Outlier-Robust Optimal Transport: Duality, Structure, and Statistical Analysis

Sloan Nietert
Cornell University

Rachel Cummings
Columbia University

Ziv Goldfeld
Cornell University

Abstract

The Wasserstein distance, rooted in optimal transport (OT) theory, is a popular discrepancy measure between probability distributions with various applications to statistics and machine learning. Despite their rich structure and demonstrated utility, Wasserstein distances are sensitive to outliers in the considered distributions, which hinders applicability in practice. We propose a new outlier-robust Wasserstein distance W_p^ε which allows for ε outlier mass to be removed from each contaminated distribution. Under standard moment assumptions, W_p^ε is shown to be minimax optimal for robust estimation under the Huber ε-contamination model. Our formulation of this robust distance amounts to a highly regular optimization problem that lends itself better for analysis compared to previously considered frameworks. Leveraging this, we conduct a thorough theoretical study of W_p^ε, encompassing robustness guarantees, characterization of optimal perturbations, regularity, duality, and statistical estimation. In particular, by decoupling the optimization variables, we arrive at a simple dual form for W_p^ε that can be implemented via an elementary modification to standard, duality-based OT solvers. We illustrate the virtues of our framework via applications to generative modeling with contaminated datasets.

1 INTRODUCTION

Discrepancy measures between probability distributions are a fundamental constituent of statistical inference, machine learning, and information theory. Among many such measures, Wasserstein distances (Villani, 2003) have recently emerged as a tool of choice for many applications. Specifically, for $p \in [1, \infty)$ and a pair of probability measures μ, ν on a metric space $(\mathcal{X}, d)$, the p-Wasserstein distance between them is[1]

$$\mathsf{W}_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \, \mathrm{d}\pi(x,y) \right)^{1/p},$$

where $\Pi(\mu, \nu)$ is the set of couplings for μ and ν. The popularity of these metrics stems from a myriad of desirable properties, including rich geometric structure, metrization of the weak topology, robustness to support mismatch, and a convenient dual form. Modern applications thereof include generative modeling (Arjovsky et al., 2017; Gulrajani et al., 2017; Tolstikhin et al., 2018), domain adaptation (Courty et al., 2014, 2016), and robust optimization (Esfahani and Kuhn, 2018; Blanchet et al., 2018; Gao and Kleywegt, 2016).

Despite their advantages, Wasserstein distances suffer from sensitivity to outliers due to the strict marginal constraints, as even a small outlier mass can contribute greatly to the distance. This has inspired a recent line of work into outlier-robust OT (Balaji et al., 2020; Mukherjee et al., 2021; Le et al., 2021), which relaxes the marginal constraints in various ways. These build upon the theory of unbalanced OT (Piccoli and Rossi, 2014; Chizat et al., 2018a,b; Liero et al., 2018; Schmitzer and Wirth, 2019) that quantifies the cost-optimal way to transform one measure into another via a combination of mass variation and transportation. We propose a new framework for outlier-robust OT that arises as the solution to a principled robust approximation problem. We conduct an in-depth theoretical study of the proposed robust distance, encompassing formal robustness guarantees, duality, characterization of primal minimizers / dual maximizers, regularity, and empirical convergence rates.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

[1] For $p = \infty$, we set $\mathsf{W}_\infty(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \|d\|_{L^\infty(\pi)}$.

1.1 Contributions

We introduce and study the *ε-outlier-robust Wasserstein distance* defined by

$$\mathsf{W}_p^\varepsilon(\mu,\nu) := \inf_{\substack{\mu',\nu' \in \mathcal{M}_+(\mathcal{X}) \\ \mu' \le \mu,\, \|\mu-\mu'\|_{\mathsf{TV}} \le \varepsilon \\ \nu' \le \nu,\, \|\nu-\nu'\|_{\mathsf{TV}} \le \varepsilon}} \mathsf{W}_p\left(\frac{\mu'}{\mu'(\mathcal{X})}, \frac{\nu'}{\nu'(\mathcal{X})}\right), \tag{1}$$

where μ' and ν' are positive measures, $\|\cdot\|_{\mathsf{TV}}$ is the total variation (TV) norm, and $\le$ denotes setwise inequality when appropriate. The minimization over μ', ν' allows outliers occupying less than fraction ε of probability mass to be omitted from consideration, after which the perturbed measures are renormalized. Compared to prior work employing TV constraints (Balaji et al., 2020; Mukherjee et al., 2021), our definition has several distinct features: (1) it is naturally derived as a robust proxy for W_p under the Huber ε-contamination model; (2) it can be reframed as an optimization problem over a highly regular domain; and, consequently, (3) it admits a simple and useful duality theory.

We show that when the clean distributions have bounded qth moments for $q > p$, W_p^ε achieves the minimax optimal robust estimation error of $\varepsilon^{1/p-1/q}$ under the Huber ε-contamination model. Moreover, our dual formulation mirrors the classic Kantorovich dual with an added penalty proportional to the range of the potential function. This provides an elementary robustification technique which can be applied to any duality-based OT solver: one needs only to compute the argmin and argmax of the discriminative potentials over the batch samples, which can be done in conjunction with existing gradient evaluations. We demonstrate this the utility of this procedure with experiments for generative modeling with contaminated datasets using Wasserstein generative adversarial networks (WGANs) (Arjovsky et al., 2017).

We also study structural properties of W_p^ε, characterizing the minimizers of (1) and maximizers of its dual, describing the regularity of the problem in ε, and drawing a connection between W_p^ε and loss trimming (Shen and Sanghavi, 2019). Finally, we study statistical aspects of W_p^ε, examining both one- and two-sample empirical convergence rates and providing additional robustness guarantees. The derived empirical convergence rates are at least as fast as the regular $n^{-1/d}$ rate for standard W_p; however, faster rates may be possible if only a small amount of high-dimensional mass is present.

1.2 Related Work

The robust Wasserstein distance[2] in (1) is closely related to the notions considered in Balaji et al. (2020) and Mukherjee et al. (2021). In Balaji et al. (2020), similar constraints are imposed with respect to (w.r.t.) general f-divergences, but the perturbed measures are restricted to probability distributions. This results in a more complex dual form (derived by invoking standard Kantorovich duality on the Wasserstein distance between perturbations) and requires optimization over a significantly larger domain. In Mukherjee et al. (2021), robustness w.r.t. the TV distance is added via a regularization term in the objective. This leads to a simple modified primal problem but the corresponding dual requires optimizing over two potentials, even when $p = 1$. Additionally, Le et al. (2021) and Nath (2020) consider robustness via Kullback-Leibler (KL) divergence and integral probability metric (IPM) regularization terms, respectively. The former focuses on Sinkhorn-based primal algorithms and the latter introduces a dual form that is distinct from ours and less compatible with existing duality-based OT computational methods. In Staerman et al. (2021), a median of means approach is used to tackle the dual Kantorovich problem from a robust statistics perspective. None of these works provide minimax error bounds for robust estimation.

The robust OT literature is intimately related to unbalanced OT theory, which addresses transport problems between measures of different mass (Piccoli and Rossi, 2014; Chizat et al., 2018a; Liero et al., 2018; Schmitzer and Wirth, 2019; Hanin, 1992). These formulations are reminiscent of the problem (1) but with regularizers added to the objective (KL being the most studied) rather than incorporated as constraints. Sinkhorn-based primal algorithms (Chizat et al., 2018b) are the standard approach to computation, and these have recently been extended to large-scale machine learning problems via minibatch methods (Fatras et al., 2021). Fukunaga and Kasai (2021) introduces primal-based algorithms for semi-relaxed OT, where marginal constraints for a single measure are replaced with a regularizer in the objective. Partial OT (Caffarelli and McCann, 2010; Figalli, 2010), where only a fraction of mass needs to be moved, is another related framework. However, Caffarelli and McCann (2010) consider a different parameterization of the problem, arriving at a distinct dual, and Figalli (2010) is mostly restricted to quadratic costs with no discussion of duality. Recently, Chapel et al. (2020) has explored partial OT for positive-unlabeled learning, but dual-based algorithms are not considered.

Notation and Preliminaries. Let $(\mathcal{X}, d)$ be a complete, separable metric space, and denote the diameter of a set $A \subset \mathcal{X}$ by $\mathrm{diam}(A) := \sup_{x,y\in A} d(x,y)$. Take $C_b(\mathcal{X})$ as the set of continuous, bounded real functions on $\mathcal{X}$, and let $\mathcal{M}(\mathcal{X})$ denote the set of signed

[2] Despite calling W_p^ε a distance, we remark that the metric structure of W_p is lost after robustification.

Radon measures on $\mathcal{X}$ equipped with the TV norm[3] $\|\mu\|_{\mathsf{TV}} := |\mu|(\mathcal{X})$. Let $\mathcal{M}_+(\mathcal{X})$ denote the space of finite, positive Radon measures on $\mathcal{X}$. The Lebesgue measure on $\mathbb{R}^d$ is designated by λ. For $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ and $p \in [1, \infty]$, we consider the standard $L^p(\mu)$ space with norm $\|f\|_{L^p(\mu)} = \left(\int |f|^p \, \mathrm{d}\mu\right)^{1/p}$, and we write $\mu \le \nu$ when $\mu(B) \le \nu(B)$ for every Borel set $B \subseteq \mathcal{X}$.

Let $\mathcal{P}(\mathcal{X}) \subset M_+(\mathcal{X})$ denote the space of probability measures on $\mathcal{X}$, and take $\mathcal{P}_p(\mathcal{X}) := \{\mu \in \mathcal{P}(\mathcal{X}) : \int d(x, x_0)^p \, \mathrm{d}\mu(x) < \infty\}$ to be those with bounded pth moment. We write $\mathcal{P}_\infty(\mathcal{X})$ for probability measures with bounded support. Given $\mu, \nu \in \mathcal{P}(\mathcal{X})$, let $\Pi(\mu, \nu)$ denote the set of their couplings, i.e., $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ such that $\pi(B \times \mathcal{X}) = \mu(B)$ and $\pi(\mathcal{X} \times B) = \nu(B)$, for every Borel set B. When $\mathcal{X} = \mathbb{R}^d$, we write the covariance matrix for $\mu \in \mathcal{P}_2(\mathcal{X})$ as $\Sigma_\mu := \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\intercal]$ where $X \sim \mu$. For $f : \mathcal{X} \to \mathbb{R}$, we define the range $\mathrm{Range}(f) := \sup_{x \in \mathcal{X}} f(x) - \inf_{x \in \mathcal{X}} f(x)$. We write $a \vee b = \max\{a, b\}$ and use $\lesssim, \gtrsim, \asymp$ to denote inequalities/equality up to absolute constants.

Recall that $\mathsf{W}_p(\mu, \nu) < \infty$, for any $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$. For any $p \in [1, \infty)$, Kantorovich duality states that

$$\mathsf{W}_p(\mu, \nu)^p = \sup_{f \in C_b(\mathcal{X})} \int f \, \mathrm{d}\mu + \int f^c \, \mathrm{d}\nu, \tag{2}$$

where the c-transform $f^c : \mathcal{X} \to \mathbb{R}$ is defined by $f^c(y) = \inf_{x \in \mathcal{X}} d(x, y)^p - f(x)$ (with respect to the cost $c(x, y) = d(x, y)^p$).

2 ROBUST ESTIMATION OF W_p

Outlier-robust OT is designed to address the fact that $\mathsf{W}_p\big((1-\varepsilon)\mu + \varepsilon\delta_x, \nu\big)$ explodes as $d(x, x_0) \to \infty$, no matter how small ε might be. More generally, one considers an adversary that seeks to dramatically alter W_p by adding small pieces of mass to its arguments. To formalize this, we fix $p \in [1, \infty)$ and consider the Huber ε-contamination model popularized in robust statistics (Huber, 1964), where a base measure $\mu \in \mathcal{P}(\mathcal{X})$ is perturbed to obtain a contaminated measure $\tilde{\mu}$ belonging to the ball

$$\begin{aligned} \mathcal{B}_\varepsilon(\mu) &:= (1-\varepsilon)\mu + \varepsilon\,\mathcal{P}(\mathcal{X}) \\ &= \big\{(1-\varepsilon)\mu + \varepsilon\alpha : \alpha \in \mathcal{P}(\mathcal{X})\big\}. \end{aligned} \tag{3}$$

The goal is to obtain a robust proxy $\hat{W} : \mathcal{P}(\mathcal{X})^2 \to \mathbb{R}$ which, for any clean distributions $\mu, \nu \in \mathcal{P}(\mathcal{X})$ with contaminated versions $\tilde{\mu} \in \mathcal{B}_\varepsilon(\mu), \tilde{\nu} \in \mathcal{B}_\varepsilon(\nu)$, achieves low error $\big|\hat{W}(\tilde{\mu}, \tilde{\nu}) - \mathsf{W}_p(\mu, \nu)\big|$. In general, this error can be unbounded, so we require that the base measures belong to some family $\mathcal{D}$ capturing distributional assumptions, e.g., bounded moments of some order. For any $\varepsilon \in [0, 1]$ and $\mathcal{D} \subseteq \mathcal{P}(\mathcal{X})$, we define the minimax robust estimation error by

$$E(\mathcal{D}, \varepsilon) := \inf_{\hat{W} : \mathcal{P}(\mathcal{X})^2 \to \mathbb{R}} \sup_{\substack{\mu, \nu \in \mathcal{D} \\ \tilde{\mu} \in \mathcal{B}_\varepsilon(\mu) \\ \tilde{\nu} \in \mathcal{B}_\varepsilon(\nu)}} \big|\hat{W}(\tilde{\mu}, \tilde{\nu}) - \mathsf{W}_p(\mu, \nu)\big|.$$

Our interest in the robust Wasserstein distance W_p^ε, as defined in (1), stems from the following theorem, characterizing it as the minimax optimal robust proxy for W_p under Huber contamination and standard moment assumptions.

Theorem 1 (Robust estimation of W_p)**.** *Fix $q > p$ and let $\mathcal{D}_q := \{\mu \in \mathcal{P}_q(\mathcal{X}) : \|d(\cdot, x)\|_{L^q(\mu)} \le M$ for some $x \in \mathcal{X}\}$ denote the family of distributions with centered qth moments uniformly bounded by an absolute constant M. Then, for $0 \le \varepsilon \le 1/4$, we have*

$$E(\mathcal{D}_q, \varepsilon) \lesssim M\varepsilon^{1/p - 1/q}. \tag{4}$$

This error rate is obtained by $\hat{W} = \mathsf{W}_p^\varepsilon$ and is optimal so long as $\mathcal{X}$ contains two points at distance $\Theta(M\varepsilon^{-1/q})$.

The distance assumption is quite mild and is satisfied, e.g., when $\mathcal{X}$ has a path connected component with diameter at least $M\varepsilon^{-1/q}$. We outline the proof below, with full details provided in Appendix A.2.

Proof Sketch. The upper bound relies crucially on the fact that $|\mathsf{W}_p^\varepsilon(\tilde{\mu}, \tilde{\nu}) - \mathsf{W}_p(\mu, \nu)| \le |\mathsf{W}_p^{O(\varepsilon)}(\mu, \nu) - \mathsf{W}_p(\mu, \nu)|$, allowing one to bound $E(\mathcal{D}_q, \varepsilon)$ uniformly over contaminated measures. To control the upper bound, we use that $\mathsf{W}_p(\mu, \nu) \le \|d(\cdot, x_0)\|_{L^q(\mu)} + \|d(\cdot, x_1)\|_{L^q(\nu)} + d(x_0, x_1)$ for $q \ge p$, which holds for any $x_0, x_1 \in \mathcal{X}$ and connects W_p to the moment bounds encoded by $\mathcal{D}_q$. The lower bound follows by a standard modulus of continuity argument which gives

$$E(\mathcal{D}, \varepsilon) \gtrsim \sup_{\substack{\mu, \tilde{\mu} \in \mathcal{D} \\ \tilde{\mu} \in \mathcal{B}_\varepsilon(\mu)}} \mathsf{W}_p(\mu, \tilde{\mu})$$

for any choice of $\mathcal{D} \subseteq \mathcal{P}(\mathcal{X})$. In the case of $\mathcal{D} = \mathcal{D}_q$, a simple choice of μ and $\tilde{\mu}$ supported on at most two points gives a matching lower bound under the distance assumption. Importantly, the hidden constants in this result are absolute and independent of $\mathcal{X}$ (in particular, they hold even when $\mathcal{X}$ is unbounded). □

We next specialize Theorem 1 to the common case of $\mathcal{X} = \mathbb{R}^d$ with base measures whose covariance matrices Σ_μ, Σ_ν have bounded spectral norms. Such measures also have bounded moments in the sense of Theorem 1, so the previous upper bound applies. The lower bound for this case (given in Appendix A.3) uses the same technique as before but with a more careful choice of measures involving a multivariate Gaussian.

[3] This definition will be convenient but omits a factor of 1/2 often present in machine learning literature.

Corollary 1 (Bounded covariance). *Fix $\mathcal{X} = \mathbb{R}^d$ and let $\mathcal{D}_2^{\mathrm{cov}} := \{\mu \in \mathcal{P}_2(\mathcal{X}) : \Sigma_\mu \preceq I_d\}$. For $p < 2$ and $0 \le \varepsilon \le 1/4$, we have*

$$E(\mathcal{D}_2^{\mathrm{cov}}, \varepsilon) \asymp \sqrt{d}\, \varepsilon^{1/p - 1/2},$$

and this optimal error rate is achieved by $\hat{W} = \mathsf{W}_p^\varepsilon$.

Remark 1 (Comparison with robust mean estimation). In the setting of mean estimation under assumptions analogous to Corollary 1 (i.e., $\Sigma_\mu \preceq I_d$, $\tilde{\mu} \in \mathcal{B}_\varepsilon(\mu)$), the optimal error rate of $\sqrt{\varepsilon}$ is *dimension-free* (Chen et al., 2018). We interpret the factor of $\sqrt{d}$ present for our W_p rate as reflecting the high-dimensional optimization inherent to the Wasserstein distance.

Remark 2 (Breakdown point). We note that the bound of $1/4$ on ε in both results can be substituted with any constant less than $1/3$. We conjecture that this breakdown point of $1/3$ can be pushed to its information-theoretic limit of $1/2$.

Remark 3 (Asymmetric contamination). The robust distance from (1) readily extends to an asymmetric distance $\mathsf{W}_p^{\varepsilon_\mu, \varepsilon_\nu}$ with distinct robustness radii, so that $\mathsf{W}_p^\varepsilon = \mathsf{W}_p^{\varepsilon,\varepsilon}$. Extensions of our main results (including Theorem 1) to this setting are presented in Appendix B.1. The one-sided version $\mathsf{W}_p^\varepsilon(\mu\|\nu) := \mathsf{W}_p^{\varepsilon,0}(\mu,\nu)$ is well-suited for applications such as generative modeling (see Section 6).

Our precise error bounds exploit the unique structure of W_p^ε and do not translate clearly to existing robust proxies for W_p. We note, however, that the TV-robustified W_p presented in Balaji et al. (2020) can be controlled and approximated to some extent by W_p^ε, via bounds presented in Appendix B.2.

3 DUALITY THEORY FOR W_p^ε

In addition to its robustness properties, W_p^ε enjoys a simple optimization structure that enables a useful duality theory. Unless stated otherwise, we henceforth assume that $\mathcal{X}$ is compact. To begin, we reformulate $\mathsf{W}_p^\varepsilon(\mu,\nu)$ as a minimization problem over Huber balls centered at μ and ν.

Proposition 1 (Mass addition). *For all $p \in [1,\infty]$ and $\mu,\nu \in \mathcal{P}(\mathcal{X})$, we have*

$$\mathsf{W}_p^\varepsilon(\mu,\nu) = (1-2\delta)^{-1/p} \inf_{\substack{\mu' \in \mathcal{B}_\delta(\mu) \\ \nu' \in \mathcal{B}_\delta(\nu)}} \mathsf{W}_p(\mu',\nu'), \qquad (5)$$

where $\delta = \varepsilon/(1+\varepsilon)$.

While the original definition (1) involves *removing* mass from the base measures and rescaling, (5) is optimizing over mass *added* to μ and ν (up to scaling). Our proof in Appendix A.4 of this somewhat surprising result relies on the symmetric nature of the OT distance objective. Roughly, instead of removing a piece mass from one measure, we may always add it to the other.

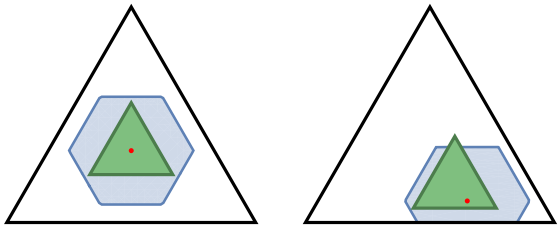

Figure 1: Huber ε-contamination balls (green) and ε-TV balls (blue), centered at distinct points (red) within the 2-dimensional simplex. The Huber balls with different centers are translates of each other, while the rightmost TV ball interacts non-trivially with the simplex boundary.

