A New Robust Partial p-Wasserstein-Based Metric for Comparing Distributions

Sharath Raghvendra [1] **Pouyan Shirzadian** [2] **Kaiyi Zhang** [2]

Abstract

The 2-Wasserstein distance is sensitive to minor geometric differences between distributions, making it a very powerful dissimilarity metric. However, due to this sensitivity, a small outlier mass can also cause a significant increase in the 2-Wasserstein distance between two similar distributions. Similarly, sampling discrepancy can cause the empirical 2-Wasserstein distance on n samples in $\mathbb{R}^2$ to converge to the true distance at a rate of $n^{-1/4}$, which is significantly slower than the rate of $n^{-1/2}$ for 1-Wasserstein distance. We introduce a new family of distances parameterized by $k \geq 0$, called k-RPW that is based on computing the partial 2-Wasserstein distance. We show that (1) k-RPW satisfies the metric properties, (2) k-RPW is robust to small outlier mass while retaining the sensitivity of 2-Wasserstein distance to minor geometric differences, and (3) when k is a constant, k-RPW distance between empirical distributions on n samples in $\mathbb{R}^2$ converges to the true distance at a rate of $n^{-1/3}$, which is faster than the convergence rate of $n^{-1/4}$ for the 2-Wasserstein distance. Using the partial p-Wasserstein distance, we extend our distance to any $p \in [1, \infty]$. By setting parameters k or p appropriately, we can reduce our distance to the total variation, p-Wasserstein, and the Lévy-Prokhorov distances. Experiments show that our distance function achieves higher accuracy in comparison to the 1-Wasserstein, 2-Wasserstein, and TV distances for image retrieval tasks on noisy real-world data sets.

1. Introduction

Given two probability distributions μ and ν with supports $\mathcal{A}$ and $\mathcal{B}$, let, for any $(a, b) \in \mathcal{A} \times \mathcal{B}$, $d(a, b)$ be the cost of moving a unit mass from a to b. A *transport plan* γ is a coupling of μ and ν, i.e., a joint distribution over the support $\mathcal{A} \times \mathcal{B}$ whose first and second marginals are μ and ν. For $p \geq 1$, consider the case where the support of μ and ν lie in a metric space $(\mathcal{X}, c)$ with a unit diameter, i.e., $c(a, b) \leq 1$ for any pair $(a, b) \in \mathcal{X} \times \mathcal{X}$ and the cost of moving unit mass from a to b is given by $d(a, b) = c(a, b)^p$. For any transport plan γ between μ and ν, the cost of γ is defined as

$$w_p(\gamma) := \left(\int_{\mathcal{X} \times \mathcal{X}} c(x, y)^p \, \mathrm{d}\gamma(x, y) \right)^{1/p}.$$

Let γ^* be a minimum-cost transport plan between μ and ν. Then, the *p-Wasserstein distance* between μ and ν is defined as $W_p(\mu, \nu) := w_p(\gamma^*)$.

The p-Wasserstein distance is a powerful metric for measuring similarities between probability distributions. Due to its numerous mathematical properties, the p-Wasserstein distance has found diverse applications including in machine learning (Chang et al., 2023; Chuang et al., 2022; Mohajerin Esfahani & Kuhn, 2018; Janati et al., 2019; Luise et al., 2018; Oquab et al., 2023; Vincent-Cuaz et al., 2021), computer vision (Backurs et al., 2020; Gupta et al., 2010; Lai et al., 2022), and natural language processing (Alvarez-Melis & Jaakkola, 2018; Huang et al., 2016; Yurochkin et al., 2019). One can estimate the p-Wasserstein distance between two unknown distributions μ and ν by simply taking n samples from each μ and ν and then computing the p-Wasserstein distance between the discrete distributions over these samples (each sample point is assigned a mass of $1/n$). For $p \in [1, \infty)$, it is well-known that as $n \to \infty$, this *empirical p-Wasserstein* distance converges to the true p-Wasserstein distance. Due to this law of weak convergence, the p-Wasserstein distance is used as a loss function in training generative models (Arjovsky et al., 2017; Genevay et al., 2018; Salimans et al., 2018).

The p-Wasserstein distance is sensitive to geometric dissimilarities between the distributions. Consider two distributions μ and $\nu = (1 - \delta)\mu + \delta\nu'$ that differ only by a mass of δ. The p-Wasserstein distance between μ and ν can be as high as $\delta^{1/p} W_p(\mu, \nu')$. Thus, as p increases, the $W_p(\mu, \nu)$ increases by a rate of $\delta^{1/p}$, making W_p more sensitive to such differences for larger values of p. The higher sensitivity of p-Wasserstein distance for $p > 1$ makes it an attractive choice as a dissimilarity metric between distributions. Con-

[1]North Carolina State University [2]Virginia Tech. Correspondence to: S. Raghvendra <skraghve@ncsu.edu>, P. Shirzadian <pshirzadian@vt.edu>, K. Zhang <kaiyiz@vt.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

sequently, it can be used in clustering (El Malki et al., 2020; Zhuang et al., 2022) and barycenter computation (Claici et al., 2018; Cuturi & Doucet, 2014; Vaskevicius & Chizat, 2023).

The higher sensitivity of p-Wasserstein distance for larger values of p also makes it susceptible to noise of two types: outliers and sampling discrepancy. Consider μ and $\nu = 0.99\mu + 0.01\nu'$ and $W_p(\mu, \nu') = 1$, i.e., we add an *outlier* mass of $\delta = 0.01$ that is placed at a distance 1 from μ. In this case, μ and ν differ in only 1% of mass and yet, the distance between μ and ν is 0.1 when $p = 2$, 0.21 when $p = 3$, and 1 when $p = \infty$. Thus, for $p \geq 2$, outliers can disproportionately increase the distance between distributions.

Similar to outliers, sampling discrepancies in empirical distributions can also contribute disproportionately to the overall p-Wasserstein distance. As a result, in 2-dimensions, the convergence rate of the empirical p-Wasserstein distance to the true distance drops to $n^{-1/2p}$ and for $p = \infty$, the empirical distance does not even converge to the real one (Fournier & Guillin, 2015). To understand this phenomenon better, consider $p = 2$ and a discrete distribution μ having two points a and b in its support, each assigned a probability mass of $1/2$. Let $c(a, b) = 1$. Consider now two sets X and Y of n samples drawn from μ and let μ_X (resp. μ_Y) be the discrete distributions with points of X (resp. Y) as the support and a mass of $1/n$ at each point in the support. Note that $\mathbb{E}[|\mu_X(a) - \mu_Y(a)|] = \Theta(1/\sqrt{n})$ and therefore, $W_2(\mu_X, \mu_Y) \approx n^{-1/4}$ and $W_p(\mu_X, \mu_Y) \approx n^{-1/2p}$. Thus, the rate of convergence for $p \geq 2$ is slower than for the case with $p = 1$. Therefore, one needs significantly more samples to get an accurate estimate of the true 2-Wasserstein distance. This restricts the use of 2-Wasserstein distance (and also other higher values of p) as a loss function in learning tasks.

One way to overcome the impact of noise from outliers or sampling discrepancy is by using the partial p-Wasserstein distance. For *α-partial p-Wasserstein distance*, one wishes to compute the cheapest cost of a transport plan that transports α mass between distributions μ and ν. Such transport plan is referred to as *α-optimal partial transport plan* (or simply *α-partial OT plan*). Given two distributions μ and $\tilde{\nu} = (1 - \delta)\nu + \delta\nu'$, and under reasonable assumptions on the outlier distribution ν', one can show that the transport plan associated with a $(1-\delta)$-partial p-Wasserstein distance will transport mass only from the inliers. This observation was used to eliminate the impact of outliers in two distributions and applied to many ML tasks (Choi et al., 2024; Le et al., 2021; Nietert et al., 2023). Most of these applications assume that the value of δ is given; see (Caffarelli & McCann, 2010; Chapel et al., 2020; Figalli, 2010; Nietert et al., 2022). Recently, Phatak et al. (2022) introduced the idea of *OT-profile*, which is a function that maps any $\alpha \in [0, 1]$ to the α-partial p-Wasserstein distance between μ and ν. They showed that this function is a non-decreasing function[1], which can be used to also identify the value of δ. All existing works that use partial p-Wasserstein distance to identify outliers are described for pairs of distribution. It is not clear how one can apply this distance on a set containing noisy distributions.

Additionally, there are two major drawbacks of using $(1-\delta)$-partial p-Wasserstein distance as a dissimilarity measure on sets of probability distributions.

- The $(1 - \delta)$-partial p-Wasserstein distance does not satisfy the triangle inequality, and
- For two distributions μ and ν that differ by a mass less than δ, the $(1 - \delta)$-partial p-Wasserstein distance will be 0, i.e., this cost is not sensitive to minor geometric differences in distributions.

In another line of work, given a parameter $\lambda > 0$, Mukherjee et al. (2021) presented a robust distance called the λ-ROBOT, which is simply the p-Wasserstein cost between μ and ν under the truncated ground distance metric $c_\lambda(a, b) = \min\{c(a, b), 2\lambda\}$ [2]. Although λ-ROBOT is a metric, it remains sensitive to outliers and sampling discrepancies. For instance, similar to the p-Wasserstein distance, a mass of δ can disproportionately increase λ-ROBOT by $2\lambda\delta^{1/p}$. Also, the convergence rate of the empirical λ-ROBOT to the true λ-ROBOT in two-dimensional space would be $2\lambda n^{-1/2p}$.

An important open question is the following:

Can we design a new metric that, for $p > 1$, retains the sensitivity of p-Wasserstein distance to minor geometric differences in the distributions, but is robust to noise?

Our Results: For any $k \geq 0$, we introduce a partial p-Wasserstein distance-based metric called (p, k)-RPW and we denote it by $\Pi_{p,k}(\cdot, \cdot)$. Our distance is simply the smallest ε such that the $(1 - \varepsilon)$-partial p-Wasserstein distance is at most $k\varepsilon$.

Our distance combines the total variation distance with the p-Wasserstein distance. Recollect that the total variation distance between μ and ν is the mass that remains after all

[1] Their function maps α to the pth power of the α-partial p-Wasserstein distance. In this paper, however, we assume that the function maps α to the partial p-Wasserstein distance and not its p^{th} power.

[2] Originally, Mukherjee et al. (2021) presented λ-ROBOT as the 1-Wasserstein distance between μ and ν under $c_\lambda(\cdot, \cdot)$. For any $p > 1$, one can extend their distance by computing the p-Wasserstein distance under $c_\lambda(\cdot, \cdot)$.

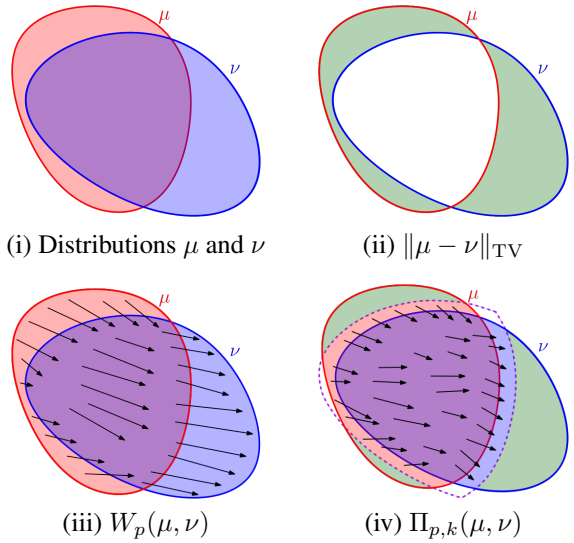


Figure 1. Interpretations of different distance functions.

the co-located mass is transported. In Figure 1 (ii), the mass inside the green region represents the TV distance between two distributions. The p-Wasserstein distance measures the cost of the cheapest transport plan that leaves no mass behind. In Figure 1 (iii), the cost of moving all the mass from the red region to the blue region represents the p-Wasserstein distance. Our distance balances the two, i.e., we find an ε such that a transport plan that leaves ε mass behind has a cost of $k\varepsilon$. In Figure 1 (iv), our distance balances the cost of moving mass from the red region to the blue region with the mass remaining inside the green region. The robustness of our distance follows from the observation that noisy mass will be part of the green region (i.e., mass that is not transported) and therefore, cannot contribute disproportionately to the cost. We establish the following properties for our distance function:

- *Metric Property:* For any choice of $p \geq 1$ and $k \geq 0$, the distance (p,k)-RPW is a metric. Unlike the $(1-\delta)$-partial p-Wasserstein distance, our distance function satisfies the triangle inequality. Furthermore, unlike the $(1-\delta)$-partial p-Wasserstein distance, where two distributions μ and ν can have a cost of 0 even if they differ by a mass of δ, for any two distributions μ and ν with $\mu \neq \nu$, $\Pi_{p,k}(\mu,\nu) > 0$. See Theorem 2.1.
- *Robust to Outliers:* Given two distributions μ and ν, adding a mass of δ to ν will change $\Pi_{p,k}(\mu,\nu)$ by at most $\pm\delta$. In other words, an outlier mass of $\delta = 0.01$ cannot increase the $\Pi_{p,k}(\mu,\nu)$ by more than 0.01. Recollect that this can be as high as 0.1 for 2-Wasserstein distance and 0.21 for 3-Wasserstein distance. See Theorem 3.1.
- *Robust to Sampling Discrepancy:* In 2 dimensions, the $(p,1)$-RPW between empirical distributions converges to the true $(p,1)$-RPW distance at a rate of $n^{-\frac{p}{4p-2}}$. In contrast, the rate of convergence of the 2-Wasserstein and the λ-ROBOT distances are $n^{-1/2p}$ and $2\lambda n^{-1/2p}$, respectively. Note that, for $p=\infty$, our distance converges at the rate of $n^{-1/4}$ whereas the ∞-Wasserstein distance does not converge. Our results extend to any dimension. For $d \geq 2$ and $p > \frac{d}{2}$, we show that the convergence rate of the empirical $(p,1)$-RPW is significantly faster than that of the p-Wasserstein distance. See Theorem 3.7.

