On Linear Optimization over Wasserstein Balls

Man-Chung Yue[1], Daniel Kuhn[2], and Wolfram Wiesemann[3]

[1]*Department of Applied Mathematics, The Hong Kong Polytechnic University,*
`manchung.yue@polyu.edu.hk`
[2]*College of Management of Technology, École Polytechnique Fédérale de Lausanne,*
`daniel.kuhn@epfl.ch`
[3]*Imperial College Business School, Imperial College London,* `ww@imperial.ac.uk`

April 15, 2020

Abstract

Wasserstein balls, which contain all probability measures within a pre-specified Wasserstein distance to a reference measure, have recently enjoyed wide popularity in the distributionally robust optimization and machine learning communities to formulate and solve data-driven optimization problems with rigorous statistical guarantees. In this technical note we prove that the Wasserstein ball is weakly compact under mild conditions, and we offer necessary and sufficient conditions for the existence of optimal solutions. We also characterize the sparsity of solutions if the Wasserstein ball is centred at a discrete reference measure. In comparison with the existing literature, which has proved similar results under different conditions, our proofs are self-contained and shorter, yet mathematically rigorous, and our necessary and sufficient conditions for the existence of optimal solutions are easily verifiable in practice.

1   Introduction

Let (X, d) be a Polish (*i.e.*, a complete and separable) metric space. We assume that (X, d) is proper, that is, for any $R > 0$ and $x_0 \in X$, the closed ball $B_R(x_0) := \{x \in X : d(x, x_0) \leq R\}$ is compact. Examples of proper metric spaces include finite-dimensional Banach spaces and complete Riemannian manifolds. We denote by $\mathcal{P}(X)$ the set of all Borel probability measures that are supported on X.

For $p \in [1, \infty)$, the p-th Wasserstein distance between $\mu_1, \mu_2 \in \mathcal{P}(X)$ is defined as

$$W_p(\mu_1, \mu_2) = \inf_{\gamma \in \Gamma(\mu_1, \mu_2)} \left(\int_{X \times X} d^p(x_1, x_2) \, \mathrm{d}\gamma(x_1, x_2) \right)^{\frac{1}{p}}, \tag{1}$$

where $\Gamma(\mu_1, \mu_2)$ is the set of all couplings of μ_1 and μ_2, that is, the set of all probability measures supported on $X \times X$ with marginals μ_1 and μ_2. Intuitively speaking, $W_p(\mu_1, \mu_2)$ measures the minimum transportation cost required to transform the mass of μ_1 into the mass of μ_2, where the transportation cost is measured according to the ground metric d.

The Wasserstein ball of radius $r > 0$ centred at the reference measure $\nu \in \mathcal{P}(X)$ is

$$\mathcal{B}_r(\nu) = \{\mu \in \mathcal{P}(X) : W_p(\mu, \nu) \leq r\}. \tag{2}$$

In this paper, we study the linear optimization problem

$$\begin{aligned} \underset{\mu}{\text{maximize}} \quad & \int_X f(x) \, \mathrm{d}\mu(x) \\ \text{subject to} \quad & \mu \in \mathcal{B}_r(\nu), \end{aligned} \tag{3}$$

where $f : X \to \mathbb{R}$ is assumed to be upper semi-continuous, and where for each $\mu \in \mathcal{B}_r(\nu)$ at least one of the integrals $\int_X [f(x)]_+ \, \mathrm{d}\mu(x)$ and $\int_X [-f(x)]_+ \, \mathrm{d}\mu(x)$ is finite. Optimization problems of the form (3) arise in distributionally robust optimization, where ambiguity averse decisions are sought that perform well under misestimations of the unknown true data generating distribution [4, 7, 10, 12, 17]. Problem (3) also emerges when one regularizes machine learning problems such as classification problems [6], clustering problems [8] or generative adversarial networks [1] against overfitting to the training samples.

The remainder of this technical note proceeds as follows. We first prove in Section 2 that the Wasserstein ball $\mathcal{B}_r(\nu)$ is weakly compact under mild conditions. We then leverage this finding in Section 3 to derive a necessary and sufficient condition for the optimal value of the optimization problem (3) to be finite. Section 4 shows that essentially the same condition is also sufficient for the optimal value of (3) to be attained. Section 5 is devoted to sparse solutions to (3) that place positive probability on finitely many atoms. To keep this note self-contained, we review basic results from measure theory and from inifinite-dimensional linear programming in the Appendices A and B, respectively.

2 Weak Compactness of the Wasserstein Ball

We say that a probability measure μ has a finite p-th moment if $\int_X d^p(x, x_0) \, \mathrm{d}\mu(x) < \infty$ for some $x_0 \in X$. The triangle inequality for d implies that the integral is finite for some $x_0 \in X$ if and only if it is finite for all $x_0 \in X$, and thus the reference point x_0 does not matter. In contrast to the literature, which commonly defines the Wasserstein distance for measures with finite p-th moment only, our definition of the Wasserstein distance may take an infinite value for measures that do not possess a finite p-th moment. We now show, however, that the Wasserstein ball $\mathcal{B}_r(\nu)$ only contains measures with a finite p-th moment whenever the reference distribution ν has a finite p-th moment.

Lemma 1. *The Wasserstein ball $\mathcal{B}_r(\nu)$ has a uniformly bounded p-th moment, that is, there exists $C > 0$ such that $\int_X d^p(x, x_0) \, \mathrm{d}\mu(x) \leq C$ for all $\mu \in \mathcal{B}_r(\nu)$, whenever ν has a finite p-th moment.*

Proof of Lemma 1. Since the reference measure ν has a finite p-th moment, there is $x_0 \in X$ and $C_0 \in \mathbb{R}$ such that $\int_X d^p(x, x_0) \, \mathrm{d}\nu(x) \leq C_0$. We claim that $\int_X d^p(x, x_0) \, \mathrm{d}\mu(x) \leq C_0 + r$

for all measures $\mu \in \mathcal{B}_r(\nu)$ in the Wasserstein ball. To see this, note that

$$\left(\int_X d^p(x, x_0) \, \mathrm{d}\mu(x) \right)^{\frac{1}{p}} = W_p(\mu, \delta_{x_0}) \leq W_p(\nu, \delta_{x_0}) + W_p(\mu, \nu)$$

$$= \left(\int_X d^p(x, x_0) \, \mathrm{d}\nu(x) \right)^{\frac{1}{p}} + W_p(\mu, \nu) \leq C_0 + r,$$

where δ_{x_0} is the Dirac measure at x_0, and where the first inequality follows from the triangle inequality of the Wasserstein distance on $\mathcal{P}(X)$ [5, Corollary 5.7]. $\qquad\square$

We are now ready to prove the weak sequential compactness of $\mathcal{B}_r(\nu)$.

