓ'_α is the left-derivative of ℓ_α. It then follows from the convexity of ℓ_α that $\ell_\alpha(x) \geqslant \ell_\alpha(x/2) + \ell'_\alpha(x/2)x/2$ for all $x > 0$, and thus,

$$\lim_{x\to\infty} \frac{\ell_\alpha(x)}{x} \geqslant \lim_{x\to\infty} \frac{1}{2}\ell'_\alpha\left(\frac{x}{2}\right) = \infty.$$

Therefore, for each $n \in \mathbb{N}$, there exists $x_n \geqslant n$ such that $\ell_\alpha(x_n) > 2^n x_n$, $n \in \mathbb{N}$. Define a random variable X such that $\mathbb{P}(X = x_n) = \frac{c}{2^n x_n}$, $n \geqslant 1$, where $c = 1/\sum_{n=1}^{\infty}(2^n x_n)^{-1}$. One can verify that $X \in L^1$ as $\mathbb{E}[X] = \sum_{n=1}^{\infty} \frac{c}{2^n} = c < \infty$, and meanwhile

$$\rho_\alpha(X) = \ell_\alpha^{-1}\left(\mathbb{E}[\ell_\alpha(X)]\right) = \ell_\alpha^{-1}\left(\sum_{n=1}^{\infty} \frac{c}{2^n x_n} \ell_\alpha(x_n)\right) \geqslant \ell_\alpha^{-1}\left(\sum_{n=1}^{\infty} c\right) = \infty,$$

where the inequality follows from $\ell_\alpha(x_n) > 2^n x_n$, $n \in \mathbb{N}$. Therefore, this yields a contradiction to that ρ_α takes finite value on L^1, and thus, (i) holds.

To see (ii), we also show it by contradiction. Suppose that there exist $\alpha_n \in A$, $n \in \mathbb{N}$, such that ρ_{α_n} converges to ess-sup. We first assert that there must exist a subsequence of $\{\alpha_n, n \in \mathbb{N}\}$, say β_n, $n \in \mathbb{N}$, and $c \in \mathbb{R}$ such that

$$\ell_{\beta_n} \text{ converges to } \infty 1_{\{(\cdot)>c\}} \text{ or } \infty 1_{\{(\cdot)\geqslant c\}} \text{ as } n \to \infty. \tag{36}$$

(Proving (36)). To see (36), for any fixed $x > 0$ and an event A with $\mathbb{P}(A) = p \in (0, 1)$, take $Y = x1_A$. By $\rho_{\alpha_n} \to$ ess-sup as $n \to \infty$, we have

$$\rho_{\alpha_n}(Y) = \ell_{\alpha_n}^{-1}(\mathbb{E}[\ell_{\alpha_n}(Y)]) = \ell_{\alpha_n}^{-1}(p\ell_{\alpha_n}(x)) \to x \text{ as } n \to \infty. \tag{37}$$

For the chosen x, denote $z_{\alpha_n} := \ell_{\alpha_n}(x)$, $n \in \mathbb{N}$. Note that if $\{z_{\alpha_n}, n \in \mathbb{N}\}$ is bounded, then there exists a subsequence α'_n of α_n, $n \in \mathbb{N}$, such that $z_{\alpha'_n} \to a$, i.e., $\ell_{\alpha_n}(x) \to a$ for some $a \in \mathbb{R}$. By the increase and convexity of $\ell_{\alpha'_n}$, (37) implies $pa = a$ and thus $a = 0$. Hence, $\lim_{n\to\infty} \ell_{\alpha_n}(x) = 0$. Take c as the supremum of x such that $\{\ell_{\alpha_n}(x), n \in \mathbb{N}\}$ is bounded. Then we have that $\lim_{n\to\infty} \ell_{\alpha_n}(x) = 0$ for $x < c$, and $\{\ell_{\alpha_n}(x), n \in \mathbb{N}\}$ is not bounded for $x > c$. We consider the following two cases.

(a) If $\{\ell_{\alpha_n}(c), n \in \mathbb{N}\}$ is not bounded, then there exist α'_n such that $\ell_{\alpha'_n}(c) \to \infty$. By the increase of $\ell_{\alpha'_n}$, we have $\lim_{n\to\infty} \ell_{\alpha'_n}(x) = \infty$ for $x \geqslant c$. Hence, $\ell_{\alpha'_n}(x)$ converges to $\infty 1_{\{(\cdot)\geqslant c\}}$ as $n \to \infty$, that is, (36) holds with $\beta_n = \alpha'_n$.

(b) If $\{\ell_{\alpha_n}(c), n \in \mathbb{N}\}$ is bounded, then by $\{\ell_{\alpha_n}(x), n \in \mathbb{N}\}$ is not bounded for $x > c$, for each $k \in \mathbb{N}$, we can find α_{n_k} such that $\ell_{\alpha_{n_k}}(x + 1/k) > k$. Then we have $\lim_{k\to\infty} \ell_{\alpha_{n_k}}(x) = \infty$ for $x > c$. Hence, $\ell_{\alpha_{n_k}}$ converges to $\infty 1_{\{(\cdot)>c\}}$ as $k \to \infty$, that is, (36) holds with $\beta_k = \alpha_{n_k}$.

Combining the above two cases, we have (36) holds.

Now with (36), let X be a random variable such that $\mathbb{P}(X = d) = p = 1 - \mathbb{P}(X = 0)$ for some

$d > c$ and $p \in (0,1)$. One can verify that

$$\lim_{n\to\infty} \rho_{\beta_n}(X) = \lim_{n\to\infty} \ell_{\beta_n}^{-1}(p\ell_{\beta_n}(d)) = \lim_{n\to\infty} \ell_{\beta_n}^{-1}(\infty) = c < d = \text{ess-sup}X,$$

yielding a contradiction to the assumption that $\rho_{\alpha_n} \to$ ess-sup and $\{\beta_n, n \in \mathbb{N}\} \subseteq \{\alpha_n, n \in \mathbb{N}\}$. Therefore, we have (ii) holds.

We thus complete the proof. □

The following lemma can be found in Theorem 4.2 of Bäuerle and Müller (2006) (see also Delbaen (2012)) which will be used in the proof of Proposition 2 and Theorem 2.

Lemma 1. *Any law-invariant convex (and thus coherent) risk measure that satisfies lower-semicontinuity in L^1 and $\rho(0) = 0$ must be consistent with increasing convex order, that is, if $X \preceq_{\text{icx}} Y$[4], then $\rho(X) \leqslant \rho(Y)$. In particular, $\mathbb{E}[X] \leqslant \rho(X)$.*

Proof of Proposition 2. (i) To show the "if" part, note that when $\mathbb{P}_1 = \mathbb{P}_2 = \mathbb{P}$, the joint distribution of (ξ, ξ) lies in the set $\Pi(\mathbb{P}_1, \mathbb{P}_2)$ where $\xi \sim \mathbb{P}$, and $\rho(\|\xi - \xi\|) = 0$. Hence, $d_\rho(\mathbb{P}_1, \mathbb{P}_2) = 0$. To show the "only if" part, suppose now that $d_\rho(\mathbb{P}_1, \mathbb{P}_2) = 0$. By Lemma 1, we have $\rho \geqslant \mathbb{E}$, and thus, $d_{\text{W}}(\mathbb{P}_1, \mathbb{P}_2) = 0$. By that the Wasserstein metric satisfies identity of indiscernibles, we have $\mathbb{P}_1 = \mathbb{P}_2$.

(ii) The symmetry follows directly from the definition.

(iii) Note that by the definition of d_ρ, for any $\varepsilon > 0$, there exist $\Pi_1 \in \Pi(\mathbb{P}_1, \mathbb{P}_2)$ and $\Pi_2 \in \Pi(\mathbb{P}_2, \mathbb{P}_3)$ such that

$$d_\rho(\mathbb{P}_1, \mathbb{P}_2) \geqslant \rho^{\Pi_1}(\|\xi_1 - \xi_2\|) - \varepsilon \quad \text{and} \quad d_\rho(\mathbb{P}_2, \mathbb{P}_3) \geqslant \rho^{\Pi_2}(\|\xi_1 - \xi_2\|) - \varepsilon.$$

By Theorem 6.10 of Kallenberg (1997), there exist $\xi_1^*, \xi_2^*, \xi_3^*$ such that (ξ_1^*, ξ_2^*) has the joint distribution Π_1 and (ξ_2^*, ξ_3^*) has the joint distribution Π_2. It follows that

$$\begin{aligned} d_\rho(\mathbb{P}_1, \mathbb{P}_2) + d_\rho(\mathbb{P}_2, \mathbb{P}_3) &\geqslant \rho(\|\xi_1^* - \xi_2^*\|) + \rho(\|\xi_2^* - \xi_3^*\|) - 2\varepsilon \\ &\geqslant \rho(\|\xi_1^* - \xi_3^*\|) - 2\varepsilon \geqslant d_\rho(\mathbb{P}_1, \mathbb{P}_3) - 2\varepsilon, \end{aligned}$$

where the second inequality follows from the subadditivity of ρ and the subadditivity of $\|\cdot\|$, and the last inequality follows from the definition of d_ρ and $\xi_i^* \sim \mathbb{P}_i$, $i = 1, 3$. By the arbitrariness of ε, we have $d_\rho(\mathbb{P}_1, \mathbb{P}_2) + d_\rho(\mathbb{P}_2, \mathbb{P}_3) \geqslant d_\rho(\mathbb{P}_1, \mathbb{P}_3)$.

[4] For two random variables X and Y, X is said to be smaller than Y with respect to increasing convex order, denoted by $X \preceq_{\text{icx}} Y$, if $\mathbb{E}[u(X)] \leqslant \mathbb{E}u(Y)$ for any increasing convex function u. It is easy to see that $X \succeq_{\text{icx}} Y$ if and only if $-Y \succeq_{\text{SSD}} -X$, i.e., $-Y$ is smaller than $-X$ in second-order stochastic dominance.

(iv) The non-negativity follows from the nonnegativity of the norm $\|\cdot\|$ and $\rho(X) \geqslant 0$ for $X \geqslant 0$. We thus complete the proof. $\square$

Proof of Proposition 3. To show the "if" part, for any $\alpha \in A$, with the representation (9), we have for any nonnegative random variable X

$$\rho_\alpha(X) = \sup_{g \in \mathcal{H}_{\rho_\alpha}} \left\{ \int_0^1 \mathrm{VaR}_\alpha(X) \mathrm{d}g(\alpha) \right\} \leqslant \sup_{g \in \mathcal{H}_{\rho_\alpha}} \|g'\|_\infty \int_0^1 \mathrm{VaR}_\alpha(X) \mathrm{d}\alpha = c_\alpha \mathbb{E}[X], \tag{38}$$

where the inequality follows from the Hölder inequality. This implies the robustness holds. By $c_{\alpha_n} \to \infty$ as $n \to \infty$, there exist $g_n \in \mathcal{H}_{\rho_{\alpha_n}}$, $n \in \mathbb{N}$, such that $\|g_n'\|_\infty > c_{\alpha_n} - 1/n$, and thus $\|g_n'\|_\infty = g_n'(1) \to \infty$ as $n \to \infty$. Note that $\|g_n'\|_\infty \in \mathbb{R}$ which implies g_n is Lipschitz continuous. We have $g_n(\alpha) \to 1_{\{\alpha=1\}}$, and thus,

$$\rho_{\alpha_n}(X) \geqslant \int_0^1 \mathrm{VaR}_\alpha(X) \mathrm{d}g_n(\alpha) \to \int_0^1 \mathrm{VaR}_\alpha(X) 1_{\{\alpha=1\}} = \text{ess-sup} X \quad \text{as } n \to \infty, \tag{39}$$

where the convergence follows from dominated convergence theorem as $\int_0^1 \mathrm{VaR}_\alpha(X) \mathrm{d}g_n(\alpha) \leqslant$ ess-sup$X$. Also, note that $\rho(c) = c$ and ρ is monotone, $\rho(X) \leqslant$ ess-sup$X$. This together with (39) implies $\rho_{\alpha_n} \to$ ess-sup as $n \to \infty$.