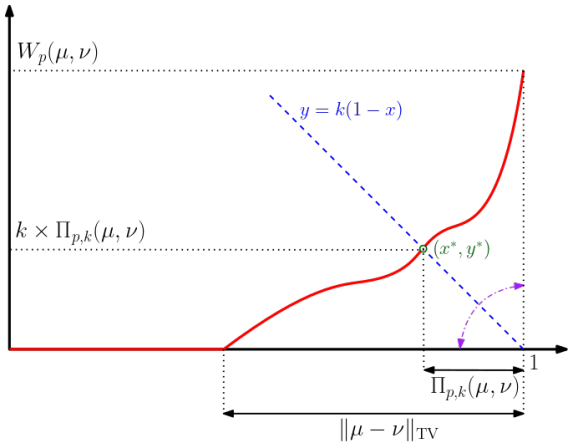


Figure 2. Interpretation of distances based on the OT-profile.

Alternatively, in Figure 2, suppose point (x^*, y^*) is the intersection point of the line $y = k(1-x)$ with the OT-profile. Then, our distance is simply $(1-x^*)$. Note that when $k=0$, our distance becomes the total variation distance. When we set k to be sufficiently large, our distance becomes $W_p(\mu,\nu)/k$. In this sense, our distance interpolates between the total variation distance and the p-Wasserstein distance. By choosing the parameters k or p correctly, we can reduce our distance to several well-known distances.

- *Relation to Lévy-Prokhorov distance:* $(\infty,1)$-RPW between any two distributions μ and ν is equal to their Lévy-Prokhorov distance. See Lemma 4.1.
- *Relation to total variation distance:* For any $p \geq 1$, $(p,0)$-RPW between any two distributions μ and ν is equal to the total variation distance between μ and ν. See Lemma 4.2.
- *Relation to p-Wasserstein distance:* For any $p \geq 1$, as $k \to \infty$, $k \times \Pi_{p,k}(\mu,\nu)$ approaches the p-Wasserstein distance. See Lemma 4.3.

In our experiments, we use our distance from a query image to rank images in a database of noisy images. Using this, we identify the top ten images in our database that are similar to

the query. For the MNIST, CIFAR-10, and COREL datasets, our distance produces a higher accuracy in comparison to the accuracy produced by the 1-Wasserstein, 2-Wasserstein, and the TV distances.

1.1. Notations.

For any distribution μ defined over a compact set $\mathcal{X}$, let $\mathcal{M}(\mu) := \int_{\mathcal{X}} \mathrm{d}\mu(x)$ denote the total mass of μ. For a metric space $(\mathcal{X}, c)$, define the diameter of $\mathcal{X}$ as $\max_{(a,b)\in\mathcal{X}\times\mathcal{X}} c(a,b)$. For any pair of distributions μ and ν defined over $(\mathcal{X}, c)$ and parameters $p \geq 1$ and $\alpha \in [0,1]$, let $W_{p,\alpha}(\mu,\nu)$ denote the α-partial p-Wasserstein distance between μ and ν.

2. Robust Partial p-Wasserstein Metric

Given two probability distributions μ and ν defined over a metric space $(\mathcal{X}, c)$ with a unit diameter and any parameters $p \geq 1$ and $k \geq 0$, we define the (p,k)-*Robust Partial p-Wasserstein distance* or simply (p,k)-*RPW* between μ and ν, denoted by $\Pi_{p,k}(\mu,\nu)$, to be the minimum value $\varepsilon \geq 0$ such that the $(1-\varepsilon)$-partial p-Wasserstein distance between μ and ν is at most $k\varepsilon$; more precisely,

$$\Pi_{p,k}(\mu,\nu) = \inf\{\varepsilon \geq 0 \mid W_{p,1-\varepsilon}(\mu,\nu) \leq k\varepsilon\}. \quad (1)$$

Alternatively, let $P = (x^*, y^*)$ be the intersection point of the OT-profile curve with the line $y = k(1-x)$. Then, (p,k)-RPW between μ and ν would be $\Pi_{p,k}(\mu,\nu) = 1 - x^*$.

We show that (p,k)-RPW distance satisfies all the metric properties. The triangle inequality is the only property for which the proof is non-trivial. We provide a sketch of the proof below; see Appendix A.1 for details.

For any three probability distributions μ, ν, and κ, suppose $\Pi_{p,k}(\mu,\kappa) = \varepsilon_1$ and $\Pi_{p,k}(\kappa,\nu) = \varepsilon_2$. Let γ_1 denote a $(1-\varepsilon_1)$-partial OT plan from μ to κ and γ_2 be a $(1-\varepsilon_2)$-partial OT plan from κ to ν. In Figure 3, the blobs in the left, middle, and right show the distributions μ, κ, and ν, respectively, and the blue (resp. red) arrows correspond to the transport plan γ_1 (resp. γ_2). Let κ_1 (resp. κ_2) be the mass of κ that is transported from μ (resp. to ν) by γ_1 (resp. γ_2) (shown in Figure 3 by the blue (resp. red) region inside the distribution κ). Define κ_c to be the distribution of mass of κ that is common to both κ_1 and κ_2 (the purple region inside the distribution κ in Figure 3). Note that the total mass of κ_1 that is not transported by γ_2 is at most ε_2; therefore,

$$\mathcal{M}(\kappa_c) \geq \mathcal{M}(\kappa_1) - \varepsilon_2 = 1 - \varepsilon_1 - \varepsilon_2. \quad (2)$$

Define μ_c (resp. ν_c) to be the distribution whose mass is transported to (resp. from) κ_c in γ_1 (resp. γ_2). In Figure 3, the distribution μ_c (resp. ν_c) is depicted by the purple region

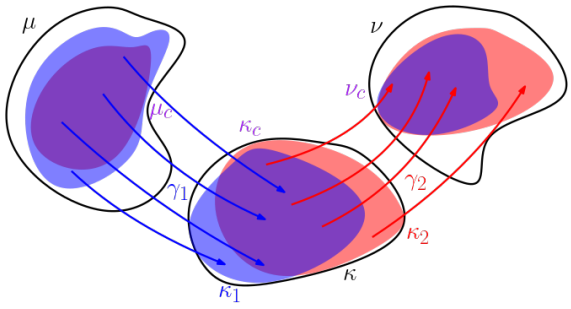


Figure 3. The triangle inequality of the RPW distance function.

inside distribution μ (resp. ν). From Equation (2),

$$\mathcal{M}(\mu_c) = \mathcal{M}(\nu_c) = \mathcal{M}(k_c) \geq 1 - \varepsilon_1 - \varepsilon_2 \quad (3)$$

Therefore, we have

$$\begin{aligned} W_{p,1-\varepsilon_1-\varepsilon_2}(\mu,\nu) &\leq W_p(\mu_c,\nu_c) \\ &\leq W_p(\mu_c,\kappa_c) + W_p(\kappa_c,\nu_c) \\ &\leq k(\varepsilon_1+\varepsilon_2). \end{aligned} \quad (4)$$

The second inequality follows from the triangle inequality of p-Wasserstein distances and the third inequality follows from the definition of our distance. Furthermore, since $\Pi_{p,k}(\mu,\nu)$ is the minimum ε with $W_{p,1-\varepsilon}(\mu,\nu) \leq k\varepsilon$, from Equation (4), $\Pi_{p,k}(\mu,\nu) \leq \varepsilon_1 + \varepsilon_2$, as desired.

Theorem 2.1. *Given a metric space $(\mathcal{X}, c)$ with a unit diameter and any parameters $p \geq 1$ and $k \geq 0$, the (p,k)-RPW distance function $\Pi_{p,k}(\cdot,\cdot)$ for all probability distributions defined over $(\mathcal{X}, c)$ is a metric.*

The following lemma highlights a useful feature of our metric, which is used in deriving its important properties.

Lemma 2.2. *Given two probability distributions μ and ν defined over a metric space $(\mathcal{X}, c)$ with a unit diameter and parameters $p \geq 1$ and $k \geq 0$, suppose $W_{p,1-\alpha}(\mu,\nu) = k\beta$ for some $\alpha, \beta \geq 0$. Then, $\Pi_{p,k}(\mu,\nu) \leq \max\{\alpha,\beta\}$. Furthermore, if $k \neq 0$, then $\Pi_{p,k}(\mu,\nu) \leq \min\{\alpha,\beta\}$.*

3. Robustness Properties

In Section 3.1, we show that an outlier mass of δ cannot increase the RPW distance by more than δ, i.e., the RPW distance is robust to outliers. In Section 3.2, we show that the rate of convergence of the empirical RPW distance to the real RPW distance is asymptotically smaller than the rate of convergence for p-Wasserstein distance. Thus, we show that the RPW distance is more robust to outliers as well as sampling discrepancies than the p-Wasserstein distance.

3.1. Robustness to Outlier Noise

For $\delta \in (0,1)$, let $\tilde{\nu} := (1-\delta)\nu + \delta\nu'$ be a noisy distribution obtained from ν contaminated with δ mass from a noise

distribution ν'. In Theorem 3.1, we show that by distorting the noise distribution ν', an adversary cannot arbitrarily change the (p,k)-RPW distance between μ and $\tilde{\nu}$.

Theorem 3.1. *For any probability distributions μ, ν, and ν' defined over a metric space $(\mathcal{X}, c)$ with a unit diameter and parameters $p \geq 1, k \geq 0$, and $\delta \in (0,1)$, let $\tilde{\nu} = (1-\delta)\nu + \delta\nu'$. Then,*

$$\Pi_{p,k}(\mu,\nu) - \delta \leq \Pi_{p,k}(\mu,\tilde{\nu}) \leq (1-\delta)\Pi_{p,k}(\mu,\nu) + \delta.$$

For a distribution μ and a noisy distribution $\tilde{\mu}$ that differs from μ by only a δ fraction of mass (i.e., $\|\mu - \tilde{\mu}\|_{TV} = \delta$), consider the following assumption:

(A1) The $(1-\frac{\delta}{10})$-partial p-Wasserstein distance between μ and $\tilde{\mu}$ is at least $\frac{1}{2}W_p(\mu,\tilde{\mu})$.

Assuming (A1), in the following lemma, we show that the (p,k)-RPW distance between μ and $\tilde{\mu}$ is proportionate to $\min\left\{\delta, \frac{1}{k}W_p(\mu,\tilde{\mu})\right\}$.

Lemma 3.2. *For a probability distribution μ defined over a metric space $(\mathcal{X}, c)$ with a unit diameter and $\delta > 0$, let $\tilde{\mu}$ be a probability distribution that differs from μ by a δ fraction of mass satisfying assumption (A1). Then, for any parameters $p \geq 1$ and $k > 0$,*

$$\Pi_{p,k}(\mu,\tilde{\mu}) = \Theta\left(\min\left\{\delta, \frac{1}{k}W_p(\mu,\tilde{\mu})\right\}\right).$$

Intuitively, if $\tilde{\mu}$ is only slightly different from μ, i.e., $W_p(\mu,\tilde{\mu}) \leq k\delta$, then the sensitivity of our metric would be similar to that of the p-Wasserstein distance. On the other hand, if $\tilde{\mu}$ is far from μ (i.e., the δ fraction of the mass of $\tilde{\mu}$ that is different from μ is an outlier noise and disproportionately increases the p-Wasserstein distance), then the sensitivity of (p,k)-RPW is bounded by δ.

3.2. Robustness to Sampling Discrepancies.

Next, we show that in the 2-dimensional Euclidean space, the rate of convergence of the empirical $(p,1)$-RPW to the true distance is $\tilde{O}(n^{-\frac{p}{4p-2}})$, which is significantly faster than the convergence rate of $\tilde{O}(n^{-\frac{1}{2p}})$ of the empirical p-Wasserstein distance (Fournier & Guillin, 2015)[3]. In particular, for $p = \infty$, the convergence rate of our metric is $\tilde{O}(n^{-\frac{1}{4}})$, whereas the empirical p-Wasserstein distance does not converge to the true distance. For simplicity in presentation, we restrict our analysis to $p = 2$. Our bounds for any $p \geq 1$ and $d \geq 2$ are stated in Theorem 3.7, whose proof is provided in Appendix A.2.