This reformulation is valuable because the updated constraint sets are simple and do not interact with the simplex boundary. Specifically, definition (3) reveals that the Huber ball $\mathcal{B}_\delta(\mu)$ is always an affine shift of $\mathcal{P}(\mathcal{X})$, with scaling independent of μ. Hence, linear optimization over $\mathcal{B}_\delta(\mu)$ is straightforward:

$$\inf_{\mu' \in \mathcal{B}_\delta(\mu)} \int f \,\mathrm{d}\mu' = (1-\varepsilon) \int f \,\mathrm{d}\mu + \delta \inf_{x \in \mathcal{X}} f(x). \qquad (6)$$

The above stands in contrast to the TV balls (i.e., sets of the form $\{\mu' \in \mathcal{P}(\mathcal{X}) : \|\mu' - \mu\|_{\mathsf{TV}} \le \delta\}$) that appear in existing robust OT formulations; these exhibit non-trivial boundary interactions as depicted in Fig. 1. Fortunately, W_p is closely tied to the linear form $\mu \mapsto \int f \,\mathrm{d}\mu$ via Kantorovich duality—a cornerstone for various theoretical derivations and practical implementations. Combining this with a minimax result, we establish a related dual form for W_p^ε.

Theorem 2 (Dual form). *For $p \in [1,\infty)$, $\varepsilon \in [0,1]$, and $\mu,\nu \in \mathcal{P}(\mathcal{X})$, we have*

$$\begin{aligned}(1-\varepsilon)\mathsf{W}_p^\varepsilon(\mu,\nu)^p &= \sup_{f \in C_b(\mathcal{X})} \int f \,\mathrm{d}\mu + \int f^c \,\mathrm{d}\nu - 2\varepsilon\|f\|_\infty \\ &= \sup_{f \in C_b(\mathcal{X})} \int f \,\mathrm{d}\mu + \int f^c \,\mathrm{d}\nu - \varepsilon \,\mathrm{Range}(f), \qquad (7)\end{aligned}$$

and the suprema are achieved by $f \in C_b(\mathcal{X})$ with $f = (f^c)^c$.

This new formulation differs from the classic dual (2) by a range penalty for the potential function. When $p = 1$, we have $f^c = -f$ and $f = (f^c)^c$ exactly when f is 1-Lipschitz. The theorem is proven in Appendix A.5, where we first apply Proposition 1 and then invoke Kantorovich duality for $\mathsf{W}_p(\mu',\nu')$, while verifying that the conditions for Sion's minimax theorem hold true.

Applying minimax gives

$$\inf_{\substack{\mu'\in\mathcal{B}_\delta(\mu)\\ \nu'\in\mathcal{B}_\delta(\nu)}} \mathsf{W}_p(\mu',\nu') = \sup_{f\in C_b(\mathcal{X})}\left[\inf_{\mu'\in\mathcal{B}_\delta(\mu)}\int f\,\mathrm{d}\mu' + \inf_{\nu'\in\mathcal{B}_\delta(\nu)}\int f^c\,\mathrm{d}\nu'\right],$$

where $\delta = \varepsilon/(1+\varepsilon)$. At this point, we employ (6), along with properties of the c-transform, to obtain the desired dual. We stress that if μ' and ν' instead varied within TV balls, the inner minimization problems would not admit closed forms due to boundary interactions.

Theorem 2 reveals an elementary procedure for robustifying the Wasserstein distance against outliers: regularize its standard Kantorovich dual w.r.t. the sup-norm of the potential function. The simplicity of this modification is its main strength. As demonstrated in Section 6, this enables adjusting popular duality-based OT solvers, e.g., Arjovsky et al. (2017), to the robust framework and opens the door for applications to generative modeling with contaminated datasets. We provide an interpretation for the maximizing potentials in Section 4. Some concrete examples for computing W_p^ε are found in Appendix B.3.

Remark 4 (TV as a dual norm)**.** Recall that $\|\cdot\|_{\mathsf{TV}}$ is the dual norm corresponding to the Banach space of measurable functions on $\mathcal{X}$ equipped with $\|\cdot\|_\infty$. An inspection of the proof of Theorem 2 reveals that our penalty scales with $\|\cdot\|_\infty$ precisely for this reason.

Finally, we describe an alternative dual form which ties robust OT to loss trimming—a popular practical tool for robustifying estimation algorithms when μ and ν have finite support (Shen and Sanghavi, 2019).

Proposition 2 (Loss trimming dual)**.** *Fix $p\in[1,\infty)$ and take μ,ν to be uniform distributions over n points each. If $\varepsilon\in[0,1]$ is a multiple of $1/n$, then*

$$\mathsf{W}_p^\varepsilon(\mu,\nu)^p = \sup_{f\in C_b(\mathcal{X})}\left(\min_{\substack{\mathcal{A}\subseteq\mathrm{supp}(\mu)\\ |\mathcal{A}|=(1-\varepsilon)n}}\frac{1}{|\mathcal{A}|}\sum_{x\in\mathcal{A}} f(x) + \min_{\substack{\mathcal{B}\subseteq\mathrm{supp}(\nu)\\ |\mathcal{B}|=(1-\varepsilon)n}}\frac{1}{|\mathcal{B}|}\sum_{y\in\mathcal{B}} f^c(y)\right).$$

The inner minimization problems above clip out the εn fraction of samples whose potential evaluations are largest. This is similar to how standard loss trimming clips out a fraction of samples that contribute most to the considered training loss.

4 STRUCTURAL PROPERTIES

We turn to structural properties of W_p^ε, exploring primal and dual optimizers, regularity of W_p^ε in ϵ, as well as an alternative (near) coupling-based primal form.

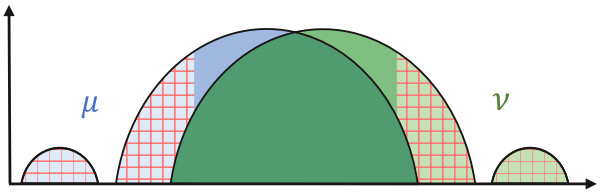

Figure 2: The gridded light blue and green regions each have mass ε, respectively, and are removed to obtain optimal μ' and ν' for W_1^ε. No mass need be removed from the dark region designating $\mu\wedge\nu$.

4.1 Primal and Dual Optimizers

We first prove that there are primal and dual optimizers satisfy certain regularity conditions.

Proposition 3 (Existence of minimizers)**.** *For $p\in[1,\infty]$ and $\mu,\nu\in\mathcal{P}(\mathcal{X})$, the infimum in (1) is achieved, and there are minimizers $\mu'\le\mu$ and $\nu'\le\nu$ such that $\mu',\nu'\ge\mu\wedge\nu$ and $\mu'(\mathcal{X})=\nu'(\mathcal{X})=1-\varepsilon$.*

We remark that the lower envelope of $\mu\wedge\nu$, illustrated in Fig. 2, is straightforward when $p=1$, since $\mathsf{W}_1(\mu,\nu)$ is a function of $\mu-\nu$. However, this conclusion is not obvious for $p>1$, and its proof in Appendix A.6 utilizes a discretization argument. For achieving the infimum, we show for $p<\infty$ that the constraint set is compact w.r.t. the classic Wasserstein topology, while the objective is clearly continuous in W_p. For $p=\infty$, we observe that the constraint set is compact and the objective is lower semicontinuous w.r.t. the topology of weak convergence over $\mathcal{P}_\infty(\mathcal{X})$.

Proposition 4 (Interpreting maximizers)**.** *If $f\in C_b(\mathcal{X})$ maximizes (7), then any $\mu',\nu'\in\mathcal{M}_+(\mathcal{X})$ minimizing (1) satisfy $\mathrm{supp}(\mu-\mu')\subseteq\mathrm{argmax}(f)$ and $\mathrm{supp}(\nu-\nu')\subseteq\mathrm{argmin}(f)$.*

Thus, the level sets of the dual potential encode the location of outliers in the original measures, as depicted in Fig. 3. In fact, optimal perturbations $\mu-\mu'$ and $\nu-\nu'$ are sometimes determined exactly by an optimal potential f, often taking the form $\mu|_{\mathrm{argmax}(f)}$ and $\nu|_{\mathrm{argmax}(f)}$ (though not always; we discuss this in Appendix A.11 along with the proof).

4.2 Regularity in Robustness Radius

We examine how W_p^ε depends on the robustness radius.

Proposition 5 (Dependence on ε)**.** *For any $p\in[1,\infty]$, $0\le\varepsilon\le\varepsilon'\le1$, and $\mu,\nu\in\mathcal{P}(\mathcal{X})$, we have*

(i) $\mathsf{W}_p^{\|\mu-\nu\|_{\mathsf{TV}}/2}(\mu,\nu)=0$, $\mathsf{W}_p^0(\mu,\nu)=\mathsf{W}_p(\mu,\nu)$;

(ii) $\mathsf{W}_p^{\varepsilon'}(\mu,\nu)\le\mathsf{W}_p^\varepsilon(\mu,\nu)$;

(iii) $\mathsf{W}_p^\varepsilon(\mu,\nu)\le\left(\frac{1-\varepsilon'}{1-\varepsilon}\right)^{\frac{1}{p}}\mathsf{W}_p^{\varepsilon'}(\mu,\nu)+4\,\mathrm{diam}(\mathcal{X})\left(\frac{\varepsilon'-\varepsilon}{1-\varepsilon}\right)^{\frac{1}{p}}$.

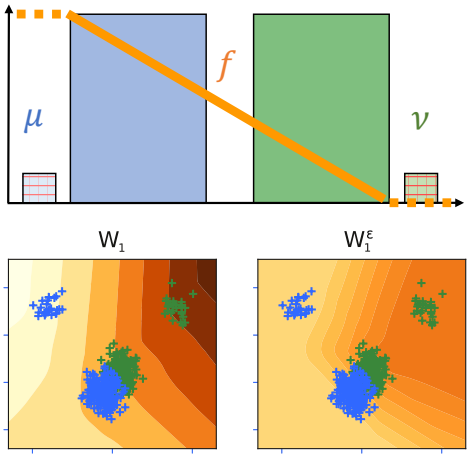

Figure 3: Optimal potentials: (top) 1D densities plotted with their optimal potential for the W_1^ε dual problem; (bottom) contour plots for optimal dual potentials to W_1 and W_1^ε between 2D Gaussian mixtures. Observe how optimal potentials for the robust dual are flat over outlier mass.

The proof is given in Appendix A.7. More precise, diameter independent, bounds in the form of (iv) are provided in the proofs of the robustness results from Section 2, but these require μ and ν to satisfy certain moment bounds that do not hold in general.

4.3 Alternative Primal Form

Mirroring the primal Kantorovich problem for classic W_p, we derive in Appendix A.8 an alternative primal form for W_p^ε in terms of (near) couplings for μ and ν.

Proposition 6 (Alternative primal form)**.** *For any $p \in [1,\infty]$ and $\mu,\nu \in \mathcal{P}(\mathcal{X})$, we have*

$$\mathsf{W}_p^\varepsilon(\mu,\nu) = \inf_{\substack{\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{X}) \\ \mu\in\mathcal{B}_\varepsilon(\pi_1),\,\nu\in\mathcal{B}_\varepsilon(\pi_2)}} \|d\|_{L^p(\pi)}, \tag{8}$$

where π_1 and π_2 are the respective marginals of π.

Remark 5 (Data privacy)**.** From this, we deduce that $\mathsf{W}_\infty^\varepsilon(\mu,\nu) \le M$ if and only if there exists a coupling (X,Y) for μ,ν such that $|X-Y| \le M$ with probability at least $1-\Theta(\varepsilon)$. In Appendix C, we state this more precisely and leverage this fact for an application to data privacy. Specifically, within the Pufferfish privacy framework, the so-called Wasserstein Mechanism (Song et al., 2017) maximizes W_∞ over certain pairs of distributions to provide a strong privacy guarantee. By substituting W_∞ with $\mathsf{W}_\infty^\varepsilon$, we reach an alternative mechanism that satisfies a slightly relaxed guarantee and can introduce significantly less noise.

Proposition 5 implies that $\lim_{\varepsilon\to 0}\mathsf{W}_p^\varepsilon = \mathsf{W}_p$, posing W_p^ε as a natural extension of W_p. Given the representation in Proposition 6, we now ask whether convergence of optimal (near) couplings also holds. A proof in Appendix A.9 provides an affirmative answer via a Γ-convergence argument.

Proposition 7 (Convergence of couplings)**.** *Fix $p \in [1,\infty]$ and $\mu,\nu \in \mathcal{P}(\mathcal{X})$. If $\varepsilon_n \searrow 0$ and $\pi_n \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$ is optimal for $\mathsf{W}_p^{\varepsilon_n}(\mu,\nu)$ via (8), for each $n\in\mathbb{N}$, then $\{\pi_n\}_{n\in\mathbb{N}}$ admits a subsequence converging weakly to an optimal coupling for $\mathsf{W}_p(\mu,\nu)$.*

Finally, we consider a case of practical importance: the discrete setting where μ and ν have finite supports. Like classic OT, computing $\mathsf{W}_p^\varepsilon(\mu,\nu)$ between discrete measures amounts to a linear program for $p<\infty$ and can be solved in polynomial time. The proof in Appendix A.10 starts from the alternative primal form of Proposition 6 and analyzes the feasible polytope when the support sizes are equal.

Proposition 8 (Finite support)**.** *Let μ and ν be uniform discrete measures over n points each. Then there exist optimal μ',ν' for $\mathsf{W}_p^\varepsilon(\mu,\nu)$ such that μ' and ν' each give mass $1/n$ to $\lfloor(1-\varepsilon)n\rfloor$ points and assign their remaining $\lceil\varepsilon n\rceil/n-\varepsilon$ mass to a single point.*

When ε is a multiple of $1/n$, the propositions says that there exist minimizers which assign equal mass to $(1-\varepsilon)n$ points, while eliminating the remaining εn that are identified as outliers.

5 STATISTICAL ANALYSIS

We now examine estimation of W_p^ε from observed data. Throughout this section, we fix $\varepsilon \le 1/2$.

5.1 Empirical Convergence Rates

In practice, we often have access only to samples from $\mu,\nu \in \mathcal{P}(\mathcal{X})$, which motivates the study of empirical convergence under W_p^ε. Consider the empirical measures $\hat\mu_n := n^{-1}\sum_{i=1}^n \delta_{X_i}$ and $\hat\nu_n := n^{-1}\sum_{i=1}^n \delta_{Y_i}$, where $X_1,\dots,X_n$ and $Y_1,\dots,Y_n$ are i.i.d. samples from μ and ν, respectively. We examine both the one- and two-sample scenarios, i.e., the speed at which $\mathsf{W}_p^\varepsilon(\mu,\hat\mu_n)$ and $\left|\mathsf{W}_p^\varepsilon(\hat\mu_n,\hat\nu_n) - \mathsf{W}_p^\varepsilon(\mu,\nu)\right|$ converge to 0 as n grows.

This section assumes that $\mathcal{X}\subset\mathbb{R}^d$ with $d>2$; extensions to the non-Euclidean case are provided in the Appendices A.13 to A.15. To state the results, we need some definitions. Recall the covering number $\mathcal{N}_\delta(\gamma,\tau)$, defined for $\delta>0, \tau\ge 0$, and $\gamma\in\mathcal{M}_+(\mathcal{X})$ as the minimum number of closed balls of radius δ needed to cover $\mathcal{X}$ up to a set A with $\gamma(A)\le\tau$. We define the lower τ-covering dimension for γ as

$$d_*^\tau(\mu) := \liminf_{\delta\to 0} \frac{\mathcal{N}_\delta(\gamma,\tau)}{-\log(\delta)}.$$

We note that lower bounds for standard W_p depend on the lower Wasserstein dimension given by $\lim_{\tau\to 0} d_*^\tau(\mu)$ (Weed and Bach, 2019). To provide meaningful bounds for W_p^ε, on the other hand, we require control of $d_*^\tau(\mu)$ for some $\tau > \varepsilon$, which can be understood as a robust notion of dimension.

Proposition 9 (One-sample rates). *Fix $p \in [1,\infty]$ and $\tau \in (\varepsilon, 1]$. If $p < d/2$, we have*

$$\mathbb{E}\left[\mathsf{W}_p^\varepsilon(\mu, \hat{\mu}_n)\right] \le C n^{-1/d}$$

for a constant C independent of n and ε. Furthermore, if $d_*^\tau(\mu) > s$, then

$$\mathsf{W}_p^\varepsilon(\mu, \gamma) \ge C' (\tau - \varepsilon)^{1/p} n^{-1/s}$$

for any γ supported on at most n points, including $\hat{\mu}_n$, where $C' > 0$ is an absolute constant.

The robust τ-covering dimension can be smaller than standard notions of intrinsic dimension when all but the outlier mass is lower dimensional. Our lower bound captures this fact. The upper bound depends on the ambient dimension d since there is no guarantee that only outlier mass is supported on a high-dimensional set[4]. Indeed, the following corollary shows that when a significant portion of mass is high-dimensional, the $n^{-1/d}$ rate is sharp.

Corollary 2 (Simple one-sample rate). *Fix $p < d/2$ and $\varepsilon \in (0, 1]$. If μ has absolutely continuous part $f\,\mathrm{d}\lambda$ with $\int f\,\mathrm{d}\lambda > \varepsilon$ and f bounded from above, then $d_*^\tau(\mu) = d$ and $\mathbb{E}\left[\mathsf{W}_p^\varepsilon(\mu, \hat{\mu}_n)\right] = \Theta(n^{-1/d})$.*

In words, if more than the ε mass that can be removed via robustification is absolutely continuous (w.r.t. Lebesgue on $\mathbb{R}^d$), then the standard $n^{-1/d}$ "curse of dimensionality" applies. Nevertheless, we conjecture that an upper bound that depends on a robust upper dimension can be derived under appropriate assumptions on μ, as discussed in Section 7. The proofs of Proposition 9 and Corollary 2 are found in Appendices A.13 and A.14, respectively.

Moving to the two-sample regime, we again have an upper bound which matches the standard rate for W_p.

Proposition 10 (Two-sample rate). *For any $\varepsilon \in [0, 1]$ and $p < d/2$, we have*

$$\mathbb{E}\left[\left|\mathsf{W}_p^\varepsilon(\hat{\mu}_n, \hat{\nu}_n)^p - \mathsf{W}_p^\varepsilon(\mu, \nu)^p\right|\right] \le C n^{-p/d}.$$

for a constant C independent of n and ε.

[4] Our upper bound holds if we substitute d with anything greater than upper Wasserstein dimension—another notion of intrinsic dimensionality defined in Weed and Bach (2019).

The proof in Appendix A.15 is a consequence of the dual form from Theorem 2, combined with standard one-sample rates for W_p. There, we also discuss obstacles to extending two-sample lower bounds for standard W_p to the robust setting.

5.2 Additional Robustness Guarantees

Finally, we provide conditions under which $\mathsf{W}_p(\mu, \nu)$ can be recovered precisely from $\mathsf{W}_p^\varepsilon(\mu, \nu)$ despite data contamination. Naturally, these conditions are stronger than those needed for approximate (minimax optimal) robust approximation, as studied in Section 2. We return to the Huber ε-contamination model, taking $\tilde{\mu} \in \mathcal{B}_\varepsilon(\mu)$ and $\tilde{\nu} \in \mathcal{B}_\varepsilon(\nu)$. One cannot hope to achieve exact recovery via W_p^ε in general, since $\mathsf{W}_p^\varepsilon(\tilde{\mu}, \tilde{\nu}) < \mathsf{W}_p(\mu, \nu)$ when $\tilde{\mu} = \mu$ and $\tilde{\nu} = \nu$ for $p < \infty$. Nevertheless, we can provide exact recovery guarantees under appropriate mass separation assumptions.

Proposition 11 (Exact recovery). *Fix $\mu, \nu \in \mathcal{P}(\mathcal{X})$, $\varepsilon \in [0, 1]$, and suppose that $\tilde{\mu} = (1-\varepsilon)\mu + \varepsilon\alpha$ and $\tilde{\nu} = (1-\varepsilon)\nu + \varepsilon\beta$, for some $\alpha, \beta \in \mathcal{P}(\mathcal{X})$. Let $S = \operatorname{supp}(\mu + \nu)$. If $d(\operatorname{supp}(\alpha), S)$, $d(\operatorname{supp}(\beta), S)$, and $d(\operatorname{supp}(\alpha), \operatorname{supp}(\beta))$ are all greater than $\operatorname{diam}(S)$, then $\mathsf{W}_p^\varepsilon(\tilde{\mu}, \tilde{\nu}) = \mathsf{W}_p(\mu, \nu)$.*

Our proof in Appendix A.16 uses an infinitesimal perturbation argument and shows that, when outliers are sufficiently far away, removing outlier mass is strictly better than inlier mass for minimizing W_p. The appendix also discussed more flexible albeit technical assumptions under which exact recovery is possible, and provides bounds for when the robustness radius does not match the contamination level ε exactly.

Proposition 11 relies on the contamination level ε being known, which is often not the case in practice. To account for this, we prove in Appendix A.17 that, under the same assumptions, there exists a principled approach for selecting the robustness radius when ε is unknown.

