Distributionally Robust Stochastic Optimization with Wasserstein Distance

Rui Gao, Anton J. Kleywegt
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205
rgao32@gatech.edu, anton@isye.gatech.edu

Distributionally robust stochastic optimization (DRSO) is an approach to optimization under uncertainty in which, instead of assuming that there is a known true underlying probability distribution, one hedges against a chosen set of distributions. In this paper we first point out that the set of distributions should be chosen to be appropriate for the application at hand, and that some of the choices that have been popular until recently are, for many applications, not good choices. We consider sets of distributions that are within a chosen Wasserstein distance from a nominal distribution, for example an empirical distribution resulting from available data. The paper points out that such a choice of sets has two advantages: (1) The resulting distributions hedged against are more reasonable than those resulting from other popular choices of sets. (2) The problem of determining the worst-case expectation over the resulting set of distributions has desirable tractability properties. We derive a dual reformulation of the corresponding DRSO problem and construct approximate worst-case distributions (or an exact worst-case distribution if it exists) explicitly via the first-order optimality conditions of the dual problem. Our contributions are five-fold. (i) We identify necessary and sufficient conditions for the existence of a worst-case distribution, which are naturally related to the growth rate of the objective function. (ii) We show that the worst-case distributions resulting from an appropriate Wasserstein distance have a concise structure and a clear interpretation. (iii) Using this structure, we show that data-driven DRSO problems can be approximated to any accuracy by robust optimization problems, and thereby many DRSO problems become tractable by using tools from robust optimization. (iv) To the best of our knowledge, our proof of strong duality is the first constructive proof for DRSO problems, and we show that the constructive proof technique is also useful in other contexts. (v) Our strong duality result holds in a very general setting, and we show that it can be applied to infinite dimensional process control problems and worst-case value-at-risk analysis.

Key words: distributionally robust optimization; data-driven decision-making; ambiguity set; worst-case distribution
MSC2000 subject classification: Primary: 90C15; secondary: 90C46
OR/MS subject classification: Primary: programming; stochastic

1. Introduction
In decision making problems under uncertainty, a decision maker wants to choose a decision $x$ from a feasible region $X$. The objective function $\Psi : X \times \Xi \rightarrow \mathbb{R}$ depends on a quantity $\xi \in \Xi$ whose value is not known to the decision maker at the time that the decision has to be made. In some settings it is reasonable to assume that $\xi$ is a random element with distribution $\mu$ supported on $\Xi$, for example, if multiple realizations of $\xi$ will be encountered. In such settings, the decision making problems can be formulated as stochastic optimization problems as follows:

$$\inf_{x \in X} \mathbb{E}_\mu[\Psi(x, \xi)].$$

We refer to Shapiro et al. [40] for a thorough study of stochastic optimization. One major criticism of the formulation above for practical applications is the requirement that the underlying distribution $\mu$ be known to the decision maker. Even if multiple realizations of $\xi$ are observed, $\mu$ still may not be known exactly, while use of a distribution different from $\mu$ may sometimes result in bad decisions. Another major criticism is that in many applications there are not multiple realizations of $\xi$ that will be encountered, for example in problems involving events that may either happen once or not happen at all, and thus the notion of a “true” underlying distribution does not apply. These criticisms motivate the notion of distributionally robust stochastic optimization (DRSO), that does
not rely on the notion of a known true underlying distribution. One chooses a set $\mathcal{M}$ of probability distributions to hedge against, and then finds a decision that provides the best hedge against the set $\mathcal{M}$ of distributions by solving the following minmax problem:

$$\inf_{x \in X} \sup_{\mu \in \mathcal{M}} \mathbb{E}_{\mu}[\Psi(x, \xi)].$$

(\text{DRSO})

Such a minmax approach has its roots in Von Neumann’s game theory and has been used in many fields such as inventory management (Scarf et al. [38], Gallego and Moon [21]), statistical decision analysis (Berger [8]), as well as stochastic optimization (Žácková [49], Dupačová [17], Shapiro and Kleywegt [41]). Recently it regained attention in the operations research literature, and sometimes is called data-driven stochastic optimization or ambiguous stochastic optimization.

A central question is: how to choose a good set of distributions $\mathcal{M}$ to hedge against? A good choice of $\mathcal{M}$ should take into account the properties of the practical application as well as the tractability of problem (DRSO). Two typical ways of constructing $\mathcal{M}$ are moment-based and distance-based. The moment-based approach considers distributions whose moments (such as mean and covariance) satisfy certain conditions (Scarf et al. [38], Delage and Ye [16], Popescu [35], Zymler et al. [51]). It has been shown that in many cases the resulting DRSO problem can be formulated as a conic quadratic or semi-definite program. However, the moment-based approach is based on the curious assumption that certain conditions on the moments are known exactly but that nothing else about the relevant distribution is known. More often in applications, either one has data from repeated observations of the quantity $\xi$, or one has no data, and in both cases the moment conditions do not describe exactly what is known about $\xi$. In addition, the resulting worst-case distributions sometimes yield overly conservative decisions (Wang et al. [46], Goh and Sim [23]). For example, Wang et al. [46] shows that for the newsvendor problem, by hedging against all the distributions with fixed mean and variance, Scarf’s moment approach yields a two-point worst-case distribution, and the resulting decision does not perform well under other more likely scenarios.

The distance-based approach considers distributions that are close, in the sense of a chosen statistical distance, to a nominal distribution $\nu$, such as an empirical distribution or a Gaussian distribution (El Ghaoui et al. [18], Calafiore and El Ghaoui [13]). Popular choices of the statistical distance are $\phi$-divergences (Bayraksan and Love [5], Ben-Tal et al. [6]), which include Kullback-Leibler divergence (Jiang and Guan [25]), Burg entropy (Wang et al. [46]), and Total Variation distance (Sun and Xu [42]) as special cases, Prokhorov metric (Erdoğan and Iyengar [19]), and Wasserstein distance (Wozabal [47, 48], Esfahani and Kuhn [20], Zhao and Guan [50]).

### 1.1. Motivation: Potential issues with $\phi$-divergence

Despite its widespread use, $\phi$-divergence has a number of shortcomings. Here we highlight some of these shortcomings. In a typical setup using $\phi$-divergence, $\Xi$ is partitioned into $B + 1$ bins represented by points $\xi^0, \xi^1, \ldots, \xi^B \in \Xi$. The nominal distribution $\nu$ associates $N_i$ observations with bin $i$. That is, the nominal distribution is given by $\nu := (N_0/N, N_1/N, \ldots, N_B/N)$, where $N := \sum_{i=0}^B N_i$. Let $\Delta_B := \{(p_0, p_1, \ldots, p_B) \in \mathbb{R}^{B+1}_+ : \sum_{j=0}^B p_j = 1\}$ denote the set of probability distributions on the same set of bins. Let $\phi: [0, \infty) \mapsto \mathbb{R}$ be a chosen convex function such that $\phi(1) = 0$, with the conventions that $0\phi(a/0) := a \lim_{t \to \infty} \phi(t)/t$ for all $a > 0$, and $0\phi(0/0) := 0$. Then the $\phi$-divergence between $\mu = (p_0, \ldots, p_B), \nu = (q_0, \ldots, q_B) \in \Delta_B$ is defined by

$$I_\phi(\mu, \nu) := \sum_{j=0}^B q_j \phi\left(\frac{p_j}{q_j}\right).$$

Let $\theta > 0$ denote a chosen radius. Then $\mathcal{M}_\phi := \{\mu \in \Delta_B : I_\phi(\mu, \nu) \leq \theta\}$ denotes the set of probability distributions given by the chosen $\phi$-divergence and radius $\theta$. The DRSO problem corresponding to the $\phi$-divergence ball $\mathcal{M}_\phi$ is then given by

$$\inf_{x \in X} \sup_{\mu \in \Delta_B} \left\{ \sum_{j=0}^B p_j \Psi(x, \xi^j) : I_\phi(\mu, \nu) \leq \theta \right\}.$$
It has been shown in Ben-Tal et al. [6] that the $\phi$-divergence ball $\mathcal{M}_\phi$ can be viewed as a statistical confidence region (Pardo [32]), and for several choices of $\phi$, the inner maximization of the problem above is tractable.

One well-known shortcoming of $\phi$-divergence balls is that, either they are not rich enough to contain distributions that are often relevant, or they they hedge against many distributions that are too extreme. For example, for some choices of $\phi$-divergence such as Kullback-Leibler divergence, if the nominal $q_i = 0$, then $p_i = 0$, that is, the $\phi$-divergence ball includes only distributions that are absolutely continuous with respect to the nominal distribution $\nu$, and thus does not include distributions with support on points where the nominal distribution $\nu$ is not supported. As a result, if $\Xi = \mathbb{R}^K$ and $\nu$ is discrete, then there are no continuous distributions in the $\phi$-divergence ball $\mathcal{M}_\phi$. Some other choices of $\phi$-divergence exhibit in some sense the opposite behavior. For example, the Burg entropy ball includes distributions with some amount of probability allowed to be shifted from $\nu$ to any other bin, with the amount of probability allowed to be shifted depending only on $\theta$ and not on how extreme the bin is. See Section 5.1 for more details regarding this potential shortcoming.

Next we illustrate another shortcoming of $\phi$-divergence that will motivate the use of Wasserstein distance.

**Example 1.** Suppose that there is an underlying true image (1b), and a decision maker possesses, instead of the true image, an approximate image (1a) obtained with a less than perfect device that loses some of the contrast. The images are summarized by their gray-scale histograms. (In fact, (1a) was obtained from (1b) by a low-contrast intensity transformation (Gonzalez and Woods [24]), by which the black pixels become somewhat whiter and the white pixels become somewhat blacker. This type of transformation changes only the gray-scale value of a pixel and not the location of a gray-scale value, and therefore it can also be regarded as a transformation from one gray-scale histogram to another gray-scale histogram.) As a result, roughly speaking, the observed histogram $\nu$ is obtained by shifting the true histogram $\mu_{\text{true}}$ inwards. Also consider the pathological image (1c) that is too dark to see many details, with histogram $\mu_{\text{pathol}}$. Suppose that the decision maker constructs a Kullback-Leibler (KL) divergence ball $\mathcal{M}_{KL} := \{ \mu \in \Delta_B : I_{KL}(\mu, \nu) \leq \theta \}$. Note that $I_{KL}(\mu_{\text{true}}, \nu) = 5.05 > I_{KL}(\mu_{\text{pathol}}, \nu) = 2.33$. Therefore, if $\theta$ is chosen small enough (less than 2.33) for $\mathcal{M}$ to exclude the pathological image (1c), then $\mathcal{M}$ will also exclude the true image (1b). If $\theta$ is chosen large enough (greater than 5.05) for $\mathcal{M}$ to include the true image (1b), then $\mathcal{M}$ also has to include the pathological image (1c), and then the resulting decision may be overly conservative due to hedging against irrelevant distributions. If an intermediate value is chosen for $\theta$ (between 2.33 and 5.05), then $\mathcal{M}$ includes the pathological image (1c) and excludes the true image (1b). In contrast, note that the Wasserstein distance $W_1$ satisfies $W_1(\mu_{\text{true}}, \nu) = 30.7 < W_1(\mu_{\text{pathol}}, \nu) = 84.0$, and thus Wasserstein distance does not exhibit the problem encountered with KL divergence (see also Example 3).

The reason for such behavior is that $\phi$-divergence does not incorporate a notion of how close two points $\xi, \xi' \in \Xi$ are to each other, for example, how likely is that observation is $\xi'$ given that the true value is $\xi$. In Example 1, $\Xi = \{0, 1, \ldots, 255\}$ represents 8-bit gray-scale values. In this case, we know that the likelihood that a pixel with gray-scale value $\xi \in \Xi$ is observed with gray-scale value $\xi' \in \Xi$ is decreasing in the absolute difference between $\xi$ and $\xi'$. However, in the definition of $\phi$-divergence, only the relative ratio $p_j/q_j$ for the same gray-scale value $j$ is taken into account, while the distances between different gray-scale values is not taken into account. This phenomenon has been observed in studies of image retrieval (Rubner et al. [37], Ling and Okada [28]).

The drawbacks of $\phi$-divergence motivates us to consider sets $\mathcal{M}$ that incorporate a notion of how close two points $\xi, \xi' \in \Xi$ are to each other. One such choice of $\mathcal{M}$ is based on Wasserstein distance. Specifically, consider any underlying metric $d$ on $\Xi$ which measures the closeness of any
two points in \( \Xi \). Let \( p \geq 1 \), and let \( \mathcal{P}(\Xi) \) denote the set of Borel probability measures on \( \Xi \). Then the Wasserstein distance of order \( p \) between two distributions \( \mu, \nu \in \mathcal{P}(\Xi) \) is defined as

\[
W_p(\mu, \nu) = \min_{\gamma \in \mathcal{P}(\Xi^2)} \left\{ \mathbb{E}_{(\xi, \zeta) \sim \gamma}^{1/p} [d_p(\xi, \zeta)] : \gamma \text{ has marginal distributions } \mu, \nu \right\}.
\]

More detailed explanation and discussion on Wasserstein distance will be presented in Section 2.

Given a radius \( \theta > 0 \), the Wasserstein ball of probability distributions \( \mathcal{M} \) is defined by

\[
\mathcal{M} := \{ \mu \in \mathcal{P}(\Xi) : W_p(\mu, \nu) \leq \theta \}.
\]

### 1.2. Related work

Wasserstein distance and the related field of optimal transport, which is a generalization of the transportation problem, have been studied in depth. In 1942, together with the linear programming problem (Kantorovich \[27\]), Kantorovich \[26\] tackled Monge’s problem originally brought up in the study of optimal transport. In the stochastic optimization literature, Wasserstein distance has been used for single stage stochastic optimization (Wozabal \[47, 48\]), and for multistage stochastic optimization (Pflug and Pichler \[34\]). The challenge for solving (DRSO) is that, the inner maximization involves a supremum over possibly an infinite dimensional space of distributions. To tackle this problem, existing works focus on the setup when \( \nu \) is the empirical distribution on a finite-dimensional space. Wozabal \[47\] transformed the inner maximization problem of (DRSO) into a finite-dimensional non-convex program, by using the fact that if \( \nu \) is supported on at most \( N \) points, then there are extreme distributions of \( \mathcal{M} \) that are supported on at most \( N + 3 \) points. Recently, using duality theory of conic linear programming (Shapiro \[39\]), Esfahani and Kuhn \[20\] and Zhao and Guan \[50\] showed that under certain conditions, the inner maximization problem of (DRSO) is actually equivalent to a finite-dimensional convex problem.

In this paper, we consider any arbitrary nominal distribution \( \nu \) on a Polish space, and study the tractability of (DRSO) via strong duality. By the time we completed the first version of this paper, we learned that Blanchet and Murthy \[9\] also considered a similar problem to ours and also obtained a strong duality result. Our focus and our approach to this problem differ from theirs in the following ways. First, we prove the strong duality result for the inner maximization of (DRSO)
using a novel, yet simple, constructive approach, in contrast with the non-constructive approaches in their work and also in Esfahani and Kuhn [20] and Zhao and Guan [50]. This enables us to establish the structural characterization of the worst-case distributions of the data-driven DRSO (Corollary 2(ii)), which improves the result of Wozabal [47] and the more recent result of Owhadi and Scovel [31] on extremal distributions of Wasserstein balls (Remark 6). It also enables us to build a close connection between DRSO and robust optimization (Corollary 2(iii)). Second, we focus on Wasserstein distance of order $p$ ($p \geq 1$), while they consider more general transport metrics in which the distance between two points $\xi, \xi' \in \Xi$ is measured by a lower semicontinuous function rather than a metric $d_p(\xi, \xi')$ as in our case. Nevertheless, our proof remains valid for such more general transport metrics (Remark 3). In the meantime, focusing on Wasserstein distance enables us to relate the condition for the existence of a worst-case distribution to the important notion of the “growth rate” of the objective function, and enables us to provide practical guidance for choosing the ambiguity set and controlling the degree of conservativeness based on the objective function (Remark 2).

1.3. Main contributions

- **General Setting.** We prove a strong duality result for DRSO problems with Wasserstein distance in a very general setting. We show that

$$\sup_{\mu \in P(\Xi)} \{ \mathbb{E}_\mu[\Psi(x, \xi)] : W_p(\mu, \nu) \leq \theta \} = \min_{\lambda \geq 0} \left\{ \lambda \theta - \int_\Xi \inf_{\xi \in \Xi} [\lambda d_p(\xi, \zeta) - \Psi(x, \xi)] \nu(d\zeta) \right\}$$

holds for any Polish space $(\Xi, d)$ and measurable function $\Psi$ (Theorem 1).

1. Both Esfahani and Kuhn [20] and Zhao and Guan [50] assume that $\Xi$ is a convex subset of $\mathbb{R}^K$ with some associated norm. The greater generality of our results enables one to consider interesting problems such as the process control problems in Sections 4.1 and 4.2, where $\Xi$ is the set of finite counting measures on $[0, 1]$, which is infinite-dimensional and non-convex.

2. Both Esfahani and Kuhn [20] and Zhao and Guan [50] assume that the nominal distribution $\nu$ is an empirical distribution, while we allow $\nu$ to be any Borel probability measure. The greater generality enables one to study problems such as the worst-case Value-at-Risk analysis in Section 4.3.

3. Both Esfahani and Kuhn [20] and Zhao and Guan [50] only consider Wasserstein distance of order $p = 1$. By considering a bigger family of Wasserstein distances, we establish the importance for DRSO problems of the notion of the “growth rate” of the objective function, which measures how fast the objective function grows compared to a polynomial of order $p$. It turns out that the growth rate of the objective function determines the finiteness of the worst-case objective value (Proposition 2), and it plays an important role in the existence conditions for the worst-case distribution (Corollary 1). This is of practical importance, since it provides guidance for choosing the proper Wasserstein distance and for controlling the degree of conservativeness based on the structure of the objective function.

- **Constructive Proof of Duality.** We prove the strong duality result using a novel, elementary, constructive approach. The results of Esfahani and Kuhn [20] and Zhao and Guan [50] and other strong duality results in the literature are based on the established Hahn-Banach theorem for certain infinite dimensional vector spaces. In contrast, our proof idea is new and is relatively elementary and straightforward: we use the weak duality result as well as the first-order optimality condition of the dual problem to construct a sequence of primal feasible solutions whose objective values converge to the dual optimal value. Our proof uses relatively elementary tools, without resorting to other “big hammers”.

Definition 1 (Push-forward Measure). Given measurable spaces \((\Xi, \mathcal{B}(\Xi))\) and \((\Xi', \mathcal{B}(\Xi'))\), a measurable function \(T: \Xi \to \Xi'\), and a measure \(\nu \in \mathcal{B}(\Xi)\), let \(T_\# \nu \in \mathcal{B}(\Xi')\) denote the push-forward measure of \(\nu\) through \(T\), defined by

\[ T_\# \nu(A) := \nu(T^{-1}(A)) = \nu\{ \zeta \in \Xi : T(\zeta) \in A \}, \quad \forall \text{ measurable sets } A \subset \Xi'. \]
That is, $T_p\nu$ is obtained by transporting ("pushing forward") $\nu$ from $\Xi$ to $\Xi'$ using the function $T$. For $i \in \{1, 2\}$, let $\pi^i: \Xi \times \Xi \mapsto \Xi$ denote the canonical projections given by $\pi^i(\xi^1, \xi^2) = \xi^i$. Then for a measure $\gamma \in \mathcal{P}(\Xi \times \Xi)$, $\pi^1\gamma \in \mathcal{P}(\Xi)$ is the $i$-th marginal of $\gamma$ given by $\pi^1\gamma(A) = \gamma(A \times \Xi)$ and $\pi^2\gamma(A) = \gamma(\Xi \times A)$.

**Definition 2 (Wasserstein distance).** The Wasserstein distance $W_p(\mu, \nu)$ between $\mu, \nu \in \mathcal{P}_p(\Xi)$ is defined by

$$W_p(\mu, \nu) := \min_{\gamma \in \mathcal{P}(\Xi \times \Xi)} \left\{ \int_{\Xi \times \Xi} d^p(\xi, \zeta) \gamma(d\xi, d\zeta) : \pi^1\gamma = \mu, \pi^2\gamma = \nu \right\}.$$  

(1)

That is, the Wasserstein distance between $\mu, \nu$ is the minimum cost (in terms of $d^p$) of redistributing mass from $\nu$ to $\mu$, which is why it is also called the "earth mover’s distance". Wasserstein distance is a natural way of comparing two distributions when one is obtained from the other by perturbations. The minimum on the right side of (1) is attained, because $d$ is non-negative, continuous and thus lower semicontinuous (Theorem 1.3 of [44]). The following example is a familiar special case of problem (1).

**Example 2 (Transportation problem).** Consider $\mu = \sum_{i=1}^M p_i \delta_{\xi^i}$ and $\nu = \sum_{j=1}^N q_j \delta_{\hat{\xi}^j}$, where $M, N \geq 1$, $p_i, q_j \geq 0$, $\xi^i, \hat{\xi}^j \in \Xi$ for all $i, j$, and $\sum_{i=1}^M p_i = \sum_{j=1}^N q_j = 1$. Then problem (1) becomes the classical transportation problem in linear programming:

$$\min_{\gamma_{ij} \geq 0} \left\{ \sum_{i=1}^M \sum_{j=1}^N d^p(\xi^i, \hat{\xi}^j) \gamma_{ij} : \sum_{j=1}^N \gamma_{ij} = p_i, \forall \, i, \sum_{i=1}^M \gamma_{ij} = q_j, \forall \, j \right\}.$$

**Example 3 (Revisiting Example 1).** Next we evaluate the Wasserstein distance between the histograms in Example 1. To evaluate $W_1(\mu_{\text{true}}, \nu)$, note that the least cost way of transporting mass from $\nu$ to $\mu_{\text{true}}$ is to move the mass outwards. In contrast, to evaluate $W_1(\mu_{\text{pathol}}, \nu)$, one has to transport mass relatively long distances from right to left (changing the gray-scale values of pixels by large amounts), resulting in a larger cost than $W_1(\mu_{\text{true}}, \nu)$. Therefore $W_1(\mu_{\text{pathol}}, \nu) > W_1(\mu_{\text{true}}, \nu)$.

Wasserstein distance has a dual representation due to Kantorovich’s duality (Theorem 5.10 in [45]):

$$W_p(\mu, \nu) = \sup_{u \in L^1(\mu), v \in L^1(\nu)} \left\{ \int_{\Xi} u(\xi) \mu(d\xi) + \int_{\Xi} v(\zeta) \nu(d\zeta) : u(\xi) + v(\zeta) \leq d^p(\xi, \zeta), \forall \, \xi, \zeta \in \Xi \right\}.$$  

(2)

where $L^1(\nu)$ represents the $L^1$ space of $\nu$-measurable (i.e., $(\mathcal{B}(\xi), \mathcal{B}(\mathbb{R}))$-measurable) functions. In addition, the set of functions under the supremum above can be replaced by $u, v \in C_b(\Xi)$, where $C_b(\Xi)$ denotes the set of continuous and bounded real-valued functions on $\Xi$. Particularly, when $p = 1$, by the Kantorovich-Rubinstein Theorem, (2) can be simplified to (see, e.g., Equation (5.11) in [45])

$$W_1(\mu, \nu) = \sup_{u \in L^1(\mu)} \left\{ \int_{\Xi} u(\xi) d(\mu - \nu)(\xi) : u \text{ is } 1\text{-Lipschitz} \right\}.$$

So for an $L$-Lipschitz function $\Psi : \Xi \mapsto \mathbb{R}$, it holds that $|E_\mu[\Psi(\xi)] - E_\nu[\Psi(\xi)]| \leq LW_1(\mu, \nu) \leq L\theta$ for all $\mu \in \mathcal{M}$.

We remark that Definition 2 and the results above can be extended to finite Borel measures. Moreover, we have the following result.

**Lemma 1.** For any finite Borel measures $\mu, \nu \in \mathcal{B}(\Xi)$ with $\mu(\Xi) \neq \nu(\Xi)$, it holds that $W_p(\mu, \nu) = \infty.$
Another important feature of Wasserstein distance is that $W_p$ metrizes weak convergence in $\mathcal{P}_p(\Xi)$ (cf. Theorem 6.9 in Villani [45]). That is, for any sequence $\{\mu_k\}_{k=1}^\infty$ of measures in $\mathcal{P}_p(\Xi)$ and $\mu \in \mathcal{P}_p(\Xi)$, it holds that $\lim_{k \to \infty} W_p(\mu_k, \mu) = 0$ if and only if $\mu_k$ converges weakly to $\mu$ and $\int_\Xi d^p(\xi, \zeta^0) \mu_k(d\xi) \to \int_\Xi d^p(\xi, \zeta^0) \mu(d\xi)$ as $k \to \infty$. Therefore, convergence in the Wasserstein distance of order $p$ implies convergence up to the $p$-th moment. Villani [45, chapter 6] discusses the advantages of Wasserstein distance relative to other distances, such as the Prokhorov metric, that metrize weak convergence.

3. Tractable Reformulation via Duality. In this section we develop a tractable reformulation by deriving its strong dual. We suppress the variable $x$ of $\Psi$ in this section, and results are interpreted pointwise for each $x$. Given $\nu \in \mathcal{P}(\Xi)$ and $\Psi \in L^1(\nu)$, for any $\theta > 0$ and $p \in [1, \infty)$, the inner maximization problem of (DRSO) is written as

$$v_p := \sup_{\mu \in \mathcal{P}(\Xi)} \int_\Xi \Psi(\xi) \mu(d\xi) = \sup_{\mu \in \mathcal{P}(\Xi)} \left\{ \int_\Xi \Psi(\xi) \mu(d\xi) : W_p(\mu, \nu) \leq \theta \right\}. \quad \text{(Primal)}$$

Our main goal is to derive its strong dual

$$v_D := \inf_{\lambda \geq 0} \left\{ \lambda \theta^p - \inf_{\xi, \zeta} \int_\Xi \lambda d^p(\xi, \zeta) - \Psi(\xi) \right\}. \quad \text{(Dual)}$$

The dual problem is a one-dimensional convex minimization problem with respect to $\lambda$, the Lagrangian multiplier of the Wasserstein constraint in the primal problem. The term $\inf_{\xi, \zeta} [\lambda d^p(\xi, \zeta) - \Psi(\xi)]$ is called Moreau-Yosida regularization of $-\Psi$ with parameter $1/\lambda$ in the literature (cf. Parikh and Boyd [33]). Its measurability with respect to $\nu$ will be established in Lemma 3(i) in Section 3.1.

3.1. General Nominal Distribution  In this subsection, we prove the strong duality result for a general nominal distribution $\nu$ on a Polish space $\Xi$. Such generality broadens the applicability of the result for (DRSO). For example, the result is useful when the nominal distribution is some distribution such as a Gaussian distribution on $\mathbb{R}^K$ (Section 4.3), or even some stochastic process (Sections 4.1 and 4.2). We begin with the weak duality, which is an application of Lagrangian weak duality.