Theorem 1 (Weak Compactness of Wasserstein Ball)**.** *The Wasserstein ball $\mathcal{B}_r(\nu)$ is weakly sequentially compact whenever the reference measure ν has a finite p-th moment.*

Proof of Theorem 1. The map $\mu \mapsto W_p(\mu, \nu)$ is lower semi-continuous with respect to the weak convergence [5, Corollary 5.3]. Since the Wasserstein ball $\mathcal{B}_r(\nu)$ is a sublevel set of this map, $\mathcal{B}_r(\nu)$ is weakly closed, that is, $\mathcal{B}_r(\nu)$ coincides with its weak closure. To see that $\mathcal{B}_r(\nu)$ is weakly compact, we show that $\mathcal{B}_r(\nu)$ is tight. We can then employ Prokhorov's Theorem (*cf.* Theorem 5 from Appendix A) to conclude that the closure of the Wasserstein ball $\mathcal{B}_r(\nu)$, which by our previous argument coincides with $\mathcal{B}_r(\nu)$, is weakly compact.

By Lemma 1, there is $C > 0$ such that $\left[\int_X d^p(x, x_0) \, \mathrm{d}\mu(x) \right]^{\frac{1}{p}} \leq C$ for all $\mu \in \mathcal{B}_r(\nu)$. To see that the Wasserstein ball $\mathcal{B}_r(\nu)$ is tight, we show that for every $\epsilon > 0$ and $\mu \in \mathcal{B}_r(\nu)$, we have $\mu \left(X \setminus B_{C/\epsilon}(x_0) \right) \leq \epsilon$ (*cf.* Definition 1 from Appendix A). Indeed, we have that

$$\mu \left(X \setminus B_{C/\epsilon}(x_0) \right) = \mu(\{x \in X \, : \, d(x, x_0) > C/\epsilon\}) \leq \frac{\int_X d(x, x_0) \, \mathrm{d}\mu(x)}{C/\epsilon}$$

$$\leq \frac{\epsilon \cdot \left[\int_X d^p(x, x_0) \, \mathrm{d}\mu(x) \right]^{\frac{1}{p}}}{C} \leq \epsilon,$$

where the first inequality is due to Markov's inequality, the second inequality is due to Jensen's inequality, and the third inequality holds by our definition of C. Note that the set $B_{C/\epsilon}(x_0)$ is compact because of the properness of (X, d). $\qquad\square$

While the result of Theorem 1 is certainly not surprising, we are not aware of any formal proof prior to [13]. Proposition 3 of that paper considers the special case of $X = \mathbb{R}^n$ and proves weak compactness using a different argument. Our result is complementary in that it covers applications where the underlying space X is not $\mathbb{R}^n$, such as the space of point processes on a compact interval, which arises in the study of on/off system control and Poisson process intensity estimation, see [7, Section 4].

3 Finiteness of the Optimal Value of Problem (3)

We provide a necessary and sufficient condition for the finiteness of the optimal value of the optimization problem (3).

Theorem 2. *Assume that ν has a finite p-th moment. Then the optimal value of problem* (3) *is finite if and only if there exist $x_0 \in X$ and $c > 0$ such that $f(x) \leq c[1 + d^p(x, x_0)]$ for all $x \in X$.*

Proof. Recall that the Wasserstein radius r is strictly positive. By [15, Theorem 6.18], there thus exists a discrete probability measure $\hat{\nu} = \sum_{i=1}^{N} \alpha_i \cdot \delta_{y_i}$, $\alpha \in \mathbb{R}_+^N$ with $\sum_{i=1}^{N} \alpha_i = 1$, that is supported on $2 \leq N < \infty$ atoms $y_1, \ldots, y_N \in X$ and that satisfies $W_p(\hat{\nu}, \nu) \leq r/2$, that is, $\hat{\nu}$ resides in the vicinity of ν. We fix this measure for the remainder of the proof.

Note also that f is real-valued and the integral of f under $\hat{\nu}$ reduces to a finite sum of real values. Since $\hat{\nu}$ is feasible in (3), the optimal value of (3) thus cannot be $-\infty$. We next prove that the boundedness of (3) implies the stated growth condition on f.

Assume for the sake of contradiction that problem (3) is bounded, but for all $x_0 \in X$ and $c > 0$ there is $x \in X$ for which $f(x) > c[1 + d^p(x, x_0)]$, in violation of the statement of the theorem. We can then construct a sequence $\{y^k\}_k \subseteq X$ such that

$$\lim_{k \to \infty} \frac{f(y^k)}{1 + d^p(y^k, y_1)} = +\infty.$$

Note that in this expression, we have chosen $x_0 = y_1$, where y_1 is the first atom of $\hat{\nu}$. In the following, we use the sequence $\{y^k\}_k$ to construct a sequence of probability measures $\{\mu^k\}_k$ from within the Wasserstein ball $\mathcal{B}_r(\nu)$ for which the integrals $\int_X f(x) \, \mathrm{d}\mu^k(x)$ diverge, in contradiction to the assumption that problem (3) is bounded.

Using the sequence $\{y^k\}_k$, we construct the sequence of probability measures

$$\mu^k = \epsilon^k \cdot \delta_{y^k} + (\alpha_1 - \epsilon^k) \cdot \delta_{y_1} + \sum_{i=2}^{N} \alpha_i \cdot \delta_{y_i} \quad \text{with} \quad \epsilon^k = \frac{\min\{\alpha_1, r^p/2^p\}}{1 + d^p(y^k, y_1)}.$$

Note that each μ^k is indeed a probability measure since $\epsilon^k \in [0, \alpha_1]$ and $\epsilon^k + (\alpha_1 - \epsilon^k) + \alpha_2 + \ldots + \alpha_N = \alpha_1 + \ldots + \alpha_N = 1$. To see that $\mu^k \in \mathcal{B}_r(\nu)$ for all k, consider the transportation plan $\gamma^k \in \mathcal{P}(X \times X)$ given by

$$\gamma^k = \epsilon^k \cdot \delta_{(y_1, y^k)} + (\alpha_1 - \epsilon^k) \cdot \delta_{(y_1, y_1)} + \sum_{i=2}^{N} \alpha_i \cdot \delta_{(y_i, y_i)}.$$

One readily verifies that $\gamma^k \in \Gamma(\hat{\nu}, \mu^k)$, as well as

$$
\begin{aligned}
W_p(\nu, \mu^k) &\leq W_p(\nu, \hat{\nu}) + W(\mu^k, \hat{\nu}) \leq \frac{r}{2} + \left[\int_{X \times X} d^p(x_1, x_2) \, \mathrm{d}\gamma^k(x_1, x_2) \right]^{\frac{1}{p}} \\
&= \frac{r}{2} + \left[\epsilon^k \cdot d^p(y_1, y^k) + (\alpha_1 - \epsilon^k) \cdot d^p(y_1, y_1) + \sum_{i=2}^{N} \alpha_i \cdot d^p(y_i, y_i) \right]^{\frac{1}{p}} \leq r.
\end{aligned}
$$

Here, the first inequality follows from the triangle inequality of the Wasserstein distance on $\mathcal{P}(X)$ [5, Corollary 5.7], the second inequality holds by construction of $\hat{\nu}$ and because $\gamma^k \in \Gamma(\hat{\nu}, \mu^k)$, and the last inequality follows from the fact that $\epsilon^k \cdot d^p(y_1, y^k) \leq r^p/2^p$.