Proposition 12 (Robustness radius for unknown ε). *Assume the setting of Proposition 11 and let D be the maximum of $d(\operatorname{supp}(\alpha), S)$, $d(\operatorname{supp}(\beta), S)$, and $d(\operatorname{supp}(\alpha), \operatorname{supp}(\beta))$. Then, for $p \in [1, \infty)$, we have*

$$\tfrac{\mathrm{d}}{\mathrm{d}\delta}\left[(1-\delta)\mathsf{W}_p^\delta(\tilde{\mu}, \tilde{\nu})^p\right] \le -D^p, \qquad \delta \in [0, \varepsilon)$$
$$\tfrac{\mathrm{d}}{\mathrm{d}\delta}\left[(1-\delta)\mathsf{W}_p^\delta(\tilde{\mu}, \tilde{\nu})^p\right] \ge -\operatorname{diam}(S)^p > -D^p, \quad \delta \in (\varepsilon, 1]$$

at the (all but countably many) points where the derivative is defined.

As $(1-\delta)\mathsf{W}_p^\delta(\tilde{\mu}, \tilde{\nu})^p$ is continuous and decreasing in δ by Proposition 5, we have identified an "elbow" in this curve located exactly at the true contamination level ε.

Next, we return to the statistical setting and show how to obtain exact recovery assuming that the fraction of

corrupted samples vanishes as $n \to \infty$. Consider i.i.d. samples $X_1, X_2, \cdots \sim \mu$ and $Y_1, Y_2, \cdots \sim \nu$, arbitrarily corrupted to obtain $\tilde{X}_1, \tilde{X}_2, \ldots$ and $\tilde{Y}_1, \tilde{Y}_2, \ldots$. We measure the level of corruption by the rates $\tau_\mu(n) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \neq \tilde{X}_i\}}$ and $\tau_\nu(n) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Y_i \neq \tilde{Y}_i\}}$. For this result, $\mathcal{X}$ may be unbounded, and only the clean distributions μ and ν need compact support. Let $\tilde{\mu}_n$ and $\tilde{\nu}_n$ denote the empirical measures associated with $\tilde{X}_1, \ldots, \tilde{X}_n$ and $\tilde{Y}_1, \ldots, \tilde{Y}_n$, respectively.

Proposition 13 (Robust consistency)**.** *Fix $p < \infty$ and suppose that $\tau_\mu(n) \vee \tau_\nu(n) = O(n^{-a})$ almost surely, for some $a > 0$. Then, setting $\varepsilon_n = n^{-b}$, for any $0 < b < a$, we have*

$$\left|\mathsf{W}_p(\mu, \nu) - \mathsf{W}_p^{\varepsilon_n}(\tilde{\mu}_n, \tilde{\nu}_n)\right| \to 0$$

as $n \to \infty$, almost surely.

That is, so long as the fraction of (potentially unbounded) corrupted samples converges to 0, there exists a schedule for robustness radii so that the correct distance is recovered. The proof in Appendix A.18 uses ideas from the previous result to alleviate potential problems arising from large corruptions.

Remark 6 (Median-of-means)**.** This consistency result mirrors that presented by Staerman et al. (2021) for a median-of-means (MoM) estimator $\mathcal{W}_{\text{MoM}}$. They produce a robust estimate for W_1 by partitioning the contaminated samples into blocks and replacing the each mean appearing in the W_1 dual with a median of block means, where the number of blocks depends on the contamination fractions. We remark that when $\tau_\mu \vee \tau_\nu = \Omega(1)$ and contaminations are stochastic—i.e., each $\tilde{X}_i$ and $\tilde{Y}_i$ are sampled from some $\tilde{\mu} \in \mathcal{B}_\varepsilon(\mu)$ and $\tilde{\nu} \in \mathcal{B}_{\varepsilon'}(\nu)$, respectively—we have $\mathcal{W}_{\text{MoM}}(\tilde{\mu}_n, \tilde{\nu}_n) \to \mathsf{W}_1(\tilde{\mu}, \tilde{\nu})$. Hence, this approach cannot provide guarantees in the vein of Section 2, since it may be that $\mathsf{W}_1(\tilde{\mu}, \tilde{\nu}) \gg \mathsf{W}_1(\mu, \nu)$.

6 APPLICATIONS

We now move to applications of the proposed robust OT framework. We first discuss computational aspects and provide an algorithm to compute W_p^ε based on its dual form. The algorithm, which requires only a minor modification to standard OT solvers, is then used for generative modeling from contaminated data.

6.1 Computation

In practice, similarity between datasets is often measured using the so-called neural network (NN) distance $\mathsf{d}_{\mathcal{F}}(\mu, \nu) = \sup_{\theta \in \Theta} \int f_\theta \, \mathrm{d}\mu - \int f_\theta \, \mathrm{d}\nu$, where $\mathcal{F} = \{f_\theta\}_{\theta \in \Theta}$ is a NN class (Arora et al., 2017). Given two batches of samples $X_1, \ldots, X_n \sim \mu$ and $Y_1, \ldots, Y_n \sim \nu$, we approximate integrals by sample means and estimate the supremum via stochastic gradient ascent. Namely, we follow the update rule

$$\theta_{t+1} \leftarrow \theta_t + \frac{\alpha_t}{n}\sum_{i=1}^n \left[\nabla_\theta f_\theta(X_i) - \nabla_\theta f_\theta(Y_i)\right].$$

When $\mathcal{F}$ approximates the class of 1-Lipschitz functions, we approach a Kantorovich dual and obtain an estimate for $\mathsf{W}_1(\mu, \nu)$, which is the core idea behind the WGAN (Arjovsky et al., 2017) (see Makkuva et al. (2020) for an extension to W_2).

By virtue of our duality theory, OT solvers as described above can be easily adapted to the robust framework. For the purpose of generative modeling, the one-sided robust distance defined by

$$\mathsf{W}_p^\varepsilon(\mu \| \nu) := \inf_{\substack{0 \le \mu' \le \mu \\ \mu'(\mathcal{X}) = 1-\varepsilon}} \mathsf{W}_p\left(\frac{\mu'}{1-\varepsilon}, \nu\right) \tag{9}$$

is most appropriate, since data may be contaminated but generated samples from the model are not. In Appendix B.1, we translate our duality result to this setting, finding that

$$\begin{aligned}(1-\varepsilon)\mathsf{W}_p^\varepsilon(\mu\|\nu)^p \\ = \sup_{f \in C_b(\mathcal{X})} \int f \, \mathrm{d}\mu + (1-\varepsilon)\int f^c \, \mathrm{d}\nu - \varepsilon \sup_{x \in \mathcal{X}}(f).\end{aligned} \tag{10}$$

This representation motivates a modified gradient update for the corresponding NN distance[5] estimate:

$$\begin{aligned}\theta_{t+1} \leftarrow \theta_t + \frac{\alpha_t}{n}\sum_{i=1}^n \left[\nabla_\theta f_\theta(X_i) - (1-\varepsilon)\nabla_\theta f_\theta(Y_i)\right] \\ - \varepsilon \nabla_\theta f_\theta(X_{i^*}), \quad i^* \in \operatorname{argmax}_i f_\theta(X_i)\end{aligned}$$

(note here that we are formally computing Clarke subgradients (Clarke, 1990)). For example, in a `PyTorch` implementation of WGAN, the modified update can be implemented with a one-line adjustment of code:

```
score = f_data.mean() - f_generated.mean() # old
score = f_data.mean() - (1-eps)*f_generated.mean()
        - eps*f_data.max() # new
```

Due to the non-convex and non-smooth nature of the objective, formal optimization guarantees seem challenging to obtain, and we defer this exploration for future work. Nevertheless, as we will see next, this approach proves quite fruitful in practice. Full experimental details for the following results and comparisons with existing work are provided in Appendix B.4, and code is provided at https://github.com/sbnietert/robust-OT. Computations were performed on a cluster machine equipped with a NVIDIA Tesla V100.

[5] An inspection of the proof of Theorem 2 reveals that a similar duality holds for NN distances, and, generally, IPMs

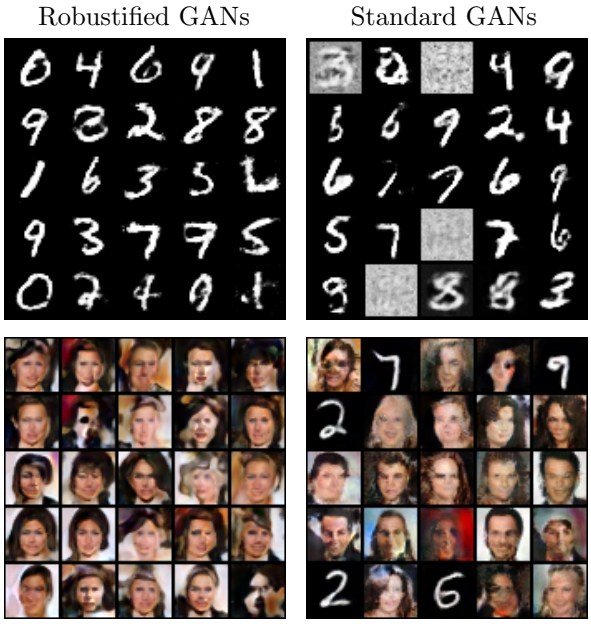

Figure 4: Top: samples generated by robustified (left) and standard (right) WGAN-GP after training on corrupted MNIST dataset. Bottom: samples generated by robustified (left) and standard (right) StyleGAN 2 after training on corrupted CelebA-HQ dataset (left).

6.2 Generative Modeling

We examine outlier-robust generative modeling using the modification suggested above. We train two WGAN with gradient penalty (WGAN-GP) models (Gulrajani et al., 2017) on a contaminated dataset with 80% MNIST data and 20% random noise, running both with a standard selection of hyper-parameters but adjusting one to compute gradient updates according to the robust objective with a selection of $\varepsilon = 0.25$. In Fig. 4 (top), we display generated samples produced by both networks after processing 125k contaminated batches of 64 samples. The effect of outliers is clearly mitigated by training with the robustified objective.

One subtlety to the above is that WGAN-GP employs a regularizer $R(f_\theta)$ penalizing large gradients, rather than enforcing a Lipschitz constraint. In Appendix B.5, we provide results demonstrating that our duality result still holds in the presence of many types of regularizers. This further motivates us to apply the robustness technique to more sophisticated GANs which incorporate additional regularization, in particular StyleGAN 2 (Karras et al., 2020). We again train two off-the-shelf models using contaminated data—this time, 80% CelebA-HQ face photos and 20% MNIST data—with one tweaked to perform gradient updates according to the robust objective with $\varepsilon = 0.25$. We present generated samples in Fig. 4 (bottom). Once again, the modified objective enables learning a model that is largely free of outliers despite being trained on a contaminated dataset.

7 CONCLUDING REMARKS

This paper introduced the outlier-robust Wasserstein distance W_p^ε, which measures proximity between probability distributions while discarding an ε-fraction of outlier mass. We conducted a theoretical study of its structural and statistical properties, covering robustness guarantees, strong duality, characteristics of optimal perturbations and dual potentials, regularity in ε, and empirical convergence. The derived dual form amounts to a simple modification of classic Kantorovich duality that regularizes the objective w.r.t. the sup-norm of the potential function. This gave rise an elementary robustification technique for duality-based OT solvers (by introducing said penalty), which enables adapting computational methods for classic W_p to compute W_p^ε. Leveraging this, we demonstrated the utility of W_p^ε for generative modeling with contaminated data.

Future research directions are abundant, both theoretical and practical. First, the derived duality can be leveraged for many high-dimensional real-world inference tasks where outlier-robustness is desired, although a large-scale empirical exploration is beyond the scope of this work. We also aim to sharpen our statistical bounds and provide finite-sample robustness guarantees. In the one-sample case, we expect a tighter upper bound that depends on an upper (robust) intrinsic dimension to hold when only a small amount of high-dimensional mass is present. Indeed, high-dimensional regions are harder to sample and we therefore expect W_p^ε to treat those as outliers in the empirical approximation setting. The following generalizations of the considered robust framework are also of interest: (i) general transportation costs (many of our structural results immediately generalize); (ii) unbounded domains; and (iii) other base and constraining distances (e.g., IPMs, NN distances, etc.).

Acknowledgements

The authors would like to thank Benjamin Grimmer for helpful conversations surrounding minimax optimization, as well as Jacob Steinhardt and Adam Sealfon for useful discussions on robust statistics. S. Nietert was supported by the National Science Foundation (NSF) Graduate Research Fellowship under Grant DGE-1650441. R. Cummings was supported in part by NSF grants CNS-1850187 and CNS-1942772 (CAREER), a Mozilla Research Grant, and a JPMorgan Chase Faculty Research Award. Z. Goldfeld was supported by NSF grants CCF-1947801 and CCF-2046018 (CAREER), and the 2020 IBM Academic Award.

References

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017.

S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning (ICML)*, 2017.

Y. Balaji, R. Chellappa, and S. Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

F. Barthe and C. Bordenave. Combinatorial optimization over two random point sets. In *Séminaire de Probabilités XLV*, volume 2078 of *Lecture Notes in Math*, pages 483–535. Springer, Cham, 2013.

J. Blanchet, K. Murthy, and F. Zhang. Optimal transport based distributionally robust optimization: Structural properties and iterative schemes. *arXiv preprint arXiv:1810.02403*, 2018.

L. A. Caffarelli and R. J. McCann. Free boundaries in optimal transport and Monge-Ampère obstacle problems. *Annals of Mathematics. Second Series*, 171(2):673–730, 2010.

L. Cao and Z. Chen. Partitions of the polytope of doubly substochastic matrices. *Linear Algebra and its Applications*, 563:98–122, 2019.

M. Cao. WGAN-GP. https://github.com/caogang/wgan-gp, 2017.

L. Chapel, M. Z. Alaya, and G. Gasso. Partial optimal tranport with applications on positive-unlabeled learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

M. Chen, C. Gao, and Z. Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932 – 1960, 2018.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018a.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314): 2563–2609, 2018b.

F. H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990.

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2014.

N. Courty, R. Flamary, D., Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2016.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, 2006.

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

K. Fatras, T. Sejourne, R. Flamary, and N. Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning (ICML)*, 2021.

A. Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195: 533–560, 2010.

T. Fukunaga and H. Kasai. Fast block-coordinate frank-wolfe algorithm for semi-relaxed optimal transport. *arXiv preprint arXiv:2103.05857*, 2021.

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

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

L. G. Hanin. Kantorovich-rubinstein norm and its application in the theory of lipschitz spaces. *Proceedings of the American Mathematical Society*, 115(2): 345–352, 1992.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1): 73–101, 1964.

T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

D. Kifer and A. Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.*, 39(1):3:1–3:36, 2014.

K. Le, H. Nguyen, Q. Nguyen, N. Ho, T. Pham, and H. Bui. On robust optimal transport: Computational complexity, low-rank approximation, and barycenter computation. *arXiv preprint arXiv:2102.06857*, 2021.

M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

A. V. Makkuva, A. Taghvaei, S. Oh, and J. D. Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning (ICML)*, 2020.

G. D. Maso. *An introduction to* Γ*-convergence*, volume 8. Springer Science & Business Media, 2012.

D. Mukherjee, A. Guha, J. Solomon, Y. Sun, and M. Yurochkin. Outlier-robust optimal transport. In *International Conference on Machine Learning (ICML)*, 2021.

J. S. Nath. Unbalanced optimal transport using integral probability metric regularization. *arXiv preprint arXiv:2011.05001*, 2020.

B. Piccoli and F. Rossi. Generalized Wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014.

F. Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

B. Schmitzer and B. Wirth. A framework for Wasserstein-1-type metrics. *Journal of Convex Analysis*, 26(2):353–396, 2019.

Y. Shen and S. Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning (ICML)*, 2019.

S. Song, Y. Wang, and K. Chaudhuri. Pufferfish privacy mechanisms for correlated data. In *ACM International Conference on Management of Data (SIGMOD)*, 2017.

G. Staerman, P. Laforgue, P. Mozharovskyi, and F. d'Alché-Buc. When OT meets MoM: robust estimation of Wasserstein distance. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

M. Talagrand. Matching random samples in many dimensions. *The Annals of Applied Probability*, 2(4):846–856, 1992. ISSN 1050-5164. URL http://links.jstor.org.proxy.library.cornell.edu/sici?sici=1050-5164(199211)2:4<846:MRSIMD>2.0.CO;2-Y&origin=MSN.

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR)*, 2018.

N. W. Varuna Jayasiri. labml.ai annotated paper implementations, 2020. URL https://nn.labml.ai/.

C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, 2003.

C. Villani. *Optimal transport: Old and new*, volume 338. Springer-Verlag, Berlin, 2009.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

W. Zhang, O. Ohrimenko, and R. Cummings. Attribute privacy: Framework and mechanisms. *arXiv preprint arXiv:2009.04013*, 2020.

Appendix A Proofs of Main Results

We begin with some further preliminaries and notation. When convenient, we write $\mathbb{E}_\mu[f(X)]$ for the expectation of $f(X)$ with $X \sim \mu$. Next, we recall the general definition of W_p between non-negative measures of equal (but potentially not unit) mass. Given two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ with $\mu(\mathcal{X}) = \nu(\mathcal{X})$, let $\Pi(\mu,\nu) = \{\mu(\mathcal{X})\pi : \pi \in \Pi(\mu/\mu(\mathcal{X}), \nu/\nu(\mathcal{X}))\}$, and, for $p \in [1,\infty]$, define

$$\mathsf{W}_p(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \|d\|_{L^p(\pi)}.$$

We further recall the two-potential version of Kantorovich duality, which states that for $p \in [1,\infty)$,

$$\mathsf{W}_p(\mu,\nu)^p = \sup_{\substack{f,g \in C_b(\mathcal{X}) \\ f(x)+g(y) \le d(x,y)^p}} \int f \,\mathrm{d}\mu + \int g \,\mathrm{d}\nu. \tag{11}$$

Finally, we define the push-forward of a measure $\mu \in \mathcal{M}_+(\mathcal{X})$ w.r.t. a measurable map $T : \mathcal{X} \to \mathcal{Y}$ by $T_\# \mu(B) := \mu(T^{-1}(B))$ for all measurable $B \subseteq \mathcal{Y}$.

A.1 Preliminary Results

We begin with a useful fact and lemma.

Fact 1. *For $p \in [1,\infty]$, $\varepsilon \in [0,1]$, and $\mu, \nu \in \mathcal{P}(\mathcal{X})$, we have $\mathsf{W}_p(\mu,\nu) \le \mathsf{W}_p(\mu - \mu \wedge \nu, \nu - \mu \wedge \nu)$.*

Proof. For any feasible coupling $\pi \in \Pi(\mu - \mu \wedge \nu, \nu - \mu \wedge \nu)$ for the second distance, we have that $\pi' = \pi + (\mathrm{Id}, \mathrm{Id})_\#(\mu \wedge \nu) \in \Pi(\mu,\nu)$ is feasible for the first distance with $\|d\|_{L^p(\pi')} = \|d\|_{L^p(\pi)}$. □

The following is a helpful rewriting of the W_p^ε primal problem.

Lemma 1. *For $p \in [1,\infty]$, $\varepsilon \in [0,1]$, and $\mu, \nu \in \mathcal{P}(\mathcal{X})$, we have*

$$\mathsf{W}_p^\varepsilon(\mu,\nu) = \inf_{\substack{\mu',\nu' \in \mathcal{P}(\mathcal{X}): \\ \mu \in \mathcal{B}_\varepsilon(\mu'),\, \nu \in \mathcal{B}_\varepsilon(\nu')}} \mathsf{W}_p(\mu',\nu'). \tag{12}$$

Proof. Here, we use that optimal perturbed measures for the original primal (1) may be taken to have mass exactly $1-\varepsilon$ (since any feasible measures may be scaled down until this is the case, without changing the original objective due to its normalization). We further observe that $\mu \in \mathcal{B}_\varepsilon(\mu')$ if and only if $(1-\varepsilon)\mu' \le \mu$. □

A.2 Proof of Theorem 1

The next lemma is the key to our upper bound, as it will allow us to provide uniform bounds on the estimation error without considering the contaminated measures.