Lemma 3.3. *For any two probability distributions μ and ν defined over a metric space with a unit diameter, suppose μ_n and ν_n are two empirical distributions of μ and ν, respectively. Then, with a high probability,*

$$|\Pi_{2,1}(\mu,\nu) - \Pi_{2,1}(\mu_n,\nu_n)| = \tilde{O}(n^{-\frac{1}{3}}).$$

Note that for any pair of distributions μ and ν and their empirical distributions μ_n and ν_n, by the triangle inequality,

$$|\Pi_{2,1}(\mu,\nu) - \Pi_{2,1}(\mu_n,\nu_n)| \leq \Pi_{2,1}(\mu,\mu_n) + \Pi_{2,1}(\nu,\nu_n).$$

Therefore, to prove Lemma 3.3, we bound $(2,1)$-RPW of any distribution μ to its empirical distribution μ_n in the following lemma.

Lemma 3.4. *Given a continuous probability distribution μ in the 2-dimensional Euclidean space and an empirical distribution μ_n of μ, $\Pi_{2,1}(\mu,\mu_n) = \tilde{O}(n^{-\frac{1}{3}})$ with a high probability.*

We begin by defining a set of notations that assist in proving Lemma 3.4. Let $\mathcal{G}$ be a grid with cell side length $n^{-\alpha}$ inside the unit square. For any cell $\square \in \mathcal{G}$, let $\mu(\square)$ denote the mass of μ inside $\square$. Define the *excess mass* of a cell $\square$ as $\text{Exc}_\mu(\square) := \max\{0, \mu(\square) - \mu_n(\square)\}$ and the excess of the grid $\mathcal{G}$, denoted by $\text{Exc}_\mu(\mathcal{G})$, as the total excess of all cells of $\mathcal{G}$, i.e., $\text{Exc}_\mu(\mathcal{G}) := \sum_{\square\in\mathcal{G}} \text{Exc}_\mu(\square)$. When μ is clear from context, we simplify notation and denote the excess of $\mathcal{G}$ by $\text{Exc}(\mathcal{G})$.

Lemma 3.5. *For any distribution μ inside the unit square, an empirical distribution μ_n of μ, and a grid $\mathcal{G}$ with cell side length $n^{-\alpha}$, $\text{Exc}_\mu(\mathcal{G}) = \tilde{O}(n^{\alpha-\frac{1}{2}})$ with a high probability.*

To better explain our proof for Lemma 3.4, we first present a weaker bound of $\tilde{O}(n^{-\frac{3}{10}})$. In Appendix A.2.2, we improve our analysis and obtain a rate of $\tilde{O}(n^{-\frac{1}{3}})$.

A weaker bound for Lemma 3.4. In the following, we construct a transport plan γ that transports all except $\tilde{O}(n^{-\frac{3}{10}})$ mass with a cost of $\tilde{O}(n^{-\frac{3}{10}})$. Having computed such transport plan, we then use Lemma 2.2 to conclude that $\Pi_{p,k}(\mu,\nu) = \tilde{O}(n^{-\frac{3}{10}})$. The details are provided next.

Define $\mathcal{G}_1$ (resp. $\mathcal{G}_2$) to be a grid with cell side length $O(n^{-\alpha_1})$ (resp. $O(n^{-\alpha_2})$) for $\alpha_1 := \frac{3}{10}$ (resp. $\alpha_2 := \frac{1}{5}$). The grids $\mathcal{G}_1$ and $\mathcal{G}_2$ are constructed in a way that their boundaries are aligned with each other. Let γ_1 be a partial transport plan that arbitrarily transports, for any cell $\square \in \mathcal{G}_1$, a mass of $\min\{\mu(\square), \mu_n(\square)\}$ from μ_n to μ. Define μ^1 (resp. μ_n^1) to be the distribution of mass of μ (resp. μ_n) that is not transported by γ_1. Let γ_2 be a transport plan that transports, for any cell $\square \in \mathcal{G}_2$, a mass of $\min\{\mu^1(\square), \mu_n^1(\square)\}$ from μ_n^1 to μ^1. Define $\gamma = \gamma_1 + \gamma_2$. This completes the construction of γ. In the appendix, instead of two grids, we consider $O(\log\log n)$ grids and obtain the bound claimed in Lemma 3.4.

The transport plan γ transports as much mass as possible inside each cell of $\mathcal{G}_2$; therefore, the total mass that is not

[3] $\tilde{O}()$ hides $\text{poly}(\log n)$ from the convergence rate.

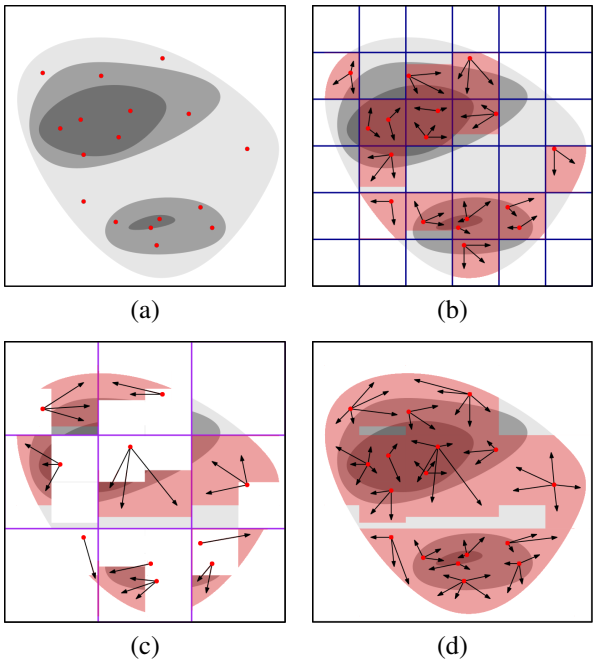


Figure 4. (a) A distribution μ (shaded gray area) and an empirical distribution μ_n (red dots), (b) γ_1 transports as much mass as possible inside each cell of $\mathcal{G}_1$, (c) for the remaining mass, γ_2 transports as much remaining mass as possible inside the cells of $\mathcal{G}_2$, and (d) the transport plan γ, which is the sum of γ_1 and γ_2.

transported by γ is equal to the excess of the grid $\mathcal{G}_2$, which from Lemma 3.5 is

$$1 - \mathcal{M}(\gamma) = \text{Exc}(\mathcal{G}_2) = \tilde{O}(n^{\alpha_2 - \frac{1}{2}}) = \tilde{O}(n^{-\frac{3}{10}}). \quad (5)$$

Next, we show that the cost $w_2(\gamma)$ of the transport plan γ is $\tilde{O}(n^{-\frac{3}{10}})$. Recall that $\gamma = \gamma_1 + \gamma_2$. In our analysis, we first show that $w_2^2(\gamma_1) = O(n^{-\frac{3}{5}})$ and then show that $w_2^2(\gamma_2) = O(n^{-\frac{3}{5}})$. Using these two bounds, we then conclude that $w_2^2(\gamma) = O(n^{-\frac{3}{5}})$, or equivalently, $w_2(\gamma) = O(n^{-\frac{3}{10}})$.

Since γ_1 transports mass between points inside the same cell of $\mathcal{G}_1$, all mass transportation in γ_1 has a squared cost of $O(n^{-2\alpha_1}) = O(n^{-\frac{3}{5}})$ and $w_2^2(\gamma_1) = O(n^{-\frac{3}{5}})$. Furthermore, by Lemma 3.5, with a high probability, the total mass of μ^1 is $\mathcal{M}(\mu^1) = \tilde{O}(n^{\alpha_1 - \frac{1}{2}}) = \tilde{O}(n^{-\frac{1}{5}})$. Since the transport plan γ_2 transports at most $\mathcal{M}(\mu^1)$ mass, each at a squared cost of $O(n^{-2\alpha_2}) = O(n^{-\frac{2}{5}})$,

$$w_2^2(\gamma_2) = \tilde{O}\big(\mathcal{M}(\mu^1) \times n^{-2\alpha_2}\big) = \tilde{O}(n^{-\frac{3}{5}}).$$

As a result, the cost of γ is

$$w_2(\gamma) = \sqrt{w_2^2(\gamma_1) + w_2^2(\gamma_2)} = \tilde{O}(n^{-\frac{3}{10}}). \quad (6)$$

Note that γ is a transport plan that transports all except $\alpha = \tilde{O}(n^{-\frac{3}{10}})$ mass (Equation (5)) and has cost $\tilde{O}(n^{-\frac{3}{10}})$ (Equation (6)). Hence,

$$W_{2,1-\alpha}(\mu, \mu_n) \leq w_2(\gamma) = \tilde{O}(n^{-\frac{3}{10}}).$$

Plugging into Lemma 2.2, $\Pi_{2,1}(\mu, \mu_n) = \tilde{O}(n^{-\frac{3}{10}})$.

In the appendix, we use an identical approach to obtain the rate of convergence for any dimension $d \geq 1$ and any parameter $p \geq 1$. The following lemma summarizes the results.

Lemma 3.6. *Given a continuous probability distribution μ in the d-dimensional Euclidean space, an empirical distribution μ_n of μ, and parameters $p \geq 1$ and $k > 0$ constant, with a high probability,*

$$\Pi_{p,k}(\mu, \mu_n) = \begin{cases} \tilde{O}(n^{-\frac{1}{d}}), & p \leq \frac{d}{2}, \\ \tilde{O}(n^{-\frac{p}{p(d+2)-d}}), & p > \frac{d}{2}. \end{cases}$$

For any pair of distributions μ and ν and their empirical distributions μ_n and ν_n, by the triangle inequality,

$$|\Pi_{p,k}(\mu, \nu) - \Pi_{p,k}(\mu_n, \nu_n)| \leq \Pi_{p,k}(\mu, \mu_n) + \Pi_{p,k}(\nu, \nu_n).$$

Combined with Lemma 3.6, we get the following corollary.

Theorem 3.7. *For two probability distributions μ and ν defined over a metric space $(\mathcal{X}, c)$ with a unit diameter, suppose μ_n and ν_n are two empirical distributions of μ and ν, respectively. Then, for any $p \geq 1$ and $k > 0$ constant, with a high probability,*

$$|\Pi_{p,k}(\mu, \nu) - \Pi_{p,k}(\mu_n, \nu_n)| = \begin{cases} \tilde{O}(n^{-\frac{1}{d}}), & p \leq \frac{d}{2}, \\ \tilde{O}(n^{-\frac{p}{p(d+2)-d}}), & p > \frac{d}{2}. \end{cases}$$

Extension to arbitrary diameter. We can extend our distance and its properties to the case where the diameter of the support is bounded by any $\Delta > 0$ as follows. Define (p, k)-RPW between distributions μ and ν to be the minimum value $\varepsilon > 0$ such that $W_{p,1-\varepsilon}(\mu, \nu) \leq k\Delta\varepsilon$. In this case, the metric property (Theorem 2.1) and the robustness to outliers (Theorem 3.1) holds without any changes, and given that the diameter Δ is a constant, our bounds for the convergence rate of the empirical (p, k)-RPW also holds as stated in Theorem 3.7.

4. Relation to Other Distances

In this section, we discuss the relation of the (p, k)-RPW metric with three other well-known distance functions, namely (i) Lévy-Prokhorov distance, (ii) total variation, and (iii) p-Wasserstein distance. In particular, we first show that the $(\infty, 1)$-RPW is the same as the Lévy-Prokhorov distance. We next show that for any $p \geq 1$, the (p, k)-RPW metric is an interpolation between total variation and the p-Wasserstein distance. More precisely, $\Pi_{p,0}(\cdot, \cdot)$ is the same as the total variation distance, and for large values of k, the (p, k)-RPW will be close to $\frac{1}{k} W_p(\cdot, \cdot)$.

Lévy-Prokhorov distance. For any two distributions μ and ν defined over a set $\mathcal{X}$ in the d-dimensional Euclidean space, let $\pi(\mu, \nu)$ denote the Lévy-Prokhorov distance of μ and ν. In the following lemma, we show that the $(\infty, 1)$-RPW metric is equal to the Lévy-Prokhorov distance. The proof of this lemma, which is provided in Appendix A.3, is similar to the approach described by Lahn et al. (2021).

Lemma 4.1. *For any pair of probability distributions μ and ν in a metric space $(\mathcal{X}, c)$ with a unit diameter, $\Pi_{\infty,1}(\mu, \nu) = \pi(\mu, \nu)$.*

Total Variation. For any pair of distributions μ and ν, let $\|\mu - \nu\|_{\text{TV}}$ denote the total variation of μ and ν. In Lemma 4.2, we show that for any $p \geq 1$, the $(p, 0)$-RPW distance between μ and ν is equal to their total variation. Intuitively, the $(p, 0)$-RPW distance measures the maximum amount of mass that can be transported from μ to ν at a 0 cost, i.e., $(p, 0)$-RPW distance is the amount of mass of μ and ν that overlap, which is the same as their total variation.

Lemma 4.2. *For any two probability distributions μ and ν in a metric space $(\mathcal{X}, c)$ with a unit diameter and any parameter $p \geq 1$, $\Pi_{p,0}(\mu, \nu) = \|\mu - \nu\|_{\text{TV}}$.*

p-Wasserstein distance. Finally, we show that for large enough values of k, the (p, k)-RPW metric would be close to $\frac{1}{k} W_p(\mu, \nu)$.