4

We thus conclude that $\{\mu^k\}_k \subseteq \mathcal{B}_r(\nu)$ as desired. To see that the integrals $\int_X f(x)\,\mathrm{d}\mu^k(x)$ diverge, we note that

$$\lim_{k\to\infty} \int_X f(x)\,\mathrm{d}\mu^k(x) = \lim_{k\to\infty}\left[\frac{\min\{\alpha_1, r^p/2^p\}}{1 + d^p(y^k, y_1)}\cdot f(y^k) + (\alpha_1 - \epsilon^k)\cdot f(y_1) + \sum_{i=2}^{N}\alpha_i \cdot f(y_i)\right].$$

The second expression diverges since $f(y^k)/[1 + d^p(y^k, y_1)] \longrightarrow \infty$ while $\min\{\alpha_1, r^p/2^p\}$ is constant in k. We have thus proved that the boundedness of problem (3) implies the stated growth condition on f.

To show that the stated growth condition on f also implies the boundedness of problem (3), fix any $\mu \in \mathcal{B}_r(\nu)$ and note that

$$\begin{aligned}
\int_X f(x)\,\mathrm{d}\mu(x) &\leq \int_X c\left[1 + d^p(x, x_0)\right]\mathrm{d}\mu(x) \\
&= c + c\cdot W_p^p(\mu, \delta_{x_0}) \\
&\leq c + c\left[W_p(\mu, \nu) + W_p(\nu, \delta_{x_0})\right]^p,
\end{aligned}$$

where the inequalities are due to the assumed growth condition on f and the triangle inequality of W_p, respectively. Since $W_p(\mu, \nu) \leq r$ as $\mu \in \mathcal{B}_r(\nu)$, we can thus uniformly bound the objective value of every measure $\mu \in \mathcal{B}_r(\nu)$ in problem (3). □

4 Existence of Optimal Solutions

We now study the existence of optimal solutions to the optimization problem (3).

Theorem 3. *Assume that ν has a finite p-th moment. If there exist $x_0 \in X$, $c > 0$ and $p' \in (0, p)$ such that $f(x) \leq c[1 + d^{p'}(x, x_0)]$ for all $x \in X$, then the optimal value of (3) is attained.*

Proof. By Theorem 1 and Weierstrass' theorem, it suffices to show that the map $\mu \mapsto \int_X f(x)\,\mathrm{d}\mu(x)$ is weakly upper semi-continuous on $\mathcal{B}_r(\nu)$. To this end, let $\{\mu^k\}_k \subseteq \mathcal{B}_r(\nu)$ be a sequence converging weakly to μ^∞. In order to show that $\limsup_{k\to\infty} \int_X f(x)\,\mathrm{d}\mu^k(x) \leq \int_X f(x)\,\mathrm{d}\mu^\infty(x)$, we may may assume w.l.o.g. that $\int_X f(x)\,\mathrm{d}\mu^k(x) > -\infty$ for all k. Since the assumptions of Theorem 2 are satisfied, the optimal value of problem (3) is finite, and hence $\int_X f(x)\,\mathrm{d}\mu^k(x)\,\mathrm{d}\mu^k(x) < \infty$ for all k as well. Since $\mathcal{B}_r(\nu)$ is weakly closed, we have $\mu^\infty \in \mathcal{B}_r(\nu)$. We need to show that for any $\epsilon > 0$ there is $k(\epsilon)$ such that

$$\int_X f(x)\,\mathrm{d}\mu^k(x) - \int_X f(x)\,\mathrm{d}\mu^\infty(x) \leq \epsilon \qquad \forall k \geq k(\epsilon).$$

Our proof relies on the construction of an auxiliary function $f_R : X \to \mathbb{R}$ whose integrals under the probability measures $\mu \in \mathcal{B}_r(\nu)$ are close to those of f but that is at the same time bounded from above by a constant involving R. To this end, fix $R > 0$ and set $f_R(x) = \min\{f(x), c(R^{p'} + 1)\}$. We then have

$$|f(x) - f_R(x)| \leq \begin{cases} c\cdot d^{p'}(x, x_0) & \text{if } d(x, x_0) \geq R, \\ 0 & \text{otherwise} \end{cases} \qquad \forall x \in X. \tag{4}$$

5

To see this, note first that $f_R(x) \le f(x)$ and thus $|f(x) - f_R(x)| = f(x) - f_R(x)$. Fix $x \in X$ and assume first that $d(x, x_0) \ge R$. If $f(x) \le c(R^{p'} + 1)$, then $f_R(x) = f(x)$ and thus $|f(x) - f_R(x)| = 0$. If $f(x) > c(R^{p'} + 1)$, on the other hand, then

$$f(x) - f_R(x) \;=\; f(x) - c(R^{p'} + 1) \;\le\; c(d^{p'}(x, x_0) + 1) - c(R^{p'} + 1) \;\le\; c \cdot d^{p'}(x, x_0),$$

where first inequality follows from the assumption made in the statement of this theorem. Assume now that $d(x, x_0) < R$. In that case, the same assumption implies that $f(x) \le c(d^{p'}(x, x_0) + 1) < c(R^{p'} + 1)$ and hence $f(x) = f_R(x)$. We thus conclude that the bound (4) indeed holds. The bound implies that for all $\mu \in \mathcal{B}_r(\nu)$, we have that

$$\left| \int_X f(x) \, \mathrm{d}\mu(x) - \int_X f_R(x) \, \mathrm{d}\mu(x) \right| \;\le\; \int_X |f(x) - f_R(x)| \, \mathrm{d}\mu(x)$$

$$\le\; c \cdot \int_{d(\cdot, x_0) \ge R} d^{p'}(x, x_0) \, \mathrm{d}\mu(x) \;\le\; \frac{c \cdot C}{R^{(p-p')}}, \quad (5)$$

where the first and second inequality follow from the triangle inequality and equation (4), respectively. The third inequality holds because

$$d^{p'}(x, x_0) \;=\; \frac{d^p(x, x_0)}{d^{p-p'}(x, x_0)} \;\le\; \frac{d^p(x, x_0)}{R^{(p-p')}} \qquad \forall x \in X : d(x, x_0) \ge R,$$

and since Lemma 1 implies that there is $C > 0$ such that

$$\int_{d(\cdot, x_0) \ge R} d^p(x, x_0) \, \mathrm{d}\mu(x) \;\le\; \int_X d^p(x, x_0) \, \mathrm{d}\mu(x) \;\le\; C \qquad \forall \mu \in \mathcal{B}_r(\nu).$$