Lemma 2. *Fix $p \in [1,\infty]$ and $\varepsilon \le 1/4$. Take $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and let $\tilde{\mu} \in \mathcal{B}_\varepsilon(\mu), \tilde{\nu} \in \mathcal{B}_\varepsilon(\nu)$. Then, we have*

$$\mathsf{W}_p^{2\varepsilon/(1-\varepsilon)}(\mu,\nu) \le \mathsf{W}_p^\varepsilon(\tilde{\mu},\tilde{\nu}) \le \mathsf{W}_p(\mu,\nu).$$

Proof. The second inequality follows directly from Lemma 1. For the other direction, we begin with $\mu_0', \nu_0' \in \mathcal{M}_+(\mathcal{X})$ feasible for the original primal (1) for $\mathsf{W}_p^\varepsilon(\tilde{\mu},\tilde{\nu})$, i.e., $\mu_0' \le \tilde{\mu}$, $\nu_0' \le \tilde{\nu}$ and $\mu_0'(\mathcal{X}), \nu_0'(\mathcal{X}) \ge 1-\varepsilon$. Then, we intersect μ_0' with $(1-\varepsilon)\mu$, intersect ν_0' with $(1-\varepsilon)\nu$, and remove up to ε additional mass from each as needed to obtain $\mu_1' \le (1-\varepsilon)\mu$ and $\nu_1' \le (1-\varepsilon)\nu$ with equal mass such that $\mathsf{W}_p(\mu_1',\nu_1') \le \mathsf{W}_p(\mu_0',\nu_0')$ and $\mu_1'(\mathcal{X}) = \nu_1'(\mathcal{X}) \ge 1 - 3\varepsilon$. Dividing both measures by $1-\varepsilon$, we obtain μ_2', ν_2' feasible for $\mathsf{W}_p^{2\varepsilon/(1-\varepsilon)}(\mu,\nu)$ such that

$$\mathsf{W}_p^{2\varepsilon/(1-\varepsilon)}(\mu,\nu) \le \mathsf{W}_p(\mu_2',\nu_2') \le (1-\varepsilon)^{-1/p}\, \mathsf{W}_p(\mu_0',\nu_0').$$

Infimizing over μ_0' and ν_0' gives $\mathsf{W}_p^{2\varepsilon/(1-\varepsilon)}(\mu,\nu) \le \mathsf{W}_p^\varepsilon(\tilde{\mu},\tilde{\nu})$, as desired. □

To complete the argument, it suffices to bound the modulus of continuity

$$\mathfrak{m}^+\left(\mathcal{D}_q, \tau\right) := \sup_{\kappa \in \mathcal{D}_q} \sup_{\substack{\kappa' \in \mathcal{P}(\mathcal{X}): \\ \kappa \in \mathcal{B}_\tau(\kappa')}} \mathsf{W}_p(\kappa, \kappa').$$

Indeed, taking $\mu, \nu, \tilde{\mu}, \tilde{\nu}$ as in the theorem and setting $\tau = 2\varepsilon/(1-\varepsilon) \leq \frac{2}{5}$, we have

$$\begin{aligned}
|\mathsf{W}_p^\varepsilon(\tilde{\mu}, \tilde{\nu}) - \mathsf{W}_p(\mu, \nu)| &\leq \mathsf{W}_p(\mu, \nu) - \mathsf{W}_p^\tau(\mu, \nu) \\
&= \sup_{\substack{\mu' \colon \mu \in \mathcal{B}_\tau(\mu') \\ \nu' \colon \nu \in \mathcal{B}_\tau(\nu')}} |\mathsf{W}_p(\mu, \nu) - \mathsf{W}_p(\mu', \nu')| \\
&\leq \sup_{\mu' \colon \mu \in \mathcal{B}_\tau(\mu')} \mathsf{W}_p(\mu, \mu') + \sup_{\nu' \colon \nu \in \mathcal{B}_\tau(\nu')} \mathsf{W}_p(\nu, \nu') \\
&\leq 2\mathfrak{m}^+\left(\mathcal{D}_q, \tau\right).
\end{aligned}$$

Hence, we conclude the upper bound with the following lemma.

Lemma 3. *For any $q > p$ and $\tau \in [0, 1/2]$, we have*

$$\mathfrak{m}^+(\mathcal{D}_q, \tau) \lesssim M\tau^{1/p - 1/q}.$$

Proof. Without loss of generality, we suppose that $M = 1$ (otherwise, one can always apply the lemma to the same metric space with distances — and hence OT measurements — shrunk by a factor of M). To control this modulus, fix any $\kappa \in \mathcal{D}_q$ and $\kappa' \in \mathcal{P}(\mathcal{X})$ such that $\kappa = (1-\tau)\kappa' + \tau\alpha$ for some $\alpha \in \mathcal{P}(\mathcal{X})$. Next, take $y_\kappa, y_\alpha, y_{\kappa'}$ minimizing $\mathbb{E}_\kappa[d(X, \cdot)^q]$, $\mathbb{E}_\alpha[d(X, \cdot)^q]$, and $\mathbb{E}_{\kappa'}[d(X, \cdot)^q]$, respectively. Then, we compute

$$\begin{aligned}
\mathsf{W}_p(\kappa, \kappa') \leq \tau^{1/p}\, \mathsf{W}_p(\kappa', \alpha) &\leq \tau^{1/p}\, \mathsf{W}_q(\kappa', \alpha) \\
&\leq \tau^{1/p} \left(\mathsf{W}_q(\kappa', \delta_{y_{\kappa'}}) + \mathsf{W}_q(\alpha, \delta_{y_\alpha}) + \mathsf{W}_q(\delta_{y_{\kappa'}}, \delta_{y_\alpha})\right) \\
&= \tau^{1/p} \left(\mathbb{E}_{\kappa'}[d(y_{\kappa'}, X)^q]^{1/q} + \mathbb{E}_\alpha[d(y_\alpha, X)^q]^{1/q} + d(y_{\kappa'}, y_\alpha)\right) \\
&\leq 3\,\tau^{1/p} \left(\mathbb{E}_{\kappa'}[d(y_{\kappa'}, X)^q] + \mathbb{E}_\alpha[d(y_\alpha, X)^q] + d(y_{\kappa'}, y_\alpha)^q\right)^{1/q}.
\end{aligned}$$

Here, the first inequality follows from Fact 1, the second uses that W_p is increasing in p, the third follows from the triangle inequality for W_q, and the final inequality is a consequence of Hölder's inequality. To bound the first two summands of the final expression, we compute

$$\begin{aligned}
1 \geq \mathbb{E}_\kappa[d(y_\kappa, X)^q] &= (1-\tau)\, \mathbb{E}_{\kappa'}[d(y_\kappa, X)^q] + \tau\, \mathbb{E}_\alpha[d(y_\kappa, X)^q] \\
&\geq \tau \left(\mathbb{E}_{\kappa'}[d(y_\kappa, X)^q] + \mathbb{E}_\alpha[d(y_\kappa, X)^q]\right) \\
&\geq \tau \left(\mathbb{E}_{\kappa'}\, d(y_{\kappa'}, X)^q] + \mathbb{E}_\alpha[d(y_\alpha, X)^q]\right),
\end{aligned}$$

where the first inequality uses that $\kappa \in \mathcal{D}_q$, the second inequality uses that $\tau \leq 1/2$, and the last uses the definitions of $y_{\kappa'}$ and y_α. For the final distance term, we begin by writing $x_{\kappa'} = \mathbb{E}_{\kappa'}[X]$ and $x_\alpha = \mathbb{E}_\alpha[X]$ and bounding

$$\begin{aligned}
1 &\geq \tau \left(\mathbb{E}_{\kappa'}\, d(y_{\kappa'}, X)^q] + \mathbb{E}_\alpha[d(y_\alpha, X)^q]\right) \\
&\geq \tau \left(d(y_{\kappa'}, x_{\kappa'})^q + d(y_\alpha, x_\alpha)^q\right) \\
&\geq \frac{\tau}{2^q} \left(d(y_{\kappa'}, x_{\kappa'}) + d(y_\alpha, x_\alpha)\right)^q,
\end{aligned}$$

where the first inequality was shown above, the second follows by Jensen's inequality, and the third by Hölder's. Next, we show

$$\begin{aligned}
1 &\geq \tau \left(\mathbb{E}_{\kappa'}\, d(y_\kappa, X)^q] + \mathbb{E}_\alpha[d(y_\kappa, X)^q]\right) \\
&\geq \tau \left(d(y_\kappa, x_{\kappa'})^q + d(y_\kappa, x_\alpha)^q\right) \\
&\geq \frac{\tau}{2^q} \left(d(y_\kappa, x_{\kappa'}) + d(y_\kappa, x_\alpha)\right)^q \\
&\geq \frac{\tau}{2^q} \left(d(x_{\kappa'}, x_\alpha)\right)^q,
\end{aligned}$$

using the previous line of logic along with the triangle inequality. Combining the two with another application of the triangle inequality gives

$$d(y_{\kappa'}, y_\alpha) \le \frac{1}{2}\tau^{-1/q} + \frac{1}{2}\tau^{-1/q} = \tau^{-1/q}.$$

All together, we obtain

$$\mathsf{W}_p(\kappa, \kappa') \le 3\,\tau^{1/p}\left(\tau^{-1} + \tau^{-1}\right)^{1/q} \le 6\,\tau^{1/p-1/q},$$

as desired. □

For the lower bound, we use the following result.

Lemma 4. *For any $\mathcal{D} \subseteq \mathcal{P}(\mathcal{X})$ and $\tau \in [0,1]$, we have $E(\mathcal{D}, \tau) \gtrsim \mathfrak{m}_-(\mathcal{D}, \tau) := \sup_{\substack{\kappa,\kappa' \in \mathcal{D} \\ \kappa \in \mathcal{B}_\tau(\kappa')}} \mathsf{W}_p(\kappa, \kappa')$.*

Proof. Fix any feasible $\kappa, \kappa' \in \mathcal{D}$ with $\kappa \in \mathcal{B}_\tau(\kappa')$, and suppose that both contaminated measures $\tilde{\mu}$ and $\tilde{\nu}$ are equal to κ. This is consistent with two cases: (1) the clean measures are $\mu = \nu = \kappa'$ (in which case $\mathsf{W}_p(\mu, \nu) = 0$); and (2) the clean measures are $\mu = \kappa$ and $\nu = \kappa'$ (in which case $\mathsf{W}_p(\mu, \nu) = \mathsf{W}_p(\kappa, \kappa')$). Hence, no matter which distance a robust proxy $\hat{W}$ assigns to $\tilde{\mu}$ and $\tilde{\nu}$, it must incur error at least $\mathsf{W}_p(\kappa, \kappa')/2$ in one of these cases, proving the lemma. □

To conclude, we fix two points $x, y \in \mathcal{X}$ such that $d(x, y) = \varepsilon^{-1/q}$ as guaranteed by the theorem condition (the same proof works when exact equality fails after shrinking ε by a constant factor). Then we apply Lemma 4 to lower bound $E(\mathcal{D}_q, \varepsilon)$ by $\frac{1}{2}\mathsf{W}_p(\kappa, \kappa')$, where $\kappa = \delta_x$ and $\kappa' = (1-\tau)\delta_x + \varepsilon\delta_y$. Indeed, both distributions belong to $\mathcal{D}_q$ (since $\mathbb{E}_{\kappa'}[d(x, X)^q] = \varepsilon d(x, y)^q = 1$) and $\mathsf{W}_p(\kappa, \kappa') = \varepsilon^{1/p} d(x, y) = \varepsilon^{1/p - 1/q}$.

A.3 Proof of Corollary 1

For the upper bound, we observe that for $\kappa \in \mathcal{D}_2^{\mathrm{cov}}$ with mean $\mathbb{E}_\kappa[X] = x_\kappa$, we have

$$\mathbb{E}[\|x_\kappa - X\|_2^2] = \mathrm{tr}(\Sigma_\kappa) \le d.$$

Hence (by scaling down $\mathbb{R}^d$ by a factor of d), we obtain the desired upper bound as an application of Theorem 1. For the other direction, we apply Lemma 4 to lower bound $E(\mathcal{D}_2^{\mathrm{cov}}, \varepsilon)$ by $\frac{1}{2}\mathsf{W}_p(\kappa, \kappa')$ where $\kappa = \delta_0$ and $\kappa' = (1-\varepsilon)\delta_0 + \varepsilon\mathcal{N}(0, I/\varepsilon)$. Indeed, both distributions belong to $\mathcal{D}_2^{\mathrm{cov}}$ (since, in case (2), $\Sigma_\nu = I$) and $\mathsf{W}_p(\mu, \nu) \ge \varepsilon^{1/p}\,\mathbb{E}_{X \sim \mathcal{N}(0, I/\varepsilon)}[\|X\|] = \Omega(\sqrt{d}\,\varepsilon^{1/p - 1/2})$, as desired.

A.4 Proof of Proposition 1

Proposition 1 (Mass addition)**.** *For all $p \in [1, \infty]$ and $\mu, \nu \in \mathcal{P}(\mathcal{X})$, we have*

$$\mathsf{W}_p^\varepsilon(\mu, \nu) = (1-2\delta)^{-1/p} \inf_{\substack{\mu' \in \mathcal{B}_\delta(\mu) \\ \nu' \in \mathcal{B}_\delta(\nu)}} \mathsf{W}_p(\mu', \nu'), \tag{5}$$

where $\delta = \varepsilon/(1+\varepsilon)$.

Proof. To begin, we rewrite the RHS as

$$\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^{1/p} \inf_{\substack{\tilde{\mu} \in \mathcal{B}_{\frac{\varepsilon}{1+\varepsilon}}(\mu) \\ \tilde{\nu} \in \mathcal{B}_{\frac{\varepsilon}{1+\varepsilon}}(\nu)}} \mathsf{W}_p(\tilde{\mu}, \tilde{\nu}) = \inf_{\substack{\mu' \ge \mu, \nu' \ge \nu \\ \mu'(\mathcal{X}), \nu'(\mathcal{X}) = 1+\varepsilon}} \mathsf{W}_p\left(\frac{\mu'}{1-\varepsilon}, \frac{\nu'}{1-\varepsilon}\right) =: \overline{\mathsf{W}}_p^\varepsilon(\mu, \nu). \tag{13}$$

First, we prove that $\mathsf{W}_p^\varepsilon(\mu, \nu) \ge \overline{\mathsf{W}}_p^\varepsilon(\mu, \nu)$. Take $\mu' \le \mu$ and $\nu' \le \nu$ optimal for the original formulation (1) of $\mathsf{W}_p^\varepsilon(\mu, \nu)$, with $\mu'(\mathcal{X}) = \nu'(\mathcal{X}) = 1 - \varepsilon$ (see Proposition 3 for existence of minimizers). Take $\pi \in \Pi(\mu', \nu')$ to be an optimal coupling for $\mathsf{W}_p(\mu', \nu')$. Then, consider the alternative perturbed measures $\mu'_+ := \mu + (\nu - \nu')$

and $\nu'_+ := \nu + (\mu - \mu')$, which are feasible for $\overline{\mathsf{W}}_p^\varepsilon(\mu,\nu)$, and define the coupling $\pi_+ \in \Pi(\mu'_+, \nu'_+)$ by $\pi_+ = \pi + (\mathrm{Id}, \mathrm{Id})_\#(\mu - \mu' + \nu - \nu')$. By construction, we have $\|d\|_{L^p(\pi_+)} = \|d\|_{L^p(\pi)}$, and so

$$\overline{\mathsf{W}}_p^\varepsilon(\mu,\nu) \le (1-\varepsilon)^{-1/p}\|d\|_{L^p(\pi_+)} = (1-\varepsilon)^{-1/p}\,\mathsf{W}_p(\mu',\nu') = \mathsf{W}_p^\varepsilon(\mu,\nu),$$

as desired.

For the other direction, consider any $\mu' \ge \mu$ and $\nu' \ge \nu$ feasible for $\overline{\mathsf{W}}_p^\varepsilon(\mu,\nu)$, and write $\mu' = \mu + \alpha, \nu' = \nu + \beta$ for $\alpha, \beta \in \mathcal{M}_+(\mathcal{X})$ with $\alpha(\mathcal{X}) = \beta(\mathcal{X}) = \varepsilon$. Take $\pi \in \mathcal{P}(\mu',\nu')$ to be an optimal coupling for $\mathsf{W}_p(\mu',\nu')$, and let $\pi(y|x)$ be the regular conditional probability distribution such that $\nu'(\cdot) = \int_{\mathcal{X}} \pi(\cdot|x)\,\mathrm{d}\mu'(x)$. Informally, we next show that the added masses α and β need not be moved during transport, since we might as well replace them with their destinations after transport. Formally, this requires a bit of labeling.

To start, we decompose ν' into $\nu' = \nu'_{\leftarrow\mu} + \nu'_{\leftarrow\alpha}$, where $\nu'_{\leftarrow\mu}(\cdot) := \int_{\mathcal{X}} \pi(\cdot|x)\,\mathrm{d}\mu(x)$ denotes the mass transported from μ to ν' via π and $\nu'_{\leftarrow\alpha}(\cdot) := \int_{\mathcal{X}} \pi(\cdot|x)\,\mathrm{d}\alpha(x)$ denotes the mass transported from α. Similarly, we decompose ν into $\nu_{\leftarrow\mu} = \nu'_{\leftarrow\mu} \wedge \nu$ and $\nu_{\leftarrow\alpha} = \nu - \nu_{\leftarrow\mu}$ and split β into $\beta_{\leftarrow\mu} = \nu'_{\leftarrow\mu} - \nu_{\leftarrow\mu}$ and $\beta_{\leftarrow\alpha} = \beta - \beta_{\leftarrow\mu}$. By this construction, we have

$$\nu = \nu_{\leftarrow\mu} + \nu_{\leftarrow\alpha}, \quad \beta = \beta_{\leftarrow\mu} + \beta_{\leftarrow\alpha}, \quad \nu'_{\leftarrow\mu} = \nu_{\leftarrow\mu} + \beta_{\leftarrow\mu}, \quad \nu'_{\leftarrow\alpha} = \nu_{\leftarrow\alpha} + \beta_{\leftarrow\alpha}.$$

Next, we arbitrarily decompose μ into $\mu_{\to\nu} + \mu_{\to\beta}$ and α into $\alpha_{\to\nu} + \alpha_{\to\beta}$ so that $\nu(\cdot) = \int_{\mathcal{X}} \pi(\cdot|x)\,\mathrm{d}(\mu_{\to\nu} + \alpha_{\to\nu})(x)$. To see that this is always possible, consider the restricted coupling $\bar\pi \in \Pi(\mu, \nu'_{\leftarrow\mu})$ defined by $\bar\pi(A \times B) = \int_A \pi(B|x)\,\mathrm{d}\mu(x)$, as well as the regular conditional probability distribution $\bar\pi(x|y)$ satisfying $\mu(\cdot) = \int_{\mathcal{X}} \bar\pi(\cdot|y)\,\mathrm{d}\nu'_{\leftarrow\mu}(y)$. We can then set $\mu_{\to\nu}(\cdot) = \int_{\mathcal{X}} \bar\pi(\cdot|y)\,\mathrm{d}\nu_{\leftarrow\mu}(y)$ and $\mu_{\to\beta} = \mu - \mu_{\to\nu}$. The same method works to construct $\alpha_{\to\nu}$ and $\alpha_{\to\beta}$.

Now, consider the alternative perturbed measures $\tilde\mu' = \mu + \nu'_{\leftarrow\alpha} = \mu_{\to\nu} + \nu'_{\leftarrow\alpha} + \mu_{\to\beta}$ and $\tilde\nu' = \nu + \mu_{\to\beta} + \beta_{\leftarrow\alpha} = \nu_{\leftarrow\mu} + \nu'_{\leftarrow\alpha} + \mu_{\to\beta}$ with $\tilde\pi \in \Pi(\tilde\mu', \tilde\nu')$ defined by

$$\begin{aligned}\tilde\pi(A \times B) &:= \int_A \pi(B|x)\,\mathrm{d}\mu_{\to\nu}(x) + (\mathrm{Id},\mathrm{Id})_\#(\nu'_{\leftarrow\alpha} + \mu_{\to\beta})(A \times B)\\ &= \int_A \pi(B|x)\,\mathrm{d}\mu_{\to\nu}(x) + (\mu_{\to\beta} + \nu'_{\leftarrow\alpha})(A \cap B).\end{aligned}$$

Note that $\tilde\mu'$ and $\tilde\nu'$ are still feasible for the mass-addition problem and $\|d\|_{L^p(\tilde\pi)} \le \mathsf{W}_p(\mu',\nu')$, implying that $\|d\|_{L^p(\tilde\pi)} = \mathsf{W}_p(\mu',\nu')$. Finally, observe that $\mu_{\to\nu} \le \mu$ and $\nu_{\leftarrow\mu} \le \nu$ are feasible for the initial mass-subtraction problem (1), so

$$\begin{aligned}\mathsf{W}_p^\varepsilon(\mu,\nu)^p &\le \mathsf{W}_p\left(\frac{\mu_{\to\nu}}{\mu_{\to\nu}(\mathcal{X})}, \frac{\nu_{\leftarrow\mu}}{\mu_{\to\nu}(\mathcal{X})}\right)^p\\ &= \frac{1}{\mu_{\to\nu}(\mathcal{X})}\,\mathsf{W}_p(\mu_{\to\nu}, \nu_{\leftarrow\mu})^p\\ &\le \frac{1}{\mu_{\to\nu}(\mathcal{X})}\|d\|^p_{L^p(\tilde\pi)}\\ &= \frac{1}{\mu_{\to\nu}(\mathcal{X})}\mathsf{W}_p(\mu',\nu')^p\\ &= \frac{1-\varepsilon}{\mu_{\to\nu}(\mathcal{X})}\overline{\mathsf{W}}_p^\varepsilon(\mu,\nu)^p\\ &\le \overline{\mathsf{W}}_p^\varepsilon(\mu,\nu)^p,\end{aligned}$$

as desired. □

A.5 Proof of Theorem 2

Theorem 2 (Dual form). *For $p \in [1,\infty)$, $\varepsilon \in [0,1]$, and $\mu,\nu \in \mathcal{P}(\mathcal{X})$, we have*

$$\begin{aligned}(1-\varepsilon)\mathsf{W}_p^\varepsilon(\mu,\nu)^p &= \sup_{f \in C_b(\mathcal{X})} \int f\,\mathrm{d}\mu + \int f^c\,\mathrm{d}\nu - 2\varepsilon\|f\|_\infty\\ &= \sup_{f \in C_b(\mathcal{X})} \int f\,\mathrm{d}\mu + \int f^c\,\mathrm{d}\nu - \varepsilon\,\mathrm{Range}(f),\end{aligned} \tag{7}$$

and the suprema are achieved by $f \in C_b(\mathcal{X})$ with $f = (f^c)^c$.