Lemma 4.3. *For any two probability distributions μ and ν over a metric space $(\mathcal{X}, c)$ with a unit diameter and any parameters $p \geq 1$ and $k > 0$, $\Pi_{p,k}(\mu, \nu) \leq \frac{1}{k} W_p(\mu, \nu) \leq \Pi_{p,k}(\mu, \nu) + k^{-\frac{p+1}{p}}$.*

5. Algorithms to Compute (p, k)-RPW

In this section, we describe two approximation algorithms for computing the (p, k)-RPW distance between two discrete distributions defined over supports of n points. The first algorithm uses a binary search on the value of $\Pi_{p,k}(\mu, \nu)$ and computes a δ-additive approximation of (p, k)-RPW (or simply *δ-close (p, k)-RPW*) for any $\delta > 0$ in $O(n^3 \log n \log \delta^{-1})$ time. Our second algorithm relies on the algorithm by (Lahn et al., 2019) (LMR algorithm) to approximate the OT-profile and computes a δ-additive approximation of our metric in $O(\frac{n^2}{\delta^p} + \frac{n}{\delta^{2p}})$ time. A high-level overview of each algorithm is presented below. See Appendix A.4 for full details.

Note that when $k = 0$, as discussed in Lemma 4.2, the $(p, 0)$-RPW is simply the total variation distance and can be computed on discrete distributions in linear time. Hence, in the following, we assume $k > 0$.

Highly-Accurate Algorithm. For any $\varepsilon \in [0, 1]$, computing the $(1 - \varepsilon)$-partial p-Wasserstein distance for discrete distributions can be done using a standard OT solver in an augmented space (Chapel et al., 2020), which takes $O(n^3 \log n)$ time (Edmonds & Karp, 1972; Orlin, 1988). Our first algorithm is based on this observation and uses a simple binary search on the value of (p, k)-RPW to obtain a δ-additive approximation in $O(n^3 \log n \log \delta^{-1})$ time.

Computing Through an Approximate OT-Profile. We can also approximate the (p, k)-RPW distance by using the LMR algorithm (Lahn et al., 2019) to approximate the OT-profile. Given two discrete probability distributions, an error parameter $\delta' > 0$, and any cost function, the LMR algorithm incrementally constructs a $(\delta')^{1/p}$-additive approximation of the OT-profile, i.e., for any $\alpha \in [0, 1]$, the LMR algorithm computes a $(\delta')^{1/p}$-additive approximation of the α-partial p-Wasserstein distance (Phatak et al., 2022).

In Lemma A.4 in the appendix, we show that computing our metric using a δ'-additive approximation of the OT-profile leads to a $\frac{2\delta'}{k}$-close (p, k)-RPW. Therefore, to compute a δ-close (p, k)-RPW distance function, we use the LMR algorithm with an error parameter $\delta' = (\frac{k\delta}{2})^p$ to approximate the OT-profile and to compute a δ-close $\Pi_{p,k}(\mu, \nu)$ in $O(\frac{n^2}{\delta^p} + \frac{n}{\delta^{2p}})$ time.

6. Experimental Results

In this section, we present the results of our experiments showing that our distance is robust to noise from outliers and sample discrepancies.

In our first experiment, we use the 1-Wasserstein distance, 2-Wasserstein distance, TV distance, $(2, 1)$-RPW, and $(2, 0.1)$-RPW to rank images from the MNIST (LeCun, 1998), CIFAR-10 (Hinton et al., 2012), and COREL datasets and measure the accuracy of the results. In our second experiment, we measure the convergence rate of empirical $(2, k)$-RPW distance to the true $(2, k)$-RPW and compare it with the convergence rate for 2-Wasserstein distance for synthetic data sets. For both experiments, we compute an additive approximation of the RPW metric using the LMR algorithm (Lahn et al., 2019).

6.1. Image Retrieval.

Following the experimental setup introduced by Rubner et al. (2000), we conduct experiments on retrieving images using $(2, 1)$-RPW and $(2, 0.1)$-RPW distances and compare their accuracy against the 1-Wasserstein, 2-Wasserstein, and TV distances. In this experiment, given a dataset of labeled images and a set of unlabeled query images, the goal is to retrieve, for each query image, a set of m similar images from the labeled dataset. The accuracy of a distance function in the image retrieval task is then computed as the ratio of the retrieved images with the correct label, averaged over all retrievals for all query images. In our experiments, we vary the value of m from 1 to 100.

Datasets. We conduct the experiments using three datasets, namely, MNIST, CIFAR-10, and COREL. In our experiments on each dataset, we randomly select $2k$ images as the labeled dataset and randomly select 50 images as the query. For each dataset, we introduce three scenarios of perturbation:

(i) (noise in datasets) In this scenario, we add random noise to the images in the datasets. For the MNIST dataset, for each image, we add a random amount of noise between 0% and 10% to a random pixel in the image. For CIFAR-10 and COREL datasets, we replaced randomly selected 10% of pixels with white pixels in each image.

(ii) (shift in datasets) In this scenario, we shift up each labeled image by 2 pixels for the MNIST dataset and increase the intensity of a random RGB channel by 20 in each image of the CIFAR-10 and COREL datasets.

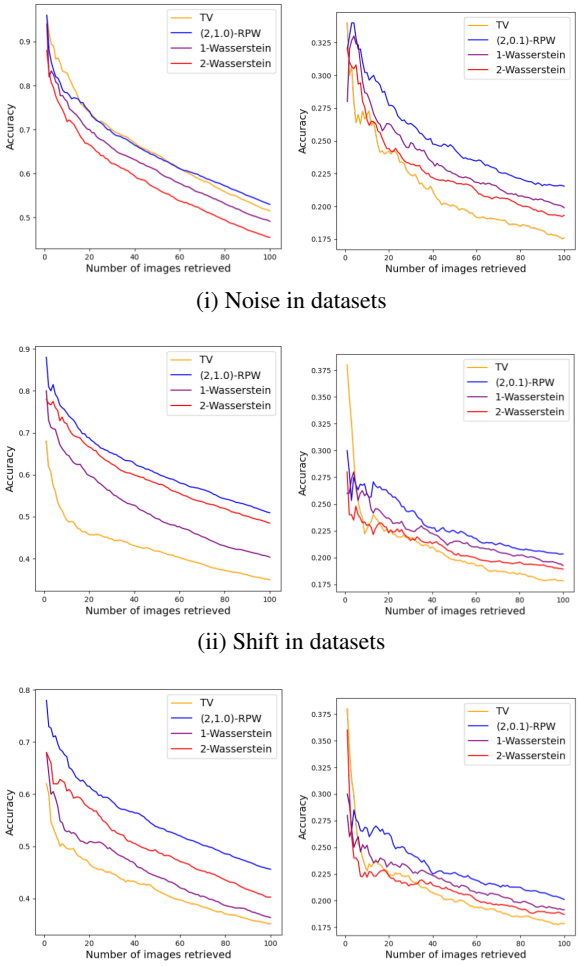
(iii) (noise and shift in datasets) In the last scenario, we introduce both random noise (scenario (i)) and random shift (scenario (ii)) to the images in datasets.

Results. Figure 5 shows the results of our experiments on MNIST and CIFAR-10 datasets. The results of our experiments on the COREL dataset are provided in Appendix B. In Figure 5, the left (resp. right) column corresponds to our experiments on the MNIST (resp. CIFAR-10) dataset in the three scenarios described above. In each plot, the horizontal axis is the number m of retrieved images for each query, and the vertical axis is the accuracy of the retrieved images. The results of our experiments suggest that our metric performs better than the 1-Wasserstein, 2-Wasserstein, and the TV distances for the task of image retrieval with perturbations.

First, for experiments on the MNIST dataset, we observe that for datasets (ii) and (iii), the accuracy achieved by the 2-Wasserstein distance is better than that of the 1-Wasserstein distance. This, we believe is due to the higher sensitivity of 2-Wasserstein distance, which enables it to detect minor differences effectively. However, for dataset (i), the presence of noise distorts 2-Wasserstein distance, causing it to underperform.

Results for MNIST dataset (i): The TV, 1-Wasserstein, and $(2,1)$-RPW distances are more robust to noise and therefore, perform better than the 2-Wasserstein distance. The $(2,1)$-RPW distance outperforms 1-Wasserstein distance. This is because it retains the higher sensitivity of 2-Wasserstein distance, which enables it to effectively detect minor differences.

Results for MNIST dataset (ii): TV distance is known to be sensitive to shifts. As a result, the accuracy of TV distance drops significantly. In contrast, the 2-Wasserstein distance and the $(2,1)$-RPW distance handle such shifts more effectively.



(i) Noise in datasets

(ii) Shift in datasets

(iii) Noise and shift in datasets

Figure 5. The results of our experiments on image retrieval on (left column) MNIST dataset and (right column) CIFAR-10 dataset.

Results for MNIST dataset (iii): The 2-Wasserstein distance is sensitive to noise and the TV distance is sensitive to shifts. Therefore, both these distances produce lower accuracy results. Recollect that the $(2,1)$-RPW distance combines the TV distance and the 2-Wasserstein distance and therefore, can handle both shifts and noise effectively. Therefore, it outperforms all distances in this setting.

Results for CIFAR-10 dataset: In the CIFAR-10 dataset, images that have the same labels may already have significant variations and shifts. Due to these variations, 2-Wasserstein and TV distances achieve lower accuracy in comparison to the 1-Wasserstein distance. The $(2,0.1)$-RPW distance, however, outperforms the 1-Wasserstein distance for this dataset.

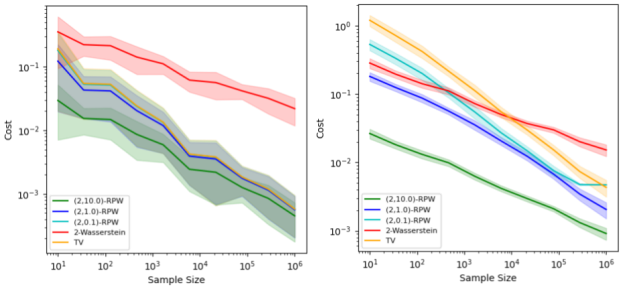


Figure 6. The convergence rate of different metrics on (left) 2-point distribution and (right) grid distribution.

6.2. Rate of Convergence

We conduct numerical experiments to compare the convergence rate of the empirical $(2, k)$-RPW metric with that of the 2-Wasserstein distance and TV distance on discrete distributions. We compute the convergence rate of each metric by drawing two sets of n i.i.d samples from a discrete distribution and compute the empirical distance between the corresponding empirical distributions.

Datasets. We employ two synthetic 2-dimensional discrete distributions, namely (i) (2-point distribution) a discrete distribution defined over 2 points a and b each with a probability $1/2$, where $\|a - b\| = 1$, and (ii) (grid distribution) a discrete distribution defined over 16 points that are placed in a grid of 4×4, where each point has a probability of $\frac{1}{16}$.

In our experiments, we vary the sample size n from 10 to 10^6. For each value of n, we conduct the experiment 10 times and take the mean distance among all 10 executions.

Results. As shown in Figure 6, experiments suggest that for both distributions, the empirical $(2, 1)$-RPW distance converges to 0 significantly faster than the 2-Wasserstein distance. We also observe that for a small value of k (e.g., $k = 0.1$), $(2, k)$-RPW is close to the TV distance, whereas for a large value of k (e.g., $k = 10$), the $(2, k)$-RPW distance values are similar to $\frac{1}{k} W_2(\cdot, \cdot)$. These results are in line with our theoretical bounds in Section 4.

7. Conclusion

In this paper, we designed a new partial p-Wasserstein-based metric called the (p, k)-RPW that is robust to outlier noise as well as sampling discrepancies but retains the sensitivity of p-Wasserstein distance in capturing the minor geometric differences in distributions. We showed that our distance interpolates between the p-Wasserstein and TV distances and inherits robustness to both noise and shifts in distribution. We also showed that, for $p = \infty$, our metric is the same as the Lévy-Prokhorov distance.

The main contribution of this paper is to introduce a new metric and derive its useful properties. Experiments suggest that the new distance is a promising alternative to TV and p-Wasserstein distances. Including this distance into many potential machine learning use cases, including as a loss function in GANs or for computing barycenters of a set of distributions, remains an open question. Designing a parallel algorithm to approximate the (p, k)-RPW distance also remains an important open question.

Acknowledgement

We would like to acknowledge Advanced Research Computing (ARC) at Virginia Tech, which provided us with the computational resources used to run the experiments. The research presented in this paper was funded by NSF CCF-2223871. We thank the anonymous reviewers for their useful feedback.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alvarez-Melis, D. and Jaakkola, T. S. Gromov-Wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*, 2018.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proc. 34th International Conference on Machine Learning*, pp. 214–223, 2017.

Backurs, A., Dong, Y., Indyk, P., Razenshteyn, I., and Wagner, T. Scalable nearest neighbor search for optimal transport. In *Proc. 37th International Conference on Machine Learning*, pp. 497–506, 2020.