We next consider the "only if" part. First note the **robustness** property implies that ρ_α takes finite value in L^1. By Corollary 2.6 and Theorem 2.9 of Rüschendorf (2013), we have ρ_α must be L^1 continuous and thus can be represented by (9) with $\mathcal{H}_{\rho_\alpha}$ being a set of Lipschitz continuous convex distortion functions with $c_\alpha = \sup_{g \in \mathcal{H}_{\rho_\alpha}} \|g'\|_\infty < \infty$. Further, by the property of **data-drivenness**, there exist $\alpha_n \in A$, $n \in \mathbb{N}$, such that $\rho_{\alpha_n} \to$ ess-sup as $n \to \infty$. By (9), for a random variable X, there exist $g_n \in \mathcal{H}_\infty^{c_{\alpha_n}}, n \in \mathbb{N}$, such that

$$\int_0^1 \mathrm{VaR}_\alpha(X) \mathrm{d}g_n(\alpha) \to \text{ess-sup} X \quad \text{as } n \to \infty.$$

This holds only if $g_n(\alpha) \to 1_{\{\alpha=1\}}$ as $n \to \infty$. Hence, we have $c_{\alpha_n} \geqslant \|g_n'\|_\infty \to \infty$ as $n \to \infty$. This completes the proof. $\square$

Proof of Proposition 4. By assumption on ρ_α, it satisfies Proposition 3 with $c_\alpha = c$, and thus, (38) holds for any nonnegative random variable X. That is, $\rho_\alpha(X) \leqslant c\mathbb{E}[X]$ for all nonnegative X. Therefore, $d_{\rho_\alpha} \leqslant c d_{\mathrm{W}}$. It follows that $d_{\mathrm{W}}\left(\mathbb{P}, \widehat{\mathbb{P}}_N\right) \leqslant \varepsilon/c$ implies $d_{\rho_\alpha}\left(\mathbb{P}, \widehat{\mathbb{P}}_N\right) \leqslant \varepsilon$, and thus,

$$\mathbb{B}_{\varepsilon/c}^{\mathrm{W}}\left(\widehat{\mathbb{P}}_N\right) \subseteq \mathbb{B}_\varepsilon^{\rho_\alpha}\left(\widehat{\mathbb{P}}_N\right). \tag{40}$$

Recall that Theorem 3.2 of Esfahani and Kuhn (2018) gives

$$\mathbb{P}\left(\mathbb{B}^{\mathrm{W}}_{\varepsilon_N^{\mathrm{EK}}(\eta)}(\widehat{\mathbb{P}}_N)\right) \geqslant 1-\eta, \quad \text{where} \ \ \varepsilon_N^{\mathrm{EK}}(\eta) = \begin{cases} \varepsilon_0^{1/m_2} & \text{if } 1 \geqslant \varepsilon_0, \\ \varepsilon_0^{1/a} & \text{if } 1 < \varepsilon_0, \end{cases} \tag{41}$$

for some constants c_1, c_2 only depending on a, A and m. Obviously, we can write $\varepsilon_N(\eta) = c\varepsilon_N^{\mathrm{EK}}(\eta)$, and thus, the set inclusion (40) and (42) imply

$$\mathbb{P}\left(\mathbb{B}^{\rho_\alpha}_{\varepsilon_N(\eta)}(\widehat{\mathbb{P}}_N)\right) \geqslant \mathbb{P}\left(\mathbb{B}^{\mathrm{W}}_{\varepsilon_N^{\mathrm{EK}}(\eta)}(\widehat{\mathbb{P}}_N)\right) \geqslant 1-\eta. \tag{42}$$

Therefore, the finite sample guarantee (12) holds for $\varepsilon_N(\eta)$, which completes the proof. □

Proof of Proposition 5. Let c_1, c_2 be constants in (42). For $N \in \mathbb{N}$, define

$$\eta_N^{(1)} = \begin{cases} c_1 e^{-c_2 N \varepsilon_N^{m_2}}, & \varepsilon_N \leqslant 1, \\ c_1 e^{-c_2 N \varepsilon_N^{a}}, & \varepsilon_N > 1, \end{cases} \quad \text{and} \quad \eta_N^{(2)} = \begin{cases} c_1 e^{-c_2 N (\varepsilon_N/c)^{m_2}}, & \varepsilon_N \leqslant c, \\ c_1 e^{-c_2 N (\varepsilon_N/c)^{a}}, & \varepsilon_N > c. \end{cases}$$

One can verify that $\varepsilon_N \leqslant 1 \leqslant c$ for N large enough, and thus,

$$\eta_N^{(i)} = c_1 e^{-c_2 k_N}, \ i=1,2, \quad \varepsilon_N^{\mathrm{EK}}(\eta_N^{(1)}) = \varepsilon_N, \quad \varepsilon_N^{\mathrm{EK}}(\eta_N^{(2)}) = \frac{\varepsilon_N}{c},$$

where $\varepsilon_N^{\mathrm{EK}}(\eta_N^{(i)})$ is defined by (42). By assumptions on k_N and ε_N, we have

$$\sum_{N=1}^{\infty} \eta_N^{(i)} < \infty, \ i=1,2, \quad \varepsilon_N^{\mathrm{EK}}(\eta_N^{(1)}) \to 0, \quad \varepsilon_N^{\mathrm{EK}}(\eta_N^{(2)}) \to 0 \ \text{ as } \ N \to \infty.$$

That is, both $\eta_N^{(i)}$, $i=1,2$ satisfy the condition of Theorem 3.6 of Esfahani and Kuhn (2018). Denote by

$$\widehat{J}_N^{(1)} := \inf_{x \in \mathbb{X}} \sup_{\mathbb{P} \in \mathbb{B}^{\mathrm{W}}_{\varepsilon_N}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[h(x,\xi)] \quad \text{and} \quad \widehat{J}_N^{(2)} := \inf_{x \in \mathbb{X}} \sup_{\mathbb{P} \in \mathbb{B}^{\mathrm{W}}_{\varepsilon_N/c}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[h(x,\xi)].$$

We have $\mathbb{P}^\infty$-almost surely $\widehat{J}_N^{(i)} \downarrow J^\star$ as $N \to \infty$. By $d_{\mathrm{W}} \leqslant d_{\rho_\alpha} \leqslant c d_{\mathrm{W}}$, which implies

$$\mathbb{B}^{\mathrm{W}}_{\varepsilon_N/c}\left(\widehat{\mathbb{P}}_N\right) \subseteq \mathbb{B}^{\rho_\alpha}_{\varepsilon_N}\left(\widehat{\mathbb{P}}_N\right) \subseteq \mathbb{B}^{\mathrm{W}}_{\varepsilon_N}\left(\widehat{\mathbb{P}}_N\right) \tag{43}$$

and thus,

$$\widehat{J}_N^{(2)} \leqslant \widehat{J}_N \leqslant \widehat{J}_N^{(1)},$$

Therefore, we have $\mathbb{P}^\infty$-almost surely $\widehat{J}_N \downarrow J^\star$ as $N \to \infty$, that is, (i) holds. The assertion (ii) can be shown by similar arguments in the proof of (ii) of Theorem 3.6 of Esfahani and Kuhn (2018). □

6.2 Proofs of Section 3

Proof of Theorem 1. Note that the distribution Π with $\Pi((\widehat{\xi}, \xi) = (\widehat{\xi}_i, y_i)) = 1/N,\ i = 1, \ldots, N$ satisfies $\Pi \in \Pi(\widehat{\mathbb{P}}_N, \mathbb{P})$. We obviously have that the problem (14) is an upper bound of the problem (15). To show their equivalence, it suffices to show that the problem (15) is also an upper bound of the problem (14). To see this, for any $\Pi \in \mathcal{M}(\Xi^2)$ satisfying $\Pi \in \Pi(\widehat{\mathbb{P}}_N, \mathbb{P})$ and $\rho^\Pi(\|\widehat{\xi} - \xi\|) \leqslant \varepsilon$, define a new joint distribution Π^* as

$$\Pi^* := \frac{1}{N}\sum_{i=1}^N \delta_{(\widehat{\xi}_i, y_i)} \quad \text{with} \quad y_i = \mathbb{E}^\Pi[\xi | \widehat{\xi} = \widehat{\xi}_i], \quad i = 1, \ldots, N,$$

δ_x is the degenerated distribution at point x. Let $(\widehat{\xi}^*, \xi^*)$ be a random vector having the distribution Π^*. Denote by $\mathbb{P}^*$ the marginal distribution of ξ^*. We have the following facts.

(i) We have $\widehat{\xi}^* \sim \widehat{\mathbb{P}}_N$, that is, $\Pi^* \in \Pi(\widehat{\mathbb{P}}_N, \mathbb{P}^*)$. By the convexity of Ξ, we have $y_i \in \Xi$, $i = 1, \ldots, N$, and thus, $\Pi^* \in \mathcal{M}(\Xi^2)$.

(ii) Denote by $c(x, y) = \|x - y\|$ which is a convex function on $\mathbb{R}^m \times \mathbb{R}^m$. We have for any increasing convex function u, it holds that $v := u \circ c$ is a convex function as for any z_1, z_2, $\lambda \in [0, 1]$, it holds that

$$u(c(\lambda z_1 + (1-\lambda)z_2)) \leqslant u(\lambda c(z_1) + (1-\lambda)c(z_2)) \leqslant \lambda u(c(z_1)) + (1-\lambda)u(c(z_2)).$$

Hence, we have

$$\begin{aligned}
\mathbb{E}^{\Pi^*}[u(\|\widehat{\xi}^* - \xi^*\|)] &= \mathbb{E}^{\Pi^*}[v(\widehat{\xi}^*, \xi^*)] \\
&= \frac{1}{N}\sum_{i=1}^N v(\widehat{\xi}_i, y_i)] \\
&= \frac{1}{N}\sum_{i=1}^N v(\mathbb{E}^\Pi[(\widehat{\xi}, \xi)|\widehat{\xi} = \widehat{\xi}_i]) \\
&\leqslant \frac{1}{N}\sum_{i=1}^N \mathbb{E}^\Pi[v(\widehat{\xi}, \xi)|\widehat{\xi} = \widehat{\xi}_i] \\
&= \mathbb{E}^\Pi[v(\widehat{\xi}, \xi)] = \mathbb{E}^\Pi[u(\|\widehat{\xi} - \xi\|)],
\end{aligned}$$

where the inequality follows from the Jensen inequality. Noting the above inequality holds for any increasing function u, which implies that

$$\|\widehat{\xi}^* - \xi^*\| \leqslant_{\rm icx} \|\widehat{\xi} - \xi\|.$$

By Lemma 1, this implies $\rho^{\Pi^*}(\|\widehat{\xi}^* - \xi^*\|) \leqslant \rho^{\Pi}(\|\widehat{\xi} - \xi\|)$, and thus, $\rho^{\Pi^*}(\|\widehat{\xi}^* - \xi^*\|) \leqslant \varepsilon$.