Proof. To start, we apply Proposition 1 and Kantorovich duality (11) to obtain

$$(1-\varepsilon)\mathsf{W}_p^\varepsilon(\mu,\nu)^p = \inf_{\substack{\alpha,\beta\in\mathcal{M}_+(\mathcal{X})\\ \alpha(\mathcal{X})=\beta(\mathcal{X})=\varepsilon}} \mathsf{W}_p\left(\mu+\alpha,\nu+\beta\right) \tag{14}$$

$$= \inf_{\substack{\alpha,\beta\in\mathcal{M}_+(\mathcal{X})\\ \alpha(\mathcal{X})=\beta(\mathcal{X})=\varepsilon}} \sup_{\substack{f,g\in C_b(\mathcal{X})\\ f(x)+g(y)\le d(x,y)^p}} \int f\,\mathrm{d}(\mu+\alpha) + \int g\,\mathrm{d}(\nu+\beta). \tag{15}$$

By compactness of $\mathcal{X}$, the infimum set is itself compact w.r.t. the weak topology. Having that, it is readily verified that the conditions for Sion's minimax theorem (convexity, continuity of objective, and compactness of infimum set) apply, giving

$$(1-\varepsilon)\mathsf{W}_p^\varepsilon(\mu,\nu)^p = \sup_{\substack{f,g\in C_b(\mathcal{X})\\ f(x)+g(y)\le d(x,y)^p}} \int f\,\mathrm{d}\mu + \int g\,\mathrm{d}\nu + \inf_{\alpha\in\mathcal{M}_+(\mathcal{X}),\alpha(\mathcal{X})=\varepsilon}\int f\,\mathrm{d}\alpha + \inf_{\beta\in\mathcal{M}_+(\mathcal{X}),\beta(\mathcal{X})=\varepsilon}\int g\,\mathrm{d}\beta$$

$$= \sup_{\substack{f,g\in C_b(\mathcal{X})\\ f(x)+g(y)\le d(x,y)^p}} \int f\,\mathrm{d}\mu + \int g\,\mathrm{d}\nu + \varepsilon\left(\inf_{x\in\mathcal{X}} f(x) + \inf_{y\in\mathcal{X}} g(y)\right).$$

Noting that replacing g with f^c preserves the constraints and can only increase the objective, we further have

$$(1-\varepsilon)\mathsf{W}_p^\varepsilon(\mu,\nu)^p = \sup_{f\in C_b(\mathcal{X})} \int f\,\mathrm{d}\mu + \int f^c\,\mathrm{d}\nu - \varepsilon\left(\sup_{x\in\mathcal{X}} f(x) - \inf_{x\in\mathcal{X}} f(x)\right)$$

$$= \sup_{f\in C_b(\mathcal{X})} \int f\,\mathrm{d}\mu + \int f^c\,\mathrm{d}\nu - \varepsilon\,\mathrm{Range}(f),$$

using the fact that $\inf_y f^c(y) = \inf_{x,y} d(x,y)^p - f(x) = -\sup_x f(x)$. The same reasoning allows us to restrict to $f = (f^c)^c$ if desired. Since adding a constant to f decreases f^c by the same constant, we are free to shift f so that the final term equals $2\varepsilon\|f\|_\infty$. Without shifting, we always have $\mathrm{Range}(f) \le 2\|f\|_\infty$, so the problem simplifies to

$$(1-\varepsilon)\mathsf{W}_p^\varepsilon(\mu,\nu)^p = \sup_{f\in C_b(\mathcal{X})} \int f\,\mathrm{d}\mu + \int f^c\,\mathrm{d}\nu - 2\varepsilon\|f\|_\infty,$$

as desired. Since $\mathsf{W}_p^\varepsilon(\mu,\nu) \le \mathsf{W}_p(\mu,\nu)$, we can assume that the supremum set is uniformly bounded with $\|f\|_\infty \le \mathsf{W}_p(\mu,\nu)/(2\varepsilon)$. Furthermore, the argument preceding (Santambrogio, 2015, Proposition 1.11) proves that the supremum set is uniformly equicontinuous. Since $\mathcal{X}$ is compact, Arzelà–Ascoli implies that the supremum is achieved.

See Appendix B.1 for an extension of this result to the asymmetric setting. □

A.6 Proof of Proposition 3

Proposition 3 (Existence of minimizers)**.** *For $p \in [1,\infty]$ and $\mu,\nu \in \mathcal{P}(\mathcal{X})$, the infimum in (1) is achieved, and there are minimizers $\mu' \le \mu$ and $\nu' \le \nu$ such that $\mu',\nu' \ge \mu\wedge\nu$ and $\mu'(\mathcal{X}) = \nu'(\mathcal{X}) = 1-\varepsilon$.*

Proof. We first prove existence via a compactness argument, and then turn to the lower envelope property $\mu',\nu' \ge \mu\wedge\nu$. This turns out to be considerably simpler to prove in the discrete setting, so we begin there and extend to the general case via a discretization argument. We already addressed the mass equality constraints in the proof of Lemma 1.

Existence: Because measures in this feasible set are positive and bounded by μ and ν, the set is tight, i.e. pre-compact w.r.t. the topology of weak convergence. If $p < \infty$, then for any sequence μ'_n, ν'_n in the infimum set,

$$\lim_{R\to\infty} \limsup_{n\to\infty} \int_{d(x_0,x)\ge R} d(x_0,x)^p\,\mathrm{d}\mu'_n(x) \le \lim_{R\to\infty} \int_{d(x_0,x)\ge R} d(x_0,x)^p\,\mathrm{d}\mu(x) = 0,$$

since $\mu \in \mathcal{P}_p(\mathcal{X})$, with the same holding for the ν'_n sequence. Thus, by the characterization of convergence in W_p given by (Villani, 2009, Theorem 6.8), the infimum set is precompact w.r.t. W_p. Finally, the mass function in the constraints is continuous w.r.t. the weak topology, so the infimum set is compact w.r.t. W_p. As the objective is clearly continuous with w.r.t. W_p, the infimum is achieved.

For $p = \infty$, we note that W_∞ is lower semicontinuous on $\mathcal{P}_\infty(\mathcal{X})$ w.r.t. the weak topology (and since supports are compact, the W_p topology, for all $p < \infty$). Indeed, for any $\mu'_n \xrightarrow{w} \mu'$, $\nu'_n \xrightarrow{w} \nu'$ in $\mathcal{P}_\infty(\mathcal{X})$, we have

$$\begin{aligned}\liminf_{n\to\infty} \mathsf{W}_\infty(\mu'_n, \nu'_n) &= \liminf_{n\to\infty} \sup_{p\to\infty} \mathsf{W}_p(\mu'_n, \nu'_n) \\ &\geq \sup_{p\to\infty} \liminf_{n\to\infty} \mathsf{W}_p(\mu'_n, \nu'_n) \\ &\geq \sup_{p\to\infty} \mathsf{W}_p(\mu', \nu') = \mathsf{W}_\infty(\mu', \nu').\end{aligned}$$

The infimum of a lower semicontinuous function over a compact set is always achieved, as desired.

Lower envelope (discrete case): Having proven the existence of minimizers $\mu' \leq \mu$ and $\nu' \leq \nu$, it remains to show that we can take $\mu', \nu' \geq \mu \wedge \nu$. We begin with the case where $\mathcal{X}$ is countable and treat measures and their mass functions interchangeably. Then, if $\mu' \geq \mu \wedge \nu$ fails to hold, there exists $x_0 \in \mathcal{X}$ such that $\mu'(x_0) < \mu(x_0)$ and $\mu'(x_0) < \nu(x_0)$. We can further assume that $\nu(x_0) = \nu'(x_0)$. Otherwise, $a := (\nu - \nu')(x_0) \wedge (\mu - \mu')(x_0)$ mass could be returned to both μ' and ν' at x_0 without increasing their transport cost. Indeed, for any $\pi \in \Pi(\mu', \nu')$, the modified plan $\pi + a \cdot \delta_{(x_0,x_0)}$ is valid for the new measures, $\mu' + a\delta_{x_0}$ and $\nu' + a\delta_{x_0}$ (still feasible by the choice of a), and attains the same cost.

Taking $\pi \in \Pi(\mu', \nu')$ optimal for $\mathsf{W}_p(\mu', \nu')$, we define the measure $\kappa \leq \mu'$ by $\kappa(x_0) := 0$, $\kappa(x) := \pi(x, x_0)$ for $x \neq x_0$. This captures the distribution of the mass that is transported away from x_0 w.r.t. π when transporting ν' to μ'. By conservation of mass, κ has total mass at least $(\nu' - \mu')(x_0) > 0$, so we can set $\tilde{\kappa} \leq \kappa$ to be a scaled-down copy of κ with total mass $(\mu \wedge \nu' - \mu')(x_0) > 0$. Now, we define an alternative perturbed measure $\tilde{\mu}' := \mu' + \tilde{\kappa}(\mathcal{X}) \cdot \delta_{x_0} - \tilde{\kappa}$. By definition, we have $0 \leq \tilde{\mu}' \leq \mu'$ with $\tilde{\mu}'(\mathcal{X}) = \mu'(\mathcal{X})$ and $\tilde{\mu}'(x_0) = (\mu \wedge \nu')(x_0)$. Furthermore, the modified plan $\tilde{\pi}$ defined by

$$\tilde{\pi}(x, y) = \begin{cases} \pi(x, y), & y \neq x_0 \\ \frac{\kappa(\mathcal{X}) - \tilde{\kappa}(\mathcal{X})}{\kappa(\mathcal{X})} \pi(x, x_0), & x \neq x_0, y = x_0 \\ \pi(x_0, x_0) + \tilde{\kappa}(\mathcal{X}), & x = y = x_0 \end{cases}$$

satisfies $\tilde{\pi} \in \Pi(\tilde{\mu}', \nu')$ and $\|d\|_{L^p(\tilde{\pi})} < \|d\|_{L^p(\pi)} = \mathsf{W}_p(\mu', \nu')$, contradicting the optimality of μ'.

Lower envelope (general case): For general $\mathcal{X}$, we will need the following lemma, which allows us to apply our discrete argument to all settings.

Lemma 5 (Dense approximation)**.** *Let $(\mathcal{Y}, \rho)$ be a separable metric space with dense subset $D \subseteq \mathcal{Y}$. For any $y \in \mathcal{Y}$, let y^λ denote a representative from D with $d(y, y^\lambda) \leq \lambda$ (which exists by separability). Similarly, for any $K \subseteq \mathcal{Y}$, define $K^\lambda = \{y^\lambda : y \in K\}$. Then, if $f : \mathcal{Y} \to \mathbb{R}$ is uniformly continuous, we have $\inf_{y\in K} f(y) = \lim_{\lambda\to 0} \inf_{y\in K^\lambda} f(y)$.*

Proof. First, if $\{y_n\}_{n\in\mathbb{N}}$ is an infimizing sequence for $\inf_{y\in K} f(y)$, then

$$\liminf_{\lambda\to 0} \inf_{y\in K^\lambda} f(y) \leq \liminf_{n\to\infty} f(y_n^{1/n}) = \lim_{n\to\infty} f(y_n) = \inf_{y\in K} f(y),$$

where the second equality relies on the uniform continuity of f. Similarly, if $\{y_n\}_{n\in\mathbb{N}}$ with $y_n \in K^{\lambda_n}$, $\lambda_n \searrow 0$, is an limiting sequence for $\limsup_{\lambda\to 0} \inf_{y\in K^\lambda} f(x)$, then we can write $y_n = x_n^{1/n}$ for $x_n \in K$ and find

$$\limsup_{\lambda\to 0} \inf_{y\in K^\lambda} f(y) = \limsup_{n\to\infty} f(y_n) = \limsup_{n\to\infty} f(x_n) \geq \inf_{y\in K} f(y),$$

where the second equality again follows from the uniform continuity of f. The two inequalities imply the lemma. □

We will essentially apply this lemma to the space of measures $\mathcal{Y} = ((1-\varepsilon)\mathcal{P}_p(\mathcal{X}), \mathsf{W}_p)$, letting D be its dense subset of discrete measures. For any $\lambda > 0$, separability of $\mathcal{X}$ implies the existence of a countable partition $\{A_i^\lambda\}_{i\in\mathbb{N}}$ of $\mathcal{X}$ such that $\operatorname{diam}(A_i^\lambda) \le \lambda$ for all $i \in \mathbb{N}$, with representatives $x_i^\lambda \in A_i^\lambda$ for all $i \in \mathbb{N}$. For any measure $\kappa \in \mathcal{Y}$, we let $\kappa^\lambda \in \mathcal{M}_+(\cup_{i\in\mathbb{N}}\{x_i^\lambda\})$ denote the discretized measure defined by $\kappa^\lambda := \sum_{i\in\mathbb{N}} \kappa(A_i^\lambda)\cdot\delta_{x_i^\lambda}$. For $p < \infty$, we have

$$\mathsf{W}_p(\kappa,\kappa^\lambda)^p \le \sum_{i\in\mathbb{N}} \mathsf{W}_p(\kappa|_{A_i^\lambda}, \kappa(A_i^\lambda)\cdot\delta_{x_i^\lambda})^p \le \sum_{i\in\mathbb{N}} \kappa(A_i^\lambda)\lambda^p = \lambda^p,$$

and so $\mathsf{W}_p(\kappa,\kappa^\lambda) \le \lambda$ (including $p=\infty$ by continuity). This discretization lifts to sets as in the lemma. Now, for any $\alpha,\beta \in \mathcal{M}_+(\mathcal{X})$ with $\alpha \le \beta$, let $K_{\alpha,\beta} = \{\kappa \in \mathcal{Y} : \alpha \le \kappa \le \beta, \kappa(\mathcal{X}) = 1-\varepsilon\}$. Importantly, this choice of discretization satisfies $K_{\alpha,\beta}^\lambda = K_{\alpha^\lambda,\beta^\lambda}$. Indeed if $\kappa \in K_{\alpha,\beta}$, then $\kappa^\lambda \in K_{\alpha^\lambda,\beta^\lambda}$ by our discretization definition, and total mass is preserved. Likewise, if $\kappa \in K_{\alpha^\lambda,\beta^\lambda}$, then we can consider

$$\tilde\kappa := \sum_{i:x_i\in\operatorname{supp}(\kappa')} \alpha|_{A_i^\lambda} + c_i(\beta-\alpha)|_{A_i^\lambda},$$

where $c_i \in [0,1]$ are chosen such that $\tilde\kappa(\{i\}) = \kappa(A_i)$ for all i. By this construction, we have $\kappa = \tilde\kappa^\lambda$ and $\tilde\kappa \in K_{\alpha,\beta}$ with $\tilde\kappa(\mathcal{X}) = \kappa(\mathcal{X}) = 1-\varepsilon$, as desired. We also observe that $\alpha^\lambda \wedge \beta^\lambda = (\alpha\wedge\beta)^\lambda$. Putting everything together, we finally obtain

$$\begin{aligned}
(1-\varepsilon)^{1/p}\,\mathsf{W}_p^\varepsilon(\mu,\nu) &= \inf_{\substack{\mu'\in K_{0,\mu}\\ \nu'\in K_{0,\nu}}} \mathsf{W}_p(\mu',\nu') && (16)\\
&= \lim_{\lambda\to 0} \inf_{\substack{\mu'\in K_{0,\mu}^\lambda\\ \nu'\in K_{0,\nu}^\lambda}} \mathsf{W}_p(\mu',\nu') && (17)\\
&= \lim_{\lambda\to 0} \inf_{\substack{\mu'\in K_{0,\mu^\lambda}\\ \nu'\in K_{0,\nu^\lambda}}} \mathsf{W}_p(\mu',\nu') && (18)\\
&= \lim_{\lambda\to 0} \inf_{\substack{\mu'\in K_{\mu^\lambda\wedge\nu^\lambda,\mu^\lambda}\\ \nu'\in K_{\mu^\lambda\wedge\nu^\lambda,\nu^\lambda}}} \mathsf{W}_p(\mu',\nu') && (19)\\
&= \lim_{\lambda\to 0} \inf_{\substack{\mu'\in K_{(\mu\wedge\nu)^\lambda,\mu^\lambda}\\ \nu'\in K_{(\mu\wedge\nu)^\lambda,\nu^\lambda}}} \mathsf{W}_p(\mu',\nu') && (20)\\
&= \lim_{\lambda\to 0} \inf_{\substack{\mu'\in K_{\mu\wedge\nu,\mu}^\lambda\\ \nu'\in K_{\mu\wedge\nu,\nu}^\lambda}} \mathsf{W}_p(\mu',\nu') && (21)\\
&= \inf_{\substack{\mu'\in K_{\mu\wedge\nu,\mu}\\ \nu'\in K_{\mu\wedge\nu,\nu}}} \mathsf{W}_p(\mu',\nu'). && (22)
\end{aligned}$$

concluding the proof. Here, (17) is an application of Lemma 5 (technically, we use the product space $\mathcal{Y}\times\mathcal{Y}$ with metric $(\alpha,\beta),(\alpha',\beta') \mapsto \max\{\mathsf{W}_p(\alpha,\alpha'),\mathsf{W}_p(\beta,\beta')\}$), (18) uses that $K_{\alpha,\beta}^\lambda = K_{\alpha^\lambda,\beta^\lambda}$, (19) is an application of the discrete result, and the remaining steps apply the same results in reverse order. The final equality shows that the lower envelope of $\mu\wedge\nu$ can be assumed even in this general setting. □

A.7 Proof of Proposition 5

Proposition 5 (Dependence on ε). *For any $p \in [1,\infty]$, $0 \le \varepsilon \le \varepsilon' \le 1$, and $\mu,\nu \in \mathcal{P}(\mathcal{X})$, we have*

(i) $\mathsf{W}_p^{\|\mu-\nu\|_{\mathrm{TV}}/2}(\mu,\nu) = 0$, $\mathsf{W}_p^0(\mu,\nu) = \mathsf{W}_p(\mu,\nu)$;

(ii) $\mathsf{W}_p^{\varepsilon'}(\mu,\nu) \le \mathsf{W}_p^\varepsilon(\mu,\nu)$;

(iii) $\mathsf{W}_p^\varepsilon(\mu,\nu) \le \left(\frac{1-\varepsilon'}{1-\varepsilon}\right)^{\frac{1}{p}} \mathsf{W}_p^{\varepsilon'}(\mu,\nu) + 4\operatorname{diam}(\mathcal{X})\left(\frac{\varepsilon'-\varepsilon}{1-\varepsilon}\right)^{\frac{1}{p}}$.

Proof. For (i), we clearly have $\mathsf{W}_p^0 = \mathsf{W}_p$, and

$$\mathsf{W}_p^{\|\mu-\nu\|_{\mathsf{TV}}/2}(\mu,\nu) \le (1 - \|\mu-\nu\|_{\mathsf{TV}}/2)^{-1/p}\,\mathsf{W}_p(\mu\wedge\nu, \mu\wedge\nu) = 0.$$

For (ii), let μ', ν' be optimal for $\mathsf{W}_p^{\varepsilon_1}(\mu,\nu)$ (in the mass removal formulation) and take $\pi \in \Pi(\mu',\nu')$ to be an optimal coupling for $\mathsf{W}_p(\mu',\nu')$. We can transform π into a feasible coupling for $\mathsf{W}_p^{\varepsilon_2}$ by restricting ourselves to the fraction $\frac{1-\varepsilon_1}{1-\varepsilon_2}$ of mass π' minimizing $\|d\|_{L^p(\pi')}$. This gives that

$$(1-\varepsilon_2)^{1/p}\,\mathsf{W}_p^{\varepsilon_2}(\mu,\nu) \le \left(\frac{1-\varepsilon_2}{1-\varepsilon_1}\right)^{1/p}(1-\varepsilon_1)^{1/p}\,\mathsf{W}_p^{\varepsilon_1}(\mu,\nu) \iff \mathsf{W}_p^{\varepsilon_2}(\mu,\nu) \le \mathsf{W}_p^{\varepsilon_1}(\mu,\nu),$$

as desired.

For (iii), we recall from Villani (2009, Theorem 6.13) that for $\alpha,\beta \in \mathcal{M}_+(\mathcal{X})$ with $\alpha(\mathcal{X}) = \beta(\mathcal{X})$, we have

$$\mathsf{W}_p(\alpha,\beta) \le 2\,\mathrm{diam}(\mathcal{X})\|\alpha-\beta\|_{\mathsf{TV}}^{1/p}.$$

Since any pair of feasible measures for $\mathsf{W}_p^{\varepsilon_2}$ are within (coordinate-wise) TV distance $\varepsilon_2 - \varepsilon_1$ from a pair of feasible measures for $\mathsf{W}_p^{\varepsilon_1}$, this bound combined with the triangle inequality for W_p gives

$$(1-\varepsilon_2)^{1/p}\,\mathsf{W}_p^{\varepsilon_2}(\mu,\nu) \ge (1-\varepsilon_1)^{1/p}\,\mathsf{W}_p^{\varepsilon_1}(\mu,\nu) - 4\,\mathrm{diam}(\mathcal{X})(\varepsilon_2-\varepsilon_1)^{1/p},$$

which can be rearranged to give the desired inequality. □

A.8 Proof of Proposition 6

Proposition 6 (Alternative primal form)**.** *For any $p \in [1,\infty]$ and $\mu,\nu \in \mathcal{P}(\mathcal{X})$, we have*

$$\mathsf{W}_p^{\varepsilon}(\mu,\nu) = \inf_{\substack{\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{X}) \\ \mu\in\mathcal{B}_\varepsilon(\pi_1),\,\nu\in\mathcal{B}_\varepsilon(\pi_2)}} \|d\|_{L^p(\pi)}, \tag{8}$$

where π_1 and π_2 are the respective marginals of π.