**Proposition 1 (Weak duality).** Consider any $\nu \in \mathcal{P}(\Xi)$ and $\Psi \in L^1(\nu)$. Then for any $p \in [1, \infty)$ and $\theta > 0$, it holds that $v_p \leq v_D$.

To prove the strong duality, we consider two separate case: $v_D = \infty$ and $v_D < \infty$. As can be seen from (Dual), if the term $-\int_\Xi \inf_{\xi, \zeta} [\lambda d^p(\xi, \zeta) - \Psi(\xi)]$ is infinite for all $\lambda \geq 0$, then $v_D = \infty$. Thus, to facilitate our analysis, we introduce the following definitions.

**Definition 3 (Regularization Operator $\Phi$).** Let $\Phi : \mathbb{R} \times \Xi \to \mathbb{R} \cup \{-\infty\}$ be given by

$$\Phi(\lambda, \zeta) := \inf_{\xi, \zeta} \{ \lambda d^p(\xi, \zeta) - \Psi(\xi) \}.$$  

**Definition 4 (Growth rate).** Define the growth rate $\kappa$ of $\Psi$ as

$$\kappa := \inf \left\{ \lambda \geq 0 : \int_\Xi \inf_{\xi, \zeta} [\lambda d^p(\xi, \zeta) - \Psi(\xi)] \nu(d\zeta) > -\infty \right\}.$$  

Particularly, if $\int_\Xi \Phi(\lambda, \zeta) \nu(d\zeta) = -\infty$ for all $\lambda \geq 0$, then $\kappa = \infty$.

By definition, for all $\lambda \geq 0$ and all $\zeta \in \Xi$, $\Phi(\lambda, \zeta) \leq -\Psi(\zeta)$. Also, for all $\lambda > \kappa$, $\Phi(\lambda, \cdot) \in L^1(\nu)$ and thus $\Phi(\lambda, \zeta) \in \mathbb{R}$ for $\nu$-almost all $\zeta$. In the sequel, for a function $f : \mathbb{R}_+ \to \mathbb{R} \cup \{-\infty\}$, we denote by dom($f$) its effective domain

$$\text{dom}(f) := \{ \lambda \geq 0 : f(\lambda) > -\infty \},$$

and denote by int(dom($f$)) the interior of dom($f$).
Proposition 2 (Strong duality with infinite optimal value). Consider any \( \nu \in \mathcal{P}(\Xi) \) and \( \Psi \in L^1(\nu) \). Let \( p \in [1, \infty) \). Suppose \( \theta > 0 \) and \( \kappa = \infty \). Then \( v_p = v_D = \infty \).

Remark 1 (Equivalent expression for \( \kappa \)). By Lemma 6 in the Appendix, for any nominal distribution
\[
\nu \in \mathcal{P}_p(\Xi) := \left\{ \mu \in \mathcal{P}(\Xi) : \int_{\Xi} d^p(\zeta, \zeta^0) \nu(d\zeta) < \infty \text{ for some } \zeta^0 \in \Xi \right\},
\]
the growth rate \( \kappa < \infty \) is equivalent to the following condition: there exists \( \zeta_0 \in \Xi \), \( L, M > 0 \) such that \( \Psi(\xi) - \Psi(\zeta_0) \leq L \cdot d^p(\xi, \zeta) + M \) for all \( \xi \in \Xi \). In addition, when \( \Xi \) is unbounded, we have that
\[
\kappa = \limsup_{\xi \in \Xi : d^p(\xi, \zeta) \to \infty} \frac{\max\{0, \Psi(\xi) - \Psi(\zeta)\}}{d^p(\xi, \zeta)}
\]
for any \( \zeta \in \Xi \).

Remark 2 (Choosing Wasserstein order \( p \)). Define
\[
p := \inf \left\{ p \geq 1 : \limsup_{d^p(\xi', \zeta^0)} \frac{\Psi(\epsilon_0) - \Psi(\xi, \zeta^0)}{d^p(\xi', \zeta^0)} < \infty \right\}.
\]
Proposition 2 suggests that a meaningful formulation of (DRSO) should be such that the Wasserstein order \( p \) is greater than or equal to \( p \). In both Esfahani and Kuhn [20] and Zhao and Guan [50] only \( p = 1 \) is considered. By considering higher orders \( p \) in our analysis, we have more flexibility to choose the ambiguity set and control the degree of conservativeness based on the information of function \( \Psi \).

The next theorem establishes the strong duality when the growth rate \( \kappa \) is finite.

Theorem 1 (Strong duality with finite optimal value). Consider any \( \nu \in \mathcal{P}(\Xi) \) and \( \Psi \in L^1(\nu) \). Let \( p \in [1, \infty) \) and \( \theta > 0 \). Suppose \( \kappa < \infty \). Then \( v_p = v_D < \infty \).

To prove this theorem, we first study some properties of the regularization operator \( \Phi \). For any \( \lambda \in \text{dom}(\int_{\Xi} \Phi(\lambda, \zeta) \nu(d\zeta)) \), define \( \overline{D}, \underline{D} : \mathbb{R} \times \Xi \to \mathbb{R} \cup \{+\infty\} \) by
\[
\overline{D}(\lambda, \zeta) := \limsup_{\delta \downarrow 0} \left\{ \sup_{\xi \in \Xi} \left\{ d^p(\xi, \zeta) : \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta \right\} \right\},
\]
\[
\underline{D}(\lambda, \zeta) := \liminf_{\delta \downarrow 0} \left\{ \inf_{\xi \in \Xi} \left\{ d^p(\xi, \zeta) : \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta \right\} \right\}.
\]

Lemma 2 (Properties of the regularization operator \( \Phi \)). Let \( (\Xi, d) \) be a Polish space. Consider any \( \nu \in \mathcal{P}(\Xi) \) and \( \Psi \in L^1(\nu) \). Let \( p \in [1, \infty) \). Suppose \( \kappa < \infty \). Then the following holds \( \nu \)-a.s.

(i) [Monotonicity] \( \Phi(\cdot, \zeta) \) is nondecreasing and concave. In addition, for \( \lambda_2 > \lambda_1 \in \text{dom}(\Phi(\cdot, \zeta)) \), it holds that \( \overline{D}(\lambda_2, \zeta) \leq \overline{D}(\lambda_1, \zeta) \leq \overline{D}(\lambda_1, \zeta) \).

(ii) [Bound] For any \( \lambda > \lambda_0 \in \text{dom}(\Phi(\cdot, \zeta)) \),
\[
(\lambda - \lambda_0) \overline{D}(\lambda, \zeta) \leq \Phi(\lambda_0, \zeta) - \Psi(\zeta).
\]

(iii) [Derivative] For any \( \lambda \in \text{int}(\text{dom}(\Phi(\cdot, \zeta))) \), the left partial derivative \( \partial \Phi(\lambda, \zeta)/\partial \lambda^– \) exists and satisfies
\[
\overline{D}(\lambda, \zeta) \leq \frac{\partial \Phi(\lambda, \zeta)}{\partial \lambda^–} \leq \lim_{\lambda_1 \uparrow \lambda} \underline{D}(\lambda_1, \zeta).
\]

For any \( \lambda \in \text{dom}(\Phi(\cdot, \zeta)) \), the right partial derivative \( \partial \Phi(\lambda, \zeta)/\partial \lambda^+ \) exists and satisfies
\[
\lim_{\lambda_2 \downarrow \lambda} \overline{D}(\lambda_2, \zeta) \leq \frac{\partial \Phi(\lambda, \zeta)}{\partial \lambda^+} \leq \underline{D}(\lambda, \zeta).
\]
LEMMA 3 (Measurability). Consider any \( \nu \in P(\Xi) \) and \( \Psi \in L^1(\nu) \). Let \( p \in [1, \infty) \).

(i) \( \Phi(\lambda, \cdot), \ D(\lambda, \cdot), \) and \( D(\lambda, \cdot) \) are \( \nu \)-measurable.

(ii) Suppose \( \kappa < \infty \). Let \( \lambda \in \text{dom}(\int_\Xi \Phi(\lambda, \zeta) \nu(d\zeta)) \) be such that \( \overline{D}(\lambda, \zeta) < \infty \). For any \( \delta, \epsilon \geq 0 \) such that the sets

\[
E(\zeta) := \{ \xi \in \Xi : \lambda d^p(\xi, \zeta) - \Psi(\lambda) \leq \Phi(\lambda, \zeta) + \delta, \ d^p(\xi, \zeta) \leq \overline{D}(\lambda, \zeta) - \epsilon \},
\]

\[
E(\zeta) := \{ \xi \in \Xi : \lambda d^p(\xi, \zeta) - \Psi(\lambda) \leq \Phi(\lambda, \zeta) + \delta, \ d^p(\xi, \zeta) \leq D(\lambda, \zeta) + \epsilon \}
\]

are non-empty for all \( \zeta \in \Xi \), there exists \( \nu \)-measurable mappings \( \overline{T}, \underline{T} : \Xi \to \Xi \) such that \( \overline{T}(\zeta) \in \overline{E}(\zeta) \) and \( \underline{T}(\zeta) \in \overline{E}(\zeta) \) for \( \nu \)-almost all \( \zeta \).

(iii) Suppose \( \kappa < \infty \). For any \( \epsilon > 0 \), and any \( \nu \)-measurable function \( M \) such that the set

\[
E(\zeta) := \{ \xi \in \Xi : \Psi(\lambda) - \Psi(\zeta) \geq (\kappa - \epsilon)d^p(\xi, \zeta), \ d^p(\xi, \zeta) \geq M(\zeta) \}
\]

is non-empty for all \( \zeta \in \Xi \), there exists a \( \nu \)-measurable mapping \( T : \Xi \to \Xi \) such that \( T(\zeta) \in \overline{E}(\zeta) \) for \( \nu \)-almost all \( \zeta \).

Proof of Theorem 1. As a consequence of weak duality (Proposition 1), it suffices to show that \( v_\nu \geq v_D \). Let \( h : \mathbb{R}_+ \to \mathbb{R} \cup \{ \infty \} \) denote the dual objective function

\[
h(\lambda) := \lambda \theta^p - \int_\Xi \Phi(\lambda, \zeta) \nu(d\zeta).
\]

By Lemma 2(i), \( h(\lambda) \) is the sum of a linear function \( \lambda \theta^p \) and an (extended real-valued) convex function \( \int_\Xi \Phi(\lambda, \zeta) \nu(d\zeta) \) on \([0, \infty)\). In addition, since \( \Phi(\lambda, \zeta) \leq -\Psi(\zeta) \), it follows that \( h(\lambda) \geq \lambda \theta^p + \int_\Xi \Psi(\zeta) \nu(d\zeta) \to \infty \) as \( \lambda \to \infty \). Thus \( h \) is a convex function on \([0, \infty)\) that goes to \( \infty \) as \( \lambda \to \infty \). By Definition 4 of \( \kappa \), either (i) \( h \) has a minimizer \( \lambda^* > \kappa \), or (ii) \( \inf_{\lambda \geq 0} h(\lambda) = \lim_{\lambda \to \kappa} h(\lambda) \). Next we consider these two cases separately. In each case, we construct a sequence of primal feasible solutions which converges to the dual optimal value by exploiting the first-order optimality condition of the dual.

- Case 1: \( h \) has a minimizer \( \lambda^* > \kappa \).

The first-order optimality conditions \( \frac{\partial}{\partial \lambda^+} h(\lambda^*) \leq 0 \) and \( \frac{\partial}{\partial \lambda^-} h(\lambda^*) \geq 0 \) at imply that

\[
\frac{\partial}{\partial \lambda^+} \left( \int_\Xi \Phi(\lambda^*, \zeta) \nu(d\zeta) \right) \leq \theta^p \leq \frac{\partial}{\partial \lambda^-} \left( \int_\Xi \Phi(\lambda^*, \zeta) \nu(d\zeta) \right). \quad (4)
\]

We verify that we can exchange the partial derivative and integration in (4). To show the first inequality, consider any decreasing sequence \( \lambda_n \downarrow \lambda^* \). Let

\[
f_n(\zeta) := \frac{\Phi(\lambda_n, \zeta) - \Phi(\lambda^*, \zeta)}{\lambda_n - \lambda^*}
\]

Since \( \Phi(\cdot, \zeta) \) is nondecreasing for all \( \zeta \), \( f_n(\zeta) \geq 0 \) for all \( \zeta \). In addition, since \( \Phi(\cdot, \zeta) \) is concave, \( f_n \leq f_{n+1} \) for all \( n \). Note that \( \lim_{n \to \infty} f_n(\zeta) = \frac{\partial}{\partial \lambda^+} \Phi(\lambda^*, \zeta) \). Thus it follows from the monotone convergence theorem that

\[
\frac{\partial}{\partial \lambda^+} \left( \int_\Xi \Phi(\lambda^*, \zeta) \nu(d\zeta) \right) = \int_\Xi \frac{\partial}{\partial \lambda^+} \Phi(\lambda^*, \zeta) \nu(d\zeta). \quad (5)
\]

To show the second inequality in (4), consider any increasing sequence \( \lambda_n \uparrow \lambda^* \) with \( \lambda_1 > \kappa \). Then by monotonicity and concavity of \( \Phi(\cdot, \zeta) \), we have that \( \overline{D}(\lambda_1, \zeta) \geq f_n(\zeta) \geq f_{n+1}(\zeta) \geq 0 \) for all \( n \). Let \( \lambda_0 \in (\kappa, \lambda_1) \). It follows from Lemma 2(ii) that

\[
f_1(\zeta) \leq \overline{D}(\lambda_1, \zeta) \leq \frac{1}{\lambda_1 - \lambda_0} (\Psi(\zeta) + \Phi(\lambda_0, \zeta)) \in L^1(\nu).
\]
Thus, applying the reverse Fatou’s lemma on the sequence \( \{f_n\}_n \) gives
\[
\frac{\partial}{\partial \lambda} \left( \int_{\Xi} \Phi(\lambda^*, \zeta) \nu(d\zeta) \right) = \int_{\Xi} \frac{\partial}{\partial \lambda} \Phi(\lambda^*, \zeta) \nu(d\zeta).
\] (6)

Using (4), (5), (6), and Lemma 2(iii), we have that
\[
\begin{align*}
\theta^p &\geq \frac{\partial}{\partial \lambda^+} \left( \int_{\Xi} \Phi(\lambda^*, \zeta) \nu(d\zeta) \right) = \int_{\Xi} \frac{\partial}{\partial \lambda^+} \Phi(\lambda^*, \zeta) \nu(d\zeta) \geq \int_{\Xi} \lim_{\lambda \uparrow \lambda^*} D(\lambda, \zeta) \nu(d\zeta) \\
\theta^p &\leq \frac{\partial}{\partial \lambda^-} \left( \int_{\Xi} \Phi(\lambda^*, \zeta) \nu(d\zeta) \right) = \int_{\Xi} \frac{\partial}{\partial \lambda^-} \Phi(\lambda^*, \zeta) \nu(d\zeta) \leq \int_{\Xi} \lim_{\lambda \downarrow \lambda^*} D(\lambda, \zeta) \nu(d\zeta).
\end{align*}
\] (7)

Observe that for any \( \lambda > \kappa \) and \( \delta, \epsilon > 0 \), the sets \( \overline{E}, \underline{E} \) in Lemma 3(ii) are well-defined. Hence there exists \( \nu \)-measurable mappings \( \overline{T}, \underline{T} : \Xi \to \Xi \) such that for \( \nu \)-almost all \( \zeta \),
\[
\begin{align*}
\overline{T}_\lambda(\zeta) &\in \left\{ \xi \in \Xi : \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta, \ d^p(\xi, \zeta) > D(\lambda, \zeta) - \epsilon \right\} \\
\underline{T}_\lambda(\zeta) &\in \left\{ \xi \in \Xi : \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta, \ d^p(\xi, \zeta) < D(\lambda, \zeta) + \epsilon \right\}.
\end{align*}
\]

In particular, for any \( \lambda_1, \lambda_2 \) with \( \kappa < \lambda_1 < \lambda^* < \lambda_2 \), it follows from (7) and Lemma 2(i) that
\[
\begin{align*}
\theta^p &\geq \int_{\Xi} \lim_{\lambda \uparrow \lambda^*} D(\lambda, \zeta) \nu(d\zeta) \geq \int_{\Xi} D(\lambda_2, \zeta) \nu(d\zeta) \geq \int_{\Xi} D(\lambda_2, \zeta) \nu(d\zeta) \geq \int_{\Xi} d^p(\overline{T}_{\lambda_2}(\zeta), \zeta) \nu(d\zeta) - \epsilon, \\
\theta^p &\leq \int_{\Xi} \lim_{\lambda \downarrow \lambda^*} D(\lambda, \zeta) \nu(d\zeta) \leq \int_{\Xi} D(\lambda_1, \zeta) \nu(d\zeta) \leq \int_{\Xi} D(\lambda_1, \zeta) \nu(d\zeta) \leq \int_{\Xi} d^p(\overline{T}_{\lambda_1}(\zeta), \zeta) \nu(d\zeta) + \epsilon.
\end{align*}
\] (8)

Based on (8), we now construct a feasible primal solution. Note that there is a \( q_\delta^*(\lambda_1, \lambda_2) \in [0, 1] \) such that
\[
q_\delta^*(\lambda_1, \lambda_2) \left[ \int_{\Xi} d^p(\overline{T}_{\lambda_1}(\zeta), \zeta) \nu(d\zeta) + \epsilon \right] + (1 - q_\delta^*(\lambda_1, \lambda_2)) \left[ \int_{\Xi} d^p(\overline{T}_{\lambda_2}(\zeta), \zeta) \nu(d\zeta) - \epsilon \right] = \theta^p.
\] (9)

Let \( q^* := \frac{\theta^p}{\theta^p + \max\{0, 1 - 2q_\delta^*(\lambda_1, \lambda_2)\}} \). Define a distribution \( \mu^*_\delta(\lambda_1, \lambda_2) \) by
\[
\mu^*_\delta(\lambda_1, \lambda_2) := q^* q_\delta^*(\lambda_1, \lambda_2) \cdot (\overline{T}_{\lambda_1})_# \nu + q^* (1 - q_\delta^*(\lambda_1, \lambda_2)) \cdot (\overline{T}_{\lambda_2})_# \nu + (1 - q^*) \nu.
\]

Then \( \mu^*_\delta(\lambda_1, \lambda_2) \) is primal feasible:
\[
W^p_p(\mu^*_\delta(\lambda_1, \lambda_2), \nu) \leq q^* q_\delta^*(\lambda_1, \lambda_2) \int_{\Xi} d^p(\overline{T}_{\lambda_1}(\zeta), \zeta) \nu(d\zeta) + q^* (1 - q_\delta^*(\lambda_1, \lambda_2)) \int_{\Xi} d^p(\overline{T}_{\lambda_2}(\zeta), \zeta) \nu(d\zeta) + q^* (\theta^p + (1 - 2q_\delta^*(\lambda_1, \lambda_2)) \epsilon) \leq \theta^p.
\]

Furthermore, recall that for all \( \zeta \in \Xi, \)
\[
\begin{align*}
\lambda_1 d^p(\overline{T}_{\lambda_1}(\zeta), \zeta) - \Phi(\lambda_1, \zeta) - \delta &\leq \Psi(\overline{T}_{\lambda_1}(\zeta)) \leq \lambda_1 d^p(\overline{T}_{\lambda_1}(\zeta), \zeta) - \Phi(\lambda_1, \zeta), \\
\lambda_2 d^p(\overline{T}_{\lambda_2}(\zeta), \zeta) - \Phi(\lambda_2, \zeta) - \delta &\leq \Psi(\overline{T}_{\lambda_2}(\zeta)) \leq \lambda_2 d^p(\overline{T}_{\lambda_2}(\zeta), \zeta) - \Phi(\lambda_2, \zeta).
\end{align*}
\]

This, together with (9), implies that
\[
\begin{align*}
v^p &\geq \int_{\Xi} \Psi(\zeta) \mu^*_\delta(\lambda_1, \lambda_2)(d\zeta) \\
= &q^* q_\delta^*(\lambda_1, \lambda_2) \int_{\Xi} \Psi(\overline{T}_{\lambda_1}(\zeta)) \nu(d\zeta) + q^* (1 - q_\delta^*(\lambda_1, \lambda_2)) \int_{\Xi} \Psi(\overline{T}_{\lambda_2}(\zeta)) \nu(d\zeta) + (1 - q^*) \int_{\Xi} \Psi(\zeta) \nu(d\zeta)
\end{align*}
\]
\[ \geq q' q^*_\delta(\lambda_1, \lambda_2) \int_{\Xi} [\lambda_1 d^p(T_{\lambda_1}(\zeta), \zeta) - \Phi(\lambda_1, \zeta) - \delta] \nu(d\zeta) \]
\[ + q' (1 - q^*_\delta(\lambda_1, \lambda_2)) \int_{\Xi} [\lambda_2 d^p(T_{\lambda_2}(\zeta), \zeta) - \Phi(\lambda_2, \zeta) - \delta] \nu(d\zeta) + (1 - q') \int_{\Xi} \Psi(\zeta) \nu(d\zeta) \]
\[ \geq q' \lambda_1 \theta^p + \int_{\Xi} \Phi(\lambda_2, \zeta) \nu(d\zeta) - q' \delta + (1 - q') \int_{\Xi} \Psi(\zeta) \nu(d\zeta). \]

Let \( \lambda_1 \uparrow \lambda^* \) and \( \lambda_2 \downarrow \lambda^* \), and apply the monotone convergence theorem on \( \int_{\Xi} \Phi(\lambda_1, \zeta) \nu(d\zeta) \) and \( \int_{\Xi} \Phi(\lambda_2, \zeta) \nu(d\zeta) \),
\[ \lim_{\lambda_1 \uparrow \lambda^*, \lambda_2 \downarrow \lambda^*} \left\{ q^*_\delta(\lambda_1, \lambda_2) \int_{\Xi} \Phi(\lambda_1, \zeta) \nu(d\zeta) + (1 - q^*_\delta(\lambda_1, \lambda_2)) \int_{\Xi} \Phi(\lambda_2, \zeta) \nu(d\zeta) \right\} = \int_{\Xi} \Phi(\lambda^*, \zeta) \nu(d\zeta), \]

hence
\[ v_P \geq q' \lambda^* \theta^p + \limsup_{\lambda_1 \uparrow \lambda^*, \lambda_2 \downarrow \lambda^*} (1 - 2q^*_\delta(\lambda_1, \lambda_2)) \epsilon - q' \int_{\Xi} \Phi(\lambda^*, \zeta) \nu(d\zeta) - q' \delta + (1 - q') \int_{\Xi} \Psi(\zeta) \nu(d\zeta). \]

Taking the limit as \( \epsilon \to 0 \), and observing that \( q^* = \frac{\theta^p}{\theta^p + \max(0, 1 - 2q^*_\delta(\lambda_1, \lambda_2))} \in (\frac{\theta^p}{\theta^p + \epsilon}, 1) \), it follows that
\[ v_P \geq \lambda^* \theta^p - \int_{\Xi} \Phi(\lambda^*, \zeta) \nu(d\zeta) - \delta. \]

Since \( \delta > 0 \) can be arbitrarily small, it follows that \( v_P \geq v_D \).

- Case 2: \( \inf_{\lambda \geq 0} h(\lambda) = \lim_{\lambda \uparrow \kappa} h(\lambda) \), and no \( \lambda > \kappa \) is dual optimal.

Then \( h \) is strictly increasing and convex on \( (\kappa, \infty) \). For any \( \lambda > \lambda_0 > \kappa \), \( h(\lambda) > h(\lambda_0) \) is equivalently written into
\[ \int_{\Xi} [\Phi(\lambda, \zeta) - \Phi(\lambda_0, \zeta)] \nu(d\zeta) < (\lambda - \lambda_0) \theta^p. \]

Consider any \( \underline{\delta} \in (0, (\lambda - \lambda_0) \theta^p - \int_{\Xi} [\Phi(\lambda, \zeta) - \Phi(\lambda_0, \zeta)] \nu(d\zeta)) \). It follows from Lemma 3 that there exists a \( \nu \)-measurable map \( T_\lambda : \Xi \to \Xi \) such that \( \lambda d^p(T_\lambda(\zeta), \zeta) - \Psi(T_\lambda(\zeta)) \leq \Phi(\lambda, \zeta) + \delta \). Also, note that \( \Phi(\lambda_0, \zeta) \leq \lambda_0 d^p(T_\lambda(\zeta), \zeta) - \Psi(T_\lambda(\zeta)) \). Thus,
\[ \Phi(\lambda, \zeta) - \Phi(\lambda_0, \zeta) \geq (\lambda - \lambda_0) d^p(T_\lambda(\zeta), \zeta) - \delta. \]

Combining this with (10) yields
\[ \int_{\Xi} d^p(T_\lambda(\zeta), \zeta) \nu(d\zeta) < \theta^p. \]

Hence, the distribution \( (T_\lambda)_\# \nu \) is primal feasible.