(iii) Denote by $\mathbb{P}_i$ the conditional distribution of ξ given $\widehat{\xi} = \widehat{\xi}_i$ when the joint distribution of $(\widehat{\xi}, \xi)$ is Π. Noting that ℓ is concave, we have $\mathbb{E}^{\mathbb{P}_i}[\ell(\xi)] \leqslant \ell(\widehat{\xi}_i)$, $i = 1, \dots, N$. It follows that

$$\mathbb{E}^{\Pi}[\ell(\xi)] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}^{\mathbb{P}_i}[\ell(\xi)] \leqslant \frac{1}{N}\sum_{i=1}^{N} \ell(\widehat{\xi}_i) = \mathbb{E}^{\Pi^*}[\ell(\xi^*)].$$

Combining the above three facts, we have that the problem (15) is also an upper bound of the problem (14), which completes the proof. □

Proof of Corollary 1. We begin by noting that given any random variable Y such that $\mathbb{P}(Y = y_i) = 1/N$, $i = 1, \dots, N$, any law-invariant convex risk measure ρ can also be written as

$$\rho(Y) = \sup_{z \in A_\rho} \frac{1}{N}\sum_{i=1}^{N} y_i z_i. \tag{44}$$

(Proving (44)**).** First note that ρ has the dual representation (Föllmer and Schied (2016))

$$\rho(Y) = \sup_{Z \in \mathcal{Z}_\rho} \mathbb{E}[YZ], \tag{45}$$

where $\mathcal{Z}_\rho = \{Z \geqslant 0 : \mathbb{E}[Z] = 1, \mathbb{E}[ZX] \leqslant \rho(X) \text{ for all } X\}$. It follows that

$$\rho(Y) = \sup_{Z \in \mathcal{Z}_\rho} \left\{ \frac{1}{N}\sum_{i=1}^{N} y_i \mathbb{E}[Z|Y = y_i] \right\} = \sup_{Z \in \mathcal{Z}_\rho} \left\{ \frac{1}{N}\sum_{i=1}^{N} y_i z_i \right\} = \sup_{z \in A_\rho^*} \left\{ \frac{1}{N}\sum_{i=1}^{N} y_i z_i \right\},$$

where $z_i = \mathbb{E}^{\mathbb{P}}[Z|Y = y_i]$, $i = 1, \dots, N$, and

$$A_\rho^* = \{z \in \mathbb{R}_+^N : \mathbb{E}[Z|Y = y_i] = z_i, Z \in \mathcal{Z}_\rho\}.$$

It remains to show that $A_\rho^* = A_\rho$. Obviously, $A_\rho \subseteq A_\rho^*$. To see the other set inclusion, take any $z \in A_\rho^*$, i.e., there exists $Z \in \mathcal{Z}_\rho$ such that $\mathbb{E}[Z|Y = y_i] = z_i$, $i = 1, \dots, N$. Define $\Delta_z = \sum_{i=1}^{N} z_i \mathbb{1}_{\{Y = y_i\}}$. Noting $\mathbb{E}[\Delta_z|Y = y_i] = z_i$, $i = 1, \dots, N$, it suffices to show $\Delta_z \in \mathcal{Z}_\rho$. For any X, define $\Delta_X = \frac{1}{N}\sum_{i=1}^N x_i \mathbb{1}_{\{Y=y_i\}}$ with $x_i = \mathbb{E}[X|Y=y_i]$, $i=1,\dots,N$. By Jensen inequality, we have for any convex function u, $\mathbb{E}u(\Delta_X) \leqslant \mathbb{E}u(X)$, that is, $\Delta_X \leqslant_{\mathrm{cx}} X$. By Lemma 1, we have

$$\begin{aligned}
\rho(X) \geqslant \rho(\Delta_X) \geqslant \mathbb{E}[\Delta_X Z] \\
&= \frac{1}{N}\sum_{i=1}^N x_i \mathbb{E}[Z|Y=y_i] \\
&= \frac{1}{N}\sum_{i=1}^N x_i z_i \\
&= \frac{1}{N}\sum_{i=1}^N z_i \mathbb{E}[X|Y=y_i] = \mathbb{E}[\Delta_z X],
\end{aligned}$$

where the second inequality follows from $Z \in \mathcal{Z}_\rho$. Hence, we have $\Delta_z \in \mathcal{Z}_\rho$, and thus, $z \in A_\rho$. Therefore, $A_\rho = A_\rho^*$, and thus, (44) holds.

Using (44), we can write the problem (15) as

$$\begin{aligned}
\sup_{y_1,\dots,y_N} \quad & \frac{1}{N}\sum_{i=1}^N \ell(y_i), \\
\text{subject to} \quad & \sup_{z\in A_\rho}\left\{\frac{1}{N}\sum_{i=1}^N \|\widehat{\xi}_i - y_i\| z_i\right\} \leqslant \varepsilon, \\
& y_i \in \Xi, \quad i = 1, ..., N,
\end{aligned} \tag{46}$$

and its Lagrange dual by

$$\begin{aligned}
&\inf_{\lambda\geqslant 0}\sup_{y_i\in\Xi} \frac{1}{N}\sum_{i=1}^N \ell(y_i) + \lambda\left(\varepsilon - \sup_{z\in A_\rho}\left\{\frac{1}{N}\sum_{i=1}^N \|\widehat{\xi}_i - y_i\| z_i\right\}\right) \\
&= \inf_{\lambda\geqslant 0}\sup_{y_i\in\Xi}\inf_{z\in A_\rho} \frac{1}{N}\sum_{i=1}^N \ell(y_i) + \lambda\left(\varepsilon - \left\{\frac{1}{N}\sum_{i=1}^N \|\widehat{\xi}_i - y_i\| z_i\right\}\right) \\
&= \inf_{\lambda\geqslant 0, z\in A_\rho}\sup_{y_i\in\Xi} \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^N \left(\ell(y_i) - \lambda\|\widehat{\xi}_i - y_i\| z_i\right) \\
&= \inf_{\lambda\geqslant 0, z\in A_\rho} \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^N \sup_{y\in\Xi}\left(\ell(y) - \lambda\|\widehat{\xi}_i - y\| z_i\right),
\end{aligned} \tag{47}$$

where the second equality follows Sion's minimax theorem, given that the objective function is convex in z and concave in y_i, and the set A_ρ is a compact set. Strong duality holds for $\varepsilon > 0$, because the slater condition can be satisfied. Strong duality also holds for $\varepsilon = 0$, because both the primal and the dual reduces to $\frac{1}{N}\sum_{i=1}^N \ell(\widehat{\xi}_i)$.

Taking $p = \lambda z$, we can write the above problem equivalently as

$$\begin{aligned}
&\inf_{\lambda \geqslant 0, \frac{p}{\lambda} \in A_\rho} \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} \sup_{y \in \Xi} \left(\ell(y) - \|\widehat{\xi}_i - y\| p_i\right) \\
=&\left\{ \begin{array}{ll} \inf_{\lambda \geqslant 0, \frac{p}{\lambda} \in A_\rho, s_i} & \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} s_i \\ \text{subject to} & \sup_{y \in \Xi} \left(\ell(y) - \|\widehat{\xi}_i - y\| p_i\right) \leqslant s_i, \quad i = 1, ..., N \end{array} \right\}.
\end{aligned}$$

In the case $\lambda = 0$, $\frac{p}{0}$ is defined as an infeasible solution for any $p \neq \mathbf{0}$, denoted by $\frac{p}{0} \notin A_\rho$, and $\frac{\mathbf{0}}{0}$ is defined as a feasible solution, denoted by $\frac{\mathbf{0}}{0} \in A_\rho$. One can easily verify that this definition is consistent with the fact that the problem (47) reduces to $\frac{1}{N}\sum_{i=1}^{N} \sup_{y \in \Xi} \ell(y)$ when $\lambda = 0$.

Let χ_Ξ denote the characteristic function of Ξ and $f(y) := \|\widehat{\xi}_i - y\| p_i$. We have from the Fenchel duality that the left-hand-side of the above constraint, i.e. $\sup_y ([\ell(y) - \chi_\Xi(y)] - f(y))$, can be replaced by its dual and thus the constraint can be equivalently written as

$$\begin{aligned}
&\inf_{u_i} [-\ell + \chi_\Xi]^*(u_i) + f^*(-u_i) \leqslant s_i, \quad i = 1, ..., N \\
\Leftrightarrow &\exists u_i, \quad [-\ell + \chi_\Xi]^*(u_i) + f^*(-u_i) \leqslant s_i, \quad i = 1, ..., N,
\end{aligned}$$

where

$$[-\ell + \chi_\Xi]^*(u_i) = \inf_{\nu} \; [-\ell]^*(u_i - \nu) + \sigma_\Xi(\nu)$$

and

$$f^*(-u_i) = \begin{cases} -\widehat{\xi}_i^\top u_i, & \|u_i\|_* \leqslant p_i, \\ \infty, & \text{o.w..} \end{cases}$$

That is, the constraint can also be equivalently written as

$$\exists u_i, \nu_i, \quad [-\ell]^*(u_i - \nu_i) + \sigma_\Xi(\nu_i) - \widehat{\xi}_i^\top u_i \leqslant s_i, \quad \|u_i\|_* \leqslant p_i. \;\; i = 1, ..., N.$$

Finally, the fact that A_ρ must be a subset of a probability simplex, i.e. satisfying $\frac{p}{\lambda} \geqslant 0$ and $\vec{1}^\top \frac{p}{\lambda} = 1$ implies that $\sum_{i=1}^{N} p_i = \lambda$ and $p_i \geqslant 0$ must hold.