Proof. Starting from formulation given by Lemma 1, we compute

$$\mathsf{W}_p^{\varepsilon}(\mu,\nu) = \inf_{\substack{\mu'\colon \mu\in\mathcal{B}_\varepsilon(\mu') \\ \nu'\colon \nu\in\mathcal{B}_\varepsilon(\nu')}} \mathsf{W}_p(\mu',\nu') = \inf_{\substack{\mu'\colon \mu\in\mathcal{B}_\varepsilon(\mu') \\ \nu'\colon \nu\in\mathcal{B}_\varepsilon(\nu') \\ \pi\in\Pi(\mu',\nu')}} \|d\|_{L^p(\pi)} = \inf_{\substack{\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{X}) \\ \mu\in\mathcal{B}_\varepsilon(\pi_1) \\ \nu\in\mathcal{B}_\varepsilon(\pi_2)}} \|d\|_{L^p(\pi)},$$

eliminating auxiliary variables μ' and ν' for the final equality. □

A.9 Proof of Proposition 7

Proposition 7 (Convergence of couplings)**.** *Fix $p \in [1,\infty]$ and $\mu,\nu \in \mathcal{P}(\mathcal{X})$. If $\varepsilon_n \searrow 0$ and $\pi_n \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$ is optimal for $\mathsf{W}_p^{\varepsilon_n}(\mu,\nu)$ via (8), for each $n \in \mathbb{N}$, then $\{\pi_n\}_{n\in\mathbb{N}}$ admits a subsequence converging weakly to an optimal coupling for $\mathsf{W}_p(\mu,\nu)$.*

Proof. In this case, because the two marginals of each π_n are bounded by $\frac{\mu}{1-\varepsilon}$ and $\frac{\nu}{1-\varepsilon}$ respectively, this sequence is tight and admits a weakly convergent subsequence converging to some $\pi_\star \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$. Furthermore, $\pi_\star(A\times\mathcal{X}) = \lim_{n\to\infty}\pi_1^n(A) = \mu(A)$, and, likewise, $\pi_\star(\mathcal{X}\times B) = \nu(B)$; hence $\pi_\star \in \Pi(\mu,\nu)$. We now recall the definition of Γ-convergence.

Definition 1 (Γ-Convergence)**.** Let $\mathcal{Y}$ be a metric space and consider a sequence $F_n : \mathcal{Y} \to \mathbb{R}\cup\{\infty\}$, $n \in \mathbb{N}$. We say that $\{F_n\}_{n\in\mathbb{N}}$ Γ-converges to $F : \mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ and write $F_n \xrightarrow{\Gamma} F$ if

(i) For every $y_n, y \in \mathcal{Y}, n \in \mathbb{N}$, with $y_n \to y$, we have $F(y) \le \liminf_{n\to\infty} F_n(y_n)$;

(ii) For any $y \in \mathcal{Y}$, there exists $y_n \in \mathcal{Y}, n \in \mathbb{N}$, with $y_n \to y$ and $F(y) \ge \limsup_{n\to\infty} F_n(y_n)$.

If $\{y_n\}_{n\in\mathbb{N}}$ is a sequence of minimizers for F_n, for each $n \in \mathbb{N}$, and $F_n \xrightarrow{\Gamma} F$, then any limit point of $\{y_n\}_{n\in\mathbb{N}}$ is a minimizer of F (Maso, 2012, Corollary 7.20). Hence, it suffices to prove the Γ-convergence of $F_n : \mathcal{P}(\mathcal{X}\times\mathcal{X}) \to \mathbb{R}\cup\{\infty\}$ to $F : \mathcal{P}(\mathcal{X}\times\mathcal{X}) \to \mathbb{R}\cup\{\infty\}$ as defined by

$$F_n(\gamma) = \begin{cases} \|d\|_{L^p(\gamma)}, & \mu \in \mathcal{B}_{\varepsilon_n}(\gamma_1), \nu \in \mathcal{B}_{\varepsilon_n}(\gamma_2) \\ \infty, & \text{o.w.} \end{cases}$$

$$F(\gamma) = \begin{cases} \|d\|_{L^p(\gamma)}, & \gamma \in \Pi(\mu,\nu) \\ \infty, & \text{o.w.} \end{cases}$$

For the lim inf inequality, we start with any sequence $\gamma_n \in \mathcal{P}(\mathcal{X}\times\mathcal{X}), n \in \mathbb{N}$, with weak limit $\gamma \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$. If $\{\gamma_n\}_{n\in\mathbb{N}}$ does not contain a subsequence with $\mu \in \mathcal{B}_{\varepsilon_n}((\gamma_n)_1), \nu \in \mathcal{B}_{\varepsilon_n}((\gamma_n)_2)$, then the claim is trivial. Otherwise, $\gamma \in \Pi(\mu,\nu)$, and the Portmanteau Theorem gives

$$F(\gamma) = \|d\|_{L^p(\gamma)} \le \liminf_{n\to\infty} \|d\|_{L^p(\gamma_n)} = F_n(\gamma_n),$$

where we used the fact that d^p is non-negative and continuous for $p < \infty$. The fact that L^p norms are continuous in $p \in [1,\infty]$ implies that the inequality holds for $p = \infty$ as well.

For the lim sup direction, fix $\gamma \in \Pi(\mu,\nu)$. Then, we must have $\mu \in \mathcal{B}_{\varepsilon_n}(\gamma_1), \nu \in \mathcal{B}_{\varepsilon_n}(\gamma_2)$ for all $n \in \mathbb{N}$, so we can consider the constant sequence $\gamma_n \equiv \gamma, n \in \mathbb{N}$, and obtain

$$F_k(\gamma_n) = \|d\|_{L^p(\gamma)} = F(\gamma),$$

which implies the desired condition. □

A.10 Proof of Proposition 8

Proposition 8 (Finite support). *Let μ and ν be uniform discrete measures over n points each. Then there exist optimal μ', ν' for $\mathsf{W}_p^\varepsilon(\mu,\nu)$ such that μ' and ν' each give mass $1/n$ to $\lfloor(1-\varepsilon)n\rfloor$ points and assign their remaining $\lceil\varepsilon n\rceil/n - \varepsilon$ mass to a single point.*

Proof. Assume first that $p < \infty$ and $\varepsilon n = k$ for some $k \in \mathbb{N}$. Suppose that μ is uniform over $\{x_1,\dots,x_n\}$ and ν is uniform over $\{y_1,\dots,y_n\}$. Define the cost matrix $C \in \mathbb{R}^{n\times n}$ with $C_{ij} = d(x_i,y_j)^p$. Then, we can rewrite definition (1) for W_p^ε as a linear program, computing

$$(1-\varepsilon)\,\mathsf{W}_p^\varepsilon(\mu,\nu)^p = \min_{\substack{\pi\in\mathbb{R}_+^{n\times n} \\ \pi\mathbb{1}\le\mathbb{1}/n \\ \pi^\intercal\mathbb{1}\le\mathbb{1}/n \\ \mathbb{1}^\intercal\pi\mathbb{1}\ge 1-\varepsilon}} \operatorname{tr}(\pi C) = \frac{1}{n} \min_{\substack{\pi\in\mathbb{R}_+^{n\times n} \\ \pi\mathbb{1}\le\mathbb{1} \\ \pi^\intercal\mathbb{1}\le\mathbb{1} \\ \mathbb{1}^\intercal\pi\mathbb{1}\ge n-k}} \operatorname{tr}(\pi C).$$

This representation relates to the problem of characterizing the extreme points of the polytope of doubly substochastic $n\times n$ matrices with entries summing to an specified integer. This polytope was studied in Cao and Chen (2019), where Theorem 4.1 shows that the extreme points of interest are exactly the partial permutation matrices of order $n-k$. This implies the existence of a coupling for $\mathsf{W}_p^\varepsilon(\mu,\nu)^p$ whose marginals are uniform over $n-k = (1-\varepsilon)n$ points (giving mass $1/n$ to each point).

When ε is not a multiple of $1/n$, the same result (Cao and Chen, 2019, Theorem 4.1) reveals that there are optimal perturbed measures that each give mass $1/n$ to $\lfloor(1-\varepsilon)n\rfloor$ points and give the remaining mass to a single point. For p sufficiently large, the set of minimizers to the above problem stabilizes to a constant set, so this argument captures the $p = \infty$ case as well. □

A.11 Proof of Proposition 4

Proposition 4 (Interpreting maximizers). *If $f \in C_b(\mathcal{X})$ maximizes (7), then any $\mu', \nu' \in \mathcal{M}_+(\mathcal{X})$ minimizing (1) satisfy* $\operatorname{supp}(\mu-\mu') \subseteq \operatorname{argmax}(f)$ *and* $\operatorname{supp}(\nu-\nu') \subseteq \operatorname{argmin}(f)$.

Proof. Take $f \in C_b(\mathcal{X})$ maximizing (7) (by Theorem 2, such f is guaranteed to exist). Take *any* $\mu' = \mu - \alpha, \nu' = \nu - \beta$ optimal for the mass removal formulation of $\mathsf{W}_p^\varepsilon(\mu,\nu)$, where $\alpha, \beta \in \mathcal{M}_+(\mathcal{X})$ with $\alpha(\mathcal{X}) = \beta(\mathcal{X}) = \varepsilon$. Of course, $\mu + \beta$ and $\nu + \alpha$ are then optimal for the mass addition formulation of $\mathsf{W}_p^\varepsilon(\mu,\nu)$. Examining the proof of Theorem 2, we see that $\mu + \beta, \nu + \alpha$ and f, f^c must be a minimax equilibrium for the problem given in (15). Consequently, we have

$$(1-\varepsilon)\mathsf{W}_p^\varepsilon(\mu,\nu)^p = \int f \,\mathrm{d}\mu + \int f^c \,\mathrm{d}\nu + \int f \,\mathrm{d}\beta + \int f^c \,\mathrm{d}\alpha \tag{23}$$

$$= \int f \,\mathrm{d}\mu + \int f^c \,\mathrm{d}\nu + \varepsilon \inf_x f(x) - \varepsilon \sup_x f(x). \tag{24}$$

Since $\inf_y f^c(y) = -\sup_x f(x)$ (and the minimizers of f^c correspond to the maximizers of f), we have that (23) is strictly less than (24) unless $\mathrm{supp}(\beta) \subseteq \mathrm{argmin}(f)$ and $\mathrm{supp}(\alpha) \subseteq \mathrm{argmax}(f)$. This suggests taking $\alpha = \mu|_{\mathrm{argmax}(f)}$ and $\beta = \nu|_{\mathrm{argmin}(f)}$, but we cannot do so in general; consider the case where μ and ν are uniform discrete measures on n points and ε is not a multiple of $1/n$. Issues with that approach also arise when μ and ν are both supported on $\mathrm{argmax}(f)$ and the optimal μ' satisfies $\mu' \geq \mu \wedge \nu$. □

A.12 Proof of Proposition 2

Proposition 2 (Loss trimming dual)**.** *Fix $p \in [1,\infty)$ and take μ, ν to be uniform distributions over n points each. If $\varepsilon \in [0,1]$ is a multiple of $1/n$, then*

$$\mathsf{W}_p^\varepsilon(\mu,\nu)^p = \sup_{f \in C_b(\mathcal{X})} \left(\min_{\substack{\mathcal{A} \subseteq \mathrm{supp}(\mu) \\ |\mathcal{A}| = (1-\varepsilon)n}} \frac{1}{|\mathcal{A}|} \sum_{x \in \mathcal{A}} f(x) + \min_{\substack{\mathcal{B} \subseteq \mathrm{supp}(\nu) \\ |\mathcal{B}| = (1-\varepsilon)n}} \frac{1}{|\mathcal{B}|} \sum_{y \in \mathcal{B}} f^c(y) \right).$$

Proof. We place no restrictions on μ and ν initially and extend to the two-sided robust distance $\mathsf{W}_p^{\varepsilon_1,\varepsilon_2}$ defined in Appendix B.1. For this result, we will apply Sion's minimax theorem to the mass-removal formulation of $\mathsf{W}_p^{\varepsilon_1,\varepsilon_2}$. Mirroring the proof of Theorem 2, we compute

$$\begin{aligned}
\mathsf{W}_p^{\varepsilon_1,\varepsilon_2}(\mu,\nu) &= \inf_{\substack{0 \leq \mu' \leq \mu \\ 0 \leq \nu' \leq \nu \\ \mu'(\mathcal{X}) = 1-\varepsilon_1 \\ \nu'(\mathcal{X}) = 1-\varepsilon_2}} \sup_{\substack{f,g \in C_b(\mathcal{X}) \\ f(x)+g(y) \leq d(x,y)^p}} \frac{1}{1-\varepsilon_1} \int f \,\mathrm{d}\mu' + \frac{1}{1-\varepsilon_2} \int g \,\mathrm{d}\nu' \\
&= \sup_{\substack{f,g \in C_b(\mathcal{X}) \\ f(x)+g(y) \leq d(x,y)^p}} \left(\frac{1}{1-\varepsilon_1} \inf_{\substack{0 \leq \mu' \leq \mu \\ \mu'(\mathcal{X}) = 1-\varepsilon_1}} \int f \,\mathrm{d}\mu' + \frac{1}{1-\varepsilon_2} \inf_{\substack{0 \leq \nu' \leq \nu \\ \nu'(\mathcal{X}) = 1-\varepsilon_2}} \int g \,\mathrm{d}\nu' \right) \\
&= \sup_{f \in C_b(\mathcal{X})} \left(\frac{1}{1-\varepsilon_1} \inf_{\substack{0 \leq \mu' \leq \mu \\ \mu'(\mathcal{X}) = 1-\varepsilon_1}} \int f \,\mathrm{d}\mu' + \frac{1}{1-\varepsilon_2} \inf_{\substack{0 \leq \nu' \leq \nu \\ \nu'(\mathcal{X}) = 1-\varepsilon_2}} \int f^c \,\mathrm{d}\nu' \right).
\end{aligned}$$

When μ and ν are uniform distributions over n points, and $\varepsilon_1, \varepsilon_2$ are multiples of $1/n$, we simplify further to

$$\mathsf{W}_p^{\varepsilon_1,\varepsilon_2}(\mu,\nu) = \sup_{f \in C_b(\mathcal{X})} \left(\frac{1}{|\mathcal{A}|} \min_{\substack{\mathcal{A} \subseteq \mathrm{supp}(\mu) \\ |\mathcal{A}| = (1-\varepsilon_1)n}} \sum_{x \in A} f(x) + \frac{1}{|\mathcal{B}|} \min_{\substack{\mathcal{B} \subseteq \mathrm{supp}(\nu) \\ |\mathcal{B}| = (1-\varepsilon_2)n}} \sum_{y \in B} f^c(y) \right),$$

as desired □

A.13 Proof of Proposition 9

Proposition 9 (One-sample rates)**.** *Fix $p \in [1,\infty]$ and $\tau \in (\varepsilon, 1]$. If $p < d/2$, we have*

$$\mathbb{E}\left[\mathsf{W}_p^\varepsilon(\mu, \hat{\mu}_n)\right] \leq C n^{-1/d}$$

for a constant C independent of n and ε. Furthermore, if $d_*^\tau(\mu) > s$, then

$$\mathsf{W}_p^\varepsilon(\mu, \gamma) \geq C' (\tau - \varepsilon)^{1/p} n^{-1/s}$$

for any γ supported on at most n points, including $\hat{\mu}_n$, where $C' > 0$ is an absolute constant.

Proof. For the upper bound, we simply use that $\mathsf{W}_p^\varepsilon(\mu, \nu) \leq \mathsf{W}_p(\mu, \nu)$ and apply well-known empirical convergence results for W_p which give $\mathbb{E}[\mathsf{W}_p(\mu, \hat{\mu}_n)] \leq Cn^{-1/d}$ for a constant C independent of n (and ε); see, e.g., Weed and Bach (2019).

For the lower bound, if $d_*(\mu) > s$, then there is $\delta_0 > 0$ such that $\mathcal{N}_\delta(\mu, \tau) \geq \delta^{-t}$ for all $\delta \leq \delta_0$. By Proposition 3, we know that $\mathsf{W}_p^\varepsilon(\mu, \gamma)$ is achieved by $\mu' \leq \mu$ and $\gamma' \leq \gamma$, with $|\operatorname{supp}(\gamma')| \leq |\operatorname{supp}(\gamma)| \leq n$. Fix $\delta = n^{-s}/2$ and take n sufficiently large so that $\delta \leq \delta_0$. Let $S = \cup_{x \in \operatorname{supp}(\gamma')} B(x, \delta/2)$, where $B(x, r) := \{y \in \mathcal{X} : d(x, y) \leq r\}$ is the ball of radius r centered at x. Since $\mathcal{N}(\mu, \tau) \geq \delta^{-t} > n$, we have that $\mu(S) \leq 1 - \tau$, which further implies $\mu'(S) \leq 1 - \tau + \varepsilon$.

For any $\pi \in \Pi(\mu', \gamma')$, we now have

$$\int d(x, y)^p \, \mathrm{d}\pi(x, y) \geq \mu'(\mathcal{X} \setminus S) \left(\frac{\delta}{2}\right)^p > (\tau - \delta) \left(\frac{\delta}{2}\right)^p = (\tau - \varepsilon) \, 4^{-p} n^{-\frac{p}{s}},$$

which gives $\mathsf{W}_p^\varepsilon(\mu, \gamma) = (1-\varepsilon)^{-1/p} \, \mathsf{W}_p(\mu', \gamma') \geq \frac{1}{4}(\tau-\varepsilon)^{1/p}(1-\varepsilon)^{-1/p} \, n^{-1/s}$. This suffices since $(1-\varepsilon)^{-1/p} \geq 1$. □

A.14 Proof of Corollary 2

Corollary 2 (Simple one-sample rate)**.** *Fix $p < d/2$ and $\varepsilon \in (0, 1]$. If μ has absolutely continuous part $f \, \mathrm{d}\lambda$ with $\int f \, \mathrm{d}\lambda > \varepsilon$ and f bounded from above, then $d_*^\tau(\mu) = d$ and $\mathbb{E}\left[\mathsf{W}_p^\varepsilon(\mu, \hat{\mu}_n)\right] = \Theta(n^{-1/d})$.*

Proof. Set $\tau = (\int f \, \mathrm{d}\lambda + \varepsilon)/2$, so that $\varepsilon < \tau < \inf f \, \mathrm{d}\lambda$. Then, for any covering of $\mathcal{X}$ with balls of radius δ except for A with $\mu(A) \leq \tau$, we have

$$\int_A f \, \mathrm{d}\lambda \leq \mu(A) \leq \tau \leq \int_{\mathcal{X}} f \, \mathrm{d}\lambda - c$$

for a positive constant $c > 0$. This gives that $\int_{\mathcal{X} \setminus A} f \, \mathrm{d}\lambda > c$, and hence $\lambda(\mathcal{X} \setminus A) \geq c/\|f\|_\infty > 0$. Since $\mathcal{X} \setminus A$ has Lebesgue measure bounded away from 0, a volumetric argument implies that the covering must contain at least $C\delta^{-d}$ balls of radius δ for some $C > 0$. This implies the desired lower covering dimension and, by the proof of the previous lower bound, the desired empirical convergence rate. □

A.15 Proof of Proposition 10

Proposition 10 (Two-sample rate)**.** *For any $\varepsilon \in [0, 1]$ and $p < d/2$, we have*

$$\mathbb{E}\left[\left|\mathsf{W}_p^\varepsilon(\hat{\mu}_n, \hat{\nu}_n)^p - \mathsf{W}_p^\varepsilon(\mu, \nu)^p\right|\right] \leq Cn^{-p/d}.$$

for a constant C independent of n and ε.