Next, we separately consider the cases \( \kappa = 0 \) and \( \kappa > 0 \). If \( \kappa = 0 \), then
\[ v_P \geq \int_{\Xi} \Psi(\xi)(T_\lambda)_\# \nu(d\xi) \geq \int_{\Xi} [\lambda d^p(T_\lambda(\zeta), \zeta) - \Phi(\lambda, \zeta) - \delta] \nu(d\zeta) \geq - \int_{\Xi} \Phi(\lambda, \zeta) \nu(d\zeta) - \delta. \]

Since \( \delta \) can be chosen arbitrarily small, it follows that \( v_P \geq - \int_{\Xi} \Phi(\lambda, \zeta) \nu(d\zeta) \) for all \( \lambda > \kappa \). Therefore
\[ v_P \geq \lim_{\lambda \downarrow 0} \left\{ - \int_{\Xi} \Phi(\lambda, \zeta) \nu(d\zeta) \right\} = \lim_{\lambda \downarrow 0} \left\{ \lambda \theta^p - \int_{\Xi} \Phi(\lambda, \zeta) \nu(d\zeta) \right\} = \inf_{\lambda \geq 0} h(\lambda) = v_D. \]
Otherwise, if $\kappa > 0$, for any $\epsilon \in (0, \kappa)$, define

$$\overline{D}(\kappa - \epsilon, \zeta) := \sup_{\xi \in \Xi} \{d^p(\xi, \zeta) : \Psi(\xi) - \Psi(\zeta) \geq (\kappa - \epsilon)d^p(\xi, \zeta)\}.$$

Then we claim that $\int_{\Xi} \overline{D}(\kappa - \epsilon, \zeta)\nu(d\zeta) = \infty$. Indeed, note that $\Psi(\xi) \leq \lambda_0 d^p(\xi, \zeta) - \Phi(\lambda_0, \zeta)$ for all $\xi \in \Xi$ and any $\lambda_0 > \kappa$, and thus

$$\int_{\Xi} \Phi(\kappa - \epsilon, \zeta)\nu(d\zeta) = \int_{\Xi} \inf_{\xi \in \Xi} \left\{ (\kappa - \epsilon)d^p(\xi, \zeta) - \Psi(\xi) : \Psi(\xi) - \Psi(\zeta) \geq (\kappa - \epsilon)d^p(\xi, \zeta) \right\} \nu(d\zeta)$$

$$\geq \int_{\Xi} \inf_{\xi \in \Xi} \left\{ -\Psi(\xi) : \Psi(\xi) - \Psi(\zeta) \geq (\kappa - \epsilon)d^p(\xi, \zeta) \right\} \nu(d\zeta)$$

$$\geq \int_{\Xi} \inf_{\xi \in \Xi} \left\{ -\lambda_0 d^p(\xi, \zeta) + \Phi(\lambda_0, \zeta) : \Psi(\xi) - \Psi(\zeta) \geq (\kappa - \epsilon)d^p(\xi, \zeta) \right\} \nu(d\zeta)$$

$$\geq -\lambda_0 \int_{\Xi} \overline{D}(\kappa - \epsilon, \zeta)\nu(d\zeta) + \int_{\Xi} \Phi(\lambda_0, \zeta)\nu(d\zeta).$$

Then $\int_{\Xi} \overline{D}(\kappa - \epsilon, \zeta)\nu(d\zeta)$ cannot be finite, since otherwise $\int_{\Xi} \Phi(\kappa - \epsilon, \zeta)\nu(d\zeta) > -\infty$, which contradicts the Definition 4 of $\kappa$. The above claim implies that for any $R > 0$, there exists $M \in L^1(\nu)$ with $\int_{\Xi} Md\nu \geq R$ such that for $\nu$-almost all $\zeta$,

$$\{ \xi \in \Xi : \Psi(\xi) - \Psi(\zeta) \geq (\kappa - \epsilon)d^p(\xi, \zeta), d^p(\xi, \zeta) \geq M(\zeta) \} \neq \emptyset.$$

Then by Lemma 3(iii), for any $\epsilon > 0$, $R > \theta^p$, there exists a $\nu$-measurable mapping $\overline{T}_x^R : \Xi \to \Xi$ such that

$$\overline{T}_x^R(\zeta) - \Psi(\zeta) \geq (\kappa - \epsilon) \cdot d^p(\overline{T}_x^R(\zeta), \zeta), \ \forall \zeta \in \Xi,$$

and that

$$\int_{\Xi} d^p(\overline{T}_x^R(\zeta), \zeta)\nu(d\zeta) > R.$$

Without loss of generality we can assume that $\int_{\Xi} d^p(\overline{T}_x^R(\zeta), \zeta)\nu(d\zeta) < \infty$, since otherwise consider $E_r := \{ \zeta \in \Xi : d^p(\overline{T}_x^R(\zeta), \zeta) \leq r \}$, then $\lim_{r \to \infty} E_r = \Xi$, and thus there exists a $r_0 > 0$ such that $R < \int_{E_{r_0}} d^p(T^\mu(\zeta), \zeta)\nu(d\zeta) < \infty$. Now we can construct a primal feasible solution

$$\mu^R_\delta(\lambda, \epsilon) := q^R_\epsilon(\overline{T}_x^R)\#\nu + (1 - q^R_\epsilon)(\overline{T}_x^R)\#\nu,$$

where $q^R_\epsilon \in (0, 1)$ is such that

$$q^R_\epsilon \int_{\Xi} d^p(T_\lambda(\zeta), \zeta)\nu(d\zeta) + (1 - q^R_\epsilon) \int_{\Xi} d^p(\overline{T}_x^R(\zeta), \zeta)\nu(d\zeta) = \theta^p.$$

Then by construction $\mu^R_\delta(\lambda, \epsilon)$ is primal feasible. Moreover,

$$v_p \geq \int_{\Xi} \Psi(\xi)\mu^R_\delta(\lambda, \epsilon)(d\xi)$$

$$= q^R_\epsilon \int_{\Xi} \Psi(\overline{T}_x^R(\zeta))\nu(d\zeta) + (1 - q^R_\epsilon) \int_{\Xi} \Psi(\overline{T}_x^R(\zeta))\nu(d\zeta)$$

$$\geq q^R_\epsilon \int_{\Xi} (\lambda d^p(\overline{T}_x^R(\zeta), \zeta) - \Phi(\lambda, \zeta) - \delta)\nu(d\zeta) + (1 - q^R_\epsilon) \int_{\Xi} (\kappa - \epsilon)d^p(\overline{T}_x^R(\zeta), \zeta) + \Psi(\zeta)\nu(d\zeta)$$

$$\geq (\kappa - \epsilon)q^R_\epsilon \int_{\Xi} d^p(\overline{T}_x^R(\zeta), \zeta)\nu(d\zeta) + (1 - q^R_\epsilon) \int_{\Xi} (\kappa - \epsilon)d^p(\overline{T}_x^R(\zeta), \zeta)\nu(d\zeta)$$

$$- q^R_\epsilon \int_{\Xi} \Phi(\lambda, \zeta)\nu(d\zeta) - q^R_\epsilon \delta + (1 - q^R_\epsilon) \int_{\Xi} \Psi(\zeta)\nu(d\zeta)$$

$$= (\kappa - \epsilon)\theta^p - q^R_\epsilon \int_{\Xi} \Phi(\lambda, \zeta)\nu(d\zeta) - q^R_\epsilon \delta + (1 - q^R_\epsilon) \int_{\Xi} \Psi(\zeta)\nu(d\zeta).$$
For any fixed \( \lambda > \kappa, \delta, \epsilon \) can be chosen arbitrarily small, and \( q^\delta \) can be arbitrarily close to 1 by choosing sufficiently large \( R \). Thus, \( v_P \geq \kappa \theta^p - \int_{\Xi} \Phi(\lambda, \zeta) \nu(d\zeta) - \delta \). Hence \( v_P \geq \inf_{\lambda > \kappa} \{ \lambda \theta^p - \int_{\Xi} \Phi(\lambda, \zeta) \nu(d\zeta) \} = v_D. \)

**Remark 3.** We remark all the above results and proofs in Section 3.1 (expect for Remark 1) remains to hold if we replace the transportation cost \( d^p(\cdot, \cdot) \) with any measurable, non-negative cost function \( c(\cdot, \cdot) \) that satisfies \( c(\xi, \zeta) = 0 \) if and only if \( \xi = \zeta \).

Next, we investigate existence conditions for worst-case distributions and their structure. In the rest of this subsection, we assume that \( \Psi \) is upper-semicontinuous, and every bounded subset in \((\Xi, d)\) is totally bounded, which is satisfied by, for example, any finite-dimensional normed space. Under this assumption, when \( \lambda > \kappa \), Lemma 2(ii) and the upper semi-continuity of \( \Psi \) imply that the set \( \arg \min_{\xi \in \Xi} \{ \lambda d^p(\xi, \zeta) - \Psi(\xi) \} \) is nonempty, and that \( \min / \max_{\xi \in \Xi} \{ d^p(\xi, \zeta) : \lambda d^p(\xi, \zeta) - \Psi(\xi) = \Phi(\lambda, \zeta) \} \) can be obtained. When \( \lambda = \kappa \) and \( \nu(\{ \xi \in \Xi : \arg \min_{\xi \in \Xi} \{ \lambda d^p(\xi, \zeta) - \Psi(\xi) \} = \emptyset \}) = 0 \), the upper semi-continuity of \( \Psi \) imply that the set \( \min_{\xi \in \Xi} \{ d^p(\xi, \zeta) : \lambda d^p(\xi, \zeta) - \Psi(\xi) = \Phi(\lambda, \zeta) \} \) can be obtained, but \( \max_{\xi \in \Xi} \{ d^p(\xi, \zeta) : \lambda d^p(\xi, \zeta) - \Psi(\xi) = \Phi(\lambda, \zeta) \} \) can be infinite. Thus, when (i) \( \lambda > \kappa \), and (ii) \( \lambda = \kappa \), and \( \nu(\{ \xi \in \Xi : \arg \min_{\xi \in \Xi} \{ \lambda d^p(\xi, \zeta) - \Psi(\xi) \} = \emptyset \}) = 0 \), the following quantity

\[
\begin{align*}
\overline{D}_0(\lambda, \zeta) &:= \max_{\xi \in \Xi} \{ d^p(\xi, \zeta) : \lambda d^p(\xi, \zeta) - \Psi(\xi) = \Phi(\lambda, \zeta) \}, \\
\underline{D}_0(\lambda, \zeta) &:= \min_{\xi \in \Xi} \{ d^p(\xi, \zeta) : \lambda d^p(\xi, \zeta) - \Psi(\xi) = \Phi(\lambda, \zeta) \},
\end{align*}
\]

are well-defined (where \( \overline{D}_0(\lambda, \zeta) \) can be infinite). Then \( \underline{D}_0(\lambda, \zeta) \) and \( \overline{D}_0(\lambda, \zeta) \) represent respectively the closest and furthest distances between \( \zeta \) and any point in \( \arg \min_{\xi \in \Xi} \{ \lambda d^p(\xi, \zeta) - \Psi(\xi) \} \). We note that \( \overline{D}_0(\lambda, \zeta) \) (resp. \( \underline{D}_0(\lambda, \zeta) \)) may not be equal to \( \overline{D}(\lambda, \zeta) \) (resp. \( \underline{D}(\lambda, \zeta) \)) as defined in (3).

**Corollary 1 (Worst-case distribution).** Consider any \( \nu \in \mathcal{P}(\Xi) \) and \( \Psi \in L^1(\nu) \). Let \( p \in [1, \infty) \) and \( \theta > 0 \). Suppose \( \kappa < \infty \). Assume \( \Psi \) is upper-semicontinuous, and that bounded subsets of \((\Xi, d)\) are totally bounded. Then the following holds:

(i) [Existence condition] A worst-case distribution exists if and only if any of the following conditions hold:

1. There exists a dual minimizer \( \lambda^* > \kappa \).
2. \( \lambda^* = \kappa > 0 \) is the unique minimizer, \( \nu(\{ \xi \in \Xi : \arg \min_{\xi \in \Xi} \{ \kappa d^p(\xi, \zeta) - \Psi(\xi) \} = \emptyset \}) = 0 \), and

\[
\int_{\Xi} \underline{D}_0(\kappa, \zeta) \nu(d\zeta) \leq \theta^p \leq \int_{\Xi} \overline{D}_0(\kappa, \zeta) \nu(d\zeta).
\]

3. \( \lambda^* = 0 \) is the unique minimizer, \( \arg \max_{\xi \in \Xi} \{ \Psi(\xi) \} \) is nonempty, and

\[
\int_{\Xi} \underline{D}_0(0, \zeta) \nu(d\zeta) \leq \theta^p.
\]

(ii) If \( \nu(\zeta \in \Xi : -\Psi(\zeta) > \inf_{\xi \in \Xi} \{ \kappa d^p(\xi, \zeta) - \Psi(\xi) \}) = 0 \), then \( \lambda^* = \kappa \) for any \( \theta > 0 \). Otherwise, there is \( \theta_0 > 0 \) such that \( \lambda^* > \kappa \) for any \( \theta < \theta_0 \).

(iii) [Structure] Whenever a worst-case distribution exists, there exists a worst-case distribution \( \mu^* \) which can be represented as a convex combination of two distributions \( \overline{T}^* \# \nu \) and \( \underline{T}^* \# \nu \), each of which is a perturbation of \( \nu \), as follows:

\[
\mu^* = p^* \overline{T}^* \# \nu + (1 - p^*) \underline{T}^* \# \nu,
\]

where \( p^* \in [0, 1] \), and \( \overline{T}^*, \underline{T}^* : \Xi \to \Xi \) satisfy

\[
\nu(\zeta \in \Xi : \overline{T}^*(\zeta), \underline{T}^*(\zeta) \notin \arg \min_{\xi \in \Xi} \{ \lambda^* d^p(\xi, \zeta) - \Psi(\xi) \}) = 0 \tag{11}
\]
(iv) If \( \Xi \) is convex, \( \Psi \) is concave, and \( d^p(\cdot, \zeta) \) is convex for all \( \zeta \in \Xi \), then there exists \( T^\ast : \Xi \to \Xi \) such that \( T^\ast \nu \) is optimal.

**Remark 4.** Compared with Corollary 4.7 in Esfahani and Kuhn [20], Corollary 1(i) provides a complete description of the necessary and sufficient conditions for the existence of a worst-case distribution. Note that Example 1 in Esfahani and Kuhn [20] corresponds to \( \lambda^\ast = \kappa = 1 \) and \( p = 1 \).

**Example 4.** We present several examples that correspond to different cases in Corollary 1(i). In all these examples, \( \Xi = [0, \infty) \), \( d(\xi, \zeta) = |\xi - \zeta| \) for all \( \xi, \zeta \in \Xi \), \( p = 1 \), \( \theta > 0 \), and \( \nu = \delta_0 \).

\[
\begin{align*}
\text{(a) } \Psi_a(\xi) &= \max\{0, \xi - a\} \\
\text{(b) } \Psi(\xi) &= \max\{0, 1 - \xi^2\} \\
\text{(c) } \Psi_{\pm}(\xi) &= 1 + \xi \pm \frac{1}{\xi + 1}
\end{align*}
\]

**Figure 2.** Examples for existence and non-existence of the worst-case distribution

1. \( \Psi_a(\xi) = \max\{0, \xi - a\} \) for some \( a \in \mathbb{R} \). It follows that \( \lambda^\ast = \kappa = 1 \).
   - If \( a \leq 0 \), then \( \arg\min_{\xi \in \Xi} \{d^p(\xi, 0) - \Psi_a(\xi)\} = [0, \infty) \), hence \( D^a(\kappa, \zeta) = 0 \) and \( D^a(\kappa, \zeta) = \infty \) satisfying condition (2.). One of the worst-case distributions is \( \mu^\ast = \delta_\theta \) with \( v_P = v_D = \theta - a \).
   - If \( a > 0 \), then \( \arg\min_{\xi \in \Xi} \{d^p(\xi, 0) - \Psi_a(\xi)\} = \{0\} \), hence \( D^a(\kappa, \zeta) = D^a(\kappa, \zeta) = 0 < \theta \), thus condition (2.) is violated. There is no worst-case distribution, but the objective value of \( \mu^\ast = (1 - \epsilon)\delta_0 + \epsilon\delta_{\theta/\epsilon} \) converges to \( v_P = v_D = \theta \) as \( \epsilon \to 0 \).

2. \( \Psi(\xi) = \max\{0, 1 - \xi^2\} \). It follows that \( \lambda^\ast = \kappa = 0 \), and \( \arg\max_{\xi \in \Xi} \Psi(\xi) = \{0\} \). Thus condition (3.) is satisfied, and the worst-case distribution is \( \mu^\ast = \delta_0 = \nu \).

3. \( \Psi_{\pm}(\xi) = 1 + \xi \pm \frac{1}{\xi + 1} \). It follows that \( \kappa = 1 \). Note that \( \Psi_{\pm}'(\xi) = 1 + \frac{1}{(\xi + 1)^2} \).
   - Note that \( \Psi_{\pm}'(\xi) < \kappa = 1 \) on \( \Xi \). Also, \( \Psi_{\pm} \) satisfies the condition in (ii), thus for all \( \theta > 0 \) it holds that \( \lambda^\ast_\pm = \kappa = 1 \) and \( \arg\min_{\xi \in \Xi} \{\lambda^\ast d^p(\xi, 0) - \Psi_{\pm}(\xi)\} = \{0\} \). There is no worst-case distribution, but the objective value of \( \mu^\ast = (1 - \epsilon)\delta_0 + \epsilon\delta_{\theta/\epsilon} \) converges to \( v_P = v_D = 2 + \theta \) as \( \epsilon \to 0 \).
   - Note that \( \Psi_{\pm}'(\xi) > \kappa = 1 \) on \( \Xi \). Also, \( \arg\min_{\lambda \geq 0} \left\{\lambda\theta - \inf_{\xi \in \Xi} \left\{\lambda\xi - \left(1 + \xi - \frac{1}{\xi + 1}\right)\right\}\right\} = \arg\min_{\lambda \geq 1} \left\{\lambda(\theta - 2\sqrt{\lambda - 1})\right\} = \left\{1 + \frac{1}{\theta + 1}\right\} \}. Thus \( \lambda^\ast_\pm > 1 = \kappa \).

**3.2. Finite-Supported Nominal Distribution** In this section, we restrict attention to the case in which the nominal distribution \( \nu = \frac{1}{N} \sum_{i=1}^N \delta_{\xi^i} \) for some \( \xi^i \in \Xi \), \( i = 1, \ldots, N \). This occurs, for example, in a data-driven setting in which the decision maker collects \( N \) observations that constitute an empirical distribution.

**Corollary 2 (Data-Driven DRSO).** Consider any \( \nu = \frac{1}{N} \sum_{i=1}^N \delta_{\xi^i} \). Let \( p \in [1, \infty) \) and \( \theta > 0 \). Then the following hold:

(i) [Strong duality] The primal problem (Primal) has a strong dual problem

\[
\begin{align*}
v_P = v_D = \inf_{\lambda \geq 0} \left\{\lambda\theta + \frac{1}{N} \sum_{i=1}^N \sup_{\xi \in \Xi} \left[\Psi(\xi) - \lambda d^p(\xi, \xi^i)\right]\right\}.
\end{align*}
\]
(ii) [Structure of the worst-case distribution] Whenever a worst-case distribution exists, there exists one which is supported on at most \(N + 1\) points and has the form

\[
\mu^* = \frac{1}{N} \sum_{i \neq i_0} \delta_{\xi_i} + \frac{p_0}{N} \delta_{\xi_{i_0}} + \frac{1 - p_0}{N} \delta_{\xi_{i_0}^*}
\]

where \(i_0 \in \{1, \ldots, N\}, \ p_0 \in [0,1], \ \xi_{i_0}, \xi_{i_0}^* \in \arg\min_{\xi \in \Xi} \{\lambda^* d^p(\xi, \xi_{i_0}) - \Psi(\xi)\}, \) and \(\xi_i \in \arg\min_{\xi \in \Xi} \{\lambda^* d^p(\xi, \xi_i) - \Psi(\xi)\} \) for all \(i \neq i_0\).

(iii) [Robust-program approximation] Suppose that there exists \(\xi_0 \in \Xi, \ L, M \geq 0\) such that \(|\Psi(\xi) - \Psi(\xi_0)| < Ld^p(\xi, \xi_0) + M\) for all \(\xi \in \Xi\). Let \(K\) be any positive integer and consider the robust optimization problem

\[
v_K := \sup_{(\xi^i_{i,k})_{i,k} \in \mathcal{M}_K} \frac{1}{NK} \sum_{i=1}^N \sum_{k=1}^K \Psi(\xi^i_{i,k}),
\]

with uncertainty set

\[
\mathcal{M}_K := \left\{(\xi^i_{i,k})_{i,k} : \frac{1}{NK} \sum_{i=1}^N \sum_{k=1}^K d^p(\xi^i_{i,k}, \xi^*_i) \leq \theta^p, \ \xi^i_{i,k} \in \Xi \ \forall \ i, k\right\}.
\]

If \(\lambda^* > \kappa\), then there exists \(L', M' > 0\), such that

\[
v_K \leq \sup_{\mu \in \mathcal{M}} \mathbb{E}_\mu[|\Psi(\xi)|] \leq v_K + \frac{M' + L'D}{NK},
\]

where \(D\) is a constant independent of \(K\). In addition, if \(\Xi\) is convex and \(\Psi\) is concave, then \(v_1 = v_P = v_D\).

Statement (ii) shows that the worst-case distribution \(\mu^*\) is a perturbation of \(\nu = \frac{1}{N} \sum_{i=1}^N \delta_{\xi_i}\), where \(N - 1\) out of the \(N\) points, \(\{\xi_i\}_{i \neq i_0}\), are perturbed with full mass \(1/N\) to a maximizer \(\xi_i^*\) respectively, while at most one point \(\xi_{i_0}^*\) is split and perturbed to two maximizers \(\xi_{i_0}^0\) and \(\xi_{i_0}^0\). (If the set of maximizers is a singleton, then there is no need to split). Using this structure, we obtain statement (iii), which suggests that the primal problem can be approximated by a robust program with uncertainty set \(\mathcal{M}_K\), which is a subset of \(\mathcal{M}\) that contains all distributions supported on \(NK\) points with equal probability \(\frac{1}{NK}\). Particularly, when \(\Psi\) is concave, such approximation is exact; and when \(\Psi\) is Lipschitz and \(p = 1\), then \(v_1\) is an \(O(1/N)\)-approximation of \(v_P = v_D\).

Remark 5. The results in Corollary 2 hold for arbitrary metric space \(\Xi\). In fact, the Polish space assumption on \(\Xi\) is only used for the measurability results in Lemma 3, which becomes trivial in finite-supported case.

Remark 6. Under compactness assumption on \(\Xi\), Wozabal [47] pointed out that to solve (Primal), it suffices to consider the set of extreme points of the Wasserstein ball \(\mathcal{M}\) contains distributions that are supported on at most \(N + 3\) points. Later in Owhadi and Scovel [31], this result was improved to \(N + 2\) for Polish space or Borel subsets of Polish space. Statement (ii) further strength these results — for arbitrary metric space (see Remark 5 above), it suffices to consider distributions that are supported on at most \(N + 1\) points, and such bound is tight as shown by Example 7 below. Moreover, the weight of the extreme distribution do not change much as compared to the nominal distribution. As can be immediately seen from the proof, the result of statement (ii) can be generalized as following. Suppose \(\nu = \frac{1}{N} \sum_{i=1}^N \nu_i \delta_{\xi_i}\), then whenever the worst-case distribution exists, there exists one of the form

\[
\sum_{i \neq i_0} \nu_i \delta_{\xi_i} + p_0 \nu_{i_0} \delta_{\xi_{i_0}} + (1 - p_0) \nu_{i_0} \delta_{\xi_{i_0}^*}.
\]
Remark 7 (Total Variation metric). By choosing the discrete metric $d(\xi, \zeta) = 1_{\{\xi \neq \zeta\}}$ on $\Xi$, the Wasserstein distance is equal to Total Variation distance (Gibbs and Su [22]), which can be used for the situation where the distance of perturbation does not matter and provides a rather conservative decision. In this case, suppose $\theta$ is chosen such that $N\theta$ is an integer, then there is no fractional point in (13) and the problem is reduced to the robust program with uncertainty set $\mathcal{M}_1$, whether $\Xi$ ($\Psi$) is convex (concave) or not.

Proof of Corollary 2.

(i) It follows directly from the proof of Theorem 1 and Proposition 2.

(ii) By Corollary 1(iii), whenever the worst-case distribution exists, there is one supported on at most $2N$ points and has the form

$$\mu^* = \frac{1}{N} \sum_{i=1}^{N} p_i^* \delta_{\xi^*_i} + (1 - p_i^*) \delta_{\xi^*_i}$$

where $p^i \in [0, 1]$, and $\xi^*_i, \xi^*_i \in \arg\min_{\xi \in \Xi} \{\lambda^* d^p(\xi, \hat{x}) - \Psi(\xi)\}$. (In fact, Corollary 1(iii) proves a stronger statement that there exists a worst-case distribution such that all $p^i$ are equal, but here we allow them to vary in order to obtain a worst-case distribution with a different form.) Given $\xi^*_i, \xi^*_i$ for all $i$, the problem

$$\max_{0 \leq p^i \leq 1} \left\{ \frac{1}{N} \sum_{i=1}^{N} p_i^* \Psi(\xi^*_i) + (1 - p_i^*) \Psi(\xi^*_i) : \frac{1}{N} \sum_{i=1}^{N} p^i d^p(\xi^*_i, \hat{x}^i) + (1 - p^i) d^p(\xi^*_i, \hat{x}^i) \leq \theta^p \right\}$$

is a linear program with $N$ variables, one equality constraint and $2N$ inequality constraints $p_i \leq 1$, $p_i \geq 1$, $i = 1, \ldots, N$. Thus according to linear programming theory, there exists an optimal solution such that among the $2N$ inequality constraints, at least $N - 1$ of them hold as equality, or equivalently, at most one $p_i$ is fractional. Therefore there exists a worst-case distribution which is supported on at most $N + 1$ points, and has the form (13).

(iii) Note that by assumption on $\Psi$ and Remark 1 we have $\kappa \leq L < \infty$. Also note that using the similar idea the above proof of (ii), the distributions $\mu^p_\delta(\lambda_1, \lambda_2, (T_{\lambda}), \nu)$ and $\mu^R_\delta(\lambda, \epsilon)$ defined in the proof of Theorem 1 can be written in the form of

$$\frac{1}{N} \sum_{i=1}^{N} p^i \delta_{\xi^1} + p^{i2} \delta_{\xi^2} + p^{i3} \delta_{\xi^3},$$

where $p^i + p^{i2} + p^{i3} = 1$. Given $\{\xi^i : 1 \leq i \leq N, 1 \leq j \leq 3\}$, the problem

$$\max_{0 \leq p^i \leq 1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{3} p^i \Psi(\xi^i) : \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{3} p^i d^p(\xi^i, \hat{x}^i) \leq \theta^p \right\}$$

is a linear program with $3N$ variables, one equality constraint and $3N$ inequality constraints $p_{ij} \leq 1$, $p_{ij} \geq 1$, $i = 1, \ldots, N, j = 1, 2, 3$. Thus according to linear programming theory, there exists an optimal solution such that among the $3N$ inequality constraints, at least $3N - 1$ of them hold as equality, or equivalently, at most one $p_{ij}$ is fractional. Hence for any $\epsilon$-optimal solution $\mu$, there exists a solution of the form

$$\mu_{\epsilon} = \frac{1}{N} \sum_{i \neq 0} \delta_{\xi^i} + \frac{p_{\epsilon}}{N} \delta_{\xi^0} + \frac{1 - p_{\epsilon}}{N} \delta_{\xi^0},$$

which yields an objective value no worse than $\mu$. Define

$$\xi^{ik} = \begin{cases} \xi^i, & 1 \leq k \leq K, \forall \ i \neq i_0, \\ \xi^0, & 1 \leq k \leq \lfloor K \rho_i \rfloor, \ i = i_0, \\ \xi^0, & \lfloor K \rho_i \rfloor < k \leq N, \ i = i_0. \end{cases}$$
Then \( \{\xi_{ik}\}_{i,k} \) belongs to \( \mathcal{M}_K \). By Lemma 2(ii), for any \( \lambda > \lambda_0 \in \text{dom}(\Phi(\cdot, \hat{\xi}^0)) \),
\[
d^p(\xi_{i_0}^0, \hat{\xi}^0) \leq \frac{1}{\lambda - \lambda_0} \left( \Phi(\lambda_0, \hat{\xi}^0) - \Phi(\hat{\xi}^0) \right) =: D.
\]
Since \( |p_x - [Kp_x]/K| < 1/K \), it follows that
\[
|v_K - \mathbb{E}_{\mu_x}[\Phi(\xi)]| \leq \frac{1}{NK} \left[ p_x - [Kp_x]/K \right] \cdot \left( \Phi(\xi_{i_0})^0 - \Phi(\xi_{i_0}^0) \right)
\]
\[
\leq \frac{1}{NK} \left( \Phi(\xi_{i_0})^0 - \Phi(\xi_{i_0}^0) \right)
\]
\[
\leq \frac{M + Ld^p(\xi_{i_0}^0, \hat{\xi}^0)}{NK}
\]
\[
\leq \frac{M + LD}{NK}.
\]
Let \( \epsilon \to 0 \) we obtain the results. \( \square \)

**Example 5 (Saddle-point Problem).** When \( \Psi(x, \xi) \) is convex in \( x \) and concave \( \xi \), \( p = 1 \), and \( d = || \cdot ||_2 \), Corollary 2(iii) shows that the DRSO (DRSO) is equivalent to a convex-concave saddle point problem
\[
\min_{x \in \mathcal{X}} \max_{(\xi^1, \ldots, \xi^N) \in Y} \frac{1}{N} \sum_{i=1}^{N} \Psi(x, \xi^i),
\]
with \( \ell_1/\ell_2 \)-norm uncertainty set
\[
Y = \left\{ (\xi^1, \ldots, \xi^N) \in \Xi^N : \sum_{i=1}^{N} ||\xi^i - \hat{\xi}^i||_2 \leq N\theta \right\}.
\]
Therefore it can be solved by the Mirror-Prox algorithm (Nemirovski [29], Nesterov and Nemirovski [30]).