Bansil, M. and Kitagawa, J. W_∞-transport with discrete target as a combinatorial matching problem. *Archiv der Mathematik*, 117(2):189–202, 2021.

Caffarelli, L. A. and McCann, R. J. Free boundaries in optimal transport and Monge-Ampere obstacle problems. *Annals of Mathematics*, pp. 673–730, 2010.

Chang, W., Shi, Y., and Wang, J. Csot: Curriculum and structure-aware optimal transport for learning with noisy labels. *ArXiv preprint arXiv:2312.06221*, 2023.

Chapel, L., Alaya, M. Z., and Gasso, G. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33: 2903–2913, 2020.

Choi, J., Choi, J., and Kang, M. Generative modeling through the semi-dual formulation of unbalanced optimal transport. *Advances in Neural Information Processing Systems*, 36, 2024.

Chuang, C.-Y., Hjelm, R. D., Wang, X., Vineet, V., Joshi, N., Torralba, A., Jegelka, S., and Song, Y. Robust contrastive learning against noisy views. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16670–16681, 2022.

Claici, S., Chien, E., and Solomon, J. Stochastic Wasserstein barycenters. In *International Conference on Machine Learning*, pp. 999–1008. PMLR, 2018.

Cuturi, M. and Doucet, A. Fast computation of Wasserstein barycenters. In *International Conference on machine learning*, pp. 685–693. PMLR, 2014.

Edmonds, J. and Karp, R. M. Theoretical improvements in algorithmic efficiency for network flow problems. *J. of the ACM*, 19(2):248–264, 1972.

El Malki, N., Cugny, R., Teste, O., and Ravat, F. Decwa: Density-based clustering using Wasserstein distance. In *Proc. 29th ACM International Conference on Information & Knowledge Management*, pp. 2005–2008, 2020.

Figalli, A. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3-4):707–738, 2015.

Genevay, A., Peyre, G., and Cuturi, M. Learning generative models with Sinkhorn divergences. *In International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

Gupta, R., Indyk, P., and Price, E. Sparse recovery for earth mover distance. In *Proc. 48th Annual Allerton Conference on Communication, Control, and Comput.*, pp. 1742–1744. IEEE, 2010.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Huang, G., Guo, C., Kusner, M. J., Sun, Y., Sha, F., and Weinberger, K. Q. Supervised word mover's distance. *Advances in neural information processing systems*, 29, 2016.

Janati, H., Cuturi, M., and Gramfort, A. Wasserstein regularization for sparse multi-task regression. In *Proc. 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1407–1416. PMLR, 2019.

Lahn, N., Mulchandani, D., and Raghvendra, S. A graph theoretic additive approximation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 13813–13823, 2019.

Lahn, N., Raghvendra, S., and Ye, J. A faster maximum cardinality matching algorithm with applications in machine learning. *Advances in Neural Information Processing Systems*, 34:16885–16898, 2021.

Lai, Z., Wang, C., Cheung, S.-c., and Chuah, C.-N. Sar: Self-adaptive refinement on pseudo labels for multiclass-imbalanced semi-supervised learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4091–4100, 2022.

Le, K., Nguyen, H., Nguyen, Q. M., Pham, T., Bui, H., and Ho, N. On robust optimal transport: Computational complexity and barycenter computation. *Advances in Neural Information Processing Systems*, 34:21947–21959, 2021.

LeCun, Y. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Luise, G., Rudi, A., Pontil, M., and Ciliberto, C. Differential properties of Sinkhorn approximation for learning with Wasserstein distance. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Mukherjee, D., Guha, A., Solomon, J. M., Sun, Y., and Yurochkin, M. Outlier-robust optimal transport. In *International Conference on Machine Learning*, pp. 7850–7860. PMLR, 2021.

Nietert, S., Goldfeld, Z., and Cummings, R. Outlier-robust optimal transport: Duality, structure, and statistical analysis. In *International Conference on Artificial Intelligence and Statistics*, pp. 11691–11719. PMLR, 2022.

Nietert, S., Goldfeld, Z., and Shafiee, S. Outlier-robust Wasserstein DRO. *ArXiv preprint arXiv:2311.05573*, 2023.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Orlin, J. A faster strongly polynomial minimum cost flow algorithm. In *Proc. 20th Annual ACM Symposium on Theory of Computing*, pp. 377–387, 1988.

Phatak, A., Raghvendra, S., Tripathy, C., and Zhang, K. Computing all optimal partial transports. In *Proc. 11th International Conference on Learning Representations*, 2022.

Rubner, Y., Tomasi, C., and Guibas, L. J. The earth mover's distance as a metric for image retrieval. *International J. of Comput. Vision*, 40(2):99, 2000.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. Improving GANs using optimal transport. *In International Conference on Learning Representations*, 2018.

Sedrakyan, H. and Sedrakyan, N. *The HM-GM-AM-QM Inequalities*, pp. 21–43. Springer International Publishing, 2018.

Vaskevicius, T. and Chizat, L. Computational guarantees for doubly entropic Wasserstein barycenters. In *Proc. 37th Conference on Neural Information Processing Systems*, 2023.

Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., and Courty, N. Semi-relaxed Gromov-Wasserstein divergence with applications on graphs. *arXiv preprint arXiv:2110.02753*, 2021.

Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., and Solomon, J. M. Hierarchical optimal transport for document representation. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhuang, Y., Chen, X., and Yang, Y. Wasserstein k-means for clustering probability distributions. *Advances in Neural Information Processing Systems*, 35:11382–11395, 2022.

A. Missing Proofs and Details

A.1. Missing Proofs of Section 2.

In this section, we provide the proofs for Theorem 2.1 and Lemma 2.2.

Theorem 2.1. *Given a metric space $(\mathcal{X}, c)$ with a unit diameter and any parameters $p \geq 1$ and $k \geq 0$, the (p,k)-RPW distance function $\Pi_{p,k}(\cdot,\cdot)$ for all probability distributions defined over $(\mathcal{X}, c)$ is a metric.*

Proof. To prove this lemma, we show that (p,k)-RPW satisfies all four properties of the metric spaces, namely (i) identity, (ii) positivity, (iii) symmetry, and (iv) triangle inequality, and conclude that it is metric.

For any probability distribution μ defined over $(\mathcal{X}, c)$, $W_p(\mu,\mu) = 0$ and therefore, $\varepsilon = 0$ satisfies the condition in Equation (1); hence, $\Pi_{p,k}(\mu,\mu) = 0$ and property (i) holds. Furthermore, if ν is another probability distribution over $(\mathcal{X}, c)$ that is distinct from μ, then $W_p(\mu,\nu) > 0$. Hence, $\varepsilon = 0$ does not satisfy the condition in Equation (1), and $\Pi_{p,k}(\mu,\nu)$, which is the smallest $\varepsilon \geq 0$ with $W_{p,1-\varepsilon}(\mu,\nu) \leq k\varepsilon$, will be positive and property (ii) holds. Additionally, for any $\varepsilon \geq 0$, the $(1-\varepsilon)$-partial p-Wasserstein distance is symmetric, i.e., $W_{p,1-\varepsilon}(\mu,\nu) = W_{p,1-\varepsilon}(\nu,\mu)$. Therefore, from Equation (1), RPW is also symmetric, i.e., $\Pi_{p,k}(\mu,\nu) = \Pi_{p,k}(\nu,\mu)$ and property (iii) holds as well.

Finally, we show that the (p,k)-RPW satisfies the triangle inequality. For any three probability distributions μ, ν, and κ, suppose $\Pi_{p,k}(\mu,\kappa) = \varepsilon_1$ and $\Pi_{p,k}(\kappa,\nu) = \varepsilon_2$. In the following, we show that $\Pi_{p,k}(\mu,\nu) \leq \varepsilon_1 + \varepsilon_2$ and conclude property (iv).

Note that if $\varepsilon_1 + \varepsilon_2 \geq 1$, then the triangle inequality holds trivially since $\Pi_{p,k}(\mu,\nu) \leq 1 \leq \varepsilon_1 + \varepsilon_2$. Therefore, we assume that $\varepsilon_1 + \varepsilon_2 < 1$. Let γ_1 denote a $(1-\varepsilon_1)$-partial OT plan from μ to κ and define $\kappa_1 := \gamma_1 \# \mu$ to be the mass of κ that is transported from μ by γ_1; here, $\#$ denotes the push-forward operation. Similarly, let γ_2 denote a $(1-\varepsilon_2)$-partial OT plan from κ to ν and define $\kappa_2 := (\gamma_2)^{-1} \# \nu$ to be the mass of κ that is transported to ν by γ_2. Then, both κ_1 and κ_2 are distributions over $\mathcal{X}$ that are dominated by κ and have masses $\mathcal{M}(\kappa_1) = 1-\varepsilon_1$ and $\mathcal{M}(\kappa_2) = 1-\varepsilon_2$.

Define κ_c to be the distribution of mass of κ that is common to both κ_1 and κ_2; more precisely, for each $x \in \mathcal{X}$,

$$\kappa_c(x) = \min\{\kappa_1(x), \kappa_2(x)\}.$$

Note that the total mass of κ_1 that is not transported by γ_2 is at most ε_2; therefore,

$$\mathcal{M}(\kappa_c) \geq \mathcal{M}(\kappa_1) - \varepsilon_2 = 1 - \varepsilon_1 - \varepsilon_2. \tag{7}$$

Define $\mu_c := (\gamma_1)^{-1} \# \kappa_c$ (resp. $\nu_c := \gamma_2 \# \kappa_c$) to be distribution dominated by μ (resp. ν) whose mass is transported to (resp. from) κ_c in γ_1 (resp. γ_2). From Equation (7),

$$\mathcal{M}(\mu_c) = \mathcal{M}(\nu_c) = \mathcal{M}(k_c) \geq 1 - \varepsilon_1 - \varepsilon_2 \tag{8}$$

By the triangle inequality of the p-Wasserstein distance,

$$W_p(\mu_c, \nu_c) \leq W_p(\mu_c, \kappa_c) + W_p(\kappa_c, \nu_c). \tag{9}$$

Furthermore,

$$W_p(\mu_c, \kappa_c) \leq w_p(\gamma_1) \leq k\varepsilon_1. \tag{10}$$

Similarly,

$$W_p(\kappa_c, \nu_c) \leq w_p(\gamma_2) \leq k\varepsilon_2. \tag{11}$$

Combining Equations (8), (9), (10), and (11),

$$W_{p,1-\varepsilon_1-\varepsilon_2}(\mu,\nu) \leq W_p(\mu_c,\nu_c) \leq W_p(\mu_c,\kappa_c) + W_p(\kappa_c,\nu_c) \leq k(\varepsilon_1+\varepsilon_2). \tag{12}$$

Since $\Pi_{p,k}(\mu,\nu)$ is the minimum ε with $W_{p,1-\varepsilon}(\mu,\nu) \leq k\varepsilon$, from Equation (12), $\Pi_{p,k}(\mu,\nu) \leq \varepsilon_1 + \varepsilon_2$, as desired. □

Lemma 2.2. *Given two probability distributions μ and ν defined over a metric space $(\mathcal{X}, c)$ with a unit diameter and parameters $p \geq 1$ and $k \geq 0$, suppose $W_{p,1-\alpha}(\mu,\nu) = k\beta$ for some $\alpha, \beta \geq 0$. Then, $\Pi_{p,k}(\mu,\nu) \leq \max\{\alpha,\beta\}$. Furthermore, if $k \neq 0$, then $\Pi_{p,k}(\mu,\nu) \leq \min\{\alpha,\beta\}$.*

Proof. Let $\delta := \Pi_{p,k}(\mu,\nu)$. We prove this lemma by considering two cases:

- If $\alpha \leq \beta$, then $W_{p,1-\beta}(\mu,\nu) \leq W_{p,1-\alpha}(\mu,\nu) = k\beta$; hence, β satisfies the condition in Equation (1), and $\Pi_{p,k}(\mu,\nu) \leq \beta = \max\{\alpha,\beta\}$. Furthermore, if $k > 0$, then $\delta \geq \alpha = \min\{\alpha,\beta\}$ because otherwise, if $\delta < \alpha$, then

$$W_{p,1-\delta}(\mu,\nu) \geq W_{p,1-\alpha}(\mu,\nu) = k\beta \geq k\alpha > k\delta,$$

  which is a contradiction.

- Otherwise, $\alpha > \beta$ and $W_{p,1-\alpha}(\mu,\nu) = k\beta < k\alpha$; hence, α satisfies the condition in Equation (1), and $\Pi_{p,k}(\mu,\nu) \leq \alpha = \max\{\alpha,\beta\}$. Additionally, if $k > 0$, then $\delta \geq \beta = \min\{\alpha,\beta\}$, since otherwise, if $\delta < \beta$, then

$$W_{p,1-\delta}(\mu,\nu) \geq W_{p,1-\beta}(\mu,\nu) \geq W_{p,1-\alpha}(\mu,\nu) = k\beta > k\delta,$$

  which is a contradiction.