Proof. To begin, we write

$$\begin{aligned}(1 - \varepsilon)\mathsf{W}_p^\varepsilon(\mu, \nu)^p &= \sup_{f \in C_b(\mathcal{X})} \mu(f) + \nu(f^c) - \operatorname{Range}(f) \\ (1 - \varepsilon)\mathsf{W}_p^\varepsilon(\hat{\mu}_n, \hat{\nu}_n)^p &= \sup_{f \in C_b(\mathcal{X})} -\hat{\mu}_n(f^c) - \hat{\nu}_n(f) - \operatorname{Range}(f),\end{aligned}$$

using that $C_b(\mathcal{X})$ is closed under negation. From here, we compute

$$\begin{aligned}(1 - \varepsilon)\left|\mathsf{W}_p^\varepsilon(\hat{\mu}_n, \hat{\nu}_n)^p - \mathsf{W}_p^\varepsilon(\mu, \nu)^p\right| &\leq \sup_{f \in C_b(\mathcal{X})} \left|\int f \, \mathrm{d}\mu + \int f^c \, \mathrm{d}\hat{\mu}_n + \int f \, \mathrm{d}\hat{\nu}_n + \int f^c \, \mathrm{d}\nu\right| \\ &\leq \mathsf{W}_p(\mu, \hat{\mu}_n)^p + \mathsf{W}_p(\nu, \hat{\nu}_n)^p.\end{aligned}$$

Hence, standard one-sample empirical convergence rates for W_p (e.g., Weed and Bach (2019)), give

$$\mathbb{E}\left[\left|\mathsf{W}_p^\varepsilon(\hat{\mu}_n,\hat{\nu}_n)^p - \mathsf{W}_p^\varepsilon(\mu,\nu)^p\right|\right] \le C(1-\varepsilon)^{-1}n^{-p/d} \le 2Cn^{-p/d},$$

for a constant C independent of n and ε. □

The standard lower bounds for two-sample empirical convergence (Talagrand, 1992; Barthe and Bordenave, 2013) do not immediately transfer to this setting. When $\mu = \nu$ are both an absolutely continuous distribution like $\mathrm{Unif}([0,1]^d)$, these can easily be adapted to provide a lower bound for

$$\begin{aligned}\inf_{\substack{\mathcal{A}\subseteq[n],|\mathcal{A}|=(1-\varepsilon)n\\ \mathcal{B}\subseteq\mathrm{supp}([n]),|\mathcal{B}|=(1-\varepsilon)n}} &\mathbb{E}\left[\mathsf{W}_p\left(\frac{1}{(1-\varepsilon)n}\sum_{i\in\mathcal{A}}\delta_{X_i},\frac{1}{(1-\varepsilon)n}\sum_{j\in\mathcal{B}}\delta_{Y_j}\right)\right]\\ &= \mathbb{E}\left[\mathsf{W}_p\left(\frac{1}{(1-\varepsilon)n}\sum_{i=1}^{(1-\varepsilon)n}\delta_{X_i},\frac{1}{(1-\varepsilon)n}\sum_{j=1}^{(1-\varepsilon)n}\delta_{Y_j}\right)\right].\end{aligned}$$

For this to be helpful for the present case, one would also need to bound

$$\begin{aligned}\Bigg|\inf_{\substack{\mathcal{A}\subseteq[n],|\mathcal{A}|=(1-\varepsilon)n\\ \mathcal{B}\subseteq[n],|\mathcal{B}|=(1-\varepsilon)n}} &\mathbb{E}\left[\mathsf{W}_p\left(\frac{1}{(1-\varepsilon)n}\sum_{i\in\mathcal{A}}\delta_{X_i},\frac{1}{(1-\varepsilon)n}\sum_{j\in\mathcal{B}}\delta_{Y_j}\right)\right]\\ &- \mathbb{E}\left[\inf_{\substack{\mathcal{A}\subseteq[n],|\mathcal{A}|=(1-\varepsilon)n\\ \mathcal{B}\subseteq[n],|\mathcal{B}|=(1-\varepsilon)n}}\mathsf{W}_p\left(\frac{1}{(1-\varepsilon)n}\sum_{i\in\mathcal{A}}\delta_{X_i},\frac{1}{(1-\varepsilon)n}\sum_{j\in\mathcal{B}}\delta_{Y_j}\right)\right]\Bigg|,\end{aligned}$$

which we defer for future work.

A.16 Proof of Proposition 11

Proposition 11 (Exact recovery). *Fix $\mu,\nu\in\mathcal{P}(\mathcal{X})$, $\varepsilon\in[0,1]$, and suppose that $\tilde{\mu} = (1-\varepsilon)\mu + \varepsilon\alpha$ and $\tilde{\nu} = (1-\varepsilon)\nu + \varepsilon\beta$, for some $\alpha,\beta\in\mathcal{P}(\mathcal{X})$. Let $S = \mathrm{supp}(\mu+\nu)$. If $d(\mathrm{supp}(\alpha),S)$, $d(\mathrm{supp}(\beta),S)$, and $d(\mathrm{supp}(\alpha),\mathrm{supp}(\beta))$ are all greater than $\mathrm{diam}(S)$, then $\mathsf{W}_p^\varepsilon(\tilde{\mu},\tilde{\nu}) = \mathsf{W}_p(\mu,\nu)$.*

Proof. Under the conditions of the proposition, we will prove that $\mu' = (1-\varepsilon)\mu$ and $\nu' = (1-\varepsilon)\nu$ are the *unique* minimizers for $\mathsf{W}_p^\varepsilon(\tilde{\mu},\tilde{\nu})$, implying that $\mathsf{W}_p^\varepsilon(\tilde{\mu},\tilde{\nu}) = \mathsf{W}_p(\mu,\nu)$. Suppose not, and take μ' and ν' optimal for (1) such that $\mu' = \mu'_g + \mu'_b$, where $\mu'_g := \mu'\wedge(1-\varepsilon)\mu$ and $\mu'_b = \mu' - \mu'_g \le \varepsilon\alpha$, $\mu'_b \ne 0$. Let $\pi\in\Pi(\mu',\nu')$ be an optimal coupling for $\mathsf{W}_p(\mu',\nu')$, and let $\pi(y|x)$ be the regular conditional probability distribution such that $\nu'(\cdot) = \int_{\mathcal{X}}\pi(\cdot|x)\,\mathrm{d}\mu'(x)$. Then we can compute

$$\begin{aligned}(1-\varepsilon)\mathsf{W}_p(\mu,\nu)^p &\ge (1-\varepsilon)\mathsf{W}_p^\varepsilon(\tilde{\mu},\tilde{\nu})^p\\ &= \mathsf{W}_p(\mu',\nu')^p\\ &= \iint d(x,y)^p\,\mathrm{d}\pi(y|x)\,\mathrm{d}\mu'(x)\\ &= \iint d(x,y)^p\,\mathrm{d}\pi(y|x)\,\mathrm{d}\mu'_g(x) + \iint d(x,y)^p\,\mathrm{d}\pi(y|x)\,\mathrm{d}\mu'_b(x)\\ &\ge (1-\varepsilon)\mathsf{W}_p(\mu,\nu)^p - \iint d(x,y)^p\,\mathrm{d}\pi(y|x)\,\mathrm{d}(\mu-\mu'_g)(x) + \iint d(x,y)^p\,\mathrm{d}\pi(y|x)\,\mathrm{d}\mu'_b(x)\\ &> (1-\varepsilon)\mathsf{W}_p(\mu,\nu)^p - (\mu-\mu'_g)(\mathcal{X})\,\mathrm{diam}(S)^p + \mu'_b(\mathcal{X})\,\mathrm{diam}(S)^p\\ &= (1-\varepsilon)\mathsf{W}_p(\mu,\nu)^p,\end{aligned}$$

a contradiction. The same argument goes through if the three relevant distances are greater than $\|d\|_{L^\infty(\pi)}$, although general conditions under which this norm is finite seem hard to obtain (unless $p=\infty$, in which case $\|d\|_{L^\infty(\pi)} = \mathsf{W}_\infty(\mu',\nu') \le \mathsf{W}_\infty(\mu,\nu)$). □

A.17 Proof of Proposition 12

Proposition 12 (Robustness radius for unknown ε)**.** *Assume the setting of Proposition 11 and let D be the maximum of $d(\mathrm{supp}(\alpha), S)$, $d(\mathrm{supp}(\beta), S)$, and $d(\mathrm{supp}(\alpha), \mathrm{supp}(\beta))$. Then, for $p \in [1, \infty)$, we have*

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}\delta}\left[(1-\delta)\mathsf{W}_p^\delta(\tilde{\mu}, \tilde{\nu})^p\right] &\leq -D^p, && \delta \in [0, \varepsilon) \\ \tfrac{\mathrm{d}}{\mathrm{d}\delta}\left[(1-\delta)\mathsf{W}_p^\delta(\tilde{\mu}, \tilde{\nu})^p\right] &\geq -\operatorname{diam}(S)^p > -D^p, && \delta \in (\varepsilon, 1] \end{aligned}$$

at the (all but countably many) points where the derivative is defined.

Proof. By Proposition 5, $(1-\delta)\mathsf{W}_p^\delta(\tilde{\mu}, \tilde{\nu})^p$ is continuous and decreasing in δ, so it must be differentiable except on a countable set. Now, if $0 \leq \delta' < \delta \leq \varepsilon$, then the proof of Proposition 11 reveals that there are optimal μ', ν' for $\mathsf{W}_p^\delta(\tilde{\mu}, \tilde{\nu})$ and μ'', ν'' for $\mathsf{W}_p^{\delta'}(\tilde{\mu}, \tilde{\nu})$ such that $\tilde{\mu} \geq \mu'' \geq \mu' \geq (1-\varepsilon)\mu$ and $\tilde{\nu} \geq \nu'' \geq \nu' \geq (1-\varepsilon)\nu$. Letting $\pi \in \Pi(\mu'', \nu'')$ be an optimal coupling for $\mathsf{W}_p(\mu'', \nu'')$ and defining $\pi(y|x)$ as before, we compute

$$\begin{aligned} (1-\delta)\mathsf{W}_p^\delta(\tilde{\mu}, \tilde{\nu})^p - (1-\delta')\mathsf{W}_p^{\delta'}(\tilde{\mu}, \tilde{\nu})^p &= \mathsf{W}_p(\mu', \nu')^p - \mathsf{W}_p(\mu'', \nu'')^p \\ &= \mathsf{W}_p(\mu', \nu')^p - \|d\|_{L^p(\pi)}^p \\ &\leq \iint d(x,y)\pi(y|x)\,\mathrm{d}\mu'(x) - \iint d(x,y)\pi(y|x)\,\mathrm{d}\mu''(x) \\ &= -\iint d(x,y)\pi(y|x)\,\mathrm{d}(\mu'' - \mu')(x) \\ &\leq -(\delta - \delta')D^p. \end{aligned}$$

Taking $\delta' \to \delta$, we find that the derivative of interest is bounded by $-D^p$ from above wherever it exists. On the other hand, if $\varepsilon \leq \delta < \delta' \leq 1$, then there are optimal μ', ν' for $\mathsf{W}_p^\delta(\tilde{\mu}, \tilde{\nu})$ and μ'', ν'' for $\mathsf{W}_p^{\delta'}(\tilde{\mu}, \tilde{\nu})$ such that $(1-\varepsilon)\mu \geq \mu' \geq \mu'' \geq 0$ and $(1-\varepsilon)\nu \geq \nu' \geq \nu''$. Letting $\pi \in \Pi(\mu'', \nu'')$ again be an optimal coupling for $\mathsf{W}_p(\mu'', \nu'')$, we compute

$$\begin{aligned} (1-\delta')\mathsf{W}_p^{\delta'}(\tilde{\mu}, \tilde{\nu})^p - (1-\delta)\mathsf{W}_p^\delta(\tilde{\mu}, \tilde{\nu})^p &= \mathsf{W}_p(\mu'', \nu'')^p - \mathsf{W}_p(\mu', \nu')^p \\ &= \|d\|_{L^p(\pi)}^p - \mathsf{W}_p(\mu', \nu') \\ &\geq \|d\|_{L^p(\pi)}^p - \|d\|_{L^p(\pi)}^p - \mathsf{W}_p(\mu' - \mu'', \nu' - \nu'')^p \\ &\geq -\operatorname{diam}(S)^p(\delta' - \delta). \end{aligned}$$

Taking $\delta' \to \delta$, we find that the derivative of interest is bounded by $-\operatorname{diam}(S)^p > -D^p$ from below, wherever it exists. □

A.18 Proof of Proposition 13

Proposition 13 (Robust consistency)**.** *Fix $p < \infty$ and suppose that $\tau_\mu(n) \vee \tau_\nu(n) = O(n^{-a})$ almost surely, for some $a > 0$. Then, setting $\varepsilon_n = n^{-b}$, for any $0 < b < a$, we have*

$$\left|\mathsf{W}_p(\mu, \nu) - \mathsf{W}_p^{\varepsilon_n}(\tilde{\mu}_n, \tilde{\nu}_n)\right| \to 0$$

as $n \to \infty$, almost surely.

Proof. Fix $p < \infty$ (a minor adjustment from the statement in the main text). By design, there exists some $n_0 \in \mathbb{N}$ such that $\varepsilon_n \geq \tau_\mu(n) \vee \tau_\nu(n)$ for all $n \geq n_0$. For such n, let μ_n and ν_n denote the empirical probability measures associated with the uncontaminated samples among $\tilde{X}_1, \ldots, \tilde{X}_n$ and $\tilde{Y}_1, \ldots, \tilde{Y}_n$, respectively. By the proof of Proposition 11, if μ' and ν' are optimal for $\mathsf{W}_p^{\varepsilon_n}(\tilde{\mu}_n, \tilde{\nu}_n)$, and $\pi \in \Pi(\mu', \nu')$ is optimal for $\mathsf{W}_p(\mu', \nu')$, then $\|d\|_{L^\infty(\pi)} \leq \operatorname{diam}(\mathrm{supp}(\mu_n + \nu_n)) \leq \operatorname{diam}(\mathrm{supp}(\mu + \nu))$. Indeed, were any mass to be transported further than this by an optimal coupling, the source or destination would have to be contaminated, and the transport cost could be strictly improved by removing this contamination instead of inlier mass. Now, let μ_g' and ν_g' denote the

restrictions of μ' and ν', respectively, to $\mathrm{supp}(\mu+\nu)$. Hence, applying the same Wasserstein-TV bound used to prove Proposition 5 (iii), we obtain

$$\begin{aligned}
(1-\varepsilon_n)\mathsf{W}_p(\mu_n,\nu_n)^p &\geq (1-\varepsilon_n)\mathsf{W}_p^{\varepsilon_n}(\tilde{\mu}_n,\tilde{\nu}_n)^p && (25)\\
&= \mathsf{W}_p(\mu',\nu')^p && (26)\\
&\geq \mathsf{W}_p(\mu'_g,\nu'_g)^p - \tau_\mu(n)\vee\tau_\nu(n)\,\|d\|^p_{L^\infty(\pi)} && (27)\\
&\geq \mathsf{W}_p(\mu'_g,\nu'_g)^p - \varepsilon_n\,\mathrm{diam}(\mathrm{supp}(\mu+\nu)) && (28)\\
&\geq (1-2\varepsilon_n)\mathsf{W}_p^{2\varepsilon_n}(\mu_n,\nu_n)^p - \varepsilon_n\,\mathrm{diam}(\mathrm{supp}(\mu+\nu)) && (29)\\
&\geq (1-2\varepsilon_n)\mathsf{W}_p(\mu_n,\nu_n)^p - C\varepsilon_n\,\mathrm{diam}(\mathrm{supp}(\mu+\nu)), && (30)
\end{aligned}$$

for an absolute constant $C>0$. Here, (25) follows by the under-estimation property of W_p^ε under Huber contamination, (25) holds since at most $\tau_\mu(n)\vee\tau_\nu(n)\leq\varepsilon_n$ fraction of samples are removed from μ' and ν' to obtain their restrictions, (29) observes that μ'_g and ν'_g are feasible for $\mathsf{W}_p^{2\varepsilon_n}(\mu_n,\nu_n)$, and (30) is an application of the bound mentioned above. Hence, $\lim_{n\to\infty}\mathsf{W}_p^{\varepsilon_n}(\tilde{\mu}_n,\tilde{\nu}_n)=\lim_{n\to\infty}\mathsf{W}_p(\mu_n,\nu_n)=\mathsf{W}_p(\mu,\nu)$ as $n\to\infty$, where the second equality follows from empirical convergence of W_p. □

Appendix B Supplementary Results and Discussion

B.1 Asymmetric Variant

First, we formally define the asymmetric robust p-Wasserstein distance with robustness radii ε_1 and ε_2 by

$$\mathsf{W}_p^{\varepsilon_1,\varepsilon_2}(\mu,\nu)=\inf_{\substack{\mu',\nu'\in\mathcal{M}_+(\mathcal{X})\\ 0\leq\mu'\leq\mu,\ \|\mu-\mu'\|_{\mathsf{TV}}\leq\varepsilon_1\\ 0\leq\nu'\leq\nu,\ \|\nu-\nu'\|_{\mathsf{TV}}\leq\varepsilon_2}}\mathsf{W}_p\left(\frac{\mu'}{\mu'(\mathcal{X})},\frac{\nu'}{\nu'(\mathcal{X})}\right),$$

naturally extending the definition in (1) so that $\mathsf{W}_p^\varepsilon=\mathsf{W}_p^{\varepsilon,\varepsilon}$ and $\mathsf{W}_p(\mu\|\nu)=\mathsf{W}_p^{\varepsilon,0}(\mu,\nu)$. Our robustness results translate immediately to this setting. Indeed, writing the asymmetric minimax error

$$E(\mathcal{D},\varepsilon_1,\varepsilon_2):=\inf_{\hat{W}:\mathcal{P}(\mathcal{X})^2\to\mathbb{R}}\sup_{\substack{\mu,\nu\in\mathcal{D}\\ \tilde{\mu}\in\mathcal{B}_{\varepsilon_1}(\mu)\\ \tilde{\nu}\in\mathcal{B}_{\varepsilon_2}(\nu)}}\left|\hat{W}(\tilde{\mu},\tilde{\nu})-\mathsf{W}_p(\mu,\nu)\right|,$$

we trivially obtain $E(\mathcal{D},\varepsilon_1,\varepsilon_2)\leq E(\mathcal{D},\varepsilon_1\vee\varepsilon_2)$ since $\mathcal{B}_\varepsilon(\mu)\subset\mathcal{B}_{\varepsilon'}(\mu)$ for $\varepsilon'\geq\varepsilon$. Moreover, it is easy to check that $E(\mathcal{D},\varepsilon_1,\varepsilon_2)\gtrsim\mathfrak{m}_-(\mathcal{D},\varepsilon_1\vee\varepsilon_2)$, as defined in Lemma 4 (using the same proof), and so our lower bounds carry over.

We next focus on obtaining dual forms for $\mathsf{W}_p^{\varepsilon_1,\varepsilon_2}$ which are relevant for applications, although most of the primal properties shown for W_p^ε translate to this case as well. Our proof of Proposition 2 is already written for this general setting. Next, although the proof of Proposition 1 was given for the symmetric distance, an identical argument reveals that the following variant holds in general:

$$\mathsf{W}_p^\varepsilon(\mu,\nu)=\left(\frac{1-\varepsilon_1\varepsilon_2}{(1-\varepsilon_1)(1-\varepsilon_2)}\right)^{1/p}\inf_{\substack{\mu'\in\mathcal{B}_{\delta_1}(\mu)\\ \nu'\in\mathcal{B}_{\delta_2}(\nu)}}\mathsf{W}_p(\mu',\nu'),$$

where $\delta_1=\frac{\varepsilon_2(1-\varepsilon_1)}{1-\varepsilon_1\varepsilon_2}$ and $\delta_2=\frac{(1-\varepsilon_2)\varepsilon_1}{1-\varepsilon_1\varepsilon_2}$. We now translate Theorem 2 to the asymmetric case, following the same argument of the original proof to obtain

$$(1-\varepsilon_1)(1-\varepsilon_2)\mathsf{W}_p^{\varepsilon_1,\varepsilon_2}(\mu,\nu)^p=\sup_{f\in C_b(\mathcal{X})}(1-\varepsilon_2)\int f\,\mathrm{d}\mu+(1-\varepsilon_1)\int f^c\,\mathrm{d}\nu+(1-\varepsilon_1)\varepsilon_2\inf_x f(x)+(1-\varepsilon_2)\varepsilon_1\sup_x f(x).$$

Note that this matches the symmetric case when $\varepsilon_1=\varepsilon_2$ and matches the one-sided dual (10) when $\varepsilon_2=0$.

B.2 Relation to TV-Robustified W_p

In Balaji et al. (2020), the authors consider

$$\mathsf{W}_p^{\varepsilon,\mathsf{TV}}(\mu,\nu):=\inf_{\substack{\mu',\nu'\in\mathcal{P}(\mathcal{X})\\ \|\mu-\mu'\|_{\mathsf{TV}},\|\nu-\nu'\|_{\mathsf{TV}}\leq\varepsilon}}\mathsf{W}_p(\mu',\nu'),$$

(although χ^2 robustification is examined more thoroughly). We have the following relationship between this quantity and our robust distance.

Proposition 14 (Relation to $\mathsf{W}_p^{\varepsilon,\mathsf{TV}}$). *For $p \in [1,\infty]$ and $\mu,\nu \in \mathcal{P}_p(\mathcal{X})$,*

$$\mathsf{W}_p^{2\varepsilon,\mathsf{TV}}(\mu,\nu) \le \mathsf{W}_p^{\varepsilon}(\mu,\nu) \le (1-\varepsilon)^{-1/p}\,\mathsf{W}_p^{\varepsilon,\mathsf{TV}}(\mu,\nu).$$

Proof. To prove $\mathsf{W}_p^{2\varepsilon,\mathsf{TV}} \le \mathsf{W}_p^{\varepsilon}$, take $\mu',\nu' \in \mathcal{P}(\mathcal{X})$ optimal for (12), with $\mu \in \mathcal{B}_\varepsilon(\mu')$ and $\nu \in \mathcal{B}_\varepsilon(\nu')$. Writing $\mu = (1-\varepsilon)\mu' + \varepsilon\alpha$, for some $\alpha \in \mathcal{P}(\mathcal{X})$, we compute $\|\mu-\mu'\|_{\mathsf{TV}} = \varepsilon\|\mu'-\alpha\|_{\mathsf{TV}} \le 2\varepsilon$; similarly, $\|\nu-\nu'\|_{\mathsf{TV}} \le 2\varepsilon$. Hence, $\mathsf{W}_p^{2\varepsilon,\mathsf{TV}}(\mu,\nu) \le \mathsf{W}_p(\mu',\nu') = \mathsf{W}_p^{\varepsilon}(\mu,\nu)$.