**Example 6 (Piecewise concave objective).** Esfahani and Kuhn [20] proves that when \( p = 1 \), \( \Xi \) is a convex subset of \( \mathbb{R}^K \) equipped with some norm \( || \cdot || \) and \( \Psi(\xi) = \max_{1 \leq j \leq J} \Psi^j(\xi) \), where \( \Psi^j \) are concave, the DRSO is equivalent to a convex program. We here show that it can be obtained as a corollary from the structure of the worst-case distribution. Indeed, using concavity of \( \Psi^j \) and Corollary 2(i), it suffices to consider distributions of the form
\[
\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} p_{ij} \delta_{\xi_{ij}}, \quad \sum_{j=1}^{J} p_{ij} = 1,
\]
where for each \( i \),
\[
\text{card}\{j : p_{ij} > 0\} \leq 2,
\]
where card represents cardinality. Relaxing the cardinality constraints yields the following problem:
\[
\sup_{p_{ij} \geq 0, \xi_{ij} \in \Xi} \left\{ \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} p_{ij} \Psi(\xi_{ij}) : \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} p_{ij} d(\xi_{ij}, \hat{\xi}^i) \leq \theta, \quad \sum_{j=1}^{J} p_{ij} = 1, \forall i \right\}.
\]
Replacing \( \xi_{ij} \) by \( \hat{\xi}^i + (\xi_{ij} - \hat{\xi}^i)/p_{ij} \), by positive homogeneity of norms and convexity-preserving property of perspective functions (cf. Section 2.3.3 in Boyd and Vandenberghe [12]), we obtain an equivalent convex program reformulation of the primal problem:
\[
\sup_{\xi_{ij} \in \mathbb{R}^K, p_{ij} \geq 0, \sum_{j} p_{ij} = 1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} p_{ij} \Psi^j(\hat{\xi}^i + \frac{\xi_{ij} - \hat{\xi}^i}{p_{ij}}) : \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} p_{ij} d(\xi_{ij}, \hat{\xi}^i) \leq \theta, \quad \hat{\xi}^i + \frac{\xi_{ij} - \hat{\xi}^i}{p_{ij}} \in \Xi, \forall i, j \right\}.
\]
So we recover Theorem 4.5 in Esfahani and Kuhn [20], which was obtained therein by a separate procedure of dualizing twice the reformulation (12).

**Example 7 (Uncertainty Quantification).** When $\Xi = \mathbb{R}^K$ and $\Psi = -\mathbb{I}_C$, where $C$ is an open set, the worst-case distribution $\mu^*$ of the problem

$$\sup_{\mu \in \mathcal{M}} \mathbb{E}_\mu[-\mathbb{I}_C(\xi)] = \min_{\mu \in \mathcal{M}} \mu(C)$$

has a clear interpretation. The worst-case distribution perturbs $\nu$ such that the set $C$ contains as little probability mass as possible, which can be achieved in a greedy fashion as follows. Suppose $\{\hat{\xi}_i\}_{i=1}^N$ are sorted such that $\hat{\xi}_1, \ldots, \hat{\xi}_i \in C$, $\hat{\xi}_{i+1}, \ldots, \hat{\xi}_N \notin C$ and satisfy $d^p(\hat{\xi}_1, \Xi \setminus C) \leq \cdots \leq d^p(\hat{\xi}_i, \Xi \setminus C)$. Then to save the total budget of perturbation $\hat{\xi}_1, \ldots, \hat{\xi}_{i_0}$ stay at the same place, and the $\hat{\xi}_i$ with small index has the priority to be transported to $\partial C$. It may happen that some point $\hat{\xi}_{i_0}$ ($i_0 \leq I$) cannot be transported to $\partial C$ with full mass $1/N$, since otherwise the Wasserstein distance constraint is violated. In this case, only partial mass (with probability $p_0/N$) is transported and the remaining stays (see Figure 3). Therefore the worst-case distribution has the form

$$\mu^* = \frac{1}{N} \sum_{i=1}^{i_0-1} \delta_{\hat{\xi}_i} + \frac{p_0}{N} \delta_{\hat{\xi}_{i_0}} + \frac{1-p_0}{N} \delta_{\hat{\xi}_{i_0}} + \frac{1}{N} \sum_{i=i_0+1}^{N} \delta_{\hat{\xi}_i}.$$

In fact, the dual optimizer $\lambda^*$ is such that

$$\xi^*_i = \arg\min_{\xi \in \Xi} \{\lambda^* d^p(\xi, \hat{\xi}_i) + \mathbb{I}_C(\xi)\} = \arg\min_{\xi \in \partial C} d^p(\xi, \hat{\xi}_i), \quad \forall i \leq I,$$

and

$$\xi^{i_0}_* = \arg\min_{\xi \in \Xi} \{\lambda^* d^p(\xi, \hat{\xi}^{i_0}) + \mathbb{I}_C(\xi)\} = \begin{cases} \hat{\xi}^{i_0} \cup \arg\min_{\xi \in \partial C} d^p(\xi, \hat{\xi}^{i_0}), & p_0 \neq 0, \\ \arg\min_{\xi \in \partial C} d^p(\xi, \hat{\xi}^{i_0}), & p_0 = 0. \end{cases}$$

Using the similar idea as above, we can prove that the worst-case probability is continuous with respect to the boundary.

**Proposition 3 (Continuity with respect to the boundary).** Let $\Xi = \mathbb{R}^K$, $\nu \in \mathcal{P}(\Xi)$, $\theta > 0$, and $\mathcal{M} = \{\mu \in \mathcal{P}(\Xi) : W_p(\mu, \nu) \leq \theta\}$. Then for any Borel set $C \subset \Xi$,

$$\inf_{\mu \in \mathcal{M}} \mu(C) = \min_{\mu \in \mathcal{M}} \mu(\text{int}(C)).$$
Now let us consider a special case when $\Xi = \{\xi^0, \ldots, \xi^{B}\}$ for some positive integer $B$. In this case, let $N_i$ be the samples that are equal to $\xi_i$, and let $q_i = N_i / N, i = 0, \ldots, B$, then the nominal distribution $\nu = \sum_{i=1}^{B} q_i \delta_{\xi_i}$. Let $\nu := (q_0, \ldots, q_B)^T \in \Delta_B$. The DRSO becomes

$$\min_{x \in X} \max_{\mu \in \Delta_B} \left\{ \sum_{i=0}^{B} p_i \Psi(x, \xi_i) : W_p(\mu, \nu) \leq \theta \right\}. \tag{15}$$

**Corollary 3.** Problem (15) has a strong dual

$$\min_{x \in X, \lambda \geq 0} \left\{ \lambda \theta p + \sum_{i=0}^{B} q_i y_i : y_i \geq \Psi(x, \xi_i) - \lambda d_P(\xi_i, \xi^j), \forall i, j = 1, \ldots, B \right\}. \tag{16}$$

For any $x$, the worst-case distribution can be computed by

$$\max_{\mu \in \Delta_B, \gamma \in R_{+} \times R_{+}} \left\{ \sum_{i=0}^{B} p_i \Psi(x, \xi_i) : \sum_{i,j} d_p(\xi_i, \xi^j) \gamma_{ij} \leq \theta p, \sum_{j} \gamma_{ij} = p_i, \forall i, \sum_{i} \gamma_{ij} = q_j, \forall j \right\}. \tag{17}$$

**Proof.** Reformulation (16) follows from Theorem 1, and (17) can be obtained using the equivalent definition of Wasserstein distance in Example 2. \qed

### 4. Applications.

In this section, we apply our results to on/off system control, intensity estimation and worst-case Value-at-Risk analysis. In the first two problem, the nominal distribution is a point process, and the corresponding underlying space $\Xi$ is the space of counting measures (sample paths), which is non-convex and infinite dimensional. In the third problem, the nominal distribution is arbitrary probability distribution on a finite dimensional space, such as Gaussian distribution. Hence, the results in Esfahani and Kuhn [20] and Zhao and Guan [50] cannot be applied.

#### 4.1. On/Off System Control.

In this problem, the decision maker faces a point process and controls a two-state (on/off) system. The point process is assumed to be exogenous, that is, the arrival times are not affected by the on/off state of the system. When the system is switched on, a cost of $c$ per unit time is incurred, and each arrival while the system is on contributes 1 unit revenue. When the system is off, no cost is incurred and no revenue is earned. The decision maker wants to choose a control to maximize the total profit during a finite time horizon. This problem is a prototype for problems in sensor network and revenue management.

In many practical settings, the decision maker does not have a probability distribution for the point process. Instead, the decision maker has observations of historical sample paths of the point process, which constitute an empirical point process. Note that if one would use the Sample Average Approximation (SAA) method with the empirical point process, it would yield a degenerate control, in which the system is switched on only at the arrival time points of the empirical point process. Consequently, if future arrival times can differ from the empirical arrival times by even a little bit, the system would be switched off and no revenue would be earned. Due to such degeneracy and instability of the SAA method, we resort to the distributionally robust approach.

To model the problem, we scale the finite time horizon to $[0, 1]$, and let

$$\Xi = \left\{ \sum_{m=1}^{M} \delta_{\xi_m} : M \in \mathbb{Z}_+, \xi_m \in [0, 1], m = 1, \ldots, M \right\}$$

be the space of finite counting measures on $[0, 1]$. Then the point processes on $[0, 1]$ are then defined by the set $\mathcal{P}(\Xi)$ of Borel probability measures on $\Xi$. To define the Wasserstein distance between to
point processes \( \mu, \nu \in \mathcal{P}(\Xi) \), we need to define the metric \( d \) on the space \( \Xi \) of counting measures. We assume that the metric \( d \) on \( \Xi \) satisfies the following conditions (note that in this subsection, when we write the \( W_1 \) distance between two Borel measures, we use the extended definition mentioned in Section 2):

(i) The metric space \((\Xi, d)\) is a Polish space.
(ii) For any \( \tilde{\eta} = \sum_{m=1}^{M} \delta_{\zeta_m} \) and \( \eta = \sum_{m=1}^{M} \delta_{\zeta_m} \), where \( M \) is a nonnegative integer and \( \{\zeta_m\}_{m=1}^{M}, \{\xi_m\}_{m=1}^{M} \subset [0,1] \), it holds that

\[
d(\eta, \tilde{\eta}) = W_1(\eta, \tilde{\eta}) = \sum_{m=1}^{M} |\xi(m) - \zeta(m)|,
\]

where \( \xi(m) \) (resp. \( \zeta(m) \)) are the order statistics of \( \xi_m \) (resp. \( \zeta_m \)).

(iii) For any Borel set \( C \subset [0,1] \), \( \theta \geq 0 \), and \( \tilde{\eta} = \sum_{m=1}^{M} \delta_{\zeta_m} \), where \( M \) is a positive integer and \( \{\zeta_m\}_{m=1}^{M} \subset [0,1] \), it holds that

\[
\inf_{\eta \in \Xi} \{ \eta(C) : d(\eta, \tilde{\eta}) = \theta \} \geq \inf_{\tilde{\eta} \in B([0,1])} \{ \tilde{\eta}(C) : W_1(\tilde{\eta}, \tilde{\eta}) \leq \theta \}.
\]

We note that condition (ii) is only imposed on \( \eta, \tilde{\eta} \in \Xi \) such that \( \eta([0,1]) = \tilde{\eta}([0,1]) \). Possible choices for \( d \) are

\[
d\left( \sum_{m=1}^{M} \delta_{\zeta_m}, \sum_{l=1}^{L} \delta_{\zeta_l} \right) = \min\{M, L\} \left( \sum_{m=1}^{M} |\xi(m) - \zeta(l)| + |M - L| \right),
\]

\[
d\left( \sum_{m=1}^{M} \delta_{\zeta_m}, \sum_{l=1}^{L} \delta_{\zeta_l} \right) = \begin{cases} \max\{M, L\}, & M \neq L, \\ \sum_{m=1}^{M} |\xi(m) - \zeta(m)|, & M = L, \end{cases}
\]

or

\[
d\left( \sum_{m=1}^{M} \delta_{\zeta_m}, \sum_{l=1}^{L} \delta_{\zeta_l} \right) = \begin{cases} +\infty, & M \neq L, \\ \sum_{m=1}^{M} |\xi(m) - \zeta(m)|, & M = L. \end{cases}
\]

These metrics are similar to the ones in Barbour and Brown [4] and Chen and Xia [14]. Given the metric \( d \), we choose the distance between two point processes \( \mu, \nu \in \mathcal{P}(\Xi) \) to be \( W_1(\mu, \nu) \) as defined in (1).

Suppose we have \( N \) sample paths \( \tilde{\eta}^i = \sum_{m=1}^{M_i} \delta_{\xi^i_m}, i = 1, \ldots, N \), where \( M_i \) is a nonnegative integer and \( \xi^i_m \in [0,1] \) for all \( i, m \). Then the nominal distribution \( \nu = \frac{1}{N} \sum_{i=1}^{N} \delta_{\tilde{\eta}^i} \), and the ambiguity set \( \\mathfrak{M} = \{ \mu \in \mathcal{P}(\Xi) : W_1(\mu, \nu) \leq \theta \} \). Let \( X \) denote the set of all functions \( x : [0,1] \to \{0,1\} \) such that \( x^{-1}(1) \) is a Borel set, where \( x^{-1}(1) := \{ t \in [0,1] : x(t) = 1 \} \). The decision maker is looking for a control \( x \in X \) that maximizes the total profit, by solving the problem

\[
v^* := \sup_{x \in X} \left\{ v(x) := -c \int_0^1 x(t) dt + \inf_{\mu \in \mathfrak{M}} \mathbb{E}_{\eta \sim \mu}[\eta(x^{-1}(1))] \right\}.
\]

We now investigate the structure of the optimal control. Let \( \text{int}(x^{-1}(1)) \) be the interior of the set \( x^{-1}(1) \) on the space \( [0,1] \) with canonical topology (and thus \( 0,1 \in \text{int}([0,1]) \)).

**Proposition 4.** For any \( \nu \in \mathcal{P}(\Xi) \) and control \( x \), it holds that

\[
\inf_{\mu \in \mathfrak{M}} \mathbb{E}_{\eta \sim \mu}[\eta(x^{-1}(1))] = \inf_{\rho \in \mathcal{P}(B([0,1]) \times \Xi)} \left\{ \mathbb{E}_{(\eta, \tilde{\eta}) \sim \rho}[\eta(\text{int}(x^{-1}(1)))] : \mathbb{E}_{(\eta, \tilde{\eta}) \sim \rho}[W_1(\eta, \tilde{\eta})] \leq \theta, \ \pi^2 \rho = \nu \right\},
\]

Suppose \( \nu = \frac{1}{N} \sum_{i=1}^{N} \delta_{\tilde{\eta}^i} \) with \( \tilde{\eta}^i = \sum_{m=1}^{M_i} \delta_{\xi^i_m} \). There exists a non-negative integer \( M \) such that

\[
v^* = \sup_{\varepsilon_j, \pi_j \in [0,1], \ \varepsilon_j < \pi_j \leq \varepsilon_{j'}, \forall 1 \leq j < j' \leq M} \left\{ v\left( \sum_{j=1}^{M} I_{[\varepsilon_j, \pi_j]} \right) := -c \sum_{j=1}^{M} (\pi_j - \varepsilon_j) + \inf_{\mu \in \mathfrak{M}} \mathbb{E}_{\eta \sim \mu}[\eta\{ \cup_{j=1}^{M} [\varepsilon_j, \pi_j] \}] \right\},
\]
Note that
\[ \inf_{\mu \in \mathcal{M}} \mathbb{E}_\mu[\eta(x^{-1}(1))] = \inf_{\gamma \in \mathcal{P}(\Xi^2)} \{ \mathbb{E}_{(\eta, \tilde{\eta}) \sim \gamma}[\eta(x^{-1}(1))]: \mathbb{E}_\gamma[d(\eta, \tilde{\eta})] \leq \theta, \, \pi_\eta^2 \gamma = \nu \}. \]

Hence (20) shows that without changing the optimal value, we can replace \( d \) by \( W_1 \) in the constraint, and enlarge the set of joint distributions from \( \mathcal{P}(\Xi^2) \) to \( \mathcal{P}(\mathcal{B}([0, 1]) \times \Xi) \). Moreover, (21) shows that it suffices to consider the set of polices of which the duration of on-state is a finite disjoint union of intervals with positive length. We next show that given a control \( \sum_{j=1}^M \mathds{1}_{[x_j, x_j]} \), the computation of worst-case point process reduces to a linear program. For every \( 1 \leq i \leq N \) and \( 1 \leq m \leq M_i \), if \( \hat{\xi}_m \in \bigcup_{j=1}^M [x_j, x_j] \), we set \( j^*_m \in \{1, \ldots, M_i\} \) to be such that \( \hat{\xi}_m \in [x_{j^*_m}, x_{j^*_m}] \); otherwise \( j^*_m = 0 \). We also set \( x_0 \) to be any real number.

**Proposition 5.** The objective \( v(\sum_{j=1}^M \mathds{1}_{[x_j, x_j]}^i) \) defined in (21) can be written as
\[
\sum_{j=1}^M -c(x_j - x_j) + \frac{1}{N} \sum_{i=1}^N \mathds{1}_{[x_j, x_j]}(\hat{\xi}_m^i) + \min \left\{ -\frac{1}{N} \sum_{i=1}^N \sum_{1 \leq m \leq M_i; j^*_m > 0} (p_m^i + \overline{p}_m^i), \right. \\
\frac{1}{N} \sum_{i=1}^N \sum_{1 \leq m \leq M_i; j^*_m > 0} (p_m^i |x_{j^*_m} - \hat{\xi}_m^i| + \overline{p}_m^i |x_{j^*_m} - \hat{\xi}_m^i|) \leq \theta \left. \right\}.
\]

Moreover, the above linear program can be solved by a greedy algorithm (see Algorithm 1), and there exists a worst-case point process that has the form
\[
\mu^*(x) = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}^* + \frac{p_0}{N} \delta_{x_0}^* + \frac{1 - p_0}{N} \delta_{\eta_0},
\]
where \( i_0 \in \{1, \ldots, N\} \), \( \eta_0^i \in \Xi \), \( \eta_0^i([0, 1]) = \tilde{\eta}^i([0, 1]) \) for all \( i \neq i_0 \), \( \eta_0^i, \eta_0^i \in \Xi \) and \( \eta_0^i([0, 1]) = \tilde{\eta}_0^i([0, 1]) \) = \( \tilde{\eta}_0^i([0, 1]) \).

**Algorithm 1** Greedy Algorithm

1. \( \tilde{\theta} \leftarrow 0 \). \( k \leftarrow 1 \). \( p_m^i \leftarrow 0 \), \( d_m^i \leftarrow \min(|x_{j^*_m} - \hat{\xi}_m^i|, |x_{j^*_m} - \hat{\xi}_m^i|) \), \( \forall i, m \).
2. Sort \( \{d_m^i\}_{1 \leq i \leq N, 1 \leq m \leq M_i} \) in increasing order, denoted by \( d_m^{i(1)} \leq d_m^{i(2)} \leq \cdots \leq d_m^{i(\sum_{i=1}^N m_i)}. \)
3. while \( \tilde{\theta} < N\theta \) do
   4. if \( d_m^i = |x_{j^*_m} - \hat{\xi}_m^i| \) then \( p_m^{i(k)^*} \leftarrow \min(1, (N\theta - \tilde{\theta})/d_m^{i(k)}). \)
   5. else \( p_m^{i(k)} \leftarrow \min(1, (N\theta - \tilde{\theta})/d_m^{i(k)}). \)
   6. end if
   7. \( k \leftarrow k + 1 \).
8. end while

**Example 8.** We illustrate our results as follows. Suppose the number of arrivals has Poisson distribution \( \text{Poisson}(\lambda) \), and given the number of arrivals, the arrival times are i.i.d. with density \( f(t) \), \( t \in [0, 1] \). Then problem (19) is \( \max_x \int_{x^{-1}(1)} [-c + \lambda f(t)] dt \), with optimal control \( x^*(t) = 1_{\{\lambda f(t) > c\}} \).

Note that \( f \equiv 1 \) corresponds to the Poisson point process with rate \( \lambda \). In this example, we instead consider \( f(t) = k[a + \sin(wt + s)] \), with \( a > 1 \) and \( k = 1/(a + (\cos(s) - \cos(w + s))/w) \). Particularly, let
Arrival time density and true optimal control

\[ f(t) \]

Figure 4. Optimal control for the true process and the DRSO

\[ w = 5\pi, \ s = \frac{5}{2}\pi, \ a = 1.1 \text{ and } c = \lambda = 10. \] Thus \( x^{*^{-1}}(1) = [0, 0.1] \cup [0.3, 0.5] \cup [0.7, 0.9]. \) In the numerical experiment, suppose we have \( N = 10 \) sample paths, each of which contains \( M_i \sim \text{Poisson}(\lambda), \ i = 1, \ldots, N, \) i.i.d. arrival time points. The optimal controls for the true process and the DRSO are shown in Figure 4. We observe that even with a relatively small number of samples, the two controls differ from each other not too much, and thus the DRSO indeed provides a good solution to the original process control problem.

4.2. Intensity Estimation for Non-homogeneous Poisson Process. Consider estimating the intensity function \( a(t) \) of a non-homogeneous Poisson process \( A(t) \) using maximum likelihood method. Given \( N \) i.i.d. sample paths \( \hat{\xi} = \sum_{m=1}^{M_i} \delta_{\hat{\xi}_m}, \ i = 1, \ldots, N, \) the log-likelihood function (see, e.g. Daley and Vere-Jones [15]) is written as

\[
\int_0^T -a(t)dt + \frac{1}{N} \sum_{i=1}^{N} \sum_{m=1}^{M_i} -\ln(a(\hat{\xi}_m)).
\]

A common practice is to partition the time horizon \([0, T]\) into several intervals, and assume \( a(t) \) is piecewise constant on each interval. Then the maximum likelihood estimator equals to the average arrival rate on each interval. Such a approach suffers from the drawback that the estimator is sensitive to the partition of intervals. If the partition is so fine that many intervals may have zero observations, then the estimator also vanishes on these intervals. On the other hand, if the partition is very coarse, then the estimator remains constant during a long interval, which may not reflect the reality. It appears that there is no systematic way to adaptively choose the partition for this sample average method. Meanwhile, distributionally robust formulation with \( \phi \)-divergence has the same problem, since the yielding estimator vanishes on intervals with zero observation.

Consider the distributionally robust formulation with Wasserstein distance

\[
\min_{a(t)} \left\{ \int_0^T a(t)dt + \max_{\mu \in \mathcal{M}} \mathbb{E}_\eta \left[ \int_0^T -\ln(a(t))\eta(dt) \right] \right\}, \tag{22}
\]

where \( \mathcal{M} \) is the same as the one in the previous subsection, namely, the Wasserstein ball centered at the empirical process. To facilitate further analysis, we choose (18) as the definition of distance between two counting measures. Our strong duality results imply that the dual reformulation of (22) is given by

\[
\min_{\lambda \geq 0} \left\{ \int_0^T a(t)dt + \lambda \theta + \frac{1}{N} \sum_{i=1}^{N} \sum_{1 \leq m \leq M_i} \sup_{\xi \in [0,T]} \left\{ -\ln(a(\xi)) - \lambda(\xi - \hat{\xi}_m) \right\} \right\}.
\]

The following proposition suggests that the optimal estimator is constant if the radius of Wasserstein ball goes to infinity.

**Proposition 6.** For sufficiently large \( \theta \), the optimal value \( a_*(t) \) is constant.
To numerically solve the problem, let us assume $a(t)$ is piecewise constant. In our numerical experiments, we assume the underlying true intensity function is given by $a(t) = 0.5 + 0.5t$ or $a(t) = 1 + \sin(\pi t)$, $t \in [0, 10]$. We fix the sample size (number of sample paths) $N = 20$ and vary the number of pieces in $\{20, 50, 100\}$. The radius $\theta$ is chosen via cross-validation method, for which half of the sample paths are used for training and the remaining are used for calibration. The out-of-sample performance is measured in terms of $L_2$ distance between the estimated intensity and true intensity. The estimation results and out-of-sample performances are shown in Figure 5 and Table 1. We observe that DRSO with Wasserstein distance has superior out-of-sample performance in all cases. The shape of the estimator from DRSO is insensitive to the fineness of the partition for the piecewise constant function. In contrast, the maximum likelihood estimator behaves terribly if we do not make the partition correctly, for example, when the number of pieces is too large.