Using the auxiliary function f_R, we can now prove the weak upper semicontinuity of the map $\mu \mapsto \int_X f(x) \, \mathrm{d}\mu(x)$ over $\mathcal{B}_r(\nu)$. Indeed, for every $\epsilon > 0$ there is $k(\epsilon)$ such that for all $k \ge k(\epsilon)$, we have

$$\int_X f(x) \, \mathrm{d}\mu^k(x) - \int_X f(x) \, \mathrm{d}\mu^\infty(x) \;\le\; \int_X f_R(x) \, \mathrm{d}\mu^k(x) - \int_X f_R(x) \, \mathrm{d}\mu^\infty(x)$$

$$+ \left| \int_X f(x) \, \mathrm{d}\mu^k(x) - \int_X f_R(x) \, \mathrm{d}\mu^k(x) \right|$$

$$+ \left| \int_X f(x) \, \mathrm{d}\mu^\infty(x) - \int_X f_R(x) \, \mathrm{d}\mu^\infty(x) \right| \;\le\; \epsilon.$$

Here, the first inequality follows from the triangle inequality. As for the second inequality, we can choose R sufficiently large such that $c \cdot C / [R^{(p-p')}] \le \epsilon/3$, which implies by equation (5) that the two absolute values are both bounded above by $\epsilon/3$. Note further that f_R is upper semicontinuous and bounded from above by construction. Lemma 3 in Appendix A thus implies that the first difference of integrals can be made smaller than $\epsilon/3$ by selecting $k(\epsilon)$ sufficiently large. We therefore conclude that $\mu \mapsto \int_X f(x) \, \mathrm{d}\mu(x)$ is weakly upper semicontinuous over $\mathcal{B}_r(\nu)$, as desired. $\qquad \square$

Note that the only difference between the conditions of Theorems 2 and 3 is that in the latter case, we require the growth condition for f to be satisfied for some $p' < p$. One can readily construct instances of problem (3) where $f(x)$ grows asymptotically as $d_p^p(x, x_0)$ for some $x_0 \in X$ but where the optimal value is still attained. On the other hand, [9, Example 4] presents an instance of (3) that satisfies the condition of Theorem 2 and for which there exists no single measure that attains the (finite) optimal value.

A necessary and sufficient condition for the existence of optimal solutions to (3) is derived in [7]. Since that condition requires the knowledge of an optimal dual solution to problem (3), however, it is difficult to verify in practice. In contrast, our result relies on a tight sufficient condition that is easily verifiable.

5 Existence of Optimal Discrete Solutions

We now show that we can restrict problem (3) to probability measures supported on $N+1$ atoms if the reference measure ν of the Wasserstein ball $\mathcal{B}_r(\nu)$ is a discrete probability measure supported on N atoms. This is frequently the case in applications, where the reference measure ν is chosen as the empirical measure on finitely many training samples.

Theorem 4 (Existence of Discrete Optimal Solutions). *Assume that ν is supported on N atoms and that the conditions of Theorem 3 hold. Then problem (3) is optimized by a probability measure that is supported on at most $N + 1$ atoms.*

Proof. By the definition of the Wasserstein distance (1) and the Wasserstein ball (2), the optimization problem (3) can be formulated as

$$
\begin{aligned}
\underset{\mu}{\text{maximize}} \quad & \int_X f(x) \, \mathrm{d}\mu(x) \\
\text{subject to} \quad & \mu \in \mathcal{P}(X) \\
& \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{X \times X} d^p(x_1, x_2) \, \mathrm{d}\gamma(x_1, x_2) \le r^p.
\end{aligned}
$$

Following [10], we can eliminate the embedded minimization over γ in the problem above and obtain the equivalent single-level problem

$$
\begin{aligned}
\underset{\mu, \gamma}{\text{maximize}} \quad & \int_X f(x) \, \mathrm{d}\mu(x) \\
\text{subject to} \quad & \mu \in \mathcal{P}(X), \quad \gamma \in \mathcal{P}(X \times X) \\
& P_1 \gamma = \mu, \quad P_2 \gamma = \nu \\
& \int_{X \times X} d^p(x_1, x_2) \, \mathrm{d}\gamma(x_1, x_2) \le r^p,
\end{aligned}
\tag{6}
$$

where $P_1\gamma$ and $P_2\gamma$ refer to the first and second marginal probability measure of γ, respectively, that is, $(P_1\gamma)(S) = \gamma(S \times X)$ and $(P_2\gamma)(S) = \gamma(X \times S)$ for any Borel subset S of X. Since $\mu = P_1\gamma$, a change of variables allows us to rewrite the objective function as

$$
\int_{X \times X} f(x_1) \, \mathrm{d}\gamma(x_1, x_2).
$$

7

We can assume that ν satisfies $\nu = \sum_{i=1}^{N} \alpha_i \cdot \delta_{y_i}$ with $\alpha \in \mathbb{R}_{+}^{N}$ and $\sum_{i=1}^{N} \alpha_i = 1$ as well as $y_1, \ldots, y_N \in X$. We can then re-express the constraint $P_2\gamma = \nu$ in (6) through

$$\int_{X \times X} \mathbb{1}_{X \times \{y_i\}}(x_1, x_2) \, \mathrm{d}\gamma(x_1, x_2) = \alpha_i \qquad \forall i = 1, \ldots, N,$$

where the indicator function satisfies $\mathbb{1}_S(x_1, x_2) = 1$ if $(x_1, x_2) \in S \subseteq X \times X$ and $\mathbb{1}_S(x_1, x_2) = 0$ otherwise. If we additionally define $Y = \{y_1, \ldots, y_N\}$ and make the normalization of γ explicit, we obtain the following equivalent reformulation of problem (6):

$$
\begin{aligned}
\underset{\gamma}{\text{maximize}} \quad & \int_{X \times Y} f(x_1) \, \mathrm{d}\gamma(x_1, x_2) \\
\text{subject to} \quad & \gamma \in \mathcal{M}_{+}(X \times Y) \\
& \int_{X \times Y} \mathrm{d}\gamma(x_1, x_2) = 1 \\
& \int_{X \times Y} \mathbb{1}_{X \times \{y_i\}}(x_1, x_2) \, \mathrm{d}\gamma(x_1, x_2) = \alpha_i \qquad \forall i = 1, \ldots, N \\
& \int_{X \times Y} d^p(x_1, x_2) \, \mathrm{d}\gamma(x_1, x_2) \leq r^p,
\end{aligned}
\tag{7}
$$

Here, $\mathcal{M}_{+}(X \times Y)$ is the set of all non-negative finite Borel measures supported on $X \times Y$, and the first integral constraint ensures that γ is indeed a probability measure residing in $\mathcal{P}(X \times Y)$. Note that any one of the $N + 1$ equality constraints in problem (7) is implied by the remaining N constraints. Hence, we can drop the normalization constraint. This seemingly redundant step ensures that we can get a discrete optimal solution with $N + 1$ support points later. Combining the aforementioned reductions, problem (7) becomes

$$
\begin{aligned}
\underset{\gamma}{\text{maximize}} \quad & \int_{X \times Y} f(x_1) \, \mathrm{d}\gamma(x_1, x_2) \\
\text{subject to} \quad & \gamma \in \mathcal{M}_{+}(X \times Y) \\
& \int_{X \times Y} \mathbb{1}_{X \times \{y_i\}}(x_1, x_2) \, \mathrm{d}\gamma(x_1, x_2) = \alpha_i \qquad \forall i = 1, \ldots, N \\
& \int_{X \times Y} d^p(x_1, x_2) \, \mathrm{d}\gamma(x_1, x_2) \leq r^p.
\end{aligned}
\tag{8}
$$

Below, we distinguish between the two cases where the optimal value of problem (8) changes (or remains unchanged) if we strengthen the last integral constraint to an equality.