□

6.3 Proofs of Section 4

Proof of Theorem 2. Denote by $\Theta_j = \max\{x \in \mathbb{R}^m : \ell_j(x) = \ell(x)\}$, $j = 1, \ldots, K$. Without loss of generality assume that Θ_j, $j = 1, \ldots, K$, are disjoint. Similar as Theorem 1, we only show that the problem (19) is also an upper bound of the problem (14). For any $\Pi \in \Pi(\widehat{\mathbb{P}}_N, \mathbb{P})$ satisfying $\Pi \in \mathcal{M}(\Xi^2)$ and $\rho^\Pi(\|\widehat{\xi} - \xi\|) \leqslant \varepsilon$, define a new joint distribution Π^* as

$$\Pi^* = \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^K p_{ij}\delta_{(\widehat{\xi}_i, \xi_{ij})}, \quad \text{with} \quad p_{ij} = \mathbb{Q}_i(\Theta_j), \quad \xi_{ij} = \mathbb{E}^{\mathbb{Q}_i}[\xi|\Theta_j] = \mathbb{E}^\Pi[\xi|\Theta_j, \widehat{\xi} = \widehat{\xi}_i], \tag{48}$$

where $\mathbb{Q}_i$ is the conditional distribution of ξ given $\widehat{\xi} = \widehat{\xi}_i$ when the joint distribution of $(\widehat{\xi}, \xi)$ is Π. Let $(\widehat{\xi}^*, \xi^*) \in \mathbb{R}^m \times \mathbb{R}^m$ be a random vector having the distribution Π^*. Denote by $\mathbb{Q}_i^*$ the marginal distributions of ξ^* conditional on $\widehat{\xi} = \widehat{\xi}_i$, $i = 1, \dots, N$.

(i) We have $\widehat{\xi}^* \sim \widehat{\mathbb{P}}_N$, that is, $\Pi^* \in \Pi(\widehat{\mathbb{P}}_N, \mathbb{P}^*)$. By the convexity of Ξ, we have $\xi_{ij} \in \Xi$, $i = 1, \dots, N$, $j = 1, \dots, K$, and thus, $\Pi^* \in \mathcal{M}(\Xi^2)$.

(ii) Similar as (ii) in the proof of Theorem 1, for any increasing convex function u, define $v := u \circ c$ is a convex function, where $c(x, y) = \|x - y\|$. Hence, we have

$$\begin{aligned}
\mathbb{E}^{\Pi^*}[u(\|\widehat{\xi}^* - \xi^*\|)] &= \mathbb{E}^{\Pi^*}[v(\widehat{\xi}^*, \xi^*)] \\
&= \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^K p_{ij} v(\widehat{\xi}_i, \xi_{ij}) \\
&= \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^K p_{ij} v(\widehat{\xi}_i, \mathbb{E}^\Pi[\xi|\Theta_j, \widehat{\xi} = \widehat{\xi}_i]) \\
&= \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^K p_{ij} v(\mathbb{E}^\Pi[(\widehat{\xi}, \xi)|\Theta_j, \widehat{\xi} = \widehat{\xi}_i]) \\
&\leqslant \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^K p_{ij} \mathbb{E}^\Pi[v(\widehat{\xi}, \xi)|\Theta_j, \widehat{\xi} = \widehat{\xi}_i] \\
&= \mathbb{E}^\Pi[v(\widehat{\xi}, \xi)] = \mathbb{E}^\Pi[u(\|\widehat{\xi} - \xi\|)],
\end{aligned}$$

where the inequality follows from the Jensen inequality. Therefore, we have $\rho^{\Pi^*}(\|\widehat{\xi}^* - \xi^*\|) \leqslant \varepsilon$.

(iii) We have

$$\mathbb{E}^{\mathbb{Q}_i}[\ell(\xi)] = \sum_{j=1}^K \int_{\Theta_j} \ell_j(\xi) d\mathbb{Q}_i = \sum_{j=1}^K p_{ij}\ell_j(\xi_{ij}) = \sum_{j=1}^K \int_{\Theta_j} \ell_j(\xi) d\mathbb{Q}_i^* \leqslant \sum_{j=1}^K \int_{\Theta_j} \ell(\xi) d\mathbb{Q}_i^* = \mathbb{E}^{\mathbb{Q}_i^*}[\ell(\xi)],$$

where the second equality follows from the linearity of ℓ_j, and the inequality follows from

$\ell_j \leqslant \ell$, $j = 1, \ldots, K$. Therefore, we have

$$\mathbb{E}^{\Pi}[\ell(\xi)] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}^{\mathbb{Q}_i}[\ell(\xi)] \leqslant \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}^{\mathbb{Q}_i^*}[\ell(\xi)] = \mathbb{E}^{\Pi^*}[\ell(\xi)].$$

Combining the above three facts, we have that the problem (14) is equivalent to

$$\begin{aligned} \sup_{p_{ij}, x_{ij}} \quad & \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}\ell(\xi_{ij}) \\ \text{subject to} \quad & \rho^{\Pi^*}(\|\widehat{\xi} - \xi\|) \leqslant \varepsilon, \\ & \Pi^*(\{(\widehat{\xi}_i, \xi_{ij})\}) = p_{ij} \geqslant 0,\ \xi_{ij} \in \Xi,\ \forall i, j, \\ & \sum_{j=1}^{K} p_{ij} = 1, \quad i = 1, \ldots, N. \end{aligned} \tag{49}$$

Note that we have shown that the original optimization problem (14) is bounded by the problem (19) which is again bounded by the problem (49) as $\ell_i \leqslant \ell$ for $i = 1, \ldots, K$. Therefore, by the equivalence between the problems (14) and (49), we have the optimization problem (14) is equivalent to the problem (19). This completes the proof. □

Proof of Theorem 3. We first consider the case that the objective function ℓ is a piecewise linear function, that is, $\ell = \max_{k=1,\ldots,K} \ell_k$, where $\ell_k(x) = a_k^\top x + b_k$, $k = 1, \ldots, K$. Without loss of generality, assume

$$\|a_1\|_* \leqslant \ldots \leqslant \|a_K\|_* \ \text{ and denote by } \ z_{ij} := \ell_j(\widehat{\xi}_i) = a_j^\top \widehat{\xi}_i + b_j,\ \forall\, i, j.$$

By Theorem 2, we have the optimization problem (22) is equivalent to

$$\begin{aligned} \sup_{t,\, p_{ij} \in \mathbb{R},\, x_{ij} \in \mathbb{R}^m} \quad & \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}(a_j^\top(x_{ij} + \widehat{\xi}_i) + b_j) \\ \text{subject to} \quad & t + \frac{1}{1-\alpha}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}(\|x_{ij}\| - t)_+ \leqslant \varepsilon, \\ & \sum_{j=1}^{K} p_{ij} = 1, \quad i = 1, \ldots, N, \quad p_{ij} \geqslant 0,\ \forall\, i, j. \end{aligned} \tag{50}$$

Note that for any $x_{ij} \in \mathbb{R}^m$, by taking $\widehat{x}_j = \arg\max_{x \in \mathbb{R}^m : \|x\|=1} a_j^\top x$ and $x_{ij}^* = \|x_{ij}\|\widehat{x}_j$, one can verify that $\|x_{ij}^*\| = \|x_{ij}\|$ and

$$a_j^\top x_{ij}^* = \|x_{ij}\| a_j^\top \widehat{x}_j = \|x_{ij}\| \|a_j\|_* \geqslant a_j^\top x_{ij}.$$

Therefore, it suffices to consider the x_{ij} that takes the form of $x_{ij} = y_{ij}\widehat{x}_j$, $y_{ij} \in \mathbb{R}_+$. By taking $x_{ij} = y_{ij}\widehat{x}_j$, we have the problem (50) is equivalent to

$$\sup_{t, p_{ij}, y_{ij} \in \mathbb{R}_+} \quad \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\|a_j\|_* + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}z_{ij} \tag{51}$$

$$\text{subject to} \quad t + \frac{1}{1-\alpha}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}(y_{ij} - tp_{ij})_+ \leqslant \varepsilon, \tag{52}$$

$$\sum_{j=1}^{K} p_{ij} = 1, \quad i = 1, \ldots, N. \tag{53}$$

For any $p_{ij} \geqslant 0$ and $y_{ij} \geqslant 0$ satisfying conditions (52) and (53), if for some i, there exists $j \neq K$ such that $y_{ij} > tp_{ij}$, then by taking $y_{ij}^* := tp_{ij}$ and $y_{iK}^* := y_{iK} + y_{ij} - tp_{ij} > y_{iK}$, one can verify that

$$t + \frac{1}{1-\alpha}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}(y_{ij}^* - tp_{ij})_+ \leqslant \varepsilon$$

and the objective function becomes larger with y_{ij}^*. Therefore, it suffices to consider the case that

$$y_{ij} \leqslant tp_{ij}, \quad j \neq K, \quad t + \frac{1}{1-\alpha}\frac{1}{N}\sum_{i=1}^{N}(y_{iK} - tp_{iK})_+ \leqslant \varepsilon.$$

In this case, we have the problem (51) is equivalent to

$$\sup_{t, p_{ij}, y_{ij} \in \mathbb{R}_+} \quad \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\|a_j\|_* + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}z_{ij} \quad \text{with} \quad z_{ij} = (a_j^\top \widehat{\xi}_i + b_j)$$

$$\text{subject to} \quad y_{ij} \leqslant tp_{ij}, \quad j < K, \quad t + \frac{1}{1-\alpha}\frac{1}{N}\sum_{i=1}^{N}(y_{iK} - tp_{iK})_+ \leqslant \varepsilon \text{ and (53)},$$

that is,

$$\sup_{t, p_{ij}, y_{ij} \in \mathbb{R}_+} \quad \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{K-1} tp_{ij}\|a_j\|_* + y_{iK}\|a_K\|_*\right) + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}z_{ij}$$

$$\text{subject to} \quad t + \frac{1}{1-\alpha}\frac{1}{N}\sum_{i=1}^{N}(y_{iK} - tp_{iK})_+ \leqslant \varepsilon \text{ and (53)}.$$

Take $y_{iK} = x_i p_{iK}$, $i = 1, \dots, N$. We can rewrite it as

$$\begin{aligned} &\sup_{t,p_{ij},x_i \in \mathbb{R}_+} \quad \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{K-1} t p_{ij}\|a_j\|_* + x_i p_{iK}\|a_K\|_*\right) + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij} z_{ij} \\ &\text{subject to} \quad t + \frac{1}{1-\alpha}\frac{1}{N}\sum_{i=1}^{N} p_{iK}(x_i - t)_+ \leqslant \varepsilon \text{ and (53).} \end{aligned}$$

For any feasible solution x_i, p_{ij}, t, $i = 1, \dots, N$, $j = 1, \dots, K$, define $x_i^* := \sum_{i=1}^N p_{iK} x_i / (\sum_{i=1}^N p_{iK}) =: \overline{x}$, $i = 1, \dots, N$. By Jensen inequality, we have $t + \frac{1}{1-\alpha}\frac{1}{N}\sum_{i=1}^N p_{iK}(\overline{x} - t)_+ \leqslant \varepsilon$ and the objective function remains unchanged. Therefore, the above optimization problem is equivalent to

$$\begin{aligned} &\sup_{t,p_{ij},\overline{x} \in \mathbb{R}_+} \quad t\frac{1}{N}\sum_{i=1}^{N}\sum_{j \neq K} p_{ij}\|a_j\|_* + \overline{x}\|a_K\|_* \frac{1}{N}\sum_{i=1}^{N} p_{iK} + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij} z_{ij} \\ &\text{subject to} \quad t + \frac{1}{1-\alpha}\frac{1}{N}\sum_{i=1}^{N} p_{iK}(\overline{x} - t)_+ \leqslant \varepsilon \text{ and (53).} \end{aligned}$$