Table 1. Out-of-sample performance of DRSO and SAA

| Pieces \ | 0.2 + 0.2t | 1 + \sin(\pi t) |
| --- | --- | --- |
| 20 | 0.394 | 2.008 |
| 50 | 0.481 | 2.122 |
| 100 | 0.544 | 2.276 |

| 20 | 6.160 | 11.766 |
| 50 | 6.591 | 11.766 |
| 100 | 11.906 | 11.766 |

4.3. Worst-case Value-at-Risk. Value-at-risk is a popular risk measure in financial industry. Given a random variable $Z$ and $\alpha \in (0, 1)$, the value-at-risk $\text{VaR}_\alpha[Z]$ of $Z$ with respect to measure $\nu$ is defined by

$$\text{VaR}_\alpha[Z] := \inf \{ t : P_\nu[Z \leq t] \geq 1 - \alpha \}.$$ 

In the spirit of El Ghaoui et al. [18], we consider the following worst-case VaR problem. Suppose we are given a portfolio consisting of $n$ assets with allocation weight $w$ satisfying $\sum_{i=1}^{N} w_i = 1$ and $w \geq 0$. Let $\xi_i$ be the (random) return rate of asset $i$, $i = 1, \ldots, n$, and $r = E[\xi]$ the vector of the
expected return rates. Assume the metric \( d \) is induced by the infinity norm \( || \cdot ||_\infty \) on \( \mathbb{R}^K \). The worst-case VaR with respect to the set of probability distributions \( \mathcal{M} \) is defined as

\[
\text{VaR}^{wc}_\alpha(w) := \min \{ q : \inf_{\mu \in \mathcal{M}} \mathbb{P}_\mu \{ -w^\top \xi \leq q \} \geq 1 - \alpha \}.
\]

**Proposition 7.** Let \( q \geq \text{VaR}_\alpha[-w^\top \xi], \theta > 0, \alpha \in (0, 1), \ w \in \{ w' \in \mathbb{R}^N : \sum_{i=1}^N w'_i = 1, \ w' \geq 0 \} \). Define

\[
\beta_0 := \min \left( 1, \frac{(\alpha - \nu(\xi : -w^\top \xi > \text{VaR}_\alpha[-w^\top \xi]))(q - \text{VaR}_\alpha[-w^\top \xi])^p}{\left| \theta - \mathbb{E}_\nu[(q + w^\top \xi)^p \mathbbm{1}_{\{ -w^\top \xi > \text{VaR}_\alpha[-w^\top \xi] \}}] \right|} \right).
\]

Then \( \inf_{\mu \in \mathcal{M}} \mathbb{P}_\mu \{ -w^\top \xi \leq q \} \geq 1 - \alpha \) is equivalent to

\[
\mathbb{E}_\nu \left[ ((q + w^\top \xi)^+)^p \mathbbm{1}_{\{ -w^\top \xi > \text{VaR}_\alpha[-w^\top \xi] \}} \right] + \beta_0 \mathbb{E}_\nu \left[ ((q + w^\top \xi)^+)^p \mathbbm{1}_{\{ -w^\top \xi = \text{VaR}_\alpha[-w^\top \xi] \}} \right] \geq \theta^p.
\]

In particular, when \( \nu \) is a continuous distribution, the condition above can be reduced to

\[
\mathbb{E}_\nu \left[ ((q + w^\top \xi)^+)^p \mathbbm{1}_{\{ -w^\top \xi \geq \text{VaR}_\alpha[-w^\top \xi] \}} \right] \geq \theta^p.
\]

**Example 9 (Worst-case VaR with Gaussian nominal distribution).** Suppose \( \nu \sim N(\mu, \Sigma) \) and consider \( p = 1 \). It follows that \( -w^\top \xi \sim N(-w^\top \bar{\mu}, w^\top \Sigma w) \) and \( \text{VaR}_\alpha[-w^\top \xi] = -w^\top \bar{\mu} + \sqrt{w^\top \Sigma w \Phi^{-1}(1 - \alpha)} \). By Proposition 7, \( \text{VaR}^{wc}_\alpha(-w^\top \xi) \) is the minimal \( q \) such that (see Figure 6)

\[
f(q) := \frac{1}{\sqrt{2\pi w^\top \Sigma w}} \int_{\text{VaR}_\alpha[-w^\top \xi]}^q (q - y) e^{-\frac{(y - w^\top \bar{\mu})^2}{2w^\top \Sigma w}} \, dy \geq \theta.
\]

Since \( f(q) \) is monotone, (23) can be solved efficiently via any one-dimensional search algorithm.

We remark that the above result indicates that finding the worst-case VaR is tractable. It should be noted that, however, finding the best allocation weight, i.e., optimizing over \( w \) is still hard, since the VaR constraint is essentially a chance-constraint.

5. **Discussions.** In this section, we discuss some advantages of Wasserstein ambiguity set. In Section 5.1, we compare the Wasserstein ambiguity set to \( \phi \)-divergence ambiguity set for newsvendor problem. In Section 5.2, we illustrate how the close connection between DRSO and robust programming (Corollary 3(iii)) can expand the tractability of DRSOs.
5.1. Newsvendor problem: a comparison to $\phi$-divergence. In this subsection, we discuss some advantages of Wasserstein ambiguity set by performing a numerical study on distributionally robust newsvendor problems, with an emphasis on the worst-case distribution.

In the newsvendor model, the decision maker has to decide the inventory level before the random demand is realized, facing both overage and underage costs. The problem can be formulated as

$$\min_{x \geq 0} \mathbb{E}_\nu [h(x - \xi)^+ + b(\xi - x)^+] ,$$

where $x$ is the decision variable for initial inventory level, $\xi$ is the random demand, and $h, b$ represent respectively the overage and underage costs per unit. We assume that $b, h > 0$, and $\xi$ is supported on $\{0, 1, \ldots, B\}$ for some positive integer $B$. Sometimes the demand data is expensive to obtain. For instance, a company is introducing a new product of which the demand data is collected by setting up pilot stores. Then the decision maker may want to consider the DRSO counterpart

$$\min_{x \geq 0} \sup_{\mu \in \Delta_B} \{ \mathbb{E}_\mu [h(x - \xi)^+ + b(\xi - x)^+] : W_\phi(\mu, \nu) \leq \theta \} .$$

Using Corollary 3, we obtain a convex programming reformulation

$$\min_{x, \lambda \geq 0} \left\{ \lambda \theta^p + \sum_{i=0}^B q_i y_i : y_i \geq \max [h(x - j), b(j - x)] - \lambda |i - j|^p, \forall 0 \leq i, j \leq B \right\} .$$

On the other hand, one may would also consider $\phi$-divergence ambiguity set (Table 2 shows some common $\phi$-divergences). As mentioned in Section 1, the worst-case distribution in $\phi$-divergence ambiguity set may be problematic. Indeed, when $\lim_{t \to \infty} \phi(t)/t = \infty$, such as $\phi_{kl}, \phi_{\text{mx}}^2$, the $\phi$-divergence ambiguity set fails to include sufficiently many relevant distributions. In fact, since $0 \phi(p_j/0) = p_j \lim_{t \to \infty} \phi(t)/t = \infty$ for all $p_j > 0$, the $\phi$-divergence ambiguity set does not include any distribution which is not absolutely continuous with respect to the nominal distribution $\nu$.

When $\lim_{t \to \infty} \phi(t)/t < \infty$, such as $\phi_{kl}, \phi_{\text{mx}}^2, \phi_{tv}$, the situation is even worse. Define $I_0 := \{1 \leq j \leq N : q_j > 0\}$ and $j_M := \arg \max_{1 \leq j \leq N} \{ \Psi(\xi^j) : q_j = 0\}$. Assume $\Psi(\xi^j)$ are different from each other, then according to Ben-Tal et al. [6] and Bayraksan and Love [5], the worst-case distribution satisfies

$$p_j^*/q_j = \partial \phi^* \left( \frac{\Psi(\xi^j) - \beta^*}{\lambda^*} \right), \forall i \in I_0,$$

$$p_j^* = 0, \forall j \notin I_0 \cup \{j_M\},$$

$$p_M^* = \begin{cases} 1 - \sum_{i \in I_0} p_j^* , & \text{if } \beta^* = \Psi(\xi^M) - \lambda^* \lim_{t \to \infty} \phi(t)/t, \\ 0 , & \text{if } \beta^* > \Psi(\xi^M) - \lambda^* \lim_{t \to \infty} \phi(t)/t, \end{cases}$$

for some $\lambda^* \geq 0$ and $\beta^* \geq \Psi(\xi^M) - \lambda^* \lim_{t \to \infty} \phi(t)/t$. (24b) suggests that the support of the worst-case distribution and that of the nominal distribution can differ by at most one point $\xi^M$. If $p_M^* > 0$, (24c) suggests that the probability mass is moved away from scenarios in $I_0$ to the worst scenario $\xi^M$. Note that in many applications where the support of $\xi$ is unknown, the choice of the underlying space $\Xi$ (and thus $\Xi^M$) may be arbitrary. Hence the worst-case behavior is sensitive to the specification of $\Xi$ and the shape of function $\Psi$.  

### Table 2. Examples of $\phi$-divergence

| Divergence | Kullback-Leibler | Burg entropy | $\chi^2$-distance | Modified $\chi^2$ | Hellinger | Total Variation |
|------------|-----------------|--------------|-------------------|-------------------|-----------|-----------------|
| $\phi(t), t \geq 0$ | $t \log t$ | $- t \log t$ | $\frac{1}{2} (t - 1)^2$ | $(t - 1)^2$ | $(\sqrt{t} - 1)^2$ | $|t - 1|$ |
| $I_0(\mu, \nu)$ | $\sum p_i \log \left( \frac{p_i}{q_i} \right)$ | $\sum q_j \log \left( \frac{q_j}{p_j} \right)$ | $\sum \frac{(p_j - q_j)^2}{p_j}$ | $\sum \frac{(p_j - q_j)^2}{q_j}$ | $\sum (\sqrt{p_j} - \sqrt{q_j})^2$ | $\sum |p_j - q_j|$ |
We perform a numerical test of which setup is similar to Wang et al. [46] and Ben-Tal et al. [6]. We set $b = h = 1$, $\bar{B} = 100$, and $N \in \{50, 500\}$ representing small and large datasets. The data is then generated from Binomial$(100, 0.5)$ and Geometric$(0.1)$ truncated on $[0, 100]$. For a fair comparison, we estimate the radius of the ambiguity set such that it covers the underlying distribution with probability greater than 95%.

When the underlying distribution is Binomial, intuitively, the symmetry of Binomial distribution and $b = h = 1$ implies that the optimal initial inventory level is close to $\bar{B}/2 = 50$, and the corresponding worst-case distribution should be similar to a mixture distribution with two modes, representing high and low demand respectively. This intuition is consistent with the solid curves in Figure (7a)(7b), representing the worst-case distribution in Wasserstein ambiguity set. In addition, their tail distributions are smooth and reasonable for both small and large datasets. In contrast, if Burg entropy is used to define the ambiguity set (dashed curves in Figure (7a)(7b)), the worst-case distribution has disconnected support, and is not symmetric. There is a large spike on the boundary 100, corresponding to the “popping” behavior mentioned in Bayraksan and Love [5]. Especially when the dataset is small, the spike is huge, which makes the solution too conservative.

When the underlying distribution is Geometric, intuitively, the worst-case distribution should have one spike for low demand and a heavy tail for high demand. Again, this is consistent with the worst-case distribution in Wasserstein ambiguity set (solid curves in Figure (7c)(7d)). While using Burg entropy (dashed curves in Figure (7c)(7d)), the tail has unrealistic spikes on the boundary. For distribution with unbounded support, the tail distribution is very sensitive to our choice of truncation threshold $B$. Hence, the conclusion for this numerical test is that Wasserstein ambiguity set is likely to yield a more reasonable, robust and realistic worst-case distribution.
5.2. Two-stage DRSO: connection with robust optimization. In Corollary 2(iii) we established the close connection between the DRSO problem and robust programming. More specifically, we show that every DRSO problem can be approximated by robust programs with rather high accuracy, which significantly enlarges the applicability of the DRSO problem. To illustrate this point, in this section we show the tractability of the two-stage linear DRSOs.

Consider the two-stage distributionally robust stochastic optimization

$$\min_{x \in \mathcal{X}} c^T x + \sup_{\mu \in \mathcal{M}} \mathbb{E}_\mu [\Psi(x, \xi)],$$

where \(\Psi(x, \xi)\) is the optimal value of the second-stage problem

$$\min_{y \in \mathbb{R}^m} \{ q(\xi)^T y : T(\xi)x + W(\xi)y \leq h(\xi) \},$$

and

$$q(\xi) = q^0 + \sum_{l=1}^s \xi_l q^l, \quad T(\xi) = T^0 + \sum_{l=1}^s \xi_l T^l, \quad W(\xi) = W^0 + \sum_{l=1}^s \xi_l W^l, \quad h(\xi) = h^0 + \sum_{l=1}^s \xi_l h^l.$$  

We assume \(p = 2\) and \(\Xi = \mathbb{R}^K\) with Euclidean distance \(d\). In general, the two-stage problem (25) is NP-hard. However, we are going to show that with tools from robust programming, we are able to obtain a tractable approximation of (25). Let \(\mathcal{M}_1 := \{ (\xi^1, \ldots, \xi^N) \in \mathbb{R}^N : \frac{1}{\sqrt{N}} \sum_{i=1}^N ||\xi^i - \hat{\xi}^i||_2 \leq \theta \sqrt{N} \}. \)

Using Theorem 2(ii) with \(K = 1\), we obtain an adjustable-robust-programming approximation

$$\min_{x \in \mathcal{X}} \left\{ c^T x + \sup_{(\xi^i)_i \in \mathcal{M}_1} \frac{1}{N} \sum_{i=1}^N \Psi(x, \xi^i) \right\} = \min_{x \in \mathcal{X}, t \in \mathbb{R}} \left\{ t \geq c^T x + \frac{1}{N} \sum_{i=1}^N q(\xi^i)^T y(\xi^i), \quad \forall (\xi^i)_i \in \mathcal{M}_1, \quad T(\xi)x + W(\xi)y(\xi) \leq h(\xi), \quad \forall \xi \in \bigcup_{i=1}^N \{ \xi' \in \Xi : ||\xi' - \hat{\xi}^i||_2 \leq \theta \sqrt{N} \} \right\},$$

where the second set of inequalities follows from the fact that \(T(\xi)x + W(\xi)y(\xi) \leq h(\xi)\) should hold for any realization \(\xi\) with positive probability for some distribution in \(\mathcal{M}_1\). Although problem (26) is still intractable in general, there has been a substantial literature on different approximations to problem (26). One popular approach is to consider the so-called affinely adjustable robust counterpart (AARC) as follows. We assume that \(y\) is an affine function of \(\xi\):

$$y(\xi) = y^0 + \sum_{l=1}^s \xi_l y^l, \quad \forall \xi \in \bigcup_{i=1}^N B^i,$$

for some \(y^0, y^l \in \mathbb{R}^m\), where \(B^i := \{ \xi' \in \Xi : ||\xi' - \hat{\xi}^i||_2 \leq \theta \sqrt{N} \}. \) Then the AARC of (26) is

$$\min_{x \in \mathcal{X}, t \in \mathbb{R}, y^i \in \mathbb{R}^m, l = 0, \ldots, s} \left\{ t \geq c^T x + \frac{1}{N} \sum_{i=1}^N \left( q^0 + \sum_{l=1}^s \xi_l^i q^l \right)^T \left( y^0 + \sum_{l=1}^s \xi_l^i y^l \right) - t \leq 0, \quad \forall (\xi^i)_i \in \mathcal{M}_1, \quad T^0 + \sum_{l=1}^s \xi_l^i T^l \right\} \left( W^0 + \sum_{l=1}^s \xi_l^i W^l \right) \left( y^0 + \sum_{l=1}^s \xi_l^i y^l \right) - \left( h^0 + \sum_{l=1}^s h^l \xi_l^i \right) \leq 0, \quad \forall \xi \in \bigcup_{i=1}^N B^i \right\}.$$

Set \(\zeta_{il} := \xi_l^i - \hat{\xi}_l^i\) for \(i = 1, \ldots, N\) and \(l = 1, \ldots, s\). In view of \(\mathcal{M}_1\), \(\zeta\) belongs to the ellipsoidal uncertainty set

$$\mathcal{U}_\zeta = \{ (\zeta_{il})_{i,l} : \sum_{i,l} \zeta_{il}^2 \leq N \theta^2 \}.$$
Set \( z = [x; t; \{y^l\}_{i=0}^s] \), and define
\[
\begin{align*}
\alpha_0(z) &:= -\left[ c^\top x + \frac{1}{N} \sum_{i=1}^N (q^0 + \sum_{l=1}^s \xi_i^0 q^l) (y^0 + \sum_{l=1}^s \xi_i^l y^l) - t \right], \\
\beta_{il}(z) &:= -\frac{(q^0 + \sum_{l=1}^s \xi_i^0 q^l) y^l + q^l (y^0 + \sum_{l=1}^s \xi_i^l y^l)}{2N}, \\
\Gamma_{il}(z) &:= -\frac{q^l y^l + q^l y^l}{2N},
\end{align*}
\]
Then the first set of constraints in (27) is equivalent to
\[
\alpha_0(z) + 2 \sum_{i,l} \beta_{il}(z) \zeta_{il} + \sum_{i} \sum_{l,l'} \Gamma_{il}(z) \zeta_{il} \zeta_{il'} \geq 0, \quad \forall (\zeta_{il})_{i,l} \in \mathcal{U}_\zeta.
\] (28)

It follows from Theorem 4.2 in Ben-Tal et al. [7] that (28) takes place if and only if there exists \( \lambda_0 \geq 0 \) such that
\[
(\alpha_0(z) - \lambda_0) v^2 + 2v \sum_{i,l} \beta_{il}(z) w_{il} + \sum_{i} \sum_{l,l'} \Gamma_{il}(z) w_{il} w_{il'} + \frac{\lambda_0}{N \theta^2} \sum_{i,l} w_{il}^2 \geq 0, \quad \forall v \in \mathbb{R}, \forall w_{il} \in \mathbb{R}, \forall i, l.
\]
Or in matrix form,
\[
\exists \lambda_0 \geq 0 : \left( \Gamma_0 \otimes I_N + \frac{\lambda_0}{N \theta} \cdot I_{sN \otimes (\beta_0)} \cdot \text{vec}(\beta_0) \right) \geq 0,
\] (29)
where \( I_N \) (resp. \( I_{sN} \)) is \( N \) (resp. \( sN \)) dimensional identity matrix, \( \otimes \) is the Kronecker product of matrices and vec is the vectorization of a matrix.

Now we reformulate the second set of constraints in (27). For all \( 1 \leq i \leq N, 1 \leq j \leq m \) and \( 1 \leq l, l' \leq s \), we set
\[
\begin{align*}
\alpha_{ij}(z) &:= -\left[ (T_j^0 + \sum_{i=1}^s \xi_i^0 T_j^i) x + (W_j^0 + \sum_{i=1}^s \xi_i^0 W_j^i) (y^0 + \sum_{i=1}^s \xi_i^l y^l) - (h_j^0 + \sum_{i=1}^s \xi_i^l h_j^l) \right], \\
\beta_{ij}(z) &:= -\frac{(T_j^0 + \sum_{i=1}^s \xi_i^0 T_j^i) x + (W_j^0 + \sum_{i=1}^s \xi_i^l W_j^i) y^l + W_j^l (y^0 + \sum_{i=1}^s \xi_i^l y^l) - h_i}{2}, \\
\Gamma_{ij}(z) &:= -\frac{W_j^l y^l + W_j^l y^l}{2}.
\end{align*}
\]
Let \( \eta^i := \xi - \xi_i^0 \) for \( 1 \leq i \leq N \). Then the second set of constraints in (27) is equivalent to
\[
\alpha_{ij}(z) + 2 \beta_{ij}(z) \eta^i + \eta^i \Gamma_j(z) \eta^j \geq 0, \quad \forall \eta^i \in \{ \eta^i \in \mathbb{R}^K : ||\eta^i||_2 \leq \theta \sqrt{N} \}, \forall 1 \leq i \leq N, 1 \leq j \leq m.
\]
Again by Theorem 4.2 in Ben-Tal et al. [7] we have further equivalence
\[
\exists \lambda_{ij} \geq 0 : \left( \Gamma_j(z) + \frac{\lambda_{ij}}{N \theta^2} \cdot I_s \cdot \beta_{ij}(z) \alpha_{ij}(z) - \lambda_{ij} \right) \geq 0, \quad \forall 1 \leq i \leq N, 1 \leq j \leq m.
\] (30)
Combining (29) and (30) we obtain the following result.

**Proposition 8.** An exact reformulation of the AARC of (26) is given by
\[
\min_{x \in X, t \in \mathbb{R}, y^l \in \mathbb{R}^n, l = 1, \ldots, s} \left\{ t : \text{(29), (30) holds} \right\}.
\]
Note that (26) is a fairly good approximation of the original two-stage DRSO problem (25) by Theorem 1. Hence, as long as the AARC of (26) is reasonably good, its semidefinite-program reformulation (8) provides a good tractable approximation of the two-stage linear DRSO (25).
6. Conclusions In this paper, we developed a constructive proof method to derive the dual reformulation of distributionally robust stochastic optimization with Wasserstein distance under a general setting. Such approach allows us to obtain a precise structural description of the worst-case distribution and connects the distributionally robust stochastic optimization to classical robust programming. Based on our results, we obtain many theoretical and computational implications. For the future work, extensions to multi-stage distributionally robust stochastic optimization will be explored.

Appendix A: Auxiliary results

Lemma 4. Consider any \( p \geq 1 \) and any \( \epsilon > 0 \). Then there exists \( C_p(\epsilon) \geq 1 \) such that
\[
(x + y)^p \leq (1 + \epsilon)x^p + C_p(\epsilon)y^p
\]
for all \( x, y \geq 0 \).

Lemma 5. Let \( \zeta^0 \in \Xi \). Then for any \( \lambda > \lambda_1 > \kappa \), there exists a constant \( C > 0 \) such that
\[
\frac{\lambda - \lambda_1}{2} \mathcal{D}^p(\lambda, \zeta) \leq \Phi(\lambda, \zeta) - \Phi(\lambda_1, \zeta^0) + \lambda_1 C d^p(\zeta, \zeta^0).
\]

Lemma 6. (i) The quantity
\[
\limsup_{\xi \in \Xi : d^p(\xi, \zeta) \to \infty} \frac{\Psi(\xi) - \Psi(\zeta)}{d^p(\xi, \zeta)}
\]
is independent of \( \zeta \).
(ii) Suppose \( \nu \in \mathcal{P}_p(\Xi) := \{ \mu \in \mathcal{P}(\Xi) : \int_{\Xi} d^p(\zeta, \zeta^0) \nu(d\zeta) < \infty \text{ for some } \zeta^0 \in \Xi \} \).

Then the growth rate \( \kappa \) defined in Definition 4 is finite if and only if there exists \( \zeta^0 \in \Xi \), \( L, M > 0 \) such that
\[
\Psi(\xi) - \Psi(\zeta^0) \leq L d^p(\xi, \zeta^0) + M, \ \forall \xi \in \Xi.
\]
(iii) When \( \kappa < \infty \), it holds that
\[
\kappa \leq \limsup_{\xi \in \Xi : d^p(\xi, \zeta) \to \infty} \frac{\Psi(\xi) - \Psi(\zeta)}{d^p(\xi, \zeta)}
\]
for any \( \zeta \in \Xi \).

Lemma 7. Let \( C \) be a Borel set in \( \Xi \) with nonempty boundary \( \partial C \). Then for any \( \epsilon > 0 \), there exists a Borel map \( T_\epsilon : \partial C \to \Xi \setminus \text{cl}(C) \) such that \( d(\xi, T_\epsilon(\xi)) < \epsilon \) for all \( \xi \in \partial C \).

Appendix B: Proofs

B.1. Proofs of Lemmas

Proof of Lemma 1. Let \( (u_0, v_0) \) be any feasible solution for the maximization problem in (2). For any \( t \in \mathbb{R} \) and any \( \xi, \zeta \in \Xi \), let \( u_t(\xi) := u_0(\xi) + t \) and \( v_t(\zeta) := v_0(\zeta) - t \). Then it follows that \( u_t(\xi) + v_t(\zeta) \leq d^p(\xi, \zeta) \) for all \( \xi, \zeta \in \Xi \), and
\[
\int_{\Xi} u_t(\xi) \mu(d\xi) + \int_{\Xi} v_t(\zeta) \nu(d\zeta) = \int_{\Xi} u_0(\xi) \mu(d\xi) + \int_{\Xi} v_0(\zeta) \nu(d\zeta) + t[\mu(\Xi) - \nu(\Xi)].
\]
Since \( \mu(\Xi) \neq \nu(\Xi) \),
\[
\sup_{t \in \mathbb{R}} \left\{ \int_{\Xi} u_t(\xi) \mu(d\xi) + \int_{\Xi} v_t(\zeta) \nu(d\zeta) \right\} = \infty,
\]
and thus \( W_p^p(\mu, \nu) = \infty \). \( \square \)
Proof of Lemma 2.

(i) For any \( \zeta \in \Xi \), \( \Phi(\cdot, \zeta) \) is the infimum of nondecreasing affine functions of \( \lambda \), \( \Phi(\lambda, \zeta) < \infty \) for all \( \lambda \geq 0 \), and \( \Phi(\lambda, \zeta) > -\infty \) for all \( \lambda > \kappa \), and thus \( \Phi(\cdot, \zeta) \) is non-decreasing and concave on \([0, \infty)\). For the second part, consider any \( \zeta \in \Xi \) and any \( \lambda_2 > \lambda_1 \) such that \( \Phi(\lambda_i, \zeta) > -\infty \) for \( i = 1, 2 \). For any \( \delta > 0 \), choose any \( \xi_i \in \Xi \) such that \( \lambda_i d^p(\xi_i, \zeta) - \Psi(\xi_i) < \Phi(\lambda_i, \zeta) + \delta \) for \( i = 1, 2 \). It follows that

\[
\lambda_2 d^p(\xi_2, \zeta) - \Psi(\xi_2) < \Phi(\lambda_2, \zeta) + \delta \\
\leq \lambda_2 d^p(\xi_1, \zeta) - \Psi(\xi_1) + \delta \\
= (\lambda_2 - \lambda_1) d^p(\xi_1, \zeta) + \lambda_1 d^p(\xi_1, \zeta) - \Psi(\xi_1) + \delta \\
< (\lambda_2 - \lambda_1) d^p(\xi_1, \zeta) + \Phi(\lambda_1, \zeta) + 2\delta \\
\leq (\lambda_2 - \lambda_1) d^p(\xi_1, \zeta) + \lambda_1 d^p(\xi_2, \zeta) - \Psi(\xi_2) + 2\delta.
\]

Hence

\[
(\lambda_2 - \lambda_1) d^p(\xi_2, \zeta) < (\lambda_2 - \lambda_1) d^p(\xi_1, \zeta) + 2\delta.
\]

Diving by \( \lambda_2 - \lambda_1 \) on both sides and let \( \delta \downarrow 0 \), and using definition (3), we obtain that \( \overline{D}(\lambda_2, \zeta) \leq \overline{D}(\lambda_1, \zeta) \leq \overline{D}(\lambda_1, \zeta) \).