□

A.2. Missing Proofs and Details of Section 3

A.2.1. Robustness to Outlier Nosie

Theorem 3.1. *For any probability distributions μ, ν, and ν' defined over a metric space $(\mathcal{X}, c)$ with a unit diameter and parameters $p \geq 1, k \geq 0$, and $\delta \in (0,1)$, let $\tilde{\nu} = (1-\delta)\nu + \delta\nu'$. Then,*

$$\Pi_{p,k}(\mu,\nu) - \delta \leq \Pi_{p,k}(\mu,\tilde{\nu}) \leq (1-\delta)\Pi_{p,k}(\mu,\nu) + \delta.$$

Proof. First, note that by the triangle inequality, $\Pi_{p,k}(\mu,\tilde{\nu}) + \Pi_{p,k}(\tilde{\nu},\nu) \geq \Pi_{p,k}(\mu,\nu)$. Furthermore, by the definition of $\tilde{\nu}$, $W_{p,1-\delta}(\nu,\tilde{\nu}) = 0$. Therefore, by Lemma 2.2, $\Pi_{p,k}(\nu,\tilde{\nu}) \leq \max\{\delta, 0\} = \delta$ and

$$\Pi_{p,k}(\mu,\tilde{\nu}) \geq \Pi_{p,k}(\mu,\nu) - \Pi_{p,k}(\nu,\tilde{\nu}) \geq \Pi_{p,k}(\mu,\nu) - \delta.$$

Define $\alpha := \Pi_{p,k}(\mu,\nu)$ and let γ be a $(1-\alpha)$-partial OT plan from μ to ν. Since $\tilde{\nu}$ is defined as $(1-\delta)\nu + \delta\nu'$, the transport plan $\gamma' := (1-\delta)\gamma$ can be seen as a $((1-\delta)(1-\alpha))$-partial transport plan from μ to $\tilde{\nu}$; therefore,

$$W_{p,(1-\delta)(1-\alpha)}(\mu,\tilde{\nu}) \leq w_p(\gamma') = (1-\delta)^{\frac{1}{p}} w_p(\gamma) \leq k(1-\delta)^{\frac{1}{p}}\alpha,$$

where the last inequality holds by the definition of the (p,k)-RPW distance. Using Lemma 2.2,

$$\Pi_{p,k}(\mu,\tilde{\nu}) \leq \max\left\{1-(1-\delta)(1-\alpha), (1-\delta)^{\frac{1}{p}}\alpha\right\} = (1-\delta)\alpha + \delta = (1-\delta)\Pi_{p,k}(\mu,\nu) + \delta.$$

□

The following lemma helps in proving Lemma 3.2.

Lemma A.1. *For two probability distributions μ and $\tilde{\mu}$ defined over a metric space $(\mathcal{X}, c)$ with a unit diameter, parameters $p \geq 1, k > 0$, and a constant $\alpha \in (0,1)$, let $\delta := \|\mu - \tilde{\mu}\|_{TV}$. Then,*

$$\min\left\{\delta(1-\alpha), \frac{1}{k}W_{p,1-\delta(1-\alpha)}(\mu,\tilde{\mu})\right\} \leq \Pi_{p,k}(\mu,\tilde{\mu}) \leq \min\left\{\delta, \frac{1}{k}W_p(\mu,\tilde{\mu})\right\}.$$

Proof. Using Lemma 2.2 on distributions μ and $\tilde{\mu}$,

$$\Pi_{p,k}(\mu,\tilde{\mu}) \geq \min\left\{\delta(1-\alpha), \frac{1}{k}W_{p,1-\delta(1-\alpha)}(\mu,\tilde{\mu})\right\}.$$

Next, since $\delta = \|\mu-\tilde{\mu}\|_{TV}$, we get $W_{p,1-\delta}(\mu,\tilde{\mu}) = 0$. Plugging into Lemma 2.2,

$$\Pi_{p,k}(\mu,\tilde{\mu}) \leq \max\{\delta, 0\} = \delta. \tag{13}$$

Furthermore, since $W_{p,1-0}(\mu,\tilde{\mu}) = W_p(\mu,\tilde{\mu})$, by Lemma 2.2,

$$\Pi_{p,k}(\mu,\tilde{\mu}) \leq \max\{0, \frac{1}{k}W_p(\mu,\tilde{\mu})\} = \frac{1}{k}W_p(\mu,\tilde{\mu}). \tag{14}$$

Combining Equations (13) and (14),

$$\Pi_{p,k}(\mu,\tilde{\mu}) \leq \min\left\{\delta, \frac{1}{k}W_p(\mu,\tilde{\mu})\right\},$$

as claimed. □

Assuming that the assumption (A1) holds for the distributions μ and $\tilde{\mu}$, by plugging $\alpha = 0.9$ in Lemma A.1, we can derive the following lemma.

Lemma 3.2. *For a probability distribution μ defined over a metric space $(\mathcal{X}, c)$ with a unit diameter and $\delta > 0$, let $\tilde{\mu}$ be a probability distribution that differs from μ by a δ fraction of mass satisfying assumption (A1). Then, for any parameters $p \geq 1$ and $k > 0$,*

$$\Pi_{p,k}(\mu,\tilde{\mu}) = \Theta\left(\min\left\{\delta, \frac{1}{k}W_p(\mu,\tilde{\mu})\right\}\right).$$

A.2.2. Robustness to Sampling Discrepancies

Lemma A.2. *For any distribution μ inside the unit d-dimensional hypercube, an empirical distribution μ_n of μ, and a grid $\mathcal{G}$ with cell side length $n^{-\alpha}$, $\mathrm{Exc}_\mu(\mathcal{G}) = \tilde{O}(n^{\frac{d\alpha}{2}-\frac{1}{2}})$ with a high probability.*

Proof. First, note that if $\alpha \geq \frac{1}{d}$, then $n^{\frac{d\alpha}{2}-\frac{1}{2}} \geq 1$, and the lemma statement holds trivially. Therefore, from now on, we assume α to be less than $\frac{1}{d}$. For any cell $\square$ of the grid $\mathcal{G}$, define $p_\square := \mu(\square)$ to be the total probability mass of μ inside $\square$, i.e., the probability that a point drawn from μ lies inside $\square$. Any cell $\square \in \mathcal{G}$ is considered a sparse cell if $p_\square \leq \frac{9\log n}{n}$, and a dense cell otherwise. Let $\mathcal{G}_\mathcal{S}$ (resp. $\mathcal{G}_\mathcal{D}$) denote the subset of sparse (resp. dense) cells of $\mathcal{G}$. For each sparse cell $\square$, $\mathrm{Exc}_\mu(\square) \leq p_\square \leq \frac{9\log n}{n}$; therefore, using $\alpha < \frac{1}{d}$, the total contribution of sparse cells to the excess of $\mathcal{G}$ is at most

$$O(|\mathcal{G}_\mathcal{S}| \times \frac{\log n}{n}) = O(n^{d\alpha} \times \frac{\log n}{n}) = \tilde{O}(n^{d\alpha-1}) = \tilde{O}(n^{\frac{d\alpha}{2}-\frac{1}{2}}).$$

Next, we analyze the excess of the dense cells. Let $X = (x_1, \ldots, x_n)$ denote the set of n samples drawn from μ that were used to construct the empirical distribution μ_n. For each dense cell $\square$, let $Y_\square$ be a random variable denoting the number of samples in X that lie inside $\square$. Using the Chernoff bound,

$$Pr\left[Y_\square \leq np_\square - 3\sqrt{np_\square \log n}\right] \leq n^{-\frac{9}{2}}.$$

In other words, for each $\square \in \mathcal{G}_\mathcal{D}$, $\mathrm{Exc}_\mu(\square) = O(\frac{1}{n}\sqrt{np_\square \log n})$ with probability at least $1 - n^{-\frac{9}{2}}$. Therefore, with probability at least $(1-n^{-9/2})^{|\mathcal{G}_\mathcal{D}|} \geq (1-n^{-9/2})^n \geq 1 - n^{-\frac{7}{2}}$, the total excess of the dense cells would be

$$\begin{aligned}\sum_{\square\in\mathcal{G}_\mathcal{D}} \mathrm{Exc}_\mu(\square) &= O\left(\sum_{\square\in\mathcal{G}_\mathcal{D}} \sqrt{\frac{p_\square \log n}{n}}\right) = O\left(\sqrt{\frac{\log n}{n}} \sum_{\square\in\mathcal{G}_\mathcal{D}} \sqrt{p_\square}\right) \\ &= O\left(\sqrt{\frac{\log n}{n}} \times \sqrt{|\mathcal{G}_\mathcal{D}|}\right) = \tilde{O}(n^{\frac{d\alpha}{2}-\frac{1}{2}}),\end{aligned}$$

where the third equality holds since $\frac{\sum_{\square\in\mathcal{G}}\sqrt{p_\square}}{|\mathcal{G}_\mathcal{D}|} \leq \sqrt{\frac{\sum_{\square\in\mathcal{G}_\mathcal{D}} p_\square}{|\mathcal{G}_\mathcal{D}|}}$ (Sedrakyan & Sedrakyan, 2018) and $\sum_{\square\in\mathcal{G}_\mathcal{D}} p_\square \leq 1$, and the last equality holds since $|\mathcal{G}_\mathcal{D}| \leq n^{d\alpha}$. □

We obtain the following lemma by simply plugging $d = 2$ in Lemma A.2.

Lemma 3.5. *For any distribution μ inside the unit square, an empirical distribution μ_n of μ, and a grid $\mathcal{G}$ with cell side length $n^{-\alpha}$, $\mathrm{Exc}_\mu(\mathcal{G}) = \tilde{O}(n^{\alpha-\frac{1}{2}})$ with a high probability.*

Improved proof of Lemma 3.4. We improve our bounds for the convergence rate of the empirical $(2,1)$-RPW by extending our approach and considering $O(\log\log n)$ grids instead of two grids. In the following, we construct a transport plan γ that transports all except $\tilde{O}(n^{-\frac{1}{3}})$ mass with a cost of $\tilde{O}(n^{-\frac{1}{3}})$. We then conclude that $\Pi_{p,k}(\mu,\nu) = \tilde{O}(n^{-\frac{1}{3}})$.

Without loss of generality, assume $n = 2^{2^h}$ for some integer $h > 0$. Let $\beta := \frac{\log n}{3\log n - 2}$. Define a set of h grids $\langle \mathcal{G}_1, \ldots, \mathcal{G}_h\rangle$, where each grid $\mathcal{G}_i$ has a side length $O(n^{-\alpha_i})$ for $\alpha_i := \frac{1}{2} - \beta\left(1 - \frac{1}{2^i}\right)$. We construct the grids in a way that their boundaries are aligned with each other.

Let $\mu^0 := \mu$ and $\mu_n^0 := \mu_n$. Starting from $i = 1$, we compute a partial transport plan γ_i from μ_n^{i-1} to μ^{i-1} that transports as much mass as possible inside each cell of $\mathcal{G}_i$. We then define μ^i (resp. μ_n^i) as the distribution of the mass of μ^{i-1} (resp. μ_n^{i-1}) that is not transported by γ_i, set $i \leftarrow i+1$, and continue the same process until we process the last grid $\mathcal{G}_h$. Define $\gamma := \sum_{i=1}^h \gamma_i$. By our construction, the transport plan γ transports as much mass as possible inside each cell of $\mathcal{G}_h$. Therefore, the total mass that is not transported by γ is equal to the excess $\mathrm{Exc}(\mathcal{G}_h)$, which by Lemma 3.5, with a high probability, is

$$\tilde{O}(n^{\alpha_h-\frac{1}{2}}) = \tilde{O}(n^{-\beta(1-\frac{1}{2^h})}) = \tilde{O}(n^{\frac{-\log n+1}{3\log n-2}}) = \tilde{O}(n^{-\frac{1}{3}+\frac{1}{3(3\log n-2)}}) = \tilde{O}(n^{-\frac{1}{3}}). \tag{15}$$

Next, we analyze the cost of γ by computing the cost of each transport plan γ_i separately. For γ_1, since each mass transportation is between points inside the same cell of $\mathcal{G}_1$ and has a squared cost of $O((n^{-\alpha_1})^2)$,

$$w_2^2(\gamma_1) = O(n^{-2\alpha_1}) = O(n^{-1+\beta}) = O(n^{-\frac{2\log n-2}{3\log n-2}}) = O(n^{-\frac{2}{3}+\frac{2}{3(3\log n-2)}}) = O(n^{-\frac{2}{3}}).$$

Furthermore, for each $i > 1$, the transport plan γ_i transports a total mass of at most $\mathcal{M}(\mu^{i-1})$, which is equal to the excess of the grid $\mathcal{G}_{i-1}$, and by Lemma 3.5 is $\tilde{O}(n^{\alpha_{i-1}-\frac{1}{2}})$. Since γ_i transports mass between points inside the same cell of $\mathcal{G}_i$, each mass transportation in γ_i has a squared cost of $O(n^{-2\alpha_i})$, and therefore,

$$w_2^2(\gamma_i) = \tilde{O}(n^{\alpha_{i-1}-\frac{1}{2}-2\alpha_i}) = \tilde{O}(n^{-1+\beta}) = \tilde{O}(n^{-\frac{2}{3}}).$$

Therefore,

$$w_2(\gamma) = \sqrt{\sum_{i=1}^h w_2^2(\gamma_i)} = \tilde{O}(\sqrt{hn^{-\frac{2}{3}}}) = \tilde{O}(n^{-\frac{1}{3}}). \tag{16}$$

By Equations (15) and (16), we have computed a transport plan γ from μ_n to μ that, with a high probability, transports all except $\tilde{O}(n^{-\frac{1}{3}})$ mass with a cost $\tilde{O}(n^{-\frac{1}{3}})$. Therefore, $\Pi_2(\mu,\mu_n) = \tilde{O}(n^{-\frac{1}{3}})$ with a high probability.