Noting that the objective function is increasing in $\overline{x}$, we have $\overline{x} \geqslant x$ holds automatically. Taking $\delta = \overline{x} - t$, e the above optimization problem is equivalent to

$$\begin{aligned} &\sup_{t,p_{ij},\overline{x} \in \mathbb{R}_+} \quad t\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}\|a_j\|_* + \delta\|a_K\|_* \frac{1}{N}\sum_{i=1}^{N} p_{iK} + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij} z_{ij} \\ &\text{subject to} \quad t + \frac{1}{1-\alpha}\frac{1}{N}\sum_{i=1}^{N} p_{iK}\delta \leqslant \varepsilon \text{ and (53).} \end{aligned}$$

Denote by $u = \frac{1}{1-\alpha}\frac{1}{N}\sum_{i=1}^N p_{iK}\delta$. We have the problem (50) is equivalent to

$$\begin{aligned} &\sup_{t,p_{ij},u \in \mathbb{R}_+} \quad t\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}\|a_j\|_* + \|a_K\|_*(1-\alpha)u + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij} z_{ij} \\ &\text{subject to} \quad t + u \leqslant \varepsilon \text{ and (53).} \end{aligned}$$

Since the objective function is increasing in both t and u, the constraint $t + u \leqslant \varepsilon$ can be replaced by $t + u = \varepsilon$ without loss of generality. Letting $u = \varepsilon - t$, we have it is equivalent to

$$\begin{aligned} &\sup_{t,p_{ij}} \ \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}(\|a_j\|_* t + z_{ij}) - t\|a_K\|_*(1-\alpha) + \|a_K\|_*(1-\alpha)\varepsilon \ \text{ subject to } \ (53) \\ &\quad = \sup_{t \in [0,\varepsilon]} \frac{1}{N}\sum_{i=1}^{N} \max_j(\|a_j\|_* t + z_{ij}) - t\|a_K\|_*(1-\alpha) + \|a_K\|_*(1-\alpha)\varepsilon. \end{aligned}$$

Note that the objective function is convex in t. We therefore have that the optimal t is either 0 or ε, and the optimal value of the problem (50) is

$$\begin{aligned}&\max\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{j}\{\|a_j\|_*\varepsilon+z_{ij}\},\ \frac{1}{N}\sum_{i=1}^{N}\max_{j}z_{ij}+\|a_K\|_*(1-\alpha)\varepsilon\right\}\\&=\max\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{e:\|e\|=1}\ell(\widehat{\xi}_i+\varepsilon e),\ \frac{1}{N}\sum_{i=1}^{N}\ell(\widehat{\xi}_i)+\|a_K\|_*(1-\alpha)\varepsilon\right\}.\end{aligned}$$

Moreover, one can verify the following statements about the worst-case distribution.

(1) If $\frac{1}{N}\sum_{i=1}^{N}\max_{e:\|e\|=1}\ell(\widehat{\xi}_i+\varepsilon e)\geqslant\frac{1}{N}\sum_{i=1}^{N}\ell(\widehat{\xi}_i)+L(1-\alpha)\varepsilon$, then the optimal solution to the problem (50) is $t^*=\varepsilon$,

$$x^*_{ij}=\widehat{\xi}_i+\varepsilon e^*_i\ \text{ with }\ e^*_i=\arg\max_{e:\|e\|=1}\ell(\widehat{\xi}_i+\varepsilon e),\quad p^*_{ij_i}=1\ \text{ for } j_i \text{ such that } e^*_i=\frac{a_{j_i}}{\|a_{j_i}\|}.$$

(2) If $\frac{1}{N}\sum_{i=1}^{N}\ell(\widehat{\xi}_i)+\|a_K\|_*(1-\alpha)\varepsilon>\frac{1}{N}\sum_{i=1}^{N}\max_{e:\|e\|=1}\ell(\widehat{\xi}_i+\varepsilon e)$, then the optimal value can be approached by the following feasible solutions: $t^*=0$,

$$x_{ij}=\widehat{\xi}_i\ \text{ for } j\neq K,\quad x_{iK}=\widehat{\xi}_i+(1-\alpha)\frac{a_K}{\|a_K\|}\frac{\varepsilon}{p_{iK}},\quad p_{iK}\downarrow 0.$$

Now, consider the general convex function ℓ. By assumption on ℓ, there exist ℓ_n,ℓ^*_n, $n\in\mathbb{N}$ such that $\ell_n\leqslant\ell\leqslant\ell^*_n$, ℓ_n and ℓ^*_n are piecewise linear functions, both ℓ_n and ℓ^*_n converge to ℓ, and $\lim_{n\to\infty}\overline{\|\nabla\ell_n\|_*}=\lim_{n\to\infty}\overline{\|\nabla\ell^*_n\|_*}=L$. Denote by $\bar{\ell}$ the worst-case value of the problem (22). By the result for piecewise linear functions, we have

$$\begin{aligned}&\max\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{e:\|e\|=1}\ell_n(\widehat{\xi}_i+\varepsilon e),\ \frac{1}{N}\sum_{i=1}^{N}\ell_n(\widehat{\xi}_i)+\overline{\|\nabla\ell_n\|_*}(1-\alpha)\varepsilon\right\}\\&\leqslant\bar{\ell}\leqslant\max\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{e:\|e\|=1}\ell^*_n(\widehat{\xi}_i+\varepsilon e),\ \frac{1}{N}\sum_{i=1}^{N}\ell^*_n(\widehat{\xi}_i)+\overline{\|\nabla\ell^*_n\|_*}(1-\alpha)\varepsilon\right\}.\end{aligned}$$

Letting $n\to\infty$, we obtain that

$$\bar{\ell}=\max\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{e:\|e\|=1}\ell(\widehat{\xi}_i+\varepsilon e),\ \frac{1}{N}\sum_{i=1}^{N}\ell(\widehat{\xi}_i)+L(1-\alpha)\varepsilon\right\}.$$

Moreover, we have the following statements about the worst-case distribution.

(1) If $\frac{1}{N}\sum_{i=1}^{N}\max_{e:\|e\|=1}\ell(\widehat{\xi}_i+\varepsilon e)\geqslant\frac{1}{N}\sum_{i=1}^{N}\ell(\widehat{\xi}_i)+L(1-\alpha)\varepsilon$, then the optimal solution to the

problem (50) is $x^* = \varepsilon$,

$$\xi_i^* = \widehat{\xi}_i + \varepsilon e_i^* \quad \text{with} \quad e_i^* = \arg \max_{e:\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e) \quad \mathbb{P}_i\text{-a.s.}, \quad i = 1, \ldots, N.$$

(2) If $\frac{1}{N}\sum_{i=1}^N \ell(\widehat{\xi}_i) + L(1-\alpha)\varepsilon > \frac{1}{N}\sum_{i=1}^N \max_{e:\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e)$, then the optimal value can be approached by the following feasible solutions: $x^* = 0$,

$$\xi_i^* = \widehat{\xi}_i, \text{ for } i = 2, \ldots, N, \quad \mathbb{P}_1(\xi_1^* = \widehat{\xi}_1 + e_n) = \frac{(1-\alpha)\varepsilon}{\|e_n\|}, \quad \mathbb{P}_1(\xi_1^* = \widehat{\xi}_1) = 1 - \frac{(1-\alpha)\varepsilon}{\|e_n\|}, \quad n \in \mathbb{N}.$$

where $e_n \in \mathbb{R}^m$, $n \in \mathbb{N}$, satisfy $\lim_{n\to\infty} \frac{\ell(\widehat{\xi}_1 + e_n) - \ell(\xi_1)}{\|e_n\|} = L$.

We thus complete the proof. □

Proof of Corollary 2. Since $\frac{1}{N}\sum_{i=1}^N \max_{e:\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e) = \frac{1}{N}\sum_{i=1}^N \max_{\|e\|\leqslant 1} \ell(\widehat{\xi}_i + \varepsilon e) > \frac{1}{N}\sum_{i=1}^N \ell(\widehat{\xi}_i)$ and $L(1-\alpha)\varepsilon$ in (25) is strictly decreasing in $\alpha \in [0,1]$, there must exist $\alpha \in (0,1)$ large enough such that

$$\sup_{\mathbb{P}\in\mathbb{B}_\varepsilon^{\mathrm{wc}}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)] = \frac{1}{N}\sum_{i=1}^N \max_{e:\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e) > \frac{1}{N}\sum_{i=1}^N \ell(\widehat{\xi}_i) + L(1-\alpha)\varepsilon = \sup_{\mathbb{P}\in\mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)].$$

In the case, the optimal value (23) is $\frac{1}{N}\sum_{i=1}^N \max_{e:\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e)$ which is attainable by the distribution $\mathbb{P}^* = \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\xi}_i + \varepsilon e_i^*}$, where $e_i^* = \arg\max_{\{e\in\mathbb{R}^m:\|e\|=1\}} \ell(\widehat{\xi}_i + \varepsilon e)$, $i = 1, \ldots, N$. This completes the proof. □

The following result shows that the worst-case expectation problem formulated based on the Wasserstein ambiguity sets $\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ is not attainable.

Proposition 7. *In the case where the loss function ℓ is a convex function satisfying $L := \sup_{x\in\mathbb{R}^m} \|\partial\ell(x)\|_*$ $< \infty$, and the sample satisfies $\|\partial\ell(\widehat{\xi}_i)\|_* < L$, $i = 1, \ldots, N$, the worst-case expectation problem*

$$\sup_{\mathbb{P}\in\mathbb{B}_\varepsilon^{\mathrm{W}}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[\ell(\xi)] \tag{54}$$

is not attainable, i.e. the worst-case distribution does not exist.