(ii) By Definition 3, for any \( \lambda_0 \in \text{dom}(\Phi(\cdot, \zeta)) \) and for \( \nu \)-almost all \( \zeta \), it holds that

\[
\Psi(\xi) \leq \lambda_0 d^p(\xi, \zeta) + \Phi(\lambda_0, \zeta), \quad \forall \xi \in \Xi.
\]

On the other hand, for every \( \xi \in \Xi \) that satisfies \( \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta \) for some \( \delta \geq 0 \), it holds that

\[
\lambda d^p(\xi, \zeta) - \Psi(\xi) - \delta \leq -\Psi(\zeta).
\]

Combining the two inequalities above yields that

\[
\lambda d^p(\xi, \zeta) + \Psi(\xi) - \delta \leq \lambda_0 d^p(\xi, \zeta) + \Phi(\lambda_0, \zeta),
\]

or equivalently,

\[
(\lambda - \lambda_0) d^p(\xi, \zeta) - \delta \leq -\Psi(\xi) + \Phi(\lambda_0, \zeta).
\]

Taking the supremum over the set \( \{\xi \in \Xi : \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta\} \) and then the limsup with \( \delta \downarrow 0 \), and using definition (3) of \( \overline{D}(\lambda, \zeta) \), we obtain that

\[
(\lambda - \lambda_0) \overline{D}(\lambda, \zeta) \leq -\Psi(\zeta) + \Phi(\lambda_0, \zeta).
\]

(iii) Consider any \( \zeta \) and any \( \lambda_2 > \lambda_1 \) with \( \Phi(\lambda_1, \zeta) > -\infty \). For any \( \delta > 0 \), choose any \( \xi_i \in \Xi \) such that \( \lambda_i d^p(\xi_i, \zeta) - \Psi(\xi_i) \leq \Phi(\lambda_i, \zeta) + \delta \) for \( i = 1, 2 \). Then

\[
\Phi(\lambda_1, \zeta) - \Phi(\lambda_2, \zeta) \leq \lambda_1 d^p(\xi_1, \zeta) - \Psi(\xi_1) - \left[\lambda_2 d^p(\xi_2, \zeta) - \Psi(\xi_2)\right] + \delta = (\lambda_1 - \lambda_2) d^p(\xi_2, \zeta) + \delta.
\]

Similarly, \( \Phi(\lambda_2, \zeta) - \Phi(\lambda_1, \zeta) \leq (\lambda_2 - \lambda_1) d^p(\xi_1, \zeta) + \delta \). It follows that

\[
d^p(\xi_2, \zeta) - \frac{\delta}{\lambda_2 - \lambda_1} \leq \frac{\Phi(\lambda_2, \zeta) - \Phi(\lambda_1, \zeta)}{\lambda_2 - \lambda_1} \leq d^p(\xi_1, \zeta) + \frac{\delta}{\lambda_2 - \lambda_1}.
\]

Then it follows from (3) that

\[
\overline{D}(\lambda_2, \zeta) \leq \frac{\Phi(\lambda_2, \zeta) - \Phi(\lambda_1, \zeta)}{\lambda_2 - \lambda_1} \leq \overline{D}(\lambda_1, \zeta).
\]
Since $\Phi(\cdot, \zeta)$ is finite-valued and concave on $(\kappa, \infty)$, the left and right derivatives $\partial \Phi(\lambda, \zeta)/\partial \lambda_{\pm}$ exist. Setting $\lambda_1 = \lambda$ and letting $\lambda_2 \downarrow \lambda$ in the inequality above, it follows that

$$\lim_{\lambda_2 \downarrow \lambda} \overline{D}(\lambda_2, \zeta) \leq \frac{\partial \Phi(\lambda, \zeta)}{\partial \lambda_+} \leq D(\lambda, \zeta).$$

Similarly, setting $\lambda_2 = \lambda$ and letting $\lambda_1 \uparrow \lambda$, it follows that

$$\bar{D}(\lambda, \zeta) \leq \frac{\partial \Phi(\lambda, \zeta)}{\partial \lambda_-} \leq \lim_{\lambda_1 \uparrow \lambda} \underline{D}(\lambda_1, \zeta).$$

\[\square\]

**Proof of Lemma 3.**

(i) By Definition 1.11 in Ambrosio et al. [2], $\nu$ has an extension, still denoted by $\nu$, such that the measure space $(\Xi, \mathcal{B}_\nu, \nu)$ is complete. Note that for any $b \in \mathbb{R}$, it holds that

$$\{ \zeta \in \Xi : \Phi(\lambda, \zeta) < b \} = \{ \zeta \in \Xi : \exists \xi \in \Xi \text{ such that } \lambda d^p(\xi, \zeta) - \Psi(\xi) < b \}
= \pi^2(\{(\xi, \zeta) \in \Xi \times \Xi : \lambda d^p(\xi, \zeta) - \Psi(\xi) < b\}).$$

Note that the set $\{(\xi, \zeta) \in \Xi \times \Xi : \lambda d^p(\xi, \zeta) - \Psi(\xi) < b\}$ on the right side is measurable. Since $(\Xi, d)$ is Polish, it follows from the measurable projection theorem (cf. Theorem 8.3.2 in Aubin and Frankowska [3]), that $\Phi(\lambda, \cdot)$ is $(\mathcal{B}_\nu, \mathcal{B}(\mathbb{R}))$-measurable.

Define functions $\overline{C}, \underline{C}$ by

$$\overline{C}(\lambda, \zeta, \delta) := \sup_{\xi \in \Xi} \{ d^p(\xi, \zeta) : \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta \}$$

$$\underline{C}(\lambda, \zeta, \delta) := \inf_{\xi \in \Xi} \{ d^p(\xi, \zeta) : \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta \}.$$

For any $b \in \mathbb{R}$ it holds that

$$\{ \zeta \in \Xi : \overline{C}(\lambda, \zeta, \delta) > b \} = \{ \zeta \in \Xi : \exists \xi \in \Xi \text{ such that } \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta, d^p(\xi, \zeta) > b \}
= \pi^2(\{(\xi, \zeta) \in \Xi \times \Xi : \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta, d^p(\xi, \zeta) > b\}),$$

and thus it follows from the measurable projection theorem that $\overline{C}(\lambda, \cdot, \delta)$ is $(\mathcal{B}_\nu, \mathcal{B}(\mathbb{R}))$-measurable. Similarly,

$$\{ \zeta \in \Xi : \underline{C}(\lambda, \zeta, \delta) < b \} = \{ \zeta \in \Xi : \exists \xi \in \Xi \text{ such that } \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta, d^p(\xi, \zeta) < b \}
= \pi^2(\{(\xi, \zeta) \in \Xi \times \Xi : \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta, d^p(\xi, \zeta) < b\}),$$

and thus $\underline{C}(\lambda, \cdot, \delta)$ is $(\mathcal{B}_\nu, \mathcal{B}(\mathbb{R}))$-measurable. Note that $\overline{D}(\lambda, \cdot) = \limsup_{\delta \downarrow 0} \overline{C}(\lambda, \cdot, \delta)$ and $\underline{D}(\lambda, \cdot) = \liminf_{\delta \downarrow 0} \underline{C}(\lambda, \cdot, \delta)$ are also $(\mathcal{B}_\nu, \mathcal{B}(\mathbb{R}))$-measurable, because measurability is preserved under limsup and liminf.

(ii) Consider any $\delta, \epsilon > 0$. Define multi-valued mappings $\overline{S}, \underline{S} : \mathbb{R}_+ \times \Xi \Rightarrow \Xi$ by

$$\overline{S}(\lambda, \zeta, \delta, \epsilon) := \{ \xi \in \Xi : \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta, d^p(\xi, \zeta) \leq \overline{D}(\lambda, \zeta) - \epsilon \},$$

$$\underline{S}(\lambda, \zeta, \delta, \epsilon) := \{ \xi \in \Xi : \lambda d^p(\xi, \zeta) - \Psi(\xi) \leq \Phi(\lambda, \zeta) + \delta, d^p(\xi, \zeta) \leq \underline{D}(\lambda, \zeta) + \epsilon \}.$$

For each $\zeta \in \Xi$, it follows from the measurability of $\Psi$ and $d^p(\cdot, \zeta)$ that $\overline{S}(\lambda, \zeta, \delta, \epsilon)$ and $\underline{S}(\lambda, \zeta, \delta, \epsilon)$ are in $\mathcal{B}(\Xi)$. Since $(\Xi, d)$ is Polish and $\nu$ is a complete finite measure, it follows from Aumann’s measurable selection theorem (see, e.g. Theorem 18.26 in Aliprantis and Border [1]) that $\nu$-measurable selections $T, \underline{T} : \Xi \Rightarrow \Xi$ exist such that $T(\zeta) \in \overline{S}(\lambda, \zeta, \delta, \epsilon)$ and $\underline{T}(\zeta) \in \underline{S}(\lambda, \zeta, \delta, \epsilon)$.

(iii) Define a multi-valued mapping $S : \mathbb{R}_+ \times \Xi \Rightarrow \Xi$ by

$$S(\epsilon, \zeta) := \{ \xi \in \Xi : \Psi(\xi) - \Psi(\zeta) \geq (\kappa - \epsilon) d^p(\xi, \zeta), d^p(\xi, \zeta) \geq M(\zeta) \}.$$

For each $\zeta \in \Xi$, it follows from the measurability of $\Psi$, $M$ and $d^p(\cdot, \zeta)$ that $S(\epsilon) \in \mathcal{B}(\Xi)$. Then using the same argument as in (ii), there exists a $\nu$-measurable selection $T : \Xi \Rightarrow \Xi$ such that $T(\zeta) \in S(\epsilon, \zeta)$. □
Proof of Lemma 4. Note that if \( x = 0 \), then the inequality holds for any \( C_p(\epsilon) \geq 1 \). Next we consider the case with \( x > 0 \), and we let \( t := y/x \). Let

\[
t_0(\epsilon) := \sup\{t > 0 : 1 + \epsilon \geq (1 + t)^p\}.
\]

Note that \( t_0(\epsilon) > 0 \). Next let

\[
C_p(\epsilon) := \max\left\{1, \sup\limits_{t \geq t_0(\epsilon)} \left(\frac{1 + t}{t} \right)^{p-1}\right\}.
\]

Note that \( C_p(\epsilon) < \infty \) because \( \lim_{t \to \infty} (1 + t)^{p-1}/t^{p-1} = 1 \). Next, consider

\[
f(t) := 1 + \epsilon + C_p(\epsilon) t^p - (1 + t)^p
\]

Note that \( f(t) \geq 0 \) for all \( t \in [0, t_0(\epsilon)] \). Also, \( f'(t) = C_p(\epsilon) p t^{p-1} - p(1 + t)^{p-1} \geq 0 \) for all \( t \in [t_0(\epsilon), \infty) \). Therefore \( f(t) \geq 0 \) for all \( t \geq 0 \), which establishes the inequality for \( x > 0 \). \( \square \)

Proof of Lemma 5. It follows from Lemma 4 with \( \epsilon := \frac{\lambda - \lambda_1}{2 \lambda_1} \) that

\[
\lambda_1 d^p(\xi, \zeta^0) \leq \frac{\lambda + \lambda_1}{2} d^p(\xi, \zeta) + \lambda_1 C_p(\epsilon) d^p(\zeta, \zeta^0)
\]

for all \( \xi, \zeta, \zeta^0 \in \Xi \). Thus

\[
\lambda d^p(\xi, \zeta) - \Psi(\xi) = \frac{\lambda - \lambda_1}{2} d^p(\xi, \zeta) - \Psi(\xi) + \frac{\lambda + \lambda_1}{2} d^p(\xi, \zeta)
\]

\[
\geq \frac{\lambda - \lambda_1}{2} d^p(\xi, \zeta) - \Psi(\xi) + \lambda_1 d^p(\xi, \zeta^0) - \lambda_1 C_p(\epsilon) d^p(\zeta, \zeta^0)
\]

\[
\geq \frac{\lambda - \lambda_1}{2} d^p(\xi, \zeta) + \Phi(\lambda_1, \zeta^0) - \lambda_1 C_p(\epsilon) d^p(\zeta, \zeta^0).
\]

Hence, for every \( \xi \in \Xi \) that satisfies \( \lambda d^p(\xi, \zeta) - \Psi(\xi) < \Phi(\lambda, \zeta) + \delta \) for some \( \delta \geq 0 \), it holds that

\[
\frac{\lambda - \lambda_1}{2} d^p(\xi, \zeta) < \Phi(\lambda, \zeta) - \Phi(\lambda_1, \zeta^0) + \lambda_1 C_p(\epsilon) d^p(\zeta, \zeta^0) + \delta.
\]

Taking the supremum over \( \xi \in \Xi \) on both sides and then the lim sup with \( \delta \downarrow 0 \), we obtain that

\[
\frac{\lambda - \lambda_1}{2} D(\lambda, \zeta) \leq \Phi(\lambda, \zeta) - \Phi(\lambda_1, \zeta^0) + \lambda_1 C_p(\epsilon) d^p(\zeta, \zeta^0). \quad \square
\]

Proof of Lemma 6.

(i) We prove this by contradiction. Suppose that for some \( \zeta^0, \zeta^1 \in \Xi \), it holds that

\[
\kappa^0 < \kappa^1 := \limsup\limits_{d(\xi, \zeta^1) \to \infty} \frac{\Psi(\xi) - \Psi(\zeta^1)}{d^p(\xi, \zeta^1)}
\]

(\( \kappa^1 = \infty \) is allowed, and the case \( \kappa^0 > \kappa^1 \) can be shown similarly). Choose any \( \epsilon \in (0, \kappa^1 - \kappa^0) \). Then there exists an \( R \) such that for all \( \xi \) with \( d(\xi, \zeta^0) > R \) it holds that

\[
\Psi(\xi) - \Psi(\zeta^1) = \Psi(\xi) - \Psi(\zeta^0) + \Psi(\zeta^0) - \Psi(\zeta^1)
\]

\[
< (\kappa^0 + \epsilon) d^p(\xi, \zeta^0) + [\Psi(\zeta^0) - \Psi(\zeta^1)]
\]

\[
\leq (\kappa^0 + \epsilon) [d(\xi, \zeta^1) + d(\zeta^1, \zeta^0)]^p + [\Psi(\zeta^0) - \Psi(\zeta^1)].
\]
It follows that
\[
\limsup_{d(\xi, \zeta^1) \to \infty} \frac{\Psi(\xi) - \Psi(\zeta^1)}{d^p(\xi, \zeta^1)} \leq \limsup_{d(\xi, \zeta^0) \to \infty} \frac{(\kappa^0 + \epsilon)[d(\xi, \zeta^1) + d(\zeta^1, \zeta^0)]^{p} + [\Psi(\zeta^0) - \Psi(\zeta^1)]}{d^p(\xi, \zeta^1)}
\]
\[
= \kappa^0 + \epsilon < \kappa^1,
\]
which is a contradiction.

(ii) (Sufficiency). If (31) holds, then
\[
\kappa^0 := \limsup_{\xi \in \Xi, d(\xi, \zeta^0) \to \infty} \frac{\Psi(\xi) - \Psi(\zeta^0)}{d^p(\xi, \zeta^0)} \leq L < \infty.
\]
Then by (i),
\[
\kappa^0 = \limsup_{\xi \in \Xi, d(\xi, \zeta) \to \infty} \frac{\Psi(\xi) - \Psi(\zeta)}{d^p(\xi, \zeta)}, \ \forall \zeta \in \Xi. \tag{32}
\]
We are going to show that \(\kappa = \kappa^0\), that is, \(\int_{\Xi} \Phi(\lambda, \zeta)\nu(d\zeta) > \infty\) for any \(\lambda > \kappa^0\).
To this end, we first show that \(\Phi(\lambda, \zeta) > \infty\) for any \(\lambda > \kappa^0\) and \(\zeta \in \Xi\). In fact, by (32), for any \(\zeta \in \Xi\), there is a \(C > 0\) such that for all \(\xi \in \Xi\) with \(d^p(\xi, \zeta) > C\), it holds that
\[
\frac{\Psi(\xi) - \Psi(\zeta)}{d^p(\xi, \zeta)} < \frac{\lambda + \kappa^0}{2},
\]
that is, \((\frac{\lambda + \kappa^0}{2})d^p(\xi, \zeta) - \Psi(\xi) > -\Psi(\zeta)\). Thus, for all \(\xi \in \Xi\) with \(d^p(\xi, \zeta) > C\), it holds that
\[
\lambda d^p(\xi, \zeta) - \Psi(\xi) = \frac{\lambda + \kappa^0}{2} d^p(\xi, \zeta) - \Psi(\xi) + \frac{\lambda - \kappa^0}{2} d^p(\xi, \zeta)
\]
\[
> -\Psi(\zeta) + \frac{\lambda - \kappa^0}{2} \cdot C > -\infty,
\]
and hence
\[
\inf_{\xi \in \Xi} \{\lambda d^p(\xi, \zeta) - \Psi(\xi) : d^p(\xi, \zeta) > C\} \geq -\Psi(\zeta) + \frac{\lambda - \kappa^0}{2} \cdot C > -\infty. \tag{33}
\]
Using (31) we have that for any \(\zeta, \xi \in \Xi\),
\[
\Psi(\xi) - \Psi(\zeta) = \Psi(\xi) - \Psi(\zeta^0) + \Psi(\zeta^0) - \Psi(\zeta)
\]
\[
\leq L d^p(\xi, \zeta^0) + \Psi(\zeta^0) - \Psi(\zeta)
\]
\[
\leq 2^{p-1} L [d^p(\xi, \zeta) + d^p(\zeta, \zeta^0)] + \Psi(\zeta^0) - \Psi(\zeta)
\]
\[
=: L' d^p(\xi, \zeta) + M(\zeta),
\]
where the second inequality follows from the elementary inequality \((a+b)^p \leq 2^{p-1}(a^p + b^p)\) for any \(a, b \geq 0\) and \(p \geq 1\), and
\[
L' := 2^{p-1} L, \quad M(\zeta) := 2^{p-1} L d^p(\zeta, \zeta^0) + \Psi(\zeta^0) - \Psi(\zeta). \tag{34}
\]
It then follows that
\[
\inf_{\xi \in \Xi} \{\lambda d^p(\xi, \zeta) - \Psi(\xi) : d^p(\xi, \zeta) \leq C\} \geq \inf_{\xi \in \Xi} \{-\Psi(\xi) : d^p(\xi, \zeta) \leq C\}
\]
\[
\geq -\Psi(\zeta) - L(\zeta)C - M(\zeta) > -\infty.
\]
Therefore, \(\Phi(\lambda, \zeta) > -\infty\) for all \(\zeta \in \Xi\) and \(\lambda > \kappa^0\).
Then we show that $\int_\Xi \Phi(\lambda, \zeta) \nu(d\zeta) > -\infty$ for any $\lambda > \kappa^0$. Let $\lambda_1 \in (\kappa^0, \lambda)$. Choose any $\zeta^0 \in \Xi$. It follows from Lemma 5 that there is a constant $C$ such that
\[
\Phi(\lambda, \zeta) \geq \frac{\lambda - \lambda_1}{2} D(\lambda, \zeta) + \Phi(\lambda_1, \zeta^0) - C d^p(\zeta, \zeta^0) \geq \Phi(\lambda_1, \zeta^0) - C d^p(\zeta, \zeta^0).
\]
Integrating over $\zeta$ with respect to $\nu$ yields that
\[
\int_\Xi \Phi(\lambda, \zeta) \nu(d\zeta) \geq \Phi(\lambda_1, \zeta^0) - C \int_\Xi d^p(\zeta, \zeta^0) \nu(d\zeta) > -\infty.
\]
Therefore we have shown that $\kappa = \kappa^0$.

(Necessity). Observe that in the proof of sufficiency, we have shown that the condition (31) is equivalent to the following condition: there exists $L' \geq 0$ and $M(\zeta) \in L^1(\nu)$ such that
\[
\Psi(\zeta) - \Psi(\zeta^0) \leq L' d^p(\zeta, \zeta^0) + M(\zeta), \quad \forall \zeta, \zeta^0 \in \Xi.
\]
Suppose $\kappa < \infty$, and we are going to prove the necessity by contradiction. Assume the above equivalent condition does not hold, then for any $\lambda \geq 0$,
\[
\inf_{\zeta \in \Xi} [\lambda d^p(\zeta, \zeta^0) - \Psi(\zeta)] \notin L^1(\nu),
\]
which, implies that $\kappa = \infty$, a contradiction.

(iii) follows directly from the proof of (ii). \hfill \square

**Proof of Lemma 7.** Since $\Xi$ is separable, $\partial C$ has a countable dense subset $\{\xi^i\}_{i=1}^\infty$. For each $\xi^i$, there exists $\xi^i' \in \Xi \setminus \text{cl}(\Xi)$ such that $\epsilon_i := 2d(\xi^i, \xi^i') < \epsilon$. Thus $\partial C = \bigcup_{i=1}^{\infty} B_{\epsilon_i}(\xi^i)$, where $B_{\epsilon_i}(\xi^i)$ is the open ball centered at $\xi^i$ with radius $\epsilon_i$. Define
\[
i^*(\xi) := \min_{i \geq 0} \{i : \xi \in B_{\epsilon_i}(\xi^i)\}, \quad \xi \in \partial C,
\]
and
\[
T_\epsilon(\xi) := \xi_{i^*(\xi)}^*, \quad \xi \in \partial C.
\]
Then $T_\epsilon$ satisfies the requirements in the lemma. \hfill \square

**B.2. Proofs of Corollaries**

**Proof of Corollary 1.**

(i) (Necessity). First we consider that there exists a dual minimizer $\lambda^* > \kappa$. Note that if $\Psi$ is upper-semicontinuous and bounded subsets of $(\Xi, d)$ are totally bounded, then in the proof of Lemma 3(ii), $C(\lambda, \zeta, \delta)$ and $\overline{C}(\lambda, \zeta, \delta)$ are well-defined for $\delta = 0$, the supremum and the infimum are attained. It then follows that $\underline{C}(\lambda, \zeta, \delta, \epsilon)$ and $\overline{C}(\lambda, \zeta, \delta, \epsilon)$ in the proof of Lemma 3(ii) are also well-defined for $\epsilon = \delta = 0$. Furthermore, we note that if $\Psi$ is upper-semicontinuous bounded subsets of $(\Xi, d)$ are totally bounded, then $\overline{D}_0(\cdot, \zeta)$ is left continuous and $\overline{D}_0(\cdot, \zeta)$ is right continuous on $(\kappa, \infty)$. Indeed, fixing $\lambda > \kappa$, let $\{\lambda_n\}_n$ be a sequence monotonically increasing to $\lambda$ with $\lambda_n \in ((\lambda + \kappa)/2, \lambda)$ for all $n$. By Lemma 2(i), it holds that
\[
\overline{D}_0(\lambda_{n+1}, \zeta) \leq \overline{D}_0(\lambda_n, \zeta) \leq \overline{D}_0((\lambda + \kappa)/2, \zeta), \quad \forall n.
\] (35)

Suppose on the contrary that there exists $\delta > 0$ such that $\lim_{n \to \infty} \overline{D}_0(\lambda_n, \zeta) > \overline{D}_0(\lambda, \zeta) + \delta$. Let
\[
\xi^0 \in \text{arg min}_{\xi \in \Xi} \left\{ d(\xi, \zeta) : \xi \in \text{arg min}_{\xi \in \Xi} \left\{ \lambda_n d^p(\xi, \zeta) - \Psi(\xi) \right\} \right\},
\]
\[
\xi^0_n \in \text{arg min}_{\xi \in \Xi} \left\{ d(\xi, \zeta) : \xi \in \text{arg min}_{\xi \in \Xi} \left\{ \lambda_n d^p(\xi, \zeta) - \Psi(\xi) \right\} \right\}.
\]
By definition we have that
\[ \lambda_n d^p(\xi^n, \zeta) - \Psi(\xi^n) \leq \lambda_n d^p(\xi^0, \zeta) - \Psi(\xi^0), \]
From (35) and totally boundedness, up to a subsequence we can assume \( \{\xi^n\} \) converges to some \( \xi^* \in \Xi \). Let \( n \to \infty \) and by lower-semicontinuity of \( -\Psi \),
\[ \lambda d^p(\xi^*, \zeta) - \Psi(\xi^*) \leq \liminf_{k \to \infty} \lambda_n d^p(\xi^n, \zeta) - \Psi(\xi^n) \leq \lambda d^p(\xi^0, \zeta) - \Psi(\xi^0), \]
thus we obtain that \( \xi^* \) is a minimizer of \( \inf_{\xi \in \Xi} \lambda d^p(\xi, \zeta) - \Psi(\xi) \), but \( d(\xi^*, \zeta) \geq D_0(\lambda, \zeta) + \delta \), which contradicts to the definition of \( D_0(\lambda, \zeta) \). Therefore we have shown that \( D_0(\lambda, \zeta) \) is left continuous with respect to \( \lambda \). Using a similar argument we can show the right continuity of \( D(\cdot, \zeta) \).

It follows from the above results and Lemma 2(i)(iii) that for \( \lambda > \kappa \), it holds that
\[ \frac{\partial \Phi(\lambda, \zeta)}{\partial \lambda^+} = D(\lambda, \zeta), \quad \frac{\partial \Phi(\lambda, \zeta)}{\partial \lambda^-} = D(\lambda, \zeta). \]

Then mimicking the proof of Theorem 1 Case 1, we define mappings
\[ T_{\lambda^*}(\zeta) \in \left\{ \xi \in \Xi : \lambda^* d^p(\xi, \zeta) - \Psi(\xi) = \Phi(\lambda^*, \zeta), \ d^p(\xi, \zeta) = D(\lambda^*, \zeta) \right\}, \]
\[ T_{\lambda^*}(\zeta) \in \left\{ \xi \in \Xi : \lambda^* d^p(\xi, \zeta) - \Psi(\xi) = \Phi(\lambda^*, \zeta), \ d^p(\xi, \zeta) = D(\lambda^*, \zeta) \right\}, \]
and define \( q \in [0, 1] \) such that
\[ q \int_{\Xi} d^p(T(\zeta), \zeta) \nu(d\zeta) + (1 - q) \int_{\Xi} d^p(\overline{T}(\zeta), \zeta) \nu(d\zeta) = \theta^p. \]
Let
\[ \mu^* := q \cdot T_# \nu + (1 - q) \cdot \overline{T}_# \nu. \] (36)
Then \( \mu^* \) is feasible and
\[ \int_{\Xi} \Psi(\xi) \mu^*(d\xi) = q \int_{\Xi} \Psi(T(\zeta)) \nu(d\zeta) + (1 - q) \int_{\Xi} \Psi(\overline{T}(\zeta)) \nu(d\zeta) \]
\[ = q \int_{\Xi} [\lambda^* d^p(T(\zeta), \zeta) - \Phi(\lambda^*, \zeta)] \nu(d\zeta) + (1 - q) \int_{\Xi} [\lambda^* d^p(\overline{T}(\zeta) - \Phi(\lambda^*, \zeta)] \nu(d\zeta) \]
\[ = \lambda^* \theta^p - \int_{\Xi} \Phi(\lambda^*, \zeta) \nu(d\zeta) = v_D. \]

Therefore \( \mu^* \) is optimal.

Next, suppose that \( \lambda^* = \kappa > 0 \) is the unique minimizer, \( \nu\left\{ \{\zeta \in \Xi : \arg\min_{\xi \in \Xi}\{\kappa d^p(\xi, \zeta) - \Psi(\xi)\} = \emptyset \right\} = 0 \), and condition (2) holds. Then the sets \( E, E \) in Lemma 3(ii) are well-defined for \( \lambda = \kappa \) and \( \delta = \epsilon = 0 \), hence there exists \( \nu \)-measurable maps \( T, \overline{T} : \Xi \to \Xi \), such that \( \kappa d^p(T(\zeta), \zeta) - \Psi(T(\zeta)) = \Phi(\kappa, \zeta) = \kappa d^p(T(\zeta), \zeta) - \Psi(T(\zeta)) \) holds \( \nu \)-a.s., and that
\[ \int_{\Xi} d^p(T(\zeta), \zeta) \nu(d\zeta) \leq \theta^p \leq \int_{\Xi} d^p(\overline{T}(\zeta), \zeta) \nu(d\zeta). \]
Then the distribution defined by (36) is optimal using the same argument.