Assume first that we can strengthen the last integral constraint to an equality without affecting the optimal value of the problem. In that case, problem (8) resembles a linear program in standard form with $N + 1$ equality constraints and infinitely many nonnegative decision variables. We should thus expect that its optimal value is attained by a basic feasible solution, that is, a solution for which at most $N + 1$ variables are strictly positive and all others vanish. In our context, such a basic feasible solution would correspond to a discrete measure γ that is supported on at most $N + 1$ atoms of $X \times Y$. To formalize

this intuition, we apply Proposition 1 from Appendix B to conclude that our variant of problem (8) is equivalent to

$$\begin{aligned}
\underset{\gamma}{\text{maximize}} \quad & \int_{X \times Y} f(x_1) \, d\gamma(x_1, x_2) \\
\text{subject to} \quad & \gamma \in \mathcal{D}_{N+1}(X \times Y) \\
& \int_{X \times Y} \mathbb{1}_{X \times \{y_i\}}(x_1, x_2) \, d\gamma(x_1, x_2) = \alpha_i \qquad \forall i = 1, \ldots, N \\
& \int_{X \times Y} d^p(x_1, x_2) \, d\gamma(x_1, x_2) = r^p,
\end{aligned} \tag{9}$$

where we have replaced the set of all non-negative finite Borel measures $\mathcal{M}_+(X \times Y)$ on $X \times Y$ with the subset of non-negative discrete measures $\mathcal{D}_{N+1}(X \times Y)$ that are supported on at most $N + 1$ points in $X \times Y$. Indeed, one readily verifies that our variant of problem (8) is an instance of problem (15) if we set $Z = X \times Y$, $\psi = f \circ \pi_1$, where $\pi_1(x_1, x_2) = x_1$ for $(x_1, x_2) \in X \times X$, $m = N + 1$, $\phi_i = \mathbb{1}_{X \times \{y_i\}}$ and $v_i = \alpha_i$, $i = 1, \ldots, N$, as well as $\phi_{N+1} = d^p$ and $v_{N+1} = r^p$. Moreover, the conditions of Proposition 1 are satisfied. Indeed, we assumed in Section 1 that at least one of the integrals $\int_X [f(x)]_+ \, d\mu(x)$ and $\int_X [-f(x)]_+ \, d\mu(x)$ is finite for each $\mu \in \mathcal{B}_r(\nu)$. Likewise, the requirement that $\int_Z |\phi_i|(z) \, d\gamma(z) < \infty$ for all $i = 1, \ldots, m$ is guaranteed by Lemma 1 and the fact that γ is a probability measure.

It is not a priori clear whether the feasible region of problem (9) is weakly compact, and thus we cannot ascertain that the optimal value of (9) is attained. If we replace the last equality in (9) with an inequality, however, we obtain the relaxation

$$\begin{aligned}
\underset{\gamma}{\text{maximize}} \quad & \int_{X \times Y} f(x_1) \, d\gamma(x_1, x_2) \\
\text{subject to} \quad & \gamma \in \mathcal{D}_{N+1}(X \times Y) \\
& \int_{X \times Y} \mathbb{1}_{X \times \{y_i\}}(x_1, x_2) \, d\gamma(x_1, x_2) = \alpha_i \qquad \forall i = 1, \ldots, N \\
& \int_{X \times Y} d^p(x_1, x_2) \, d\gamma(x_1, x_2) \leq r^p
\end{aligned} \tag{10}$$

whose optimal value is attained by Lemma 2 below. Note that the optimal value of (10) is sandwiched by the optimal values of (8) and (9). Since the optimal values of (8) and (9) coincide, we conclude that problem (10) must have the same optimal value as well.

Let $\gamma^\star$ be an optimal solution to problem (10). Since $\gamma^\star \in \mathcal{D}_{N+1}(X \times Y)$, there are $(x_1^i, x_2^i) \in X \times Y$, $i = 1, \ldots, N + 1$, as well as $\beta \in \mathbb{R}_+^{N+1}$ such that $\gamma^\star = \sum_{i=1}^{N+1} \beta_i \cdot \delta_{(x_1^i, x_2^i)}$ and $\sum_{i=1}^{N+1} \beta_i = 1$. By construction, $\mu^\star = P_1 \gamma^\star$ satisfies $\mu^\star \in \mathcal{D}^{N+1}(X)$, $\int_X f(x) \, d\mu^\star(x) = \sup(10) = \sup(3)$, as well as

$$\begin{aligned}
W_p(\mu^\star, \nu) \quad &= \quad \inf_{\gamma \in \Gamma(\mu^\star, \nu)} \left(\int_{X \times X} d^p(x_1, x_2) \, d\gamma(x_1, x_2) \right)^{\frac{1}{p}} \\
&\leq \quad \left(\int_{X \times X} d^p(x_1, x_2) \, d\gamma^\star(x_1, x_2) \right)^{\frac{1}{p}} \quad \leq \quad r,
\end{aligned}$$

which implies that $\mu^\star \in \mathcal{B}_r(\nu)$. We thus conclude that $\mu^\star$ is an $N+1$-point distribution in $\mathcal{B}_r(\nu)$ that optimizes problem (3), as desired.

Assume now that strengthening the last integral constraint of problem (8) to an equality changes the optimal value of (8). We claim that in this case, any optimal solution $\mu^\star$ to problem (3) satisfies the strict inequality $W_p(\mu^\star, \nu) < r$. Indeed, assume to the contrary that there is an optimal solution $\mu^\star$ to (3) that satisfies $W_p(\mu^\star, \nu) = r$. By [15, Theorem 4.1], the distance $W_p(\mu^\star, \nu)$ is attained by some minimizer $\gamma^\star \in \mathcal{P}(X \times X)$ of (1), and one readily verifies that $\gamma^\star$ would constitute a feasible solution to problem (8) that satisfies the last integral constraint as equality and that attains the optimal value of (8). This, however, contradicts our assumption that the optimal value of (8) changes if we strengthen the last integral constraint to an equality.