We extend the same approach to any dimension $d \geq 2$ and any $p \geq 1$ in Lemma 3.6.

Lemma 3.6. *Given a continuous probability distribution μ in the d-dimensional Euclidean space, an empirical distribution μ_n of μ, and parameters $p \geq 1$ and $k > 0$ constant, with a high probability,*

$$\Pi_{p,k}(\mu,\mu_n) = \begin{cases} \tilde{O}(n^{-\frac{1}{d}}), & p \leq \frac{d}{2}, \\ \tilde{O}(n^{-\frac{p}{p(d+2)-d}}), & p > \frac{d}{2}. \end{cases}$$

Proof. In Lemma 4.3, we show that for any $p \geq 1$ and $k > 0$, $\Pi_{p,k}(\mu,\nu) \leq \frac{1}{k} W_p(\mu,\nu)$. Therefore, for any constant $k > 0$, the convergence rate of the empirical (p,k)-RPW is upper-bounded by the convergence rate of the empirical p-Wasserstein distance. Fournier & Guillin (2015) showed that when $p \leq \frac{d}{2}$, the p-Wasserstein distance achieves a convergence rate of $\tilde{O}(n^{-\frac{1}{d}})$. Therefore, our metric also achieves a convergence rate of $\tilde{O}(n^{-\frac{1}{d}})$ in this case, proving the bound claimed in the lemma statement for $p \leq \frac{d}{2}$. In the remaining of this proof, we prove our bounds for $p > \frac{d}{2}$.

Let $h = \log_{\frac{2p}{d}} \log_2 n$. We assume that h is an integer. Let $\beta := \frac{\log n}{(p+\frac{2p}{d}-1)\log n - p}$. Define a set of h grids $\langle \mathcal{G}_1, \ldots, \mathcal{G}_h \rangle$, where each grid $\mathcal{G}_i$ has a cell side length of $n^{-\alpha_i}$ for

$$\alpha_i := \frac{1}{d} - \frac{2p}{d^2}\beta\left(1 - (\frac{d}{2p})^i\right).$$

Following the approach discussed in Section 3.2, we construct h transport plans $\gamma_1, \ldots, \gamma_h$ and define a transport plan $\gamma := \sum_{i=1}^{h} \gamma_i$ to be a transport plan that transports total mass of $\min\{\mu(\square), \mu_n(\square)\}$ inside each cell $\square \in \mathcal{G}_h$. By Lemma A.2, the total free mass with respect to γ would be

$$1 - \mathcal{M}(\gamma) = \mathrm{Exc}(\mathcal{G}_h) = \tilde{O}\left(n^{\frac{d\alpha_h}{2} - \frac{1}{2}}\right) = \tilde{O}\left(n^{\frac{1}{2} - \frac{p}{d}\beta + \frac{p}{d}\beta\cdot(\frac{d}{2p})^h - \frac{1}{2}}\right) = \tilde{O}\left(n^{-\frac{\log_2 n - 1}{(d+2-\frac{d}{p})\log n - d}}\right)$$
$$= \tilde{O}\left(n^{-\frac{1}{d+2-\frac{d}{p}} + \frac{1-\frac{d}{d+2-\frac{d}{p}}}{(d+2-\frac{d}{p})\log n - d}}\right) = \tilde{O}\left(n^{-\frac{p}{(d+2)p-d}}\right). \tag{17}$$

Next, we bound the cost of γ by analyzing the cost of each transport plan $\gamma_i, i \in [1, h]$ separately. For γ_1, each mass transportation is inside a cell of $\mathcal{G}_1$ and has a pth power cost of $\tilde{O}(n^{-p\alpha_1})$; hence,

$$w_p^p(\gamma_1) = \tilde{O}(n^{-p\alpha_1}) = \tilde{O}\left(n^{-\frac{p}{d} + \frac{2p^2}{d^2}\beta\left(1-\frac{d}{2p}\right)}\right) = \tilde{O}\left(n^{-\frac{p^2}{(d+2)p-d}}\right). \tag{18}$$

Finally, for each $i > 1$, the total mass transported by γ_i is equal to the excess of $\mathcal{G}_{i-1}$, which by Lemma 3.5 is $\tilde{O}(n^{\frac{d\alpha_{i-1}}{2} - \frac{1}{2}})$. Each mass transportation in γ_i is between points within the same cell of $\mathcal{G}_i$ and thus has a pth power cost of $\tilde{O}(n^{-p\alpha_i})$. Therefore,

$$w_p^p(\gamma_i) = \tilde{O}(n^{\frac{d\alpha_{i-1}}{2} - \frac{1}{2} - p\alpha_i}) = \tilde{O}\left(n^{-\frac{p^2}{(d+2)p-d}}\right). \tag{19}$$

Adding the cost of all transport plans,

$$w_p(\gamma) = \left(\sum_{i=1}^{h} w_p^p(\gamma_i)\right)^{1/p} = \tilde{O}\big((n^{-\frac{p^2}{(d+2)p-d}}\log\log n)^{1/p}\big) = \tilde{O}(n^{-\frac{p}{(d+2)p-d}}). \tag{20}$$

Combining Equations (17) and (20), $\Pi_p(\mu, \mu_n) = \tilde{O}(n^{-\frac{p}{(d+2)p-d}})$. □

A.3. Missing Proofs of Section 4.

To prove Lemma 4.1, we begin by showing in Lemma A.3 that when μ and ν are discrete distributions, $\Pi_{\infty,1}(\mu, \nu) = \pi(\mu, \nu)$. We then use Lemma A.3 to show that the same also holds for continuous distributions.

Lemma A.3. *For any pair of discrete probability distributions μ and ν in a metric space $(\mathcal{X}, c)$ with a unit diameter, $\Pi_{\infty,1}(\mu, \nu) = \pi(\mu, \nu)$.*

Proof. To prove this lemma, we first show that $\Pi_{\infty,1}(\mu, \nu) \leq \pi(\mu, \nu)$. We then show that $\pi(\mu, \nu) \leq \Pi_{\infty,1}(\mu, \nu)$ and conclude the lemma statement.

Let $\delta := \pi(\mu, \nu)$ and suppose A and B denote the support of μ' and ν', respectively. Define the δ-disc graph G_δ between points in A and B to be a bipartite graph where for each pair $(a, b) \in A \times B$ with $c(a, b) \leq \delta$, there is an edge between a and b in G_δ. For any set S of vertices of G_δ, let $\mathcal{N}(S)$ denote the set of neighbors of S in G_δ. By the definition of the Lévy-Prokhorov distance, for any set of points $S \subseteq A$, $\mu(S) \leq \nu(\mathcal{N}(S)) + \delta$. Similarly, for any subset $T \subseteq B$, $\nu(T) \leq \mu(\mathcal{N}(T)) + \delta$.

We prove that $\Pi_{\infty,1}(\mu, \nu) \leq \pi(\mu, \nu)$ by showing that the maximum transport plan γ on the δ-disc graph G_δ transports a total mass of at least $1 - \delta$. In this case, since all edges of γ has a cost of at most δ, $w_\infty(\gamma) \leq \delta$. Hence, the $(1 - \delta)$-partial ∞-Wasserstein distance from μ to ν would be at most δ and $\Pi_{\infty,1}(\mu, \nu) \leq \delta$, as claimed.

Consider a bipartite graph G'_δ obtained from G_δ by adding a fake vertex b', where b' has a mass of δ and is connected to all points of A with a cost δ. For any subset $S \subseteq A$ (resp. $T \subseteq B$), let $\mathcal{N}'(S)$ (resp. $\mathcal{N}'(T)$) denote the set of neighbors of S (resp. T) in G'_δ and suppose $\mu'(S)$ (resp. $\nu'(T)$) denotes the total mass of points in S (resp. T) for any subset $S \subseteq A$ (resp. $T \subseteq B \cup \{b'\}$). By construction, for any subset $S \subseteq A$, $\mu'(S) \leq \nu'(\mathcal{N}'(S))$ and similarly, for any subset $T \subseteq B$, $\nu'(T) \leq \mu'(\mathcal{N}'(T))$; hence, by the extension of Hall's marriage theorem (Bansil & Kitagawa, 2021), there exists a transport plan γ' on the graph G'_δ that transports all mass of the points in A to the points in $B \cup \{b'\}$. Let γ denote the transport plan obtained from γ' after removing the fake vertex b' and all mass transportation to b'. The transport plan γ transports at least $1-\delta$ mass and has a cost $w_\infty(\gamma') \leq \delta$. Therefore, if γ transports a total mass of $1-\delta'$ for some $\delta' \leq \delta$,

$$W_{\infty,1-\delta}(\mu,\nu) \leq W_{\infty,1-\delta'}(\mu,\nu) \leq w_\infty(\gamma) \leq \delta,$$

which means that $\Pi_{\infty,1}(\mu,\nu) \leq \delta = \pi(\mu,\nu)$. We next show that $\Pi_{\infty,1}(\mu,\nu) \geq \pi(\mu,\nu)$ in a similar way and conclude that $\Pi_{\infty,1}(\mu,\nu) = \pi(\mu,\nu)$.

Let $\delta := \Pi_{\infty,1}(\mu,\nu)$. Let γ be a $(1-\delta)$-partial OT plan from μ to ν and let G_δ be a δ-disk graph on $A \cup B$. Since $w_\infty(\gamma) \leq \delta$, all mass transportation by γ has a cost at most δ, i.e., all edges carrying a positive mass in γ are present in G_δ. For any subset $S \subseteq A$, let μ_S be the distribution of the mass of μ on the points in S. Define $\nu_S := \gamma \# \mu_S$ to be the subset of mass of ν that is transported from μ_S according to γ, and let $T_S \subseteq B$ be the support of ν_S. Recall that all edges carrying a positive mass in γ are present in G_δ; therefore, all points in T_S are neighbors of S in G_δ, i.e., $T \subseteq \mathcal{N}(S)$ and $\nu(T_S) \leq \nu(\mathcal{N}(S))$. Furthermore, since γ is a $(1-\delta)$-partial OT plan, the total mass of μ_S that is not transported by γ is at most δ, and hence,

$$\mu(S) \leq \nu(T_S) + \delta \leq \nu(\mathcal{N}(S)) + \delta.$$

One can also show that for each subset $T \subseteq B$, $\nu(T) \leq \mu(\mathcal{N}(T)) + \delta$ using an identical argument. Therefore, by the definition of the Lévy-Prokhorov distance, $\pi(\mu,\nu) \leq \delta = \Pi_{\infty,1}(\mu,\nu)$. □

In the following, we show that for any pair of (continuous) probability distributions μ and ν and any $\varepsilon > 0$,

$$|\Pi_{\infty,1}(\mu,\nu) - \pi(\mu,\nu)| \leq \varepsilon \tag{21}$$

and conclude that $\Pi_{\infty,1}(\mu,\nu) = \pi(\mu,\nu)$ for any pair of probability distributions (discrete or continuous).

Define $\square$ to be a unit d-dimensional hypercube containing the set $\mathcal{X}$. Let $\mathcal{G}$ be a grid of cell side length $\frac{\varepsilon}{4\sqrt{d}}$ that partitions $\square$ into smaller cells. Using the grid $\mathcal{G}$, we construct two discrete distributions μ^ε and ν^ε as follows. Let $\mathcal{G}_\mathcal{X}$ denote the subset of cells of $\mathcal{G}$ that intersects the set $\mathcal{X}$. For each cell $\xi \in \mathcal{G}_\mathcal{X}$, we pick an arbitrary point r_ξ inside $\xi \cap \mathcal{X}$ as the representative point of ξ. Let $\mathcal{R} := \bigcup_{\xi \in \mathcal{G}} \{r_\xi\}$. Define μ^ε (resp. ν^ε) as a discrete distribution over $\mathcal{R}$ that assigns, for each $\xi \in \mathcal{G}_\mathcal{X}$, a mass of $\mu(\xi)$ (resp. $\nu(\xi)$) to its representative point r_ξ. This completes the construction of μ^ε and ν^ε. Note that by Lemma 4.1,

$$\Pi_{\infty,1}(\mu^\varepsilon,\nu^\varepsilon) = \pi(\mu^\varepsilon,\nu^\varepsilon). \tag{22}$$

Furthermore, $W_\infty(\mu,\mu^\varepsilon) \leq \frac{\varepsilon}{4}$, since there is a transport plan that transports the mass of μ inside each cell $\xi \in \mathcal{G}$ to the mass of μ^ε at r_ξ and each mass transportation has a cost at most $\frac{\varepsilon}{4}$. From Lemma 2.2, $\Pi_{\infty,1}(\mu,\mu^\varepsilon) \leq \frac{\varepsilon}{4}$. Similarly, $\Pi_{\infty,1}(\nu,\nu^\varepsilon) \leq \frac{\varepsilon}{4}$. Therefore, using the triangle inequality,

$$|\Pi_{\infty,1}(\mu,\nu) - \Pi_{\infty,1}(\mu^\varepsilon,\nu^\varepsilon)| \leq \Pi_{\infty,1}(\mu,\mu^\varepsilon) + \Pi_{\infty,1}(\nu,\nu^\varepsilon) \leq \frac{\varepsilon}{2}. \tag{23}$$

One can also show in a similar way that

$$|\pi(\mu,\nu) - \pi(\mu^\varepsilon,\nu^\varepsilon)| \leq \frac{\varepsilon}{2}. \tag{24}$$

We conclude Equation (21) by combining Equations (22), (23), and (24).