Proof. By Theorem 3, the optimal value of the problem (54) is $\frac{1}{N}\sum_{i=1}^N \ell(\widehat{\xi}_i) + L\varepsilon$. We can rewrite the problem (54) as

$$\sup_{\mathbb{Q}_1,\ldots,\mathbb{Q}_N} \quad \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}^{\mathbb{Q}_i}[\ell(\widehat{\xi}_i+\xi_i)] \tag{55}$$

$$\text{subject to} \quad \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}^{\mathbb{Q}_i}[\|\xi_i\|]=\varepsilon, \tag{56}$$

where $\mathbb{Q}_i$ is the distribution of ξ_i, $i=1,\ldots,N$. By the convexity of ℓ and the definition of L, we have for any $\mathbb{Q}_1,\ldots,\mathbb{Q}_N$, there exist random variables $y_1,\ldots,y_N$ such that

$$\ell(\widehat{\xi}_i+\xi_i)=\ell(\widehat{\xi}_i)+\partial\ell(y_i)^\top\xi_i\leqslant\ell(\widehat{\xi}_i)+\|\partial\ell(y_i)\|_*\|\xi_i\|\leqslant\ell(\widehat{\xi}_i)+L\|\xi_i\| \;\; \mathbb{Q}_i \text{ a.s..}$$

Therefore, for any $\mathbb{Q}_1,\ldots,\mathbb{Q}_N$ satisfying (56), we have

$$\mathbb{E}^{\mathbb{Q}_i}[\ell(\widehat{\xi}_i+\xi_i)]\leqslant\ell(\widehat{\xi}_i)+L\mathbb{E}^{\mathbb{Q}_i}[\|\xi_i\|]$$

and the inequality reduces to equality only if $\ell(\widehat{\xi}_i+\xi_i)=\ell(\widehat{\xi}_i)+L\|\xi_i\|$ a.s. $\mathbb{Q}_i$, $i=1,\ldots,N$. Therefore, we have

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}^{\mathbb{Q}_i}[\ell(\widehat{\xi}_i+\xi_i)]\leqslant\frac{1}{N}\sum_{i=1}^{N}\ell(\widehat{\xi}_i)+L\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}^{\mathbb{Q}_i}[\|\xi_i\|]=\frac{1}{N}\sum_{i=1}^{N}\ell(\widehat{\xi}_i)+L\varepsilon,$$

and the first inequality reduces to equality only if $\ell(\widehat{\xi}_i+\xi_i)=\ell(\widehat{\xi}_i)+L\|\xi_i\|$ a.s. $\mathbb{Q}_i$, $i=1,\ldots,N$. By the assumption that the sample satisfies $\|\partial\ell(\widehat{\xi}_i)\|_*<L$, $i=1,\ldots,N$, we have $\ell(\widehat{\xi}_i+\xi_i)=\ell(\widehat{\xi}_i)+L\|\xi_i\|$ a.s. $\mathbb{Q}_i$, $i=1,\ldots,N$, can not happen simultaneously. Therefore, we have

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}^{\mathbb{Q}_i}[\ell(\widehat{\xi}_i+\xi_i)]<\frac{1}{N}\sum_{i=1}^{N}\ell(\widehat{\xi}_i)+L\varepsilon,$$

which means that $(\mathbb{Q}_1,\ldots,\mathbb{Q}_N)$ is not the optimal solution and thus, any feasible solution is not the optimal solution. □

Proof of Corollary 3. By assumption, we have that f_1 is a non-increasing convex function on $\mathbb{R}$, f_2 is non-decreasing convex function on $\mathbb{R}$, and one can verify that the value of (24) with $\ell(\xi)=f(x^\top\xi)$ is

$$\max_{\|e\|=1} f(x^\top(\widehat{\xi}_i+\varepsilon e))=\max\left\{f_1(x^\top\widehat{\xi}_i-\varepsilon\|x\|_*),\; f_2(x^\top\widehat{\xi}_i+\varepsilon\|x\|_*)\right\},$$

which is a convex function in $x \in \mathbb{R}^m$. Further, we have the value of (25) with $\ell(\xi) = f(x^\top \xi)$ is

$$\sup_{\mathbb{P} \in \mathbb{B}^{\mathrm{W}}_{(1-\alpha)\varepsilon}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{P}}[f(x^\top \xi)] = \frac{1}{N} \sum_{i=1}^{N} f(x^\top \widehat{\xi}_i) + \mathrm{Lip}(f) \|x\|_* (1-\alpha)\varepsilon.$$

Substituting them to (23), we obtain the result. □

Proof of Theorem 4. By Theorem 2, we have the problem (29) is equivalent to

$$\begin{aligned} \sup_{p_{ij}, \xi_{ij}} \quad & \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p_{ij} (a_j^\top \xi_{ij} + b_j) \\ \text{subject to} \quad & e_\alpha^\Pi (\|\widehat{\xi} - \xi\|) \leqslant \varepsilon, \\ & \Pi((\widehat{\xi}, \xi) = (\widehat{\xi}_i, \xi_{ij})) = p_{ij} \geqslant 0, \;\; \xi_{ij} \in \Xi, \;\; \forall i, j, \\ & \sum_{j=1}^{K} p_{ij} = 1, \quad i = 1, \ldots, N. \end{aligned} \tag{57}$$

Note that $e_\alpha(X) \leqslant \varepsilon$ is equivalent to $e_\alpha(X - \varepsilon) \leqslant 0$ as e_α satisfies translation invariance. By the monotonicity of $x \mapsto \alpha \mathbb{E}[(X-x)_+] - (1-\alpha)\mathbb{E}[(X-x)_-]$, we have $e_\alpha(X - \varepsilon) \leqslant 0$ if and only if $\alpha \mathbb{E}(X - \varepsilon)_+ \leqslant (1-\alpha)\mathbb{E}(\varepsilon - X)_+$, that is, $(2\alpha - 1)\mathbb{E}(X - \varepsilon)_+ \leqslant (1-\alpha)(\varepsilon - \mathbb{E}X)$. We have the constraint $e_\alpha^\Pi(\|\widehat{\xi} - \xi\|) \leqslant \varepsilon$ is equivalent to

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p_{ij} (\|\xi_{ij} - \widehat{\xi}_i\| - \varepsilon)_+ \leqslant \frac{1-\alpha}{2\alpha - 1} \left[\varepsilon - \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p_{ij} \|\xi_{ij} - \widehat{\xi}_i\| \right].$$

Therefore, we have the problem (57) is equivalent to

$$\begin{aligned} \sup_{p_{ij}, \xi_{ij}} \quad & \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p_{ij} (a_j^\top \xi_{ij} + b_j) \\ \text{subject to} \quad & \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p_{ij} (\|\xi_{ij} - \widehat{\xi}_i\| - \varepsilon)_+ \leqslant \frac{1-\alpha}{2\alpha - 1} \left[\varepsilon - \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p_{ij} \|\xi_{ij} - \widehat{\xi}_i\| \right], \\ & \sum_{j=1}^{K} p_{ij} = 1, \;\; i = 1, \ldots, N, \quad \xi_{ij} \in \Xi, \;\; \forall i, j. \end{aligned} \tag{58}$$

Substituting $y_{ij} = p_{ij}(\xi_{ij} - \widehat{\xi}_i)$ for $i = 1, \ldots, N$, $j = 1, \ldots, K$, into the problem (58) yields (30) by standard computation.

To derive the alternative minimization formulation, we begin by dualizing the first constraint in (30)

$$\inf_{\lambda\geqslant 0}\left\{\begin{array}{ll}\sup_{p_{ij},y_{ij}} & \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}(a_j^\top y_{ij}+\bar{z}_{ij}p_{ij})+\lambda\left(\beta\varepsilon-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\left((\|y_{ij}\|-\varepsilon p_{ij})_+ +\beta\|y_{ij}\|\right)\right)\\ \text{subject to} & \sum_{j=1}^{K}p_{ij}=1,\ i=1,...,N,\ p_{ij}\geqslant 0,\ \widehat{\xi}_i+\frac{y_{ij}}{p_{ij}}\in\Xi,\ \ \forall i,j.\end{array}\right\} \tag{59}$$

where $\beta=\frac{1-\alpha}{2\alpha-1}$ and $\bar{z}_{ij}=(a_j^\top\widehat{\xi}_i+b_j)$. Strong duality holds, i.e. (59) = (30), because (30) is a convex optimization problem that satisfies the slater condition when $\varepsilon>0$. Strong duality holds also for $\varepsilon=0$, because both the primal and the dual reduces to the same problem. Replacing y_{ij} with $p_{ij}(\xi_{ij}-\widehat{\xi}_i)$, we have

$$\begin{aligned}&\inf_{\lambda\geqslant 0}\lambda\beta\varepsilon+\frac{1}{N}\sum_{i=1}^{N}\left\{\begin{array}{ll}\sup_{p_{ij},\xi_{ij}} & \sum_{j=1}^{K}p_{ij}\left[\left(a_j^\top\xi_{ij}+b_j\right)-\lambda\left(\left(\|\xi_{ij}-\widehat{\xi}_i\|-\varepsilon\right)_+ +\beta\|\xi_{ij}-\widehat{\xi}_i\|\right)\right]\\ \text{subject to} & \sum_{j=1}^{K}p_{ij}=1,\ p_{ij}\geqslant 0,\ \xi_{ij}\in\Xi,\ \ \forall j.\end{array}\right\}\\ &=\inf_{\lambda\geqslant 0}\lambda\beta\varepsilon+\frac{1}{N}\sum_{i=1}^{N}\left\{\begin{array}{ll}\sup_{\xi_{ij}} & \max_{j\in\{1,...,K\}}\left[\left(a_j^\top\xi_{ij}+b_j\right)-\lambda\left(\left(\|\xi_{ij}-\widehat{\xi}_i\|-\varepsilon\right)_+ +\beta\|\xi_{ij}-\widehat{\xi}_i\|\right)\right]\\ \text{subject to} & \xi_{ij}\in\Xi,\ \ \forall j.\end{array}\right\}\\ &=\left\{\begin{array}{ll}\inf_{\lambda\geqslant 0,s_i} & \lambda\beta\varepsilon+\frac{1}{N}\sum_{i=1}^{N}s_i\\ \text{subject to} & \max_{j\in\{1,...,K\}}\left[\sup_{\xi\in\Xi}\left(a_j^\top\xi+b_j\right)-\lambda\left(\left(\|\xi-\widehat{\xi}_i\|-\varepsilon\right)_+ +\beta\|\xi-\widehat{\xi}_i\|\right)\right]\leqslant s_i,\ i=1,...,N\end{array}\right\}.\end{aligned}$$

The constraint is equivalent to

$$\sup_{\xi\in\Xi}\left(a_j^\top\xi+b_j\right)-\lambda\left(\left(\|\xi-\widehat{\xi}_i\|-\varepsilon\right)_+ +\beta\|\xi-\widehat{\xi}_i\|\right)\leqslant s_i,\ i=1,...,N,\ j=1,...,K. \tag{60}$$

Let $f(\xi)=f_1(\xi)+f_2(\xi)$, where $f_1(\xi):=\lambda\left(\|\xi-\widehat{\xi}_i\|-\varepsilon\right)_+$ and $f_2(\xi)=\lambda\beta\|\xi-\widehat{\xi}_i\|$. Using Fenchel duality, we can substitute the left-hand-side of (60) by its dual and arrive at

$$\exists u_{ij}:\ \ [-\ell_j+\chi_\Xi]^*(u_{ij})+f^*(-u_{ij})\leqslant s_i,\ i=1,...,N,\ j=1,...,K, \tag{61}$$

where $\ell_j(\xi)=a_j^\top\xi+b_j$ and

$$[-\ell_j+\chi_\Xi]^*(u_{ij})=b_j+\sigma_\Xi(u_{ij}+a_j).$$

To derive the conjugate f^*, we derive first the conjugate f_1^* and f_2^*. Note that $f_1(\xi)=h\left(g(\xi)\right)$ where $h(u)=\max(0,u)$ is nondecreasing, and $g(\xi)=\lambda(\|\xi-\widehat{\xi}_i\|-\varepsilon)$. Assuming first that $u_1\neq 0$

and $\lambda > 0$, we have

$$\begin{aligned}
f_1^*(u_1) =& \inf_{z\geqslant 0} \left\{(zg)^*(u_1) + h^*(z)\right\} && (62)\\
=& \inf_{0\leqslant z\leqslant 1} (zg)^*(u_1) && (63)\\
=& \inf_{0\leqslant z\leqslant 1} \sup_{\xi}\ u_1^\top \xi - z\lambda(\|\xi - \widehat{\xi}_i\| - \varepsilon)\\
=& \inf_{0\leqslant z\leqslant 1} \sup_{y}\ u_1^\top (y + \widehat{\xi}_i) - z\lambda(\|y\| - \varepsilon) && (64)\\
=& u_1^\top \widehat{\xi}_i + \inf_{0< z\leqslant 1} z\lambda \left\{ \sup_y \left(\frac{1}{z\lambda}u_1\right)^\top y - \|y\| \right\} + z\lambda\varepsilon && (65)\\
=& u_1^\top \widehat{\xi}_i + \inf_{0< z\leqslant 1,\ z\lambda\geqslant \|u_1\|_*} z\lambda\varepsilon && (66)\\
=& \begin{cases} u_1^\top \widehat{\xi}_i + \|u_1\|_*\varepsilon, & \|u_1\|_* \leqslant \lambda, \\ \infty, & \text{o.w.}, \end{cases} && (67)
\end{aligned}$$

where (62) follows the property of conjugate functions, (63) follows

$$h^*(z) = \begin{cases} 0, & 0 \leqslant z \leqslant 1, \\ \infty, & \text{o.w.}, \end{cases}$$