We now claim that the Wasserstein ball $\mathcal{B}_r(\nu)$ must contain a Dirac measure that places all probability mass on a global maximizer of f. Assume to the contrary that $\mathcal{B}_r(\nu)$ does not contain such a Dirac measure. In that case, we must have $\arg\max_{x \in X} f(x) \neq \emptyset$, and any optimal solution $\mu^\star$ to (3) must be supported on $\arg\max_{x \in X} f(x)$. Indeed, if that was not the case, there would be $\hat{x} \in X$ with $f(\hat{x}) > \int_X f(x) \, d\mu^\star$. Consider now all convex combinations $\mu(\lambda) = \lambda \cdot \delta_{\hat{x}} + (1-\lambda) \cdot \mu^\star$, $\lambda \in [0,1]$. Since the map $\lambda \mapsto W_p^p(\mu(\lambda), \nu)$ is finite, convex and lower semi-continuous on $\lambda \in [0,1]$, see [5, Corollary 5.3], it is continuous on the entire interval. For sufficiently small λ, $\mu(\lambda)$ is therefore feasible in (3) and attains a larger objective value than $\mu^\star$, thus violating the optimality of $\mu^\star$. Consider now any Dirac distribution $\delta_{x'}$ supported on $x' \in \arg\max_{x \in X} f(x)$. This Dirac distribution must be contained in $\mathcal{B}_r(\nu)$, for otherwise we could again form convex combinations μ' between $\mu^\star$ and $\delta_{x'}$ that are optimal in (3) and that satisfy $W_p(\mu', \nu) = r$, in contradiction to our earlier finding. We thus conclude that the Wasserstein ball $\mathcal{B}_r(\nu)$ contains a Dirac measure that places all probability mass on a global maximizer of f, and this Dirac measure clearly constitutes an optimal discrete solution to (3). $\qquad\square$

The following technical lemma is used in the proof of Theorem 4.

Lemma 2. *Assume that the assumptions of Theorem 4 hold. If the optimal value of problem (10) is finite, then it is attained.*

Proof. Since problem (10) has a finite optimal value, there is a sequence of feasible solutions $\{\gamma^k\}_k$ to (10) that attains the optimal value of (10) asymptotically. We prove the lemma by constructing from $\{\gamma^k\}_k$ a solution $\gamma^\star$ that is feasible in (10) and that attains the optimal value of (10). To simplify the exposition, we assume in the representation of the reference measure ν that $\alpha_i > 0$ for all i and that $y_i \neq y_j$ for all $i \neq j$; both conditions can always be satisfied by reducing the number of atoms N if necessary.

By going over to a subsequence if necessary, we may assume w.l.o.g. that every measure γ^k of the sequence $\{\gamma^k\}_k$ can be represented as

$$\gamma^k = \sum_{i=1}^{N-1} \alpha_i \cdot \delta_{(x_i^k, y_i)} + (\alpha_N - \beta_k) \cdot \delta_{(x_N^k, y_N)} + \beta_k \cdot \delta_{(x_{N+1}^k, y_N)}$$

for some $x_1^k, \ldots, x_{N+1}^k \in X$ and $\beta_k \in [0, \alpha_N]$. Indeed, every γ^k is supported on $N+1$ atoms from $X \times Y$ since $\gamma^k \in \mathcal{D}_{N+1}(X \times Y)$. Moreover, the first integral constraint

10

in (10) implies that $\gamma^k(X \times \{y_i\}) = \alpha_i$ for all $i = 1, \dots, N$. Since the reference measure ν has N atoms but γ^k is supported on $N + 1$ atoms, there is an atom y_i of the reference measure ν whose probability mass α_i is split across two atoms in γ^k (with one of them possibly having zero probability mass). While a different atom y_i may be split for different γ^k, we can again go over to a subsequence if necessary to ensure that the same atom y_i is split in all measures of $\{\gamma^k\}_k$. Moreover, we can assume w.l.o.g. that the split atom is y_N, which gives rise to the two atoms (x_N^k, y_N) and (x_{N+1}^k, y_N) for every γ^k; this can always be ensured by reordering the atoms of ν if necessary. In the remainder of this proof we argue that a subsequence of $\{\gamma^k\}_k$ converges weakly to a measure $\gamma^\star$ that is feasible in (10) and that achieves the optimal value of (10).

We first show that $\{\gamma^k\}_k$ has a subsequence for which $\{x_i^k\}_k$ converges to some $x_i^\star \in X$ for every $i = 1, \dots, N-1$. Indeed, assume to the contrary that $\{x_1^k\}_k$ does not have a convergent subsequence. Since the space (X, d) is assumed to be proper, this is only possible if $\{x_1^k\}_k$ diverges, that is, if $d(x_1^k, y_1) \longrightarrow \infty$. In that case, the associated transportation cost $\alpha_1 \cdot d(x_1^k, y_1)$ would also diverge since $\alpha_1 > 0$. Thus, the last constraint in problem (10) would be violated, which contradicts the assumed feasibility of each member of the sequence $\{\gamma^k\}_k$. Iteratively applying the same argument to $\{x_i^k\}_k$, $i = 2, \dots, N-1$, allows us to replace $\{\gamma^k\}_k$ with a subsequence such that $\{x_i^k\}_k \longrightarrow x_i^\star$ for all $i = 1, \dots, N-1$.

Consider now the sequences $\{x_N^k\}_k$ and $\{x_{N+1}^k\}_k$. By going over to a subsequence if necessary, we may assume w.l.o.g. that $\{\beta_k\}_k$ converges to $\beta^\star \in [0, \alpha_N]$ since the interval $[0, \alpha_N]$ is compact. If $\beta^\star \in (0, \alpha_N)$, then the argument from the previous paragraph applies equally to $\{x_N^k\}_k$ and $\{x_{N+1}^k\}_k$, that is, by iteratively going over to subsequences we may assume that $\{x_i^k\}_k \longrightarrow x_i^\star$ also for $i = N, N+1$. In this case, $\{\gamma^k\}_k$ converges weakly to

$$\gamma^\star = \sum_{i=1}^{N-1} \alpha_i \cdot \delta_{(x_i^\star, y_i)} + (\alpha_N - \beta^\star) \cdot \delta_{(x_N^\star, y_N)} + \beta^\star \cdot \delta_{(x_{N+1}^\star, y_N)}.$$

The measure $\gamma^\star$ resides in $\mathcal{D}_{N+1}$, and it satisfies the first integral constraint in (10) by construction. It satisfies the second integral constraint since the metric d is continuous in the topology that it generates. Moreover, $\gamma^\star$ attains the optimal value of (10) since

$$
\begin{aligned}
\int_{X \times Y} f(x_1) \, \mathrm{d}\gamma^\star(x_1, x_2) &= \sum_{i=1}^{N-1} \alpha_i \cdot f(x_i^\star) + (\alpha_N - \beta^\star) \cdot f(x_N^\star) + \beta^\star \cdot f(x_{N+1}^\star) \\
&\geq \limsup_{k \to \infty} \left[\sum_{i=1}^{N-1} \alpha_i \cdot f(x_i^k) + (\alpha_N - \beta_k) \cdot f(x_N^k) + \beta_k \cdot f(x_{N+1}^k) \right] \\
&= \limsup_{k \to \infty} \left[\int_{X \times Y} f(x_1) \, \mathrm{d}\gamma^k(x_1, x_2) \right] = \sup (10). \quad (11)
\end{aligned}
$$

Here, the first and the penultimate identity hold by the definitions of $\gamma^\star$ and γ^k, respectively, and the last identity follows from the fact that $\{\gamma^k\}_k$ attains the optimal value of problem (10) asymptotically. The inequality in (11) follows from the upper semi-continuity of f and the fact that $x_i^k \longrightarrow x_i^\star$, $i = 1, \dots, N+1$, as well as $\beta_k \longrightarrow \beta^\star$.