For the opposite inequality, take any $\mu',\nu' \in \mathcal{P}(\mathcal{X})$ with $\|\mu-\mu'\|_{\mathsf{TV}}, \|\nu-\nu'\|_{\mathsf{TV}} \le \varepsilon$. Then, we can write $\mu' = \mu + \frac{\varepsilon}{2}(\alpha_+ - \alpha_-)$ and $\nu' = \nu + \frac{\varepsilon}{2}(\beta_+ - \beta_-)$, where $\alpha_+,\alpha_-,\beta_+,\beta_- \in \mathcal{P}(\mathcal{X})$. Next, set

$$\tilde{\mu}' = \frac{1}{1+\varepsilon}\mu + \frac{\varepsilon}{1+\varepsilon}\left(\frac{\alpha_+ + \beta_-}{2}\right) \in \mathcal{B}_{\frac{\varepsilon}{1+\varepsilon}}(\mu), \qquad \tilde{\nu}' = \frac{1}{1+\varepsilon}\nu + \frac{1}{1+\varepsilon}\left(\frac{\beta_+ + \alpha_-}{2}\right) \in \mathcal{B}_{\frac{\varepsilon}{1+\varepsilon}}(\nu),$$

and observe that

$$\mathsf{W}_p(\tilde{\mu}',\tilde{\nu}')^p = \frac{1}{1+\varepsilon}\mathsf{W}_p\left(\mu + \frac{\varepsilon}{2}(\alpha_+ + \beta_-), \nu + \frac{\varepsilon}{2}(\beta_+ + \alpha_-)\right)^p \le \frac{1}{1+\varepsilon}\mathsf{W}_p(\mu',\nu'),$$

where the last inequality follows by considering the transport plan which leaves β_- and α_- stationary while otherwise following an optimal plan for $\mathsf{W}_p(\mu',\nu')$. Finally, noting that $\tilde{\mu}'$ and $\tilde{\nu}'$ are feasible for the mass-addition formulation (5), we obtain

$$\mathsf{W}_p^{\varepsilon}(\mu,\nu) \le \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^{1/p}\mathsf{W}_p(\tilde{\mu}',\tilde{\nu}') \le (1-\varepsilon)^{-1/p}\,\mathsf{W}_p(\mu',\nu').$$

Infimizing over all such μ' and ν' gives $\mathsf{W}_p^{\varepsilon}(\mu,\nu) \le (1-\varepsilon)^{-1/p}\,\mathsf{W}_p^{\varepsilon,\mathsf{TV}}(\mu,\nu)$. □

B.3 Examples

On the real line, $\mathsf{W}_p^{\varepsilon}$ is often simple enough to compute, although it does not lend itself to many closed-form solutions. One useful class of examples is that of μ and ν which are absolutely continuous with respect to Lebesgue and whose supports have disjoint convex hulls, e.g., when μ is supported on the negative numbers and ν is supported on the positive numbers. In this case, due to the structure of optimal couplings for real distributions, we know that the optimal μ' and ν' simply cut off the extreme ends of the original measures. Assuming that the support of μ is to the left of that of ν, we find that $\mu' = \mu|_{[a,\infty)}$ and $\nu' = \nu|_{(-\infty,b]}$ for constants a,b such that $\mu'(\mathcal{X}) = \nu'(\mathcal{X}) = 1-\delta$. For example, if $\mu = \mathrm{Exp}(\lambda)$ and $\nu = (-\mathrm{Id})_\#\mu$, then $\mathsf{W}_\infty(\mu,\nu) = \infty$ but $\mathsf{W}_\infty^{\varepsilon}(\mu,\nu) \le 2\ln(1/\varepsilon)/\lambda$.

B.4 Experiment Details and Comparison with Existing Work

Full code is available on GitHub at https://github.com/sbnietert/robust-OT. We stress that we never had to adjust hyper-parameters or loss computations from their defaults, only adding the additional robustness penalty and procedures to corrupt the datasets. The implementation of WGAN-GP used for the robust GAN experiments in the main text was based on a standard PyTorch implementation (Cao, 2017), as was that of StyleGAN 2 (Varuna Jayasiri, 2020). The images presented in Fig. 4 were generated without any manual filtering after a predetermined number of batches (125k, batch size 64, for WGAN-GP; 100k, batch size 32, for StyleGAN 2). Training for the WGAN-GP and StyleGAN 2 experiments took 5 hours and 20 hours of compute, respectively, on a cluster machine equipped with a NVIDIA Tesla V100 and 14 CPU cores.

In addition to the experiments provided in the main text, we also compared our robustification method for GANs with the existing techniques of Balaji et al. (2020) and Staerman et al. (2021) (noting that the GAN described in Mukherjee et al. (2021) does not appear to scale to high-dimensional image data). Using publicly available code for these projects, we trained these robust GAN variants with their provided default options (using the continuous weighting scheme, recommended robustness strength, and DCGAN architecture for that of Balaji et al. (2020)) using the 50k CIFAR-10 training set of 32x32 images contaminated with 2632 images with uniform

	Robust GAN Method			
Epoch	Ours	Balaji et al. (2020)	Staerman et al. (2021)	Control (Balaji et al., 2020)
20k	38.61	60.92	58.19	59.19
60k	22.08	35.49	49.64	39.08
100k	18.87	30.44	40.81	31.40

Table 1: FID scores between uncontaminated CIFAR-10 image dataset and datasets generated by various robust GANs during training.

Figure 5: Samples generated by various robust GANs after 100k epochs of training.

noise in each pixel (giving a contamination fraction of 0.05). In the table and figure above, the control GAN is that of Balaji et al. (2020) without its robustification weights enabled, and our robust GAN is obtained by adding our robustness penalty with $\varepsilon = 0.07$ (two lines of code, as described in Section 6) to this control. Each GAN was trained for 100k epochs, taking approximately 12 hours of compute on a cluster machine equipped with a NVIDIA V100 GPU and 14 CPU cores. In addition to the raw generated images presented in Fig. 5, we also computed Frechet inception distance (FID) scores (Heusel et al., 2017) to track progress during training, provided in Table 1. Our method performs favorably without tuning, but more detailed empirical study is needed to separate the impact of various hyperparameters (for example, the poor relative performance of Staerman et al. (2021) may be due in part to its distinct architecture).

B.5 Presence of Regularization

Consider the Banach space $C_b(\mathcal{X})$ equipped with the uniform topology. Given a convex family of functions $\mathcal{F} \subseteq C_b(\mathcal{X})$ and a convex, lower semicontinuous regularizer $R : \mathcal{F} \to \mathbb{R}_+$, consider the statistical discrepancy measure between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ defined by

$$d_{\mathcal{F},R}(\mu,\nu) = \sup_{f\in\mathcal{F}} \int f \,\mathrm{d}(\mu-\nu) - R(f).$$

Then, we can verify that the conditions of Sion's minimax theorem apply as in Theorem 2 to obtain

$$\begin{aligned}\inf_{\substack{\mu'\geq\mu,\mu'(\mathcal{X})=1+\varepsilon\\ \nu'\geq\nu,\nu'(\mathcal{X})=1+\varepsilon}} d_{\mathcal{F},R}(\mu',\nu') &= \inf_{\alpha,\beta\in\mathcal{P}(\mathcal{X})} \sup_{f\in\mathcal{F}} \int f \,\mathrm{d}(\mu-\nu) + \varepsilon \int f \,\mathrm{d}(\alpha-\beta) + R(f)\\ &= \sup_{f\in\mathcal{F}} \int f \,\mathrm{d}(\mu-\nu) + \varepsilon \inf_{\alpha,\beta\in\mathcal{P}(\mathcal{X})} \int f \,\mathrm{d}(\alpha-\beta) + R(f)\\ &= \sup_{f\in\mathcal{F}} \int f \,\mathrm{d}(\mu-\nu) - \varepsilon \,\mathrm{Range}(f) - R(f)\end{aligned}$$

As an example, this applies when $\mathcal{F}$ is a neural-net family and $R(f) = \|\nabla f\|^p_{L^p(\kappa)}$ for some κ. For the case of WGAN-GP, where $R(f) = \lambda \mathbb{E}_{x\sim\kappa}\left[(\|\nabla f(X)\|_2 - 1)^2\right]$, convexity may fail to hold, but the final expression above still serves as a non-trivial lower bound for the robustified discrepancy measure. In this particular setting, near maximizers $f \in \mathcal{F}$ are approximately 1-Lipschitz for sufficiently large λ, suggesting that our original theory should apply up to some approximation error. We defer formal guarantees for non-convex regularizers for future work.

Appendix C Applications to Pufferfish Privacy

Pufferfish privacy (PP) (Kifer and Machanavajjhala, 2014; Song et al., 2017; Zhang et al., 2020) is a general privacy framework, which generalizes the popular notion of differential privacy (Dwork et al., 2006). For a measurable data space $\mathcal{X}$, the framework consists of three components: (i) a collection $\mathcal{S}$ of measurable subsets of $\mathcal{X}$ called secrets (e.g., facts about the database that one might wish to hide); (ii) a set of discriminative pairs $\mathcal{Q} \subseteq \mathcal{S} \times \mathcal{S}$ (i.e., pairs of secret events (S, T) that should be indistinguishable); and (iii) a class of potential data distributions $\Theta \subseteq \mathcal{P}(\mathcal{X})$. For any $\theta \in \Theta$, let $X_\theta \sim \theta$ denote the dataset sampled from that distribution. A randomized algorithm satisfies PP w.r.t. $(\mathcal{S}, \mathcal{Q}, \Theta)$ if all pairs of secrets in $\mathcal{Q}$ are indistinguishable as follows.

Definition 2 (Pufferfish Privacy (Kifer and Machanavajjhala, 2014))**.** A (possibly randomized) mechanism $\mathsf{M} : \mathcal{X} \to \mathcal{Y}$ is (ε, δ)-Pufferfish private in a framework $(\mathcal{S}, \mathcal{Q}, \Theta)$ if for all data distributions $\theta \in \Theta$, secret pairs $(S, T) \in \mathcal{Q}$ with $\mathbb{P}(X_\theta \in S) > 0$ and $\mathbb{P}(X_\theta \in T) > 0$, and measurable $Y \subseteq \mathcal{Y}$, we have

$$\mathbb{P}\big(\mathsf{M}(X_\theta) \in Y \big| S\big) \leq e^\varepsilon \mathbb{P}\big(\mathsf{M}(X_\theta) \in Y \big| T\big) + \delta,$$

where $\mathbb{P}$ is the underlying joint probability measure.

Remark 7 (Special cases)**.** A notable special case of PP is differential privacy (Dwork et al., 2006), where $\mathcal{S}$ is the set of all possible databases, $\mathcal{Q}$ contains all pairs that differ in a single entry, and $\Theta = \mathcal{P}(\mathcal{X})$ (i.e., no distributional assumptions are made, and privacy is guaranteed in the worst case). Another special case is *attribute privacy* (Zhang et al., 2020), which protects global statistical properties of data attributes. In this case, $\mathcal{S}$ is the value of a function evaluated on the data, $\mathcal{Q}$ contains pairs of function values, and Θ captures assumptions on how the data were sampled and correlations across attributes in the data.

A general approach for publishing the value of a function $f : \mathcal{X} \to \mathbb{R}$ on a dataset X_θ with $(\varepsilon, 0)$-PP is the *Wasserstein Mechanism* (Song et al., 2017). This mechanism computes the true value $f(X_\theta)$, and then adds Laplace noise that scales with the maximal W_∞ distance between conditional distributions of $f(X_\theta)$ given any pair of secrets from $\mathcal{Q}$. Denoting this maximized distance by W, the mechanism outputs $\mathsf{M}(X_\theta) = f(X_\theta) + Z$, where $Z \sim \mathrm{Lap}(W/\varepsilon)$, which can be shown to be $(\varepsilon, 0)$-PP for any framework $(\mathcal{S}, \mathcal{Q}, \Theta)$; see Algorithm 1 with $\delta = 0$. The privacy guarantee relies on the fact that $\mathsf{W}_\infty(\mu, \nu) \leq \varepsilon$ if and only if there exists a coupling (A, B) of μ and ν such that $|A - B| \leq \varepsilon$. If we are instead interested in (ε, δ)-PP, we only need $|A - B| \leq \varepsilon$ to hold with probability $1 - \delta$. We next show that this property is characterized precisely by W_∞^δ.

Lemma 6 (Coupling property)**.** *If $\mathsf{W}_\infty^\delta(\mu, \nu) \leq \varepsilon$, then there exists a coupling (X, Y) for μ, ν such that $|X - Y| \leq \varepsilon$ with probability at least $1 - 2\delta$. Conversely, if there exists such a coupling, then $\mathsf{W}_\infty^{2\delta}(\mu, \nu) \leq \varepsilon$.*

Proof. If $\mathsf{W}_\infty^\delta(\mu, \nu) \leq \varepsilon$, then there exists $\nu', \nu' \in \mathcal{P}(\mathcal{X})$ such that $\mu \in \mathcal{B}_\delta(\mu')$, $\nu \in \mathcal{B}_\delta(\nu')$, and $\mathsf{W}_\infty(\mu', \nu') \leq \varepsilon$. By the gluing lemma, there exists a joint distribution for (X, X', Y, Y') with $X \sim \mu, X' \sim \mu', Y \sim \nu, Y \sim \nu'$ such that $\mathbb{P}(X \neq X'), \mathbb{P}(Y \neq Y') \leq \delta$ and $|X' - Y'| \leq \varepsilon$. By a union bound, $|X - Y| \leq \varepsilon$ with probability at least $1 - 2\delta$.

On the other hand, if there exists a coupling $\pi = (1 - \delta)\pi_g + \delta\pi_b \in \Pi(\mu, \nu)$ for $\pi_g, \pi_b \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ such that $\|d\|_{L^\infty(\pi_g)} \leq \varepsilon$ (i.e., restating the converse assumption), then taking marginals gives $\mu \in \mathcal{B}_\delta((\pi_g)_1)$ and $\nu \in \mathcal{B}_\delta((\pi_g)_2)$. Hence, π_g is feasible for the (near) coupling problem Proposition 6, and $\mathsf{W}_\infty^\delta(\mu, \nu) \leq \varepsilon$. □

We can use this fact to define a Robust Wasserstein Mechanism that achieves (ε, δ)-PP.

Theorem 3 (Robust Wasserstein Mechanism)**.** *For any PP framework $(\mathcal{S}, \mathcal{Q}, \Theta)$ and $\delta \leq 1/2$, the Robust Wasserstein Mechanism from Algorithm 1 is $(\varepsilon, 2\delta)$-PP.*

Algorithm 1 Robust Wasserstein Mechanism $(f, \{\mathcal{S}, \mathcal{Q}, \Theta\}, \varepsilon, \delta)$

for all $(S, T) \in \mathcal{Q}$ and $\theta \in \Theta$ such that $\mathbb{P}(X_\theta \in S) > 0$ and $\mathbb{P}(X_\theta \in T) > 0$
  Set $\mu^\theta_{|S} := \mathcal{L}\big(f(X_\theta)\big|S\big)$ and $\mu^\theta_{|T} := \mathcal{L}\big(f(X_\theta)\big|T\big)$
Calculate $W_\delta = \sup_{\substack{(S,T)\in\mathcal{Q}, \theta\in\Theta \\ \mathbb{P}(X_\theta\in S)>0, \mathbb{P}(X_\theta\in T)>0}} \mathsf{W}_\infty^\delta\big(\mu^\theta_{|S}, \mu^\theta_{|T}\big)$.
Sample $Z \sim \mathrm{Lap}(W_\delta/\varepsilon)$.
Return $\mathsf{M}(X_\theta) = f(X_\theta) + Z$

Proof. The proof follows that of the original Wasserstein mechanism established in Song et al. (2017). We will write p_X to denote the probability density function of an absolutely continuous random variable X. Let M denote the robust Wasserstein mechanism and consider the distributions $\mu^\theta_{|S}, \mu^\theta_{|T}$ associated with a secret pair $(S,T) \in \mathcal{Q}$ such that $\mathbb{P}(X_\theta \in S)$ and $\mathbb{P}(X_\theta \in T)$ are positive. Because $\mathsf{W}_p^\varepsilon(\mu^\theta_{|S}, \mu^\theta_{|T}) \le W_\delta$, we know that there exists a coupling $\pi \in \Pi(\mu^\theta_{|S}, \mu^\theta_{|T})$ as guaranteed by Lemma 6. We decompose $\pi = (1-2\delta)\pi_g + 2\delta\pi_b$ into pieces so that $\pi_g \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$ has the W_δ distance guarantee and π_b is arbitrary. Then, for any measurable $Y \subseteq \mathcal{Y}$, we have

$$\begin{aligned}
p_{\mathsf{M}(X_\theta)|S}(w) &= \int p_{f(X_\theta)|S}(a)\, p_Z(w-a)\,\mathrm{d}a \\
&= \int p_{f(X_\theta)|S}(a) e^{-\varepsilon|w-a|/W_\delta}\,\mathrm{d}a \\
&= \int e^{-\varepsilon|w-a|/W_\delta}\,\mathrm{d}\pi(a,b) \\
&\le (1-2\delta)\int e^{-\varepsilon|w-a|/W_\delta}\,\mathrm{d}\pi_g(a,b) + 2\delta,
\end{aligned}$$

where we have simply used the definition of the mechanism and the fact that $e^{-\varepsilon|w-a|/W_\delta} \le 1$. Continuing, we compute

$$\begin{aligned}
p_{\mathsf{M}(X_\theta)|S}(w) &\le \int_{\{(a,b):|a-b|\le W_\delta\}} e^{-\varepsilon|w-a|/W_\delta}\,\mathrm{d}\pi_g(a,b) + 2\delta \\
&\le e^\varepsilon \int_{\{(a,b):|a-b|\le W_\delta\}} e^{-\varepsilon|w-b|/W_\delta}\,\mathrm{d}\pi_g(a,b) + 2\delta \\
&\le e^\varepsilon \int e^{-\varepsilon|w-b|/W_\delta}\,\mathrm{d}\pi(a,b) + 2\delta \\
&\le e^\varepsilon \int e^{-\varepsilon|w-b|/W_\delta}\, p_{f(X_\theta)|T}(b)\,\mathrm{d}b + 2\delta \\
&\le e^\varepsilon p_{\mathsf{M}(X_\theta)|T}(w) + 2\delta,
\end{aligned}$$

using the distance property of the coupling, performing a change of variables, and using the definition of the mechanism. Since the previous argument holds for any secret pair, we conclude that the mechanism is $(\varepsilon, 2\delta)$-PP private. □

This result suggests that W_∞^δ naturally generalizes the relation between W_∞ and $(\varepsilon,0)$-PP to the (ε,δ)-PP regime. In applications where a positive δ is tolerable, the Robust Wasserstein Mechanism allows a significantly lower scale of noise relative to the original Wasserstein Mechanism from Song et al. (2017) — i.e., $W_\delta \ll W$ for some Pufferfish frameworks — corresponding to a substantial improvement in accuracy. To demonstrate this, consider the following toy example.

Example 1. *Consider a toy scenario where a bank has 100 customers of two types, A and B, with at most 10 of type B. The incomes of Type A and Type B customers are distributed according to μ_A and μ_B, respectively. Suppose that incomes are independent conditioned on the customer types, and the bank wishes to publish the combined income of its customers while keeping the count of B users private. The corresponding PP framework has $\mathcal{S}$ as the set of possible number of type B customers (between 1 and 10), $\mathcal{Q}$ contains all distinct pairs from $\mathcal{S}$, and each $\theta \in \Theta$ gives the joint distribution over incomes given a selection of types.*

In order to publish the sum of incomes while guaranteeing (ε,δ)-PP, the Robust Wasserstein Mechanism dictates that it suffices to add Laplace noise at scale $\mathsf{W}_\infty^\delta(\mu_A^{*100}, \mu_A^{*90} * \mu_B^{*10})/\varepsilon$, where γ^{*k} denotes a k-fold convolution of γ with itself. Assume all incomes are less than \$100k, except Type B customers who have probability δ' of being millionaires, with $1-(1-\delta')^{10} \ll \delta$. In this case, the classic Wasserstein Mechanism would require noise at scale $\sim 10^8/\varepsilon$, while the δ-robust mechanism only needs noise at scale $\sim 10^7/\varepsilon$. Thus, utilizing W_∞^δ and its connection to (ε,δ)-PP allows injecting an order of magnitude less noise.

Remark 8 ((ε,δ)-PP and W_p)**.** Using the robust Wasserstein distance we can further connect PP and classic W_p. This may be practically advantageous, as finite-order Wasserstein distances are often easier to compute than W_∞. First note that Markov's inequality implies that $\mathsf{W}_\infty^\varepsilon(\mu,\nu) = O\big(\mathsf{W}_p(\mu,\nu)/\delta^{1/p}\big)$. Hence, the mechanism that adds noise at scale $W_p/(\varepsilon\delta^{1/p})$, where W_p is a uniform bound on $\mathsf{W}_p(\mu^\theta_{|S}, \mu^\theta_{|T})$, is (ε,δ)-Pufferfish private.