Third, suppose that \( \lambda^* = 0 \) is the unique minimizer, \( \arg\max_{\xi \in \Xi}\{\Psi(\xi)\} \) is nonempty, and condition (3) holds. Then the sets \( \overline{E}, E \) in Lemma 3(ii) are well-defined for \( \lambda = 0 \) and \( \delta = \epsilon = 0 \), hence there exists \( \nu \)-measurable map \( T : \Xi \to \Xi \), such that \( T(\zeta) \in \arg\max_{\xi \in \Xi}\{\Psi(\xi)\} \) holds \( \nu \)-a.s., and that
\[ \int_{\Xi} d^p(T(\zeta), \zeta) \nu(d\zeta) \leq \theta^p. \]
Define \(\mu := T_{\#} \nu\). It follows that
\[
\int_{\Xi} \Psi(\xi) \mu(d\xi) = \int_{\Xi} \Psi(T(\xi)) \nu(d\xi) = \max_{\xi \in \Xi} \Psi(\xi) = v_D.
\]
Therefore we have shown the “if” part.

(Sufficiency). Let \(\mu\) be a primal feasible solution, and let \((\xi, \zeta)\) be a random vector with marginal distributions \(\mu\) and \(\nu\). Let \(\gamma_\zeta\) be a conditional distribution of \(\xi\) given \(\zeta\) such that \(\int_{\Xi^2} d\nu(\xi, \zeta) \gamma_\zeta(d\xi) \leq \theta^p\). Then the weak duality implies for any \(\lambda \geq 0\),
\[
\int_{\Xi} \Psi(\xi) \mu(d\xi) = \int_{\Xi^2} [\Psi(\xi) - \lambda d\nu(\xi, \zeta)] \gamma_\zeta(d\xi) \nu(d\zeta) + \int_{\Xi^2} \lambda d\nu(\xi, \zeta) \gamma_\zeta(d\xi) \nu(d\zeta)
\leq -\int_{\Xi} \Phi(\lambda, \zeta) \nu(d\zeta) + \lambda \theta^p.
\]
Hence, to make the inequality holds as equality (and thus \(\mu\) is a worst-case distribution and \(\lambda\) is dual optimal), it must hold that
(a) \(\arg \min_{\xi \in \Xi} \{\lambda d\nu(\xi, \zeta) - \Psi(\xi)\}\) is non-empty for \(\nu\)-almost every \(\zeta\).
(b) For \(\nu\)-almost every \(\zeta\), the conditional distribution \(\gamma_\zeta\) of \(\xi\) should be supported on the set \(\arg \min_{\xi \in \Xi} \{\lambda d\nu(\xi, \zeta) - \Psi(\xi)\}\).
(c) \(\lambda \cdot (\theta^p - \int_{\Xi^2} d\nu(\xi, \zeta) \gamma_\zeta(d\xi) \nu(d\zeta)) = 0\).

Now suppose all conditions in Corollary 1(i) fail to hold. This could happen when
(1) \(\lambda^* = \kappa > 0\) and (2) fails to hold.
(2) \(\lambda^* = 0\) and (3) fails to hold.
Let us first consider (1). Suppose \(\mu\) is an optimal solution. If \(\int_{\Xi} D_0(\kappa, \zeta) \nu(d\zeta) < \theta^p\), then together with (a)(b) we have that
\[
\int_{\Xi} \int_{\Xi} d\nu(\xi, \zeta) \gamma_\zeta(d\xi) \nu(d\zeta) \leq \int_{\Xi} D_0(\kappa, \zeta) \nu(d\zeta) < \theta^p,
\]
which contracts to (c). If \(\theta^p < \int_{\Xi} D_0(\kappa, \zeta) \nu(d\zeta)\), by Lemma 2(ii)(iii) and Lemma 3, we have that
\[
\frac{\partial}{\partial \lambda} \int_{\Xi} \Phi(\lambda, \zeta) \nu(d\zeta) = \int_{\Xi} D_0(\kappa, \zeta) \nu(d\zeta) > \theta^p.
\]
Therefore, there exists \(\lambda > \kappa\), such that \(\lambda \theta^p - \int_{\Xi} \Phi(\lambda, \zeta) \nu(d\zeta) < \kappa \theta^p - \int_{\Xi} \Phi(\kappa, \zeta) \nu(d\zeta)\), and thus \(\kappa\) cannot be a dual minimizer, a contradiction.

Next we consider (2), i.e., \(\lambda^* = 0\) and \(\int_{\Xi} D_0(0, \zeta) \nu(d\zeta) > \theta^p\). Suppose \(\mu\) is optimal, namely, \(\int_{\Xi} \Psi(\xi) \mu(d\xi) = \max_{\xi \in \Xi} \Psi(\xi)\). Then (a)(b) and \(\int_{\Xi} D_0(0, \zeta) \nu(d\zeta) > \theta^p\) imply that \(\int_{\Xi^2} d\nu(\xi, \zeta) \gamma_\zeta(d\xi) \nu(d\zeta) > \theta^p\), thus we arrive at a contradiction.

(ii) If \(-\Psi(\xi) \leq \inf_{\xi \in \Xi} \{\kappa d\nu(\xi, \zeta) - \Psi(\xi)\}\) \(\nu\)-almost everywhere, i.e., \(\Psi(\xi) - \Psi(\zeta) \leq \kappa d\nu(\xi, \zeta)\), then for any \(\lambda > \kappa\), \(\Phi(\lambda, \zeta) = \Psi(\zeta)\). Hence the dual optimal solution \(\lambda^* = \kappa\).

Otherwise there exists a set \(E \subset \Xi\) with \(\nu(E) > 0\) such that \(\Psi(\zeta) > \Phi(\kappa, \zeta)\) for all \(\zeta \in E\), and thus \(\int_{\Xi} \Psi(\xi) \nu(d\zeta) > \int_{\Xi} \Phi(\kappa, \zeta) \nu(d\zeta)\). Then by continuity (follows from concavity) of \(\int_{\Xi} \Phi(\cdot, \zeta) \nu(d\zeta)\), there exists \(\lambda_0 > \kappa\) such that \(\int_{\Xi} \Psi(\xi) \nu(d\zeta) > \int_{\Xi} \Phi(\lambda_0, \zeta) \nu(d\zeta)\). For such \(\lambda_0\), using the upper-semicontinuity and totally boundedness assumptions and Lemma 2, there exists a \(\nu\)-measurable map \(T_{\lambda_0} : \Xi \to \Xi\), such that \(\lambda_0 d\nu(T_{\lambda_0}(\zeta), \zeta) - \Psi(T_{\lambda_0}(\zeta)) = \Phi(\lambda_0, \zeta)\), and
\[
eq \int_{\Xi} d\nu(T_{\lambda_0}(\zeta), \zeta) > 0,
\]
since otherwise \(\int_{\Xi} \Psi(\xi) \nu(d\zeta) = \int_{\Xi} \Phi(\lambda_0, \zeta) \nu(d\zeta)\). Choose \(\theta < \epsilon^{1/p}\), then we claim that \(\lambda = \kappa\) is cannot be optimal. Indeed, according to Case 2 in the proof of Theorem 1, \(\lambda^* = \kappa\) implies
\[ \int_{\Xi} d^p(T^p(\zeta), \zeta) \nu(d\zeta) < \theta^p < \epsilon \text{ for all } \lambda > \kappa, \text{ and in particular, for } \lambda = \lambda_0. \] Thus we arrive at a contradiction.

(iii) This directly follows from the fact that in the proof for necessity in (i), in each case we construct an optimal solution with the the structure described as in (iii).

(iv) Note that the concavity of \( \Psi \) implies that \( \kappa < \infty \). Let \( \mu = qT_\#^p \nu + (1 - q) \cdot T_\#^q \nu \) be an optimal solution as described in (i). We define \( T : \Xi \to \Xi \) by

\[ T(\zeta) := q \cdot T(\zeta) + (1 - q) \cdot T(\zeta). \]

Then it follows from \( \Xi \) being convex that \( T(\zeta) \in \Xi \) for all \( \zeta \in \Xi \). It follows from \( d^p(\cdot, \zeta) \) being convex for all \( \zeta \in \Xi \) that

\[ W^p_p(T_\#^p \nu, \nu) \leq \int_{\Xi} d^p(T(\zeta), \zeta) \nu(d\zeta) \]

\[ = \int_{\Xi} d^p(q \cdot T(\zeta) + (1 - q) \cdot T(\zeta), \zeta) \nu(d\zeta) \]

\[ \leq q \int_{\Xi} d^p(T(\zeta), \zeta) \nu(d\zeta) + (1 - q) \int_{\Xi} d^p(T(\zeta), \zeta) \nu(d\zeta) \]

\[ = W^p_p(\mu, \nu) \leq \theta^p. \]

Also, it follows from \( \Psi \) being concave that

\[ \int_{\Xi} \Psi(T(\zeta)) \nu(d\zeta) = \int_{\Xi} \Psi(q \cdot T(\zeta) + (1 - q) \cdot T(\zeta)) \nu(d\zeta) \]

\[ \geq q \int_{\Xi} \Psi(T(\zeta)) \nu(d\zeta) + (1 - q) \int_{\Xi} \Psi(T(\zeta)) \nu(d\zeta) \]

\[ = \mathbb{E}_{\mu^*}[\Psi] = v_D. \]

Hence \( T_\#^p \nu \) is an optimal solution. \( \square \)

**B.3. Proofs of Propositions**

*Proof of Proposition 1.* Let \( \mathcal{Q} := \{ \mu \in \mathcal{P}(\Xi) : W^p_p(\mu, \nu) < \infty \} \). For any \( \mu \in \mathcal{Q} \), let \( \gamma^\mu \in \mathcal{P}(\Xi^2) \) denote a minimizer in the definition (1) of \( W^p_p(\mu, \nu) \). Then by tower property of conditional probability,

\[ \int_{\Xi^2} \Psi(\xi) \mu(d\xi) = \int_{\Xi^2} \Psi(\xi) \gamma^\mu(\xi) \cdot \nu(d\xi) \cdot \nu(d\zeta), \]

and

\[ W^p_p(\mu, \nu) = \int_{\Xi^2} d^p(\xi, \zeta) \gamma^\mu(\xi) \cdot \nu(d\xi) \cdot \nu(d\zeta) = \int_{\Xi^2} d^p(\xi, \zeta) \gamma^\mu(\xi, \zeta) \cdot \nu(d\xi), \]

where \( \gamma^\mu(\cdot, \cdot) \) denotes the conditional distribution of \( \xi \) given \( \zeta \) when the joint distribution of \( (\xi, \zeta) \) is \( \gamma \). Then the primal problem can be written as

\[ v_P = \sup_{\mu \in \mathcal{Q}} \left\{ \int_{\Xi^2} \Psi(\xi) \gamma^\mu(\xi) \cdot \nu(d\xi) : \int_{\Xi^2} d^p(\xi, \zeta) \gamma^\mu(\xi, \zeta) \cdot \nu(d\xi) \leq \theta^p \right\}. \]

Next we show that

\[ v_P \leq \sup_{\mu \in \mathcal{Q}} \inf_{\lambda \geq 0} \left\{ \int_{\Xi^2} \Psi(\xi) \gamma^\mu(\xi) \cdot \nu(d\xi) \cdot \lambda \left( \theta^p - \int_{\Xi^2} d^p(\xi, \zeta) \gamma^\mu(\xi, \zeta) \cdot \nu(d\xi) \right) \right\}. \]  

(37)
If $\int_{E} \Psi(\xi) \mu(d\xi) < \infty$ for all $\mu \in \mathcal{M}$, then for any $\mu \in \mathcal{P}(\Xi) : W_{p}^{\mu}(\mu, \nu) \leq \theta$ it holds that

$$\inf_{\lambda \geq 0} \left\{ \int_{\Xi} \Psi(\xi) \gamma_{\xi}(d\xi) \nu(d\zeta) + \lambda \left( \theta^{p} - \int_{\Xi} d^{p}(\xi, \zeta) \gamma_{\zeta}^{\mu}(d\xi) \nu(d\zeta) \right) \right\} = \int_{\Xi} \Psi(\xi) \gamma_{\xi}(d\xi) \nu(d\zeta),$$

and for any $\mu \in \mathcal{Q} \setminus \mathcal{M}$ it holds that

$$\inf_{\lambda \geq 0} \left\{ \int_{\Xi} \Psi(\xi) \gamma_{\xi}(d\xi) \nu(d\zeta) + \lambda \left( \theta^{p} - \int_{\Xi} d^{p}(\xi, \zeta) \gamma_{\zeta}^{\mu}(d\xi) \nu(d\zeta) \right) \right\} = -\infty.$$ 

Thus the objective functions in (Primal) and the right side of (37) are the same for all $\mu \in \mathcal{Q}$, and therefore (37) holds as an equality.

Otherwise, if $\int_{E} \Psi(\xi) \mu(d\xi) = \infty$ for some $\mu \in \mathcal{Q}$, then for any $\lambda \geq 0$, we have that

$$\int_{\Xi} \Psi(\xi) \gamma_{\xi}(d\xi) \nu(d\zeta) + \lambda \left( \theta^{p} - \int_{\Xi} d^{p}(\xi, \zeta) \gamma_{\zeta}^{\mu}(d\xi) \nu(d\zeta) \right) = \infty,$$

because $\int_{\Xi} d^{p}(\xi, \zeta) \gamma_{\zeta}^{\mu}(d\xi) \nu(d\zeta) = W_{p}^{\mu}(\mu, \nu) = \infty$, and thus (37) holds. Therefore we conclude that

$$v_{P} \leq \inf_{\lambda \geq 0} \left\{ \lambda \theta^{p} + \sup_{\mu \in \mathcal{Q}} \left\{ \int_{\Xi} \left[ \Psi(\xi) - \lambda d^{p}(\xi, \zeta) \right] \gamma_{\zeta}^{\mu}(d\xi) \nu(d\zeta) \right\} \right\} \leq \inf_{\lambda \geq 0} \left\{ \lambda \theta^{p} + \int_{\Xi} \sup_{\xi \in \Xi} \left[ \Psi(\xi) - \lambda d^{p}(\xi, \zeta) \right] \nu(d\zeta) \right\} = v_{D}. \quad \square$$

**Proof of Proposition 2.** If $\kappa = \infty$, then for any $n > 0$, we have that

$$\phi^{n}(\zeta) := \sup_{\xi \in \Xi} \{ \Psi(\xi) - \Psi(\zeta) - n \cdot d^{p}(\xi, \zeta) \} \notin L^{1}(\nu).$$

Hence, for any $n > 0$, there exists $E \subset \Xi$ with $\nu(E) > 0$, such that $\phi^{n}(\zeta) > n$ for all $\zeta \in E$. Observe that $\phi^{n}(\zeta) = -\Phi(n, \zeta) - \Psi(\zeta)$. Thus by Lemma 3, there exists a $\nu$-measurable mapping $T^{n} : E \to \Xi$ such that

$$T^{n}(\zeta) \in \{ \zeta \in \Xi : \Psi(\xi) - \Psi(\zeta) \geq n \cdot d^{p}(\xi, \zeta) + n/2 \}.$$

For $r = 1, 2, \ldots$, consider the set

$$E_{r} := \{ \zeta \in E : d^{p}(T^{n}(\zeta), \zeta) \leq r \}.$$

Then $\lim_{r \to \infty} E_{r} = E$ and $\lim_{r \to \infty} \nu(E_{r}) = \nu(E)$. Hence there exists a $r_{0}$ such that $\nu(E_{r_{0}}) > 0$, and that $\int_{E_{r_{0}}} d^{p}(T^{n}(\zeta), \zeta) \nu(d\zeta) < \infty$. Set $T^{n}$ to be the restriction of $T^{n}$ on $E_{r_{0}}$.

Now define a distribution

$$\mu^{n} := p \cdot T^{n}_{\#} \nu + (1 - p) \cdot \nu,$$

where

$$p := \min \left( 1, \frac{\theta^{p}}{\int_{E_{r_{0}}} d^{p}(T^{n}(\zeta), \zeta) \nu(d\zeta)} \right).$$

Then $\mu^{n}$ is a primal feasible solution, and that

$$\int_{\Xi} \Psi d\mu^{n} - \int_{\Xi} \Psi d\nu \geq p \cdot \left( n \int_{E_{r_{0}}} d^{p}(T^{n}(\zeta), \zeta) \nu(d\zeta) + n/2 \right) \geq \min(n \theta^{p}, n/2).$$

Let $n \to \infty$ we conclude that $v_{P} = \infty = v_{D}. \quad \square$
Proof of Proposition 3. Since $-\mathbb{1}\{\xi \in \text{int}(C)\}$ is upper semicontinuous and binary-valued, by Corollary 1 the worst-case distribution of $\min_{\mu \in \mathcal{M}} \mu(\text{int}(C))$ exists. Thus it suffices to show that for any $\epsilon > 0$, there exists $\mu \in \mathcal{M}$ such that $\mu(C) \leq \min_{\mu \in \mathcal{M}} \mu(\text{int}(C)) + \epsilon$. Observe that there exists an optimal transportation plan $\gamma_0$ such that

$$\text{supp} \gamma_0 \subset (\text{supp} \nu \times \text{supp} \nu) \cup ((\text{supp} \nu \cap \text{int}(C)) \times \partial C).$$

Set $\mu_0 := \pi^2 \gamma_0$, then $\mu_0$ is an optimal solution for $\min_{\mu \in \mathcal{M}} \mu(\text{int}(C))$.

If $\mu_0(\partial C) = 0$, there is nothing to show, so we assume that $\mu_0(\partial C) > 0$. We first consider the case $\nu(\text{int}(C)) = 0$ (and thus $\mu_0$ can be chosen to be $\nu$ and the worst-case value is 0). By Lemma 7, we can define a Borel map $T_\epsilon$ which maps each $\xi \in \partial C$ to some $\xi' \in \Xi \setminus \text{cl}(C)$ with $d(\xi, \xi') < \epsilon \in (0, \theta)$ and is an identity mapping elsewhere. We further define a distribution $\mu_\epsilon$ by

$$\mu_\epsilon(A) := \mu_0(A \setminus \partial C) + \mu_0(\{\xi \in \partial C : T_\epsilon(\xi) \in A\}), \quad \text{for all Borel set } A \subset \Xi.$$

Then $W_p(\mu_\epsilon, \mu_0) = W_p(\mu_\epsilon, \nu) \leq \epsilon \leq \theta$ and $\mu_\epsilon(C) = \mu_0(\text{int}(C))$.

Now let us consider $\nu(\text{int}(C)) > 0$. For any $\epsilon \in (0, \theta)$, we define a distribution $\mu'_\epsilon$ by

$$\mu'_\epsilon(A) := \mu_0(A \cap \text{int}(C)) + \frac{\epsilon}{\theta} \left( \gamma_0(\{A \cap \text{int}(C)\} \times \partial C) + \nu(A \cap \partial C) \right) + \left(1 - \frac{\epsilon}{\theta}\right) \mu_0(\{\xi \in \partial C : T_\epsilon(\xi) \in A\}) + \mu_0(A \setminus \text{cl}(C)), \quad \text{for all Borel set } A \subset \Xi.$$

Then

$$\mu'_\epsilon(C) = \mu_0(\text{int}(C)) + \frac{\epsilon}{\theta} [\mu_0(\partial C) - \nu(\partial C) + \nu(\partial C)] + 0 + 0 \leq \mu_0(\text{int}(C)) + \frac{\epsilon}{\theta}.$$

Note that $W_p(\mu_0, \nu) = \int_{\text{int}(C) \times \partial C} d^p(\xi, \zeta) \gamma_0(\{d\xi, d\zeta\})$, it follows that

$$W_p(\mu'_\epsilon, \nu) \leq \left(1 - \frac{\epsilon}{\theta}\right) \int_{\text{int}(C) \times \partial C} d^p(\xi, \zeta) \gamma_0(\{d\xi, d\zeta\}) + \left(1 - \frac{\epsilon}{\theta}\right) \epsilon + 0 \leq \theta.$$

Hence the proof is completed. \hfill \Box

Proof of Proposition 4. Observe that

$$\inf_{\mu \in \mathcal{M}} \mathbb{E}_{\eta \sim \mu}[\eta(x^{-1}(1))] = \inf_{\mu \in \mathcal{P}(\Xi)} \{\mathbb{E}_{\eta \sim \mu}[\eta(x^{-1}(1))] : \min_{\gamma \in \Gamma(\mu, \nu)} \mathbb{E}_{\gamma}[d(\eta, \tilde{\eta})] \leq \theta\} = \inf_{\gamma \in \mathcal{P}(\Xi^2)} \{\mathbb{E}_{(\eta, \tilde{\eta}) \sim \gamma}[\eta(x^{-1}(1))] : \mathbb{E}_{\gamma}[d(\eta, \tilde{\eta})] \leq \theta, \pi^2_\gamma = \nu\}.$$  \hspace{1cm} (38)

For any $\gamma \in \mathcal{P}(\Xi^2)$, denote by $\gamma_{\tilde{\eta}}$ the conditional distribution of $\tilde{\eta} := d(\eta, \tilde{\eta})$ given $\tilde{\eta}$, and by $\gamma_{\tilde{\eta}, \tilde{\theta}}$ the conditional distribution of $\eta$ given $\tilde{\eta}$ and $\tilde{\theta}$. Using tower property of conditional probability, we have that for any $\gamma \in \mathcal{P}(\Xi^2)$ with $\pi^2_\gamma = \nu$,

$$\mathbb{E}_{(\eta, \tilde{\eta}) \sim \gamma}[\eta(x^{-1}(1))] = \mathbb{E}_{\tilde{\eta} \sim \nu} \left[ \mathbb{E}_{\tilde{\theta} \sim \gamma_{\tilde{\eta}}}[\mathbb{E}_{\eta \sim \gamma_{\tilde{\eta}, \tilde{\theta}}}[\eta(x^{-1}(1))]] \right],$$

and

$$\mathbb{E}_{\gamma}[d(\eta, \tilde{\eta})] = \mathbb{E}_{\tilde{\eta} \sim \nu} \left[ \mathbb{E}_{\tilde{\theta} \sim \gamma_{\tilde{\eta}}}[\tilde{\theta}] \right].$$
Observe that the right-hand side of the second equation above does not depend on \( \gamma_{\tilde{\eta}, \tilde{\theta}} \). Thereby (38) can be reformulated as

\[
\inf_{\mu \in \mathfrak{M}} \mathbb{E}_{\eta \sim \mu} [\eta(x^{-1}(1))] = \inf_{\{\gamma_{\tilde{\eta}}, \tau_{\tilde{\eta}, \tilde{\theta}}\} \sim \rho} \left\{ \mathbb{E}_{\tilde{\eta} \sim \gamma_{\tilde{\eta}}} \left[ \mathbb{E}_{\tilde{\eta} \sim \gamma_{\tilde{\eta}}} [\eta(x^{-1}(1))] \right] : \mathbb{E}_{\tilde{\eta} \sim \gamma_{\tilde{\eta}}} [\tilde{\theta}] \leq \theta \right\}
\]

(39)

Indeed, let \( \rho \) be any feasible solution of the right-hand side of (40). We denote by \( \rho_{\tilde{\eta}} \) the conditional distribution of \( \tilde{\theta} := W_1(\tilde{\eta}, \tilde{\theta}) \) given \( \tilde{\eta} \) and by \( \rho_{\tilde{\eta}, \tilde{\theta}} \) the conditional distribution of \( \tilde{\eta} \) given \( \tilde{\eta} \) and \( \tilde{\theta} \). When \( \tilde{\eta} = 0 \) (i.e., no arrival) or \( \tilde{\theta} = 0 \), set \( \tilde{\eta} = 0 \) and \( \tilde{\eta}_{\tilde{\eta}, \tilde{\theta}} = \tilde{\eta} \), that is, we choose \( \tilde{\eta} \) and \( \tilde{\eta}_{\tilde{\eta}, \tilde{\theta}} \) be such that \( \eta = \tilde{\eta} \). When \( \tilde{\eta} \neq 0 \) and \( \tilde{\theta} > 0 \), applying Corollary 1 (Example 7) and Proposition 3 to the problem \( \min_{\tilde{\eta} \in B([0,1])} \{ \tilde{\eta}(x^{-1}(1)) : W_1(\tilde{\eta}, \tilde{\eta}) \leq \tilde{\theta} \} \), we have that for any \( \epsilon > 0 \), there exists an \( \epsilon \)-optimal solution \( \tilde{\eta} \) of the form

\[
\tilde{\eta} = \sum_{i=1}^{\tilde{\eta}((0,1])} \delta_{\tilde{\xi}_i} + 1 - \tilde{\eta}(0,1])\delta_{\tilde{\xi}_0}
\]

where \( 1 \leq i_0 \leq \tilde{\eta}((0,1]) \), \( p_{\tilde{\eta}, \tilde{\theta}} \in [0,1] \), and \( \tilde{\xi}_i \in [0,1] \) for all \( i \neq i_0 \) and \( \tilde{\xi}_i \in [0,1] \). Define

\[
\eta_{\tilde{\eta}, \tilde{\theta}}^\pm := \sum_{i=1}^{\tilde{\eta}((0,1])} \delta_{\tilde{\xi}_i} + \delta_{\tilde{\xi}_i}
\]

It follows that \( \eta_{\tilde{\eta}, \tilde{\theta}}^\pm((0,1]) = \tilde{\eta}((0,1]) \), and

\[
p_{\tilde{\eta}, \tilde{\theta}} \eta_{\tilde{\eta}, \tilde{\theta}}^+(x^{-1}(1)) + (1 - p_{\tilde{\eta}, \tilde{\theta}}) \eta_{\tilde{\eta}, \tilde{\theta}}^-(x^{-1}(1)) \leq \epsilon + \min_{\tilde{\eta} \in B([0,1])} \{ \tilde{\eta}(x^{-1}(1)) : W_1(\tilde{\eta}, \tilde{\eta}) \leq \tilde{\theta} \},
\]

(41)

and

\[
p_{\tilde{\eta}, \tilde{\theta}} W_1(\eta_{\tilde{\eta}, \tilde{\theta}}^+, \tilde{\eta}) + (1 - p_{\tilde{\eta}, \tilde{\theta}}) W_1(\eta_{\tilde{\eta}, \tilde{\theta}}^-, \tilde{\eta}) \leq \tilde{\theta}.
\]

(42)