Assume now that $\beta^\star = 0$; the case where $\beta^\star = \alpha_N$ is symmetric. A similar argument as before implies that by going over to a subsequence if necessary, we may assume w.l.o.g. that

11

$x_N^k \longrightarrow x_N^\star$. In this case, the sequence $\{\gamma^k\}_k$ converges weakly to

$$\gamma^\star = \sum_{i=1}^{N} \alpha_i \cdot \delta_{(x_i^\star, y_i)},$$

even though $\{x_{N+1}^k\}_k$ may not converge. Again, $\gamma^\star \in \mathcal{D}_{N+1}$, and $\gamma^\star$ satisfies the first integral constraint in problem (10). As for the second integral constraint in (10), we have

$$
\begin{aligned}
\int_{X \times Y} d^p(x_1, x_2) \, d\gamma^\star(x_1, x_2) &= \sum_{i=1}^{N} \alpha_i \cdot d^p(x_i^\star, y_i) \\
&\leq \limsup_{k \longrightarrow \infty} \left[\sum_{i=1}^{N-1} \alpha_i \cdot d^p(x_i^k, y_i) + (\alpha_N - \beta_k) \cdot d^p(x_N^k, y_N) + \beta_k \cdot d^p(x_{N+1}^k, y_N) \right] \\
&= \limsup_{k \longrightarrow \infty} \left[\int_{X \times Y} d^p(x_1, x_2) \, d\gamma^k(x_1, x_2) \right] \leq r^p.
\end{aligned}
$$

Here, the first identity holds by definition of $\gamma^\star$. The inequality follows from the fact that the sum of $\sum_{i=1}^{N-1} \alpha_i \cdot d^p(x_i^k, y_i)$ and $(\alpha_N - \beta_k) \cdot d^p(x_N^k, y_N)$ converges to the expression $\sum_{i=1}^{N} \alpha_i \cdot d^p(x_i^\star, y_i)$ on the left-hand side and the last expression $\beta_k \cdot d^p(x_{N+1}^k, y_N)$ is non-negative by construction. The last line, finally, follows from the definition of γ^k and the feasibility of γ^k in (10). We thus conclude that $\gamma^\star$ is feasible in problem (10).

To see that $\gamma^\star$ attains the optimal value of problem (10), we observe that

$$
\begin{aligned}
\int_{X \times Y} f(x_1) \, d\gamma^\star(x_1, x_2) &= \sum_{i=1}^{N} \alpha_i \cdot f(x_i^\star) \\
&\geq \limsup_{k \longrightarrow \infty} \left[\sum_{i=1}^{N-1} \alpha_i \cdot f(x_i^k) + (\alpha_N - \beta_k) \cdot f(x_N^k) \right] \\
&\geq \limsup_{k \longrightarrow \infty} \left[\sum_{i=1}^{N-1} \alpha_i \cdot f(x_i^k) + (\alpha_N - \beta_k) \cdot f(x_N^k) + \beta_k \cdot f(x_{N+1}^k) \right] \\
&= \limsup_{k \longrightarrow \infty} \left[\int_{X \times Y} f(x_1) \, d\gamma^k(x_1, x_2) \right] = \sup (10). \quad (12)
\end{aligned}
$$

Here, the identities can be justified as in (11). The first inequality holds since f is upper semi-continuous, $x_i^k \longrightarrow x_i^\star$, $i = 1, \ldots, N$, and $\beta_k \longrightarrow 0$. We claim that

$$\limsup_{k \longrightarrow \infty} \left[\beta_k \cdot f(x_{N+1}^k) \right] \leq 0, \qquad (13)$$

which proves the second inequality in (12). To show this, fix an arbitrary $\epsilon > 0$ and choose $R > 0$ large enough such that

$$\frac{r^p}{R^p} \cdot c \left[1 + (R + d(y_N, x_0))^{p'} \right] \leq \epsilon, \qquad (14)$$

which exists since $p' < p$. Note that (14) is monotone in R, that is, if it is satisfied for some $R > 0$, it is also satisfied for all $R' > R$. Now select $k(\epsilon) \in \mathbb{N}$ large enough such that $\beta_k \cdot \sup_{x \in X}\{f(x) \,:\, d(x, y_N) \leq R\} \leq \epsilon$ for all $k \geq k(\epsilon)$. Such a $k(\epsilon)$ exists because f is upper semi-continuous and hence its supremum over any compact set is finite. We claim that

$$\beta_k \cdot f(x_{N+1}^k) \leq \epsilon \qquad \forall k \geq k(\epsilon).$$

Fix any $k \geq k(\epsilon)$. If $d(x_{N+1}^k, y_N) \leq R$, then the claim follows immediately from our choice of $k(\epsilon)$. If $d(x_{N+1}^k, y_N) > R$, on the other hand, then the claim holds since

$$
\begin{aligned}
\beta_k \cdot f(x_{N+1}^k) \;\; &\leq \;\; \frac{r^p}{d^p(x_{N+1}^k, y_N)} \cdot c\left[1 + d^{p'}(x_{N+1}^k, x_0)\right] \\
&\leq \;\; \frac{r^p}{d^p(x_{N+1}^k, y_N)} \cdot c\left[1 + \left(d(x_{N+1}^k, y_N) + d(y_N, x_0)\right)^{p'}\right] \\
&\leq \;\; \frac{r^p}{R^p} \cdot c\left[1 + (R + d(y_N, x_0))^{p'}\right] \;\; \leq \;\; \epsilon.
\end{aligned}
$$

Indeed, the first inequality holds since γ^k satisfies the last integral constraint of problem (10) and f satisfies the growth condition of Theorem 4. The second inequality follows from the triangle inequality. The last two inequalities, finally, follow from (14) and the fact that $d(x_{N+1}^k, y_N) > R$. We thus conclude that (13) indeed holds. $\qquad\square$

One may wonder whether the result of Theorem 4 can be strengthened further to the existence of optimal solutions to (3) that are supported on fewer than $N+1$ atoms. While this is possible for specific instances (for example, if $f(x)$ is concave), one can construct instances of problem (3) where the optimal value is only attained by measures supported on at least $N + 1$ atoms [9, Example 5].