(65) invokes the observation that $z\lambda > 0$ can be assumed without the loss of generality because $\lambda > 0$ and $z = 0$ cannot be optimal in (64) (in which case (64) equals to ∞ unless $u_1 = 0$), (66) follows the conjugate of the norm $\|y\|$, and finally (67) is because the optimal z must satisfy $z\lambda = \|u_1\|_*$ for any $\|u_1\|_* \leqslant \lambda$ and does not exist otherwise. From (64), one can see that in the case $u_1 = 0$, we have $f_1^*(0) = \inf_{0\leqslant z\leqslant 1} \sup_y -z\lambda(\|y\| - \varepsilon) = \inf_{0\leqslant z\leqslant 1} z\lambda\varepsilon = 0$, which is consistent with (67).

We also have

$$f_2^*(u_2) = \begin{cases} u_2^\top \widehat{\xi}_i, & \|u_2\|_* \leqslant \lambda\beta, \\ \infty, & \text{o.w.}, \end{cases}$$

We thus have the conjugate $f^*(-u)$

$$
\begin{aligned}
f^*(-u) &= \inf_{u_1+u_2=-u} f_1^*(u_1) + f_2^*(u_2) \\
&= \left\{ \begin{array}{ll} \inf_{u_1,u_2} & u_1^\top \widehat{\xi}_i + \|u_1\|_* \varepsilon + u_2^\top \widehat{\xi}_i \\ & u_1 + u_2 = -u, \ \|u_1\|_* \leqslant \lambda, \ \|u_2\|_* \leqslant \lambda\beta \end{array} \right\} \\
&= \left\{ \begin{array}{ll} \inf_{u_1} & (-u)^\top \widehat{\xi}_i + \|u_1\|_* \varepsilon \\ & \|u_1\|_* \leqslant \lambda, \ \|u + u_1\|_* \leqslant \lambda\beta \end{array} \right\}.
\end{aligned}
$$

Thus, the constraint (61) can be equivalently formulated as

$$
\begin{aligned}
& \exists u_{ij} : \ b_j + \sigma_\Xi(u_{ij} + a_j) + \left\{ \begin{array}{ll} \inf_{u_1} & (-u_{ij})^\top \widehat{\xi}_i + \|u_1\|_* \varepsilon \\ & \|u_1\|_* \leqslant \lambda, \ \|u_{ij} + u_1\|_* \leqslant \lambda\beta \end{array} \right\} \leqslant s_i, \ i = 1, ..., N, \ j = 1, ..., K. \\
\Leftrightarrow & \exists u_{ij}, v_{ij} : \begin{array}{l} b_j + \sigma_\Xi(u_{ij} + a_j) + (-u_{ij})^\top \widehat{\xi}_i + \|v_{ij}\|_* \varepsilon \leqslant s_i, \\ \|v_{ij}\|_* \leqslant \lambda, \ \|u_{ij} + v_{ij}\|_* \leqslant \lambda\beta, \end{array} \quad i = 1, ..., N, \ j = 1, ..., K.
\end{aligned}
$$

We thus arrive at the final formulation. Note that in the case $\lambda = 0$, (60) can be directly reduced to $[-\ell_j + \chi_\Xi]^*(0) \leqslant s_i, \ i = 1, ..., N, \ j = 1, ..., K$ by the property of convex conjugate, which is consistent with the final formulation. □

Proof of Theorem 5. We first show (33) for piecewise linear function, that is, $\ell(x) = \max_{j=1,\ldots,K}\{a_j^\top x + b_j\}$. Without loss of generality, assume that $\|a_1\|_* \leqslant \ldots \leqslant \|a_K\|_*$ and Denote by $z_{ij} := \ell_j(\widehat{\xi}_i) = a_j^\top \widehat{\xi}_i + b_j, \ \forall \ i, j$. By Theorem 4 for $\Xi = \mathbb{R}^m$, letting $x_{ij} = \|y_{ij}\| \ \forall i, j$, we have the problem (30) is equivalent to

$$
\begin{aligned}
\sup_{p_{ij}, x_{ij} \in \mathbb{R}_+} \quad & \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^K p_{ij}(\|a_j\|_* x_{ij} + z_{ij}) \\
\text{subject to} \quad & \alpha \sum_{i=1}^N \sum_{j=1}^K p_{ij}(x_{ij} - \varepsilon)_+ \leqslant (1-\alpha) \sum_{i=1}^N \sum_{j=1}^K p_{ij}(x_{ij} - \varepsilon)_-, \\
& \sum_{j=1}^K p_{ij} = 1, \quad i = 1, \ldots, N, \quad p_{ij} \geqslant 0, \ \forall i, j.
\end{aligned} \tag{68}
$$

For each $i = 1, \ldots, N$, we have the following two observations.

(a) If $x_{ij} < \varepsilon$ and $x_{ik} < \varepsilon$ for some $j < k$, then taking $x_{ij}^* = x_{ij} - \delta \geqslant 0$ and $x_{ik}^* = x_{ik} + \delta p_{ij}/p_{ik} \leqslant \varepsilon$ for any $\delta > 0$ yields a feasible solution and a larger value of the objective function.

(b) If $x_{ij} \geqslant \varepsilon$ and $x_{ik} \geqslant \varepsilon$ for some $j < k$, then taking $x^*_{ij} = \varepsilon$ and $x^*_{ik} = x_{ik} + (x_{ij} - \varepsilon)p_{ij}/p_{ik}$ yields a feasible solution and a larger value of the objective function.

Combining the above two observations, for each $i = 1, \dots, N$, we have either one of the following two cases holds:

(i) $x_{ij} \leqslant \varepsilon$ for all $j = 1, \dots, K$. In this case, there exists j_i such that $x_{i1} = \ldots = x_{ij_i} = 0 \leqslant x_{ij_i+1} \leqslant x_{ij_i+2} = \cdots = x_{iK} = \varepsilon$.

(ii) $x_{iK} > \varepsilon$ and $x_{ij} \leqslant \varepsilon$ for all $j = 1, \dots, K-1$. In this case, there exists j_i such that $x_{i1} = \ldots = x_{ij_i} = 0 \leqslant x_{ij_i+1} \leqslant x_{ij_i+2} = \cdots = x_{iK-1} = \varepsilon$. We then have the following two cases.

(ii.a) If $\alpha\|a_K\| \geqslant (1-\alpha)\|a_{j_i+1}\|$, then letting $x^*_{iK} = x_{iK} + (1-\alpha)\delta p_{ij_i+1}$ and $x^*_{ij_i+1} = x_{ij_i+1} - \alpha\delta p_{iK} \geqslant 0$ for any $\delta > 0$ yields a feasible solution and a larger value of the objective function.

(ii.b) If $\alpha\|a_K\| < (1-\alpha)\|a_{j_i+1}\|$, then letting $x^*_{iK} = x_{iK} - (1-\alpha)\delta p_{ij_i+1}$ and $x^*_{ij_i+1} = x_{ij_i+1} + \alpha\delta p_{iK} \leqslant \varepsilon$ for any $\delta > 0$ yields a feasible solution and a larger value of the objective function.

Therefore, combining the above two cases (ii.a) and (ii.b), we have that the case (ii) reduces to either case (i) or $x_{ij_i+1} = 0$.

We can rephrase the above two cases (i) and (ii) as follows: For each $i = 1, \dots, N$, we have either one of the following two cases holds.

(i) $x_{ij} \leqslant \varepsilon$ for all $j = 1, \dots, K$. In this case, there exists j_i such that $x_{i1} = \ldots = x_{ij_i} = 0 \leqslant x_{ij_i+1} \leqslant x_{ij_i+2} = \cdots = x_{iK} = \varepsilon$.

(ii) $x_{iK} > \varepsilon$ and $x_{ij} \leqslant \varepsilon$ for all $j = 1, \dots, K-1$. In this case, there exists j_i such that $x_{i1} = \ldots = x_{ij_i} = 0 \leqslant x_{ij_i+1} = \cdots = x_{iK-1} = \varepsilon$.

Moreover, note that for any feasible x_{ij}, $i = 1, \dots, N$, $j = 1, \dots, K$, define

$$x^*_{ij} = \begin{cases} \frac{\sum_{i=1}^N p_{ij} x_{ij} 1_{\{x_{ij} \leqslant \varepsilon\}}}{\sum_{i=1}^N p_{ij} 1_{\{x_{ij} \leqslant \varepsilon\}}} =: y_j & \text{if } x_{ij} \leqslant \varepsilon, \\ \frac{\sum_{i=1}^N p_{ij} x_{ij} 1_{\{x_{ij} > \varepsilon\}}}{\sum_{i=1}^N p_{ij} 1_{\{x_{ij} > \varepsilon\}}} =: z_j & \text{if } x_{ij} > \varepsilon. \end{cases}$$

We have

$$\sum_{i=1}^N \sum_{j=1}^K p_{ij}(x^*_{ij} - \varepsilon)_+ = \sum_{i=1}^N \sum_{j=1}^K p_{ij}(x_{ij} - \varepsilon)_+ \quad \text{and} \quad \sum_{i=1}^N \sum_{j=1}^K p_{ij}(x^*_{ij} - \varepsilon)_- = \sum_{i=1}^N \sum_{j=1}^K p_{ij}(x_{ij} - \varepsilon)_-,$$

and

$$\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}(\|a_j\|_* x_{ij}^* + z_{ij}) = \sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}(\|a_j\|_* x_{ij} + z_{ij}).$$