Define \( \tilde{\gamma}_{\tilde{\eta}} \) and \( \tilde{\gamma}_{\tilde{\eta}, \tilde{\theta}} \) by

\[
\tilde{\gamma}_{\tilde{\eta}}(C) := \int_0^\infty \left[ p_{\tilde{\eta}, \tilde{\theta}} \mathbb{I} \{ W_1(\eta_{\tilde{\eta}, \tilde{\theta}}^+, \tilde{\eta}) \in C \} + (1 - p_{\tilde{\eta}, \tilde{\theta}}) \mathbb{I} \{ W_1(\eta_{\tilde{\eta}, \tilde{\theta}}^-, \tilde{\eta}) \in C \} \right] \rho_{\tilde{\eta}}(d\tilde{\theta}), \quad \forall \text{ Borel set } C \subset [0, \infty),
\]

and

\[
\tilde{\gamma}_{\tilde{\eta}, \tilde{\theta}}(A) := \int_0^\infty \int_\Xi \left[ p_{\tilde{\eta}, \tilde{\theta}} \mathbb{I} \{ \eta_{\tilde{\eta}, \tilde{\theta}}^+ \in A, W_1(\eta_{\tilde{\eta}, \tilde{\theta}}^+, \tilde{\eta}) = \tilde{\theta} \} + (1 - p_{\tilde{\eta}, \tilde{\theta}}) \mathbb{I} \{ \eta_{\tilde{\eta}, \tilde{\theta}}^- \in A, W_1(\eta_{\tilde{\eta}, \tilde{\theta}}^-, \tilde{\eta}) = \tilde{\theta} \} \right] \rho_{\tilde{\eta}, \tilde{\theta}}(d\eta) \rho_{\tilde{\eta}}(d\tilde{\theta}), \quad \forall \text{ Borel set } A \subset \Xi.
\]
Then \( \{\widetilde{\gamma}_1, \{\tilde{\gamma}_i, \tilde{\gamma}_i\}_i \} \) is a feasible solution to the left-hand side of (40). Indeed, by condition (ii), we have \( d(\eta_{\tilde{\gamma}_i}, \eta) = W_1(\eta_{\tilde{\gamma}_i}, \eta) \), hence (42) implies that \( p_{\tilde{\gamma}_i}d(\eta_{\tilde{\gamma}_i}, \eta) + (1 - p_{\tilde{\gamma}_i})d(\eta_{\tilde{\gamma}_i}, \eta) \leq \theta \). Then taking expectation on both sides,

\[
\mathbb{E}_{\eta \sim \nu} \left[ \mathbb{E}_{\tilde{\gamma}_i \sim \eta} \left[ P_{\eta_{\tilde{\gamma}_i}}(\tilde{\eta}) \right] \right] = \int_{\Xi} \int_{\rho \in \mathcal{P}(\mathcal{B}([0, 1]) \times \Xi)} \mathbb{E}_{\eta \sim \nu} \left[ \mathbb{E}_{\tilde{\gamma}_i \sim \eta} \left[ P_{\eta_{\tilde{\gamma}_i}}(\tilde{\eta}) \right] \right] \rho(d\tilde{\eta}) \tilde{\eta}(d\eta) \leq \theta,
\]

hence \( \{\widetilde{\gamma}_1, \{\tilde{\gamma}_i, \tilde{\gamma}_i\}_i \} \) is feasible. Similarly, taking expectation on both sides of (41), we have that

\[
\mathbb{E}_{\eta \sim \nu} \left[ \mathbb{E}_{\tilde{\gamma}_i \sim \eta} \left[ P_{\eta_{\tilde{\gamma}_i}}(\tilde{\eta}) \right] \right] \leq \epsilon + \mathbb{E}_{\nu}[\tilde{\eta}(x^{-1}(1))].
\]

Let \( \epsilon \to 0 \), we obtain that

\[
\inf_{\{\gamma_i \}_i} \left\{ \mathbb{E}_{\tilde{\gamma}_i \sim \eta} \left[ \mathbb{E}_{\eta \sim \gamma_i} \left[ P_{\eta_{\tilde{\gamma}_i}}(\tilde{\eta}) \right] \right] \right\} : \mathbb{E}_{\tilde{\gamma}_i \sim \eta} \left[ \mathbb{E}_{\eta \sim \gamma_i} \left[ P_{\eta_{\tilde{\gamma}_i}}(\tilde{\eta}) \right] \right] \leq \theta \}
\]

\[
\leq \inf_{\rho \in \mathcal{P}(\mathcal{B}([0, 1]) \times \Xi)} \left\{ \mathbb{E}_{\tilde{\gamma}_i \sim \eta} \left[ \mathbb{E}_{\eta \sim \gamma_i} \left[ P_{\eta_{\tilde{\gamma}_i}}(\tilde{\eta}) \right] \right] \right\} : \mathbb{E}_{\tilde{\gamma}_i \sim \eta} \left[ \mathbb{E}_{\eta \sim \gamma_i} \left[ P_{\eta_{\tilde{\gamma}_i}}(\tilde{\eta}) \right] \right] \leq \theta,
\]

To show the opposite direction of the above inequality, observe that \( \inf_{\eta \in \Xi} \left\{ \eta(x^{-1}(1)) : d(\eta, \tilde{\eta}) = \theta \right\} = \inf_{\eta \in \Xi} \left\{ \eta(x^{-1}(1)) : d(\eta, \tilde{\eta}) = \theta \right\} \).

Hence

\[
\inf_{\{\gamma_i \}_i} \left\{ \mathbb{E}_{\tilde{\gamma}_i \sim \eta} \left[ \mathbb{E}_{\eta \sim \gamma_i} \left[ P_{\eta_{\tilde{\gamma}_i}}(\tilde{\eta}) \right] \right] \right\} : \mathbb{E}_{\tilde{\gamma}_i \sim \eta} \left[ \mathbb{E}_{\eta \sim \gamma_i} \left[ P_{\eta_{\tilde{\gamma}_i}}(\tilde{\eta}) \right] \right] \leq \theta \}
\]

\[
\leq \inf_{\rho \in \mathcal{P}(\mathcal{B}([0, 1]) \times \Xi)} \left\{ \mathbb{E}_{\tilde{\gamma}_i \sim \eta} \left[ \mathbb{E}_{\eta \sim \gamma_i} \left[ P_{\eta_{\tilde{\gamma}_i}}(\tilde{\eta}) \right] \right] \right\} : \mathbb{E}_{\tilde{\gamma}_i \sim \eta} \left[ \mathbb{E}_{\eta \sim \gamma_i} \left[ P_{\eta_{\tilde{\gamma}_i}}(\tilde{\eta}) \right] \right] \leq \theta)
\]

Let \( \{\{\gamma_i \}_i \{\eta_{\tilde{\gamma}_i} \tilde{\gamma}_i, \} \} \) be a feasible solution of the right-hand side of (43). Then the joint distribution \( \rho \in \mathcal{P}(\mathcal{B}([0, 1]) \times \Xi) \) defined by

\[
\rho(B) := \int_{\mathcal{B}([0, 1])} \int_{\Theta} 1\{\eta_{\tilde{\gamma}_i} \tilde{\gamma}_i, \in \pi^1(B)\} \gamma_i(d\theta) \nu(d\tilde{\eta}), \quad \forall \text{ Borel set } B \subset \mathcal{B}([0, 1]) \times \Xi
\]

is a feasible solution of the right-hand side of (40). By condition (iii), we have that

\[
\inf_{\eta \in \Xi} \left\{ \eta(x^{-1}(1)) : d(\eta, \tilde{\eta}) = \theta \right\} \geq \inf_{\eta \in \mathcal{B}([0, 1])} \left\{ \tilde{\eta}(x^{-1}(1)) : W_1(\tilde{\eta}, \tilde{\eta}) \leq \theta \right\},
\]

and thus \( \mathbb{E}_{\tilde{\gamma}_i \sim \eta} \left[ \inf_{\eta \in \Xi} \left\{ \eta(x^{-1}(1)) : d(\eta, \tilde{\eta}) = \theta \right\} \right] \geq \mathbb{E}_{\tilde{\gamma}_i \sim \rho} [\tilde{\eta}(x^{-1}(1))]. \) Therefore we prove the opposite direction and (40) holds. Together with (39), we obtain (20). It then follows that it suffices to only consider policy \( x \) such that \( x^{-1}(1) \) is an open set. Then by Corollary 1 (Example 7), the problem \( \min_{\tilde{\eta} \in \mathcal{B}([0, 1])} \left\{ \tilde{\eta}(x^{-1}(1)) : W_1(\tilde{\eta}, \tilde{\eta}) \leq \theta \right\} \) admits a worst-case distribution \( \eta_{\tilde{\gamma}_i} \tilde{\gamma}_i \) and let \( \lambda_{\tilde{\gamma}_i} \tilde{\gamma}_i \) be the associated dual optimizer. Let \( \tilde{\Xi} := \{\xi \}_{t=1, \ldots, M} \). We claim that it suffices to further restrict attention to those policies \( x \) such that each connected component of \( x^{-1}(1) \) contains at least one point in \( \tilde{\Xi} \). Indeed, suppose there exists a connected component \( C_0 \) of \( x^{-1}(1) \) such that \( C_0 \cap \tilde{\Xi} \phi \). Then for every \( \zeta \in \text{ supp } \tilde{\eta}, \arg \min_{\zeta \in \Xi} \left\{ \text{L}(x^{-1}(1)) + [\xi - \zeta] \right\} \notin C_0, \) and thus \( \eta_{\tilde{\gamma}_i} \tilde{\gamma}_i(x^{-1}(1)) = \eta_{\tilde{\gamma}_i} \tilde{\gamma}_i(x^{-1}(1) \setminus C_0). \) Hence, \( x' := \text{L}(x^{-1}(1) \setminus C_0) \) achieves a higher objective value \( v(x') \) than \( v(x) \) and so \( x \) cannot be optimal. We finally conclude that there exists \( \{E_x, \mathcal{X}_j\}_{j=1}^M \), where \( M \leq \text{card}(\tilde{\Xi}) \), such that (21) holds.

\[ \square \]
Proof of Proposition 5. Using Corollary 2 and Proposition 4, we have that
\[
v\left(\sum_{j=1}^{M} I_{[\xi_j, \bar{\xi}_j]} \right) = \min_{\xi^i_m \in [0,1], 0 \leq p^i \leq 1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} \left[ -c(\bar{\xi}_j - \xi_j) + p^i \eta^i \{[\xi_j, \bar{\xi}_j] \} + (1 - p^i) \bar{\eta}^i \{[\xi_j, \bar{\xi}_j] \} \right] : \frac{1}{N} \sum_{i=1}^{N} \left[ p^i W_1(\eta^i, \bar{\eta}^i) + (1 - p^i) W_1(\eta^i, \bar{\eta}^i) \right] \leq \theta \right\}.
\]
By the equivalent definition of one-dimensional Wasserstein distance [43], for \( \eta^i = \sum_{j=1}^{M_i} \delta_{\hat{\xi}_j} \), we have that \( W_1(\eta^i, \bar{\eta}^i) = \min_{\xi^i_m \in [0,1]} \sum_{m=1}^{M_i} |\xi^i_m - \hat{\xi}_j\| \), where the minimum is taken over all \( M_i \)-permutations. Hence
\[
v\left(\sum_{j=1}^{M} I_{[\xi_j, \bar{\xi}_j]} \right) = \min_{\xi^i_m \in [0,1], 0 \leq p^i \leq 1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[ p^i \sum_{m=1}^{M_i} \sum_{j=1}^{M} I_{[\xi_j, \bar{\xi}_j]}(\xi^i_m) + (1 - p^i) \sum_{m=1}^{M_i} \sum_{j=1}^{M} I_{[\xi_j, \bar{\xi}_j]}(\hat{\xi}_j) \right] : \frac{1}{N} \sum_{i=1}^{N} \left[ p^i \sum_{m=1}^{M_i} \xi^i_m - \hat{\xi}_j\| + (1 - p^i) \sum_{m=1}^{M_i} \hat{\xi}_j\| \right] \leq \theta \right\}.
\]
Using Example 6, we have that
\[
v\left(\sum_{j=1}^{M} I_{[\xi_j, \bar{\xi}_j]} \right) = \min \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[ p^i \sum_{m=1}^{M_i} \sum_{j=1}^{M} I_{[\xi_j, \bar{\xi}_j]}(\hat{\xi}_j) - (p^i m_j + p m_j) \right] + (1 - p^i) \sum_{m=1}^{M_i} \sum_{j=1}^{M} I_{[\xi_j, \bar{\xi}_j]}(\xi^i_m) - (p^i m_j + p m_j) \right] : \frac{1}{N} \sum_{i=1}^{N} \left[ p^i m_j + p m_j \right] \leq 1, \sum_{i=1}^{M} \left[ p^{i} m_j + p m_j \right] \leq 1, \forall i, t \},
\]
where the minimum is taken over all \( p^i, p^i m_j, p m_j, p^i m_j, p m_j \in [0,1] \). Replacing \( p^i m_j + (1 - p^i) p^i m_j \) by \( p m_j \), and \( p^i m_j + (1 - p^i) p^i m_j \) by \( p m_j \), and noticing that at optimality, \( p m_j, p^i m_j > 0 \) only if \( \hat{\xi}_j \in [x_{j-}, \bar{\xi}_j] \), and at most one of \( \{p m_j, p^i m_j\}_{i,t,j} \) can be fractional, we obtain the result.

Proof of Proposition 6. The dual optimizer of the inner maximization problem of (22) is zero when \( \theta \) is sufficiently large, whence the worst-case value of the inner maximization problem equals \( \sup_{\xi \in [0,1]} \ln(a(\xi)) \). Then the overall objective function equals
\[
\int_{0}^{T} a(t) dt + \sup_{\xi \in [0,1]} - \ln(a(\xi)).
\]
Set \( b = \int_{0}^{T} a(t) dt \). Then due to the second term \( \sup_{\xi \in [0,1]} - \ln(a(\xi)) \), the solution \( a(t) = b/T \) yields an objective value no larger than \( a(t) \). Hence we complete the proof.

Proof of Proposition 7. Define \( C_w := \{\xi : -w^T \xi < q\} \) for all \( w \). Similar to Example 7, there exists a worst-case distribution \( \mu^* \) which attains the infimum \( \inf_{\mu \in \mathcal{M}} \mathbb{P}_{\mu}\{ -w^T \xi < q \} \) and there exists maps
such that for each \( \zeta \in \text{supp } \nu \), it holds that \( T^*(\zeta), \bar{T}^*(\zeta) \in \{ \zeta \} \cup \arg \min_{\xi \in \mathbb{C}_w} \| \xi - \zeta \|^p_{\infty} \). With this in mind, let \( \gamma^* \) be the optimal transport plan between \( \nu \) and \( \mu^* \), and let

\[
t^* := \nu \text{-ess sup}_{\zeta \in \Xi} \left\{ \min_{\xi \in \Xi \setminus \mathbb{C}_w} \| \xi - \zeta \|^p_{\infty} : \zeta \neq T^*(\zeta) \right\}.
\]

So \( t^* \) is the longest distance of transportation among all the points that are transported. (We note that infinity is allowed in the definition of \( t^* \), however, as will be shown, this violates the probability bound.) Then \( \mu^* \) transports all the points in \( \text{supp } \nu \cap \{ \xi : q - t^* < -w^T\xi < b \} \), and possibly a fraction of mass \( \beta^* \in [0,1] \) in \( \text{supp } \nu \cap \{ \xi : -w^T\xi = q - t^* \} \). Also note that by Hölder’s inequality, the distance between two hyperplanes \( \{ \xi : -w^T\xi = s \} \), and \( \{ \xi : -w^T\xi = s' \} \) equals to \( |s - s'|/\|w\|_1 = |s - s'| \). Using this characterization, let us define a probability measure \( \nu_w \) on \( \mathbb{R} \) by

\[
\nu_w \{ (-\infty, s) \} := \nu \{ \xi : -w^T\xi < s \}, \quad \forall s \in \mathbb{R},
\]

then using the changing of measure, the total distance of transportation can be computed by

\[
\int_{(\Xi \setminus \mathbb{C}_w) \times \mathbb{C}_w} d\nu \gamma^*(d\xi, d\zeta) = \int_{(q - t^*)^+} (q - s)^p \nu_w(ds) + \beta^* \nu_w(\{ q - t^* \}) t^p \leq \theta^p.
\]

On the other hand, using property of marginal expectation and the characterization of \( \gamma^* \),

\[
\mu^*(\mathbb{C}_w) = \int_{\mathbb{C}_w \times \Xi} \gamma^*(d\xi, d\zeta)
\]

\[
= \nu(\mathbb{C}_w) - \int_{(\Xi \setminus \mathbb{C}_w) \times \mathbb{C}_w} \gamma^*(d\xi, d\zeta)
\]

\[
= 1 - \nu_w([q, \infty)) - \beta^* \nu_w(\{ q - t^* \}) + \nu_w(\{ q - t^*, q \})
\]

\[
= 1 - \nu_w(q - t^*, \infty) - \beta^* \nu_w(\{ q - t^* \}).
\]

Thereby the condition \( \inf_{\mu \in \mathcal{M}} \mu(\mathbb{C}_w) \geq 1 - \alpha \) is equivalent to

\[
\beta^* \nu_w(\{ q - t^* \}) + \nu_w(q - t^*, \infty) \leq \alpha.
\]

Now consider the quantity

\[
J := \int_{(\text{VaR}_{\alpha}[-w^T\xi])^+} (q - s)^p \nu_w(ds) + \beta_0 \nu_w(\{ \text{VaR}_{\alpha}[-w^T\xi] \}) \nu(\mathbb{C}_w) \leq \theta^p.
\]

If \( J < 0 \), due to the monotonicity in \( t^* \) of the right-hand side of (46), either \( q - t^* \) \( \leq \text{VaR}_{\alpha}[-w^T\xi] \) or \( q - t^* \) \( \geq \text{VaR}_{\alpha}[-w^T\xi] \) and \( \beta^* \) \( > \beta_0 \). But in both cases (47) is violated. On the other hand if \( J \geq 0 \), again by monotonicity, either \( q - t^* \) \( > \text{VaR}_{\alpha}[-w^T\xi] \), or \( q - t^* \) \( > \text{VaR}_{\alpha}[-w^T\xi] \) and \( \beta^* \leq \beta_0 \) and thus (47) is satisfied. Therefore we complete the proof. \( \square \)

**Appendix C: Selecting Radius \( \theta \)** We mainly use a classical result on Wasserstein distance from Bolley et al. [10]. Let \( \nu_N \) be the empirical distribution of \( \xi \) obtained from the underlying distribution \( \nu_0 \). In Theorem 1.1 (see also Remark 1.4) of Bolley et al. [10], it is shown that

\[
P\{ W_1(\nu_N, \nu_0) > \theta \} \leq C(\theta) e^{-\frac{1}{2}N/\theta^2}
\]

for some constant \( \lambda \) dependent on \( \nu_0 \), and \( C \) dependent on \( \theta \). Since their result holds for general distributions, we here simplify it for our purpose and explicitly compute the constants \( \lambda \) and \( C \). For a more detailed analysis, we refer the reader to Section 2.1 in Bolley et al. [10].
Noticing that by assumption \( \text{supp } \nu_0 \subset [0, \bar{B}] \), the truncation step in Bolley et al. [10] is no longer needed, thus the probability bound (2.12) (see also (2.15)) of Bolley et al. [10] is reduced to
\[
P\{W_1(\nu_N, \nu_0) > \theta\} \leq \max \left( 8e^{\frac{\bar{B}}{\delta}}, 1 \right) \frac{N(\frac{\delta}{2})}{e^{-\frac{\lambda}{2} N(\theta-\delta)^2}}
\]
for some constant \( \lambda > 0 \), \( \delta \in (0, \theta) \), where \( e \) is the natural logarithm, and \( N(\frac{\delta}{2}) \) is the minimal number of balls need to cover the support of \( \xi \) by balls of radius \( \delta/2 \) and in our case, \( N(\frac{\delta}{2}) = \bar{B}/\delta \). Now let us compute \( \lambda \). By Theorem 1.1 of Bolley et al. [10], \( \lambda \) is the constant appeared in the Talagrand inequality
\[
W_1(\mu, \nu_0) \leq \sqrt{\frac{2}{\lambda} I_{\phi_{kl}}(\mu, \nu_0)},
\]
where the Kullback-Leibler divergence of \( \mu \) with respect to \( \nu \) is defined by \( I_{\phi_{kl}}(\mu, \nu_0) = +\infty \) if \( \mu \) is not absolutely continuous with respect to \( \nu_0 \), otherwise \( I_{\phi_{kl}}(\mu, \nu_0) = \int f \log f \, d\nu_0 \), where \( f \) is the Radon-Nikodym derivative \( d\mu/d\nu_0 \). Corollary 4 in Bolley and Villani [11] shows that \( \lambda \) can be chosen as
\[
\lambda = \left[ \inf_{c^0 \in \Xi, \alpha > 0} \left( \frac{1}{1 + \log \int e^{\alpha d^2(\xi, \cdot)} \nu(d\xi)} \right) \right]^{-1},
\]
which can be estimated from data. Finally, we obtain a concentration inequality
\[
P\{W_1(\nu_N, \nu_0) > \theta\} \leq \max \left( 8e^{\frac{\bar{B}}{\delta}}, 1 \right) \frac{\bar{B}}{\delta} e^{-\frac{\lambda}{2} N(\theta-\delta)^2}.
\] (48)

In the numerical experiment, we choose \( \delta \) to make the right-hand side of (48) as small as possible, and \( \theta \) is chosen such that the right-hand side of (48) is equal to 0.05.

**Acknowledgments.** The authors would like to thank Wilfrid Gangbo, David Goldberg, Alex Shapiro, and Weijun Xie for several stimulating discussions, and Yuan Li for providing the image (Figure (1b)).

**References**

[1] Aliprantis CD, Border K (2006) *Infinite Dimensional Analysis: A Hitchhiker’s Guide* (Springer Science & Business Media).

[2] Ambrosio L, Fusco N, Pallara D (2000) *Functions of bounded variation and free discontinuity problems*, volume 254 (Oxford: Clarendon Press).

[3] Aubin JP, Frankowska H (2009) *Set-valued analysis* (Springer Science & Business Media).

[4] Barbour AD, Brown TC (1992) Stein’s method and point process approximation. *Stochastic Processes and their Applications* 43(1):9–31.

[5] Bayraksan G, Love DK (2015) Data-driven stochastic programming using phi-divergences. *Tutorials in Operations Research*.

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

[7] Ben-Tal A, Goryashko A, Guslitzer E, Nemirovski A (2004) Adjustable robust solutions of uncertain linear programs. *Mathematical Programming* 99(2):351–376.

[8] Berger JO (1984) The robust bayesian viewpoint. *Studies in Bayesian Econometrics and Statistics : In Honor of Leonard J. Savage., Edited by Joseph B. Kadane* 4(2):63–124.

[9] Blanchet J, Murthy KR (2016) Quantifying distributional model risk via optimal transport. *arXiv preprint arXiv:1604.01446* .

[10] Bolley F, Guillin A, Villani C (2007) Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields* 137(3-4):541–593.
[11] Bolley F, Villani C (2005) Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. Annales de la Faculté des sciences de Toulouse: Mathématiques, volume 14, 331–352.
[12] Boyd S, Vandenberghe L (2004) Convex optimization (Cambridge university press).
[13] Calafiore GC, El Ghaoui L (2006) On distributionally robust chance-constrained linear programs. Journal of Optimization Theory and Applications 130(1):1–22.
[14] Chen LH, Xia A (2004) Stein’s method, palm theory and poisson process approximation. Annales of Probability 2545–2569.
[15] Daley DJ, Vere-Jones D (2003) An Introduction to the Theory of Point Processes Volume I: Elementary Theory and Methods (Springer), second edition, ISBN 0-387-95541-0.
[16] Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations Research 58(3):595–612.
[17] Dupačová J (1987) The minimax approach to stochastic programming and an illustrative application. Stochastics: An International Journal of Probability and Stochastic Processes 20(1):73–88.
[18] El Ghaoui L, Oks M, Oustry F (2003) Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. Operations Research 51(4):543–556.
[19] Erdoğan E, Iyengar G (2006) Ambiguous chance-constrained problems and robust optimization. Mathematical Programming 107(1-2):37–61.
[20] Esfahani PM, Kuhn D (2015) Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. arXiv preprint arXiv:1505.05116.
[21] Gallego G, Moon I (1993) The distribution free newsboy problem: review and extensions. Journal of the Operational Research Society 825–834.
[22] Gibbs AL, Su FE (2002) On choosing and bounding probability metrics. International statistical review 70(3):419–435.
[23] Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. Operations Research 58(4-part-1):902–917.
[24] Gonzalez RC, Woods RE (2006) Digital Image Processing (3rd Edition) (Prentice-Hall, Inc.).
[25] Jiang R, Guan Y (2015) Data-driven chance constrained stochastic program. Mathematical Programming 1–37.
[26] Kantorovich LV (1942) On the translocation of masses. Dokl. Akad. Nauk SSSR, volume 37, 199–201.
[27] Kantorovich LV (1960) Mathematical methods of organizing and planning production. Management Science 6(4):366–422.
[28] Ling H, Okada K (2007) An efficient earth mover’s distance algorithm for robust histogram comparison. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29(5):840–853.
[29] Nemirovski A (2004) Prox-method with rate of convergence O(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM Journal on Optimization 15(1):229–251.
[30] Nesterov Y, Nemirovsky A (2013) On first-order algorithms for l 1/nuclear norm minimization. Acta Numerica 22:509–575.
[31] Owhadi H, Scovel C (2015) Extreme points of a ball about a measure with finite support. arXiv preprint arXiv:1504.06745.
[32] Pardo L (2005) Statistical inference based on divergence measures (CRC Press).
[33] Parikh N, Boyd SP (2014) Proximal algorithms. Foundations and Trends in Optimization 1(3):127–239.
[34] Pflug GC, Pichler A (2014) Multistage stochastic optimization (Springer).
[35] Popescu I (2007) Robust mean-covariance solutions for stochastic optimization. Operations Research 55(1):98–112.
[36] Rockafellar RT, Wets RJB (2009) Variational analysis, volume 317 (Springer Science & Business Media).
[37] Rubner Y, Tomasi C, Guibas LJ (2000) The earth mover’s distance as a metric for image retrieval. International Journal of Computer Vision 40(2):99–121.
[38] Scarf H, Arrow K, Karlin S (1958) A min-max solution of an inventory problem. Studies in the Mathematical Theory of Inventory and Production 10:201–209.
[39] Shapiro A (2001) On duality theory of conic linear problems. *Semi-infinite programming*, 135–165 (Springer).

[40] Shapiro A, Dentcheva D, et al. (2014) *Lectures on stochastic programming: modeling and theory*, volume 16 (SIAM).

[41] Shapiro A, Kleywegt A (2002) Minimax analysis of stochastic problems. *Optimization Methods and Software* 17(3):523–542.

[42] Sun H, Xu H (2015) Convergence analysis for distributionally robust optimization and equilibrium problems. *Mathematics of Operations Research*.

[43] Vallender S (1974) Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications* 18(4):784–786.

[44] Villani C (2003) *Topics in optimal transportation*. Number 58 (American Mathematical Soc.).

[45] Villani C (2008) *Optimal transport: old and new*, volume 338 (Springer Science & Business Media).

[46] Wang Z, Glynn PW, Ye Y (2015) Likelihood robust optimization for data-driven problems. *Computational Management Science* 1–21.

[47] Wozabal D (2012) A framework for optimization under ambiguity. *Annals of Operations Research* 193(1):21–47.

[48] Wozabal D (2014) Robustifying convex risk measures for linear portfolios: A nonparametric approach. *Operations Research* 62(6):1302–1315.

[49] Žáčková J (1966) On minimax solutions of stochastic linear programming problems. *Časopis pro pěstování matematiky* 91(4):423–430.

[50] Zhao C, Guan Y (2015) Data-driven risk-averse stochastic optimization with Wasserstein metric. Available on Optimization Online.

[51] Zymler S, Kuhn D, Rustem B (2013) Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming* 137(1-2):167–198.