The sparsity of optimal solutions to problem (3) has been investigated by several authors. To our best knowledge, the first result in this direction is [16], which employs the Kantorovich-Rubinstein and the Richter-Rogosinski theorems to prove that if (3) is solvable, then it is solved by a measure that is supported on at most $N + 3$ atoms. Subsequently, [11] showed that, if problem (3) is solvable, there are indeed optimal solutions that are only supported on at most $N + 2$ atoms. The sharp characterization of optimal measures supported on at most $N + 1$ atoms has been first derived in [7]. In contrast to our result, the authors do not employ the Richter-Rogosinski theorem or results remiscent of those in Appendix B. Instead, they rely on the first-order optimality conditions of the problem dual to (3). While this allows them to provide further insights into the structure of optimal solutions, their proof is substantially more difficult to verify than ours.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint 1701.07875*, 2018.

[2] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2nd edition, 1992.

[3] P. Billingsley. *Probability and Measure*. John Wiley & Sons, 3rd edition, 1995.

[4] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[5] P. Clément and W. Desch. Wasserstein metric and subordination. *Studia Mathematica*, 1(189):35–52, 2008.

[6] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems 28*, pages 2053–2061. 2015.

[7] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint 1604.02199*, 2016.

[8] N. Ho, X. L. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung. Multilevel clustering via Wasserstein means. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1501–1509, 2017.

[9] D. Kuhn, P. Mohajerin Esfahani, V. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

[10] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[11] H. Owhadi and C. Scovel. Extreme points of a ball about a measure with finite support. *Communications in Mathematical Sciences*, 15(1):77–96, 2017.

[12] G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

[13] A. Pichler and H. Xu. Quantitative stability analysis for minimax distributionally robust risk optimization. *Mathematical Programming*, Available Online, 2018.

[14] I. Pinelis. On the extreme points of moments sets. *Mathematical Methods of Operations Research*, 83(3):325–349, 2016.

[15] C. Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2008.

[16] D. Wozabal. A Framework for Optimization under Ambiguity. *Annals of Operations Research*, 193(1):21–47, 2012.

[17] C. Zhao and Y. Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262–267, 2018.

Appendix A   Auxiliary Measure-Theoretic Results

We review some well-known facts from measure theory that we use to prove our results. We first recall a connection between the notions of tightness and weak sequential compactness of collections of probability measures.

Definition 1. *A collection $\mathcal{S} \subseteq \mathcal{P}(X)$ of probability measures is tight if for any $\epsilon > 0$, there exists a compact subset $B \subseteq X$ such that $\mu(X \setminus B) \leq \epsilon$ for all $\mu \in \mathcal{S}$.*

Definition 2. *A sequence $\{\mu^k\}_k \subseteq \mathcal{P}(X)$ of probability measures converges weakly to $\mu^\infty \in \mathcal{P}(X)$ if for any bounded and continuous function g on X, we have*

$$\lim_{k \longrightarrow \infty} \int_X g(x) \, \mathrm{d}\mu^k \;=\; \int_X g(x) \, \mathrm{d}\mu^\infty.$$

Definition 3. *A collection $\mathcal{S} \subseteq \mathcal{P}(X)$ of probability measures is weakly sequentially compact if every sequence in $\mathcal{S}$ has a subsequence that converges weakly to an element of $\mathcal{S}$.*

The concepts of tightness and weak sequential compactness are connected by Prokhorov's Theorem, see for example [2, Theorem 5.1].

Theorem 5 (Prokhorov's Theorem)**.** *A collection $\mathcal{S} \subseteq \mathcal{P}(X)$ of probability measures is tight if and only if the closure of $\mathcal{S}$ is weakly sequentially compact in $\mathcal{P}(X)$.*

The following lemma, which is excerpted from the Portmanteau Theorem (see for example [3, Problem 29.1(c)]), provides a useful characterization of weak convergence.

Lemma 3. *A sequence $\{\mu^k\}_k \subseteq \mathcal{P}(X)$ of probability measures converges weakly to $\mu^\infty \in \mathcal{P}(X)$ if and only if for any upper bounded and upper semi-continuous function g on X, we have*

$$\limsup_{k \longrightarrow \infty} \int_X g(x) \, \mathrm{d}\mu^k(x) \;\leq\; \int_X g(x) \, \mathrm{d}\mu^\infty(x).$$

Appendix B   Basic Feasible Solutions in Infinite-Dimensional Linear Programming

It is well-known that if a finite-dimensional linear program with m equality constraints has an optimal solution, then there must be an optimal basic feasible solution with at most m non-zero entries. An infinite-dimensional analogue of this fact is proved in [14, Corollary 5 and Proposition 6(v)]. To state this result, let Z be a topological space, let $\mathcal{M}_+(Z)$ be the set of non-negative finite Borel measures supported on Z, and let $\psi, \phi_1, \ldots, \phi_m : Z \to \mathbb{R}$ be Borel functions as well as $v \in \mathbb{R}^m$. Consider now the optimization problem

$$
\begin{aligned}
\underset{\gamma}{\text{maximize}} \quad & \int_Z \psi(z) \, \mathrm{d}\gamma(z) \\
\text{subject to} \quad & \gamma \in \mathcal{M}_+(Z) \\
& \int_Z \phi_i(z) \, \mathrm{d}\gamma(z) = v_i \quad \forall i = 1, \ldots, m,
\end{aligned}
\tag{15}
$$

and denote by $\mathcal{F}$ the feasible region of (15) and by $\text{ext}(\mathcal{F})$ the set of extreme points of $\mathcal{F}$.

Proposition 1. *Suppose that for all $\gamma \in \mathcal{F}$, at least one of the integrals $\int_X [\psi(z)]_+ \, \mathrm{d}\gamma(z)$ and $\int_X [-\psi(z)]_+ \, \mathrm{d}\gamma(z)$ is finite and that $\int_Z |\phi_i|(z) \, \mathrm{d}\gamma(z) < \infty$ for all $i = 1, \ldots, m$. If*

$$
\sup \left\{ \int_Z \psi(z) \, \mathrm{d}\gamma(z) : \gamma \in \mathcal{F} \right\} \;=\; \sup \left\{ \int_Z \psi(z) \, \mathrm{d}\gamma(z) : \gamma \in \text{ext}(\mathcal{F}) \right\},
\tag{16}
$$

then it holds that

$$
\sup \left\{ \int_Z \psi(z) \, \mathrm{d}\gamma(z) : \gamma \in \mathcal{F} \right\} \;=\; \sup \left\{ \int_Z \psi(z) \, \mathrm{d}\gamma : \gamma \in \mathcal{F} \cap \mathcal{D}_m(Z) \right\},
$$

where $\mathcal{D}_m(Z)$ is the set of non-negative discrete measures supported on at most m points in Z. Furthermore, if $\mathcal{F} \subseteq \mathcal{P}(Z)$ and Z is Hausdorff, then the condition (16) is satisfied.