That is, p_{ij}, x_{ij}^*'s satisfy the constraint and the objective function remains unchanged. This combined with the above observation, i.e., $x_{ij} \leqslant \varepsilon$ for all $j < K$, we have the optimal solution x_{ij} must satisfy: $x_{iK} \geqslant \varepsilon$ for all $i = 1, \dots, N$, $x_{ij} \leqslant \varepsilon$ for all $i = 1, \dots, N$ and $j = 1, \dots, K-1$, and moreover, there exists k which is independent of i such that $x_{i1} = \cdots = x_{ik} = 0 \leqslant x_{ik+1} = \cdots = x_{K-1} = \varepsilon$. Therefore, the problem (68) is equivalent to

$$\begin{aligned} \sup_{p_{ij}, x_{ij} \in \mathbb{R}_+, k} \quad & \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=k+1}^{K-1} p_{ij}\|a_j\|_*\varepsilon + p_{iK}\|a_K\|_* x_{iK} + \sum_{j=1}^{K} p_{ij} z_{ij}\right) && (69)\\ \text{subject to} \quad & \alpha \sum_{i=1}^{N} p_{iK}(x_{iK} - \varepsilon) \leqslant (1-\alpha)\sum_{i=1}^{N}\sum_{j=1}^{k} \varepsilon p_{ij}, && (70)\\ & \sum_{j=1}^{K} p_{ij} = 1, \quad i = 1, \dots, N, \quad k = 1, \dots, K. && (71) \end{aligned}$$

Since the objective function is increasing in x_{iK}, it suffices to consider the case that the inequality of (70) is an equality. Substituting $\sum_{i=1}^{N} p_{iK} x_{iK} = \sum_{i=1}^{N} \varepsilon p_{iK} + \beta \sum_{i=1}^{N}\sum_{j=1}^{k} \varepsilon p_{ij}$ into the objective function, we have the problem (69) is equivalent to

$$\sup_{p_{ij} \in \mathbb{R}_+,\ k} \ \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{k} \beta p_{ij}\|a_K\|_*\varepsilon + \sum_{j=k+1}^{K-1} p_{ij}\|a_j\|_*\varepsilon + p_{iK}\|a_K\|_*\varepsilon + \sum_{j=1}^{K} p_{ij} z_{ij}\right] \quad \text{s.t. } (71). \qquad (72)$$

Note that for any feasible p_{ij}, the optimal k must be the largest index of j such that $\|a_j\|_* \leqslant \beta\|a_K\|_*$, i.e.,

$$k^* = \max\{j = 1, \dots, K-1 : \|a_j\|_* \leqslant \beta\|a_K\|_*\}.$$

Hence, the problem (72) is equivalent to

$$\sup_{p_{ij} \in \mathbb{R}_+} \ \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{k^*} p_{ij}\big(\beta\|a_K\|_*\varepsilon + z_{ij}\big) + \sum_{j=k^*+1}^{K} p_{ij}(\|a_j\|_*\varepsilon + z_{ij})\right] \quad \text{s.t. } \sum_{j=1}^{K} p_{ij} = 1, \ \forall\, i,$$

whose optimal value equals to

$$\frac{1}{N}\sum_{i=1}^{N} \max\left\{\big(\beta\|a_K\|_*\varepsilon + z_{ij}\big)_{j=1}^{k^*},\ (\|a_j\|_*\varepsilon + z_{ij})_{j=k^*+1}^{K}\right\} =: \frac{1}{N}\sum_{i=1}^{N} \mathcal{E}_i. \qquad (73)$$

Note that $\|a_j\|_*\varepsilon + z_{ij} \leqslant \beta\|a_K\|_*\varepsilon + z_{ij}$, $j = 1, \ldots, k^*$, and thus,

$$\mathcal{E}_i \geqslant \max_{j=1,\ldots,K}\{\|a_j\|_*\varepsilon + z_{ij}\} = \max_{\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e).$$

Also, note that $\|a_j\|_*\varepsilon + z_{ij} > \beta\|a_K\|_*\varepsilon + z_{ij}$, $j = k^* + 1, \ldots, K$, and thus,

$$\mathcal{E}_i \geqslant \beta\|a_K\|_*\varepsilon + \max_{j=1,\ldots,K}\{\|a_j\|_*\varepsilon + z_{ij}\} = \beta\|a_K\|_*\varepsilon + \ell(\widehat{\xi}_i).$$

Therefore, we have

$$(73) \geqslant \frac{1}{N}\sum_{i=1}^{N} \max\left\{\ell(\widehat{\xi}_i) + \beta\|a_K\|_*\varepsilon,\ \max_{\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e)\right\}.$$

On the other hand, obviously we have

$$\max_{j=1,\ldots,k^*} \big(\beta\|a_K\|_*\varepsilon + z_{ij}\big) \leqslant \beta\|a_K\|_*\varepsilon + \ell(\widehat{\xi}_i) \quad \text{and} \quad \max_{j=k^*+1,\ldots,K} \big(\|a_j\|_*\varepsilon + z_{ij}\big) \leqslant \max_{\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e).$$

Hence, we have

$$(73) = \frac{1}{N}\sum_{i=1}^{N} \max\left\{\ell(\widehat{\xi}_i) + \beta\|a_K\|_*\varepsilon,\ \max_{\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e)\right\}.$$

Therefore, we have shown (33) for piecewise linear functions. For general convex loss function ℓ, by similar arguments in Theorem 3, we have the optimal value is (33). □

Proof of Corollary 4. Since $\beta = (1-\alpha)/\alpha \to 0$ as $\alpha \to 1$ and $\ell(\widehat{\xi}_i) < \max_{\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e)$ for all $i = 1, \ldots, N$, there exists $\alpha \in (1/2, 1)$ large enough such that $\ell(\widehat{\xi}_i) + \beta L\varepsilon < \max_{\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e)$ for $i = 1, \ldots, N$. In this case, (33) in Theorem 5 reduces to

$$\frac{1}{N}\sum_{i=1}^{N} \max_{\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e).$$

One can verify that this optimal value is attained by

$$\mathbb{P}_\varepsilon = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i + \varepsilon e_i} \quad \text{with} \quad e_i = \arg\max_{\|e\|=1} \ell(\widehat{\xi}_i + \varepsilon e), \quad i = 1, \ldots, N.$$

We thus complete the proof. □

Proof of Corollary 5. The proof is similar to that of Corollary 3 and thus omitted. □

References

Acerbi, C. and Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking and Finance*, **26**(7), 1487–1503.

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

Bäuerle, N. and Müller, A. (2006). Stochastic orders and risk measures: Consistency and bounds. *Insurance: Mathematics and Economics*, **38**(1), 132–148.

Bellini, F., Klar, B., Müller, A., and Gianin, E. R. (2014). Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, **54**, 41–48.

Bellini, F. and Bernardino, E.D. (2017) Risk management with expectiles, *The European Journal of Finance*, **23**(6), 487–506.

Bertsimas, D., Doan, X.V., Natarajan, K., Teo, C.P. (2010) Models for minimax stochastic linear optimization problems with risk aversion *Mathematics of Operations Research* **35**(3), 580–602.

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

Blanchet, J., Kang, Y., Murthy, K. (2019). Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, **56**(3), 830–857.

Birghila, C., Aigner, M. and Engelke, S. (2022) Distributionally robust tail bounds based on Wasserstein distance and f-divergence. https://arxiv.org/abs/2106.06266.

Carlsson, J. G., Behroozi, M., Mihic, K. (2018) Wasserstein distance and the distributionally robust TSP. *Operations Research*, **66**(6), 1603–1624.

Castagnoli, E., Cattelan, G., Maccheroni, F., Tebaldi, C. and Wang, R. (2022). Star-shaped risk measures. *Operations Research*.

Cheung, K. C and Lo, A. (2013). Characterizations of counter-monotonicity and upper comonotonicity by (tail) convex order. *Insurance: Mathematics and Economics*, **53**(2), 334–342.

Csiszár, I. (1964). Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, **8**, 85–108.

de Haan, L. and Ferreira, A., 2006. *Extreme Value Theory: An Introduction.* In: Springer Series in Operations Research and Financial Engineering, Springer, New York.

Delbaen, F. (2002). Coherent risk measures on general probability spaces. In *Advances in finance and stochastics* (pp. 1–37). Springer, Berlin, Heidelberg.

Delbaen, F. (2012). *Monetary Utility Functions.* Osaka University Press.

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

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

Embrechts, P., Puccetti, G., Rušchendorf, L., Wang, R. and Beleraj, A. (2014). An academic response to 36 Basel 3.5. Risks, 2(1), 25-48.

Föllmer, H. and Schied, A. (2016). *Stochastic Finance. An Introduction in Discrete Time.* (Fourth Edition.) Walter de Gruyter, Berlin.

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

Frittelli, M. and Gianin, E. R. (2005) Law invariant convex risk measures. In *Advances in mathematical economics* (pp. 33–46). Springer, Tokyo.

Gao, R., Chen, X., Kleywegt, A. J. (2020) Wasserstein Distributionally Robust Optimization and Variation Regularization. *arxiv.org/abs/1712.06050.*

Gao, R. and Kleywegt, A. J. (2022) Distributionally Robust Stochastic Optimization with Wasserstein Distance. *Mathematics of Operations Research0.*

Gneiting, T. (2011). Making and Evaluating Point Forecast. *Journal of the American Statistical Association*, **106**(494): 746–762.

Hu, Z. and Hong, L. J. (2013). Kullback-Leibler divergence constrained distributionally robust optimization. Available at Optimization Online, 1695-1724.

Kallenberg, O. (1997). Foundations of modern probability (Vol. 2). New York: springer.

Jouini, E., Schachermayer, W. and Touzi, N. (2006). Law invariant risk measures have the Fatou property. In *Advances in mathematical economics* (pp. 49-71). Springer, Tokyo.

Kantorovich, L.V., Rubinshtein, G.S. (1958) On a space of totally additive functions. Vestn. Leningr. Univ. **13**, 52–59.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., Shafieezadeh-Abadeh, S. (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Operations research & Management Science in the age of analytics*, 130–166.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**(1), 79–86.

Kusuoka, S. (2001). On law invariant coherent risk measures. In *Advances in mathematical economics* (pp. 83–95). Springer, Tokyo.

Krätschmer, V., Schied, A. and Zähle, H. (2014). Comparative and quantitiative robustness for law-invariant 27 risk measures. *Finance and Stochastics*, **18**(2), 271–295.

Mao, T. and Wang, R. (2015). On aggregation sets and lower-convex sets. Journal of Multivariate Analysis, 138, 170-181.

McNeil, A. J., Frey, R. and Embrechts, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools.* (Revised Edition.) Princeton University Press, Princeton, NJ.

Morimoto, T. (1963). Markov processes and the H-theorem. *Journal of the Physical Society of Japan*, **18**(3), 328–331.

Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems: Volume I: Theory* (Vol. 1). Springer Science & Business Media.

Rockafellar, R.T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, **26**(7), 1443–1471.

Rüschendorf, L. (2013). Law Invariant Convex Risk Measures on L_d^p and Optimal Mass Transportation. In *Mathematical Risk Analysis* (pp. 189–221). Springer, Berlin, Heidelberg.

Shafieezadeh-Abadeh, S., Kuhn, D., Esfahani, P.M. (2019) Regularization via mass transportation. *Journal of Machine Learning Research*, **20**(103), 1–68

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer Series in Statistics.

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

Sinha, A., Namkoong, H., Volpi, R., Duchi, J. (2020) Certifying Some Distributional Robustness with Principled Adversarial Training. *arXiv:1710.10571.*

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