Lemma 4.2. *For any two probability distributions μ and ν in a metric space $(\mathcal{X},c)$ with a unit diameter and any parameter $p \geq 1$, $\Pi_{p,0}(\mu,\nu) = \|\mu-\nu\|_{\mathrm{TV}}$.*

Proof. We prove this lemma by first showing that $\Pi_{p,0}(\mu,\nu) \leq \|\mu-\nu\|_{\mathrm{TV}}$ and then showing that $\|\mu-\nu\|_{\mathrm{TV}} \leq \Pi_{p,0}(\mu,\nu)$.

Let $\mathcal{P}(\mathcal{X})$ denote the set of all probability distributions defined over the compact set $\mathcal{X}$. Nietert et al. (2023) showed that one can rewrite the $(1-\varepsilon)$-partial p-Wasserstein distance between μ and ν as

$$W_{p,1-\varepsilon}(\mu,\nu) = \inf_{\mu' \in \mathcal{P}(\mathcal{X}):\|\mu-\mu'\|_{\mathrm{TV}} \leq \varepsilon} W_p(\mu',\nu). \tag{25}$$

Define $\delta = \|\mu - \nu\|_{\mathrm{TV}}$. Plugging $\varepsilon = \delta$ in Equation (25),

$$W_{p,1-\delta}(\mu,\nu) = 0. \tag{26}$$

Therefore, by Lemma 2.2,

$$\Pi_{p,0}(\mu,\nu) \leq \max\{0,\delta\} = \delta = \|\mu - \nu\|_{\mathrm{TV}}. \tag{27}$$

Next, let $\delta' = \Pi_{p,0}(\mu,\nu)$. By definition of the $(p,0)$-RPW, $W_{p,1-\delta'}(\mu,\nu) \leq 0 \times \delta' = 0$ (since the parameter k is set to 0), and since the partial p-Wasserstein distance is non-negative, $W_{p,1-\delta'}(\mu,\nu) = 0$. Therefore,

$$0 = W_{p,1-\delta'}(\mu,\nu) = \inf_{\mu' \in \mathcal{P}(\mathcal{X}):\|\mu-\mu'\|_{\mathrm{TV}} \leq \delta'} W_p(\mu',\nu). \tag{28}$$

Let μ^* be the distribution realizing the infimum in Equation (28). Then, $W_p(\mu^*,\nu) = 0$, and by the metric properties of the p-Wasserstein distance, $\mu^* = \nu$; hence,

$$\|\mu - \nu\|_{\mathrm{TV}} = \|\mu - \mu^*\|_{\mathrm{TV}} \leq \delta' = \Pi_{p,0}(\mu,\nu). \tag{29}$$

Combining Equations (27) and (29), $\Pi_{p,0}(\mu,\nu) = \|\mu - \nu\|_{\mathrm{TV}}$. □

Lemma 4.3. *For any two probability distributions μ and ν over a metric space $(\mathcal{X},c)$ with a unit diameter and any parameters $p \geq 1$ and $k > 0$, $\Pi_{p,k}(\mu,\nu) \leq \frac{1}{k} W_p(\mu,\nu) \leq \Pi_{p,k}(\mu,\nu) + k^{-\frac{p+1}{p}}$.*

Proof. Let $\delta' := W_p(\mu,\nu)$. In this case,

$$W_{p,1-\min\{1,\frac{\delta'}{k}\}}(\mu,\nu) \leq W_p(\mu,\nu) = k \times \frac{\delta'}{k}.$$

Therefore, by Lemma 2.2, $\Pi_{p,k}(\mu,\nu) \leq \max\{\min\{1,\frac{\delta'}{k}\},\frac{\delta'}{k}\} = \frac{1}{k} W_p(\mu,\nu)$.

We next show that $\frac{1}{k} W_p(\mu,\nu) \leq \Pi_{p,k}(\mu,\nu) + k^{-\frac{p+1}{p}}$. Note that the inequality holds trivially for any $k \leq 1$, since $k^{-\frac{p+1}{p}} \geq \frac{1}{k} \geq \frac{1}{k} W_p(\mu,\nu)$. We therefore assume that $k > 1$. Let $\delta = \Pi_{p,k}(\mu,\nu)$. Since the $(1-\frac{1}{k})$-partial p-Wasserstein distance is at most 1, by Lemma 2.2, $\delta \leq \max\{\frac{1}{k}, \frac{1}{k} W_{p,1-\frac{1}{k}}(\mu,\nu)\} \leq \frac{1}{k}$. Let γ be a $(1-\delta)$-partial OT plan. Since the underlying metric space has a unit diameter, the remaining δ mass of μ and ν with respect to γ can be transported at a cost at most δ; therefore,

$$W_p(\mu,\nu) \leq \left(w_p^p(\gamma) + \delta\right)^{1/p} \leq \left((k\delta)^p + \frac{1}{k}\right)^{1/p} \leq k\delta + k^{-1/p}.$$

Equivalently, $\frac{1}{k} W_p(\mu,\nu) \leq \Pi_{p,k}(\mu,\nu) + k^{-\frac{p+1}{p}}$. □

A.4. Missing Details of Section 5.

In this section, we provide the details of the algorithms mentioned in Section 5.

Highly-Accurate Algorithm. In this algorithm, we obtain an approximation of our metric by a simple guessing procedure as follows. Starting from an initial guess $g_1 = 0.5$ for the value of our metric, at any step i of our algorithm and for any guess value $g_i \geq 0$, define $w_i := W_{p,1-g_i}(\mu,\nu)$. If $w_i \leq k g_i$, then by Lemma 2.2, $w_i \leq \Pi_{p,k}(\mu,\nu) \leq g_i$, i.e., our guess value is large and we set $g_{i+1} \leftarrow g_i - 2^{-(i+1)}$. Otherwise, $w_i > k g_i$, and in this case, by Lemma 2.2, $g_i \leq \Pi_{p,k}(\mu,\nu) < w_i$, i.e., our guess value is small and we set $g_{i+1} \leftarrow g_i + 2^{-(i+1)}$. Note that at any step i, $|g_i - \Pi_{p,k}(\mu,\nu)| \leq 2^{-i}$. Therefore, to obtain a δ-additive approximation of the (p,k)-RPW, the algorithm returns the guess value g_i when $2^{-i} \leq \delta$. This completes the description of our algorithm.

We next analyze the running time of this algorithm. Computing the $(1-g_i)$-partial p-Wasserstein distance can be done using a standard OT solver in an augmented space (Chapel et al., 2020), which takes $O(n^3 \log n)$ time (Edmonds & Karp, 1972; Orlin, 1988). The total number of iterations of our algorithm is $O(\log \delta^{-1})$ and therefore, our algorithm runs in $O(n^3 \log n \log \delta^{-1})$ time.

Computing Through an Approximate OT-Profile. In this part, we show that an approximation of the OT-profile can be used to approximate our metric. We then conclude that one can use the LMR algorithm to obtain such approximations of the OT-profile and to obtain a δ-additive approximation of the RPW distance in $O(\frac{n^2}{\delta^p} + \frac{n}{\delta^{2p}})$ time.

For a $\delta' \in (0,1]$, $p \geq 1$, and $\alpha \in [0,1]$, let $\overline{W}_{p,\alpha}(\mu,\nu)$ denote a δ'-close α-partial p-Wasserstein distance, i.e., $W_{p,\alpha}(\mu,\nu) \leq \overline{W}_{p,\alpha}(\mu,\nu) \leq W_{p,\alpha}(\mu,\nu) + \delta'$. Define

$$\overline{\Pi}_{p,k}(\mu,\nu) = \min\{\varepsilon \geq 0 \mid \overline{W}_{p,1-\varepsilon}(\mu,\nu) \leq k\varepsilon\}$$

to be the (p,k)-RPW distance function when computed using the approximate partial p-Wasserstein distances. In the following lemma, we show that $\overline{\Pi}_{p,k}(\mu,\nu)$ is a $\frac{2\delta'}{k}$-additive approximation of $\Pi_{p,k}(\mu,\nu)$.

Lemma A.4. *For any pair of distributions μ and ν in a metric space $(\mathcal{X},c)$ with a unit diameter and any parameters $p \geq 1$, $k > 0$, and $\delta' > 0$,*

$$\Pi_{p,k}(\mu,\nu) \leq \overline{\Pi}_{p,k}(\mu,\nu) \leq \Pi_{p,k}(\mu,\nu) + \frac{2\delta'}{k}.$$

Proof. Let $\overline{\delta} := \overline{\Pi}_{p,k}(\mu,\nu)$. By definition,

$$W_{p,1-\overline{\delta}}(\mu,\nu) \leq \overline{W}_{p,1-\overline{\delta}}(\mu,\nu) \leq k\overline{\delta}. \tag{30}$$

Therefore, $\Pi_{p,k}(\mu,\nu) \leq \overline{\delta} = \overline{\Pi}_{p,k}(\mu,\nu)$. Next, let $\delta := \Pi_{p,k}(\mu,\nu)$. By definition,

$$\overline{W}_{p,1-\delta}(\mu,\nu) \leq W_{p,1-\delta}(\mu,\nu) + \delta' \leq k\delta + \delta'. \tag{31}$$

By properties of the partial p-Wasserstein distance,

$$\overline{W}_{p,1-\delta-\frac{2\delta'}{k}}(\mu,\nu) \leq W_{p,1-\delta-\frac{2\delta'}{k}}(\mu,\nu) + \delta' \leq W_{p,1-\delta}(\mu,\nu) + \delta' \leq \overline{W}_{p,1-\delta}(\mu,\nu) + \delta'. \tag{32}$$

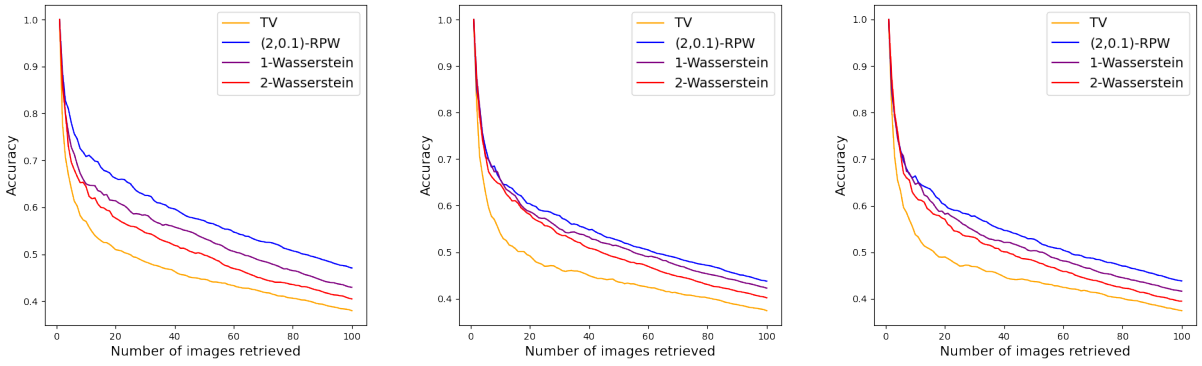
Combining Equations (31) and (32),

$$\overline{W}_{p,1-\delta-\frac{2\delta'}{k}}(\mu,\nu) \leq \overline{W}_{p,1-\delta}(\mu,\nu) + \delta' \leq k(\delta + \frac{2\delta'}{k}).$$

Therefore, $\overline{\Pi}_{p,k}(\mu,\nu) \leq \delta + \frac{2\delta'}{k} = \Pi_{p,k}(\mu,\nu) + \frac{2\delta'}{k}$. □

B. Additional Experiment Results of Section 6

In this section, we present the results of our experiments on the COREL dataset for the task of image retrieval.



(i) Noise in datasets (ii) Shift in datasets (iii) Noise and shift in datasets

Figure 7. The results of our experiments on image retrieval on the COREL dataset.

Similar to the CIFAR-10 dataset, the COREL dataset also consists of color images, where images with the same labels may have significant variations and shifts. Due to these variations, 2-Wasserstein and TV distances achieve lower accuracy in comparison to the 1-Wasserstein distance. The $(2,0.1)$-RPW distance, however, outperforms the 1-Wasserstein distance for this dataset.