Asymptotics of smoothed Wasserstein distances in the small noise regime

Author name withheld

Abstract

We study the behavior of the Wasserstein-2 distance between discrete measures μ and ν in $\mathbb{R}^d$ when both measures are smoothed by small amounts of Gaussian noise. This procedure, known as *Gaussian-smoothed optimal transport*, has recently attracted attention as a statistically attractive alternative to the unregularized Wasserstein distance. We give precise bounds on the approximation properties of this proposal in the small noise regime, and establish the existence of a phase transition: we show that, if the optimal transport plan from μ to ν is unique and a perfect matching, there exists a critical threshold such that the difference between $W_2(\mu, \nu)$ and the Gaussian-smoothed OT distance $W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ scales like $\exp(-c/\sigma^2)$ for σ below the threshold, and scales like σ above it. These results establish that for σ sufficiently small, the smoothed Wasserstein distance approximates the unregularized distance exponentially well.

1 Introduction: optimal transport

Optimal Transport (OT) has seen a recent surge of applications in machine learning, in areas such as generative modeling [ACB17, GPC18], image processing [PKD07, RTG00, FCVP17], and domain adaptation [CFT14, CFTR17]. A natural statistical question raised by these applications is to estimate the OT distances with samples. These distances, known as the Wasserstein distances, are defined by

$$W_p^p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^p d\pi(x, y)\,,$$

where $\Pi(\mu, \nu)$ denotes the set of joint measures with marginals μ and ν, known as *transport plans*. It is well known that plug-in estimators for this quantity, obtained by replacing μ and ν with empirical measures consisting of i.i.d. samples, have performance in high dimensions, with rates of convergence typically of order $n^{-p/d}$ [Dud69, BLG14, DY95, FG15, MNW21] when $d > 2p$. Moreover, minimax lower bounds show that this curse of dimensionality is unavoidable in general [SP18, NWR19].

The existence of the curse of dimensionality for OT has led to a series of proposals to obtain better rates of convergence by imposing additional structural assumptions—such as latent low-dimensionality [NWR19] or smoothness [SUL$^+$18, NWB19]—or by replacing W_p by a better-behaved surrogate, such as an entropy-regularized version with much better statistical and computational properties [Cut13, GCB$^+$19, RW18, MNW19, AWR17].

A particularly intriguing option, developed by [GG20], consists in *smoothing* the Wasserstein distance by adding Gaussian noise. The following result shows the statistical benefits of this approach.

Proposition 1.1 ([GGNWP20])**.** *For $d > 1$ and $\sigma > 0$, denote by $\mathcal{N}_\sigma$ the centered Gaussian measure on $\mathbb{R}^d$ with covariance $\sigma^2 I_d$. For any compactly supported probability measure μ in $\mathbb{R}^d$,*

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

let $x_1, x_2, \ldots, x_n$ be i.i.d. samples from μ, and define the empirical measure

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta(x_i).$$

Then there exists a constant $c = c(\mu, \sigma)$ such that

$$\mathbb{E}W_2(\hat{\mu}_n * \mathcal{N}_\sigma, \mu * \mathcal{N}_\sigma) \leq cn^{-1/2}.$$

[GG20] call this framework *Gaussian-smoothed optimal transport* (GOT), and follow up work has shown that it possesses significant statistical benefits, with fast rates of convergence and clean limit laws [ZCR21, GBG+18, GGK20].

To leverage the beneficial properties of the GOT framework, it is necessary to understand how well the smoothed distance $W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ approximates the standard Wasserstein distance $W_2(\mu, \nu)$. An application of the triangle inequality shows that

$$|W_2(\mu, \nu) - W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)| \lesssim \sigma\,. \tag{1}$$

Indeed, the triangle inequality implies $|W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) - W_2(\mu, \nu)| \leq W_2(\mu, \mu * \mathcal{N}_\sigma) + W_2(\nu, \nu * \mathcal{N}_\sigma)$ and the latter two terms are of order at most σ. In general, this upper bound is unimprovable, as we show below. On the other hand, it can also be very loose: if μ is a translation of ν, then $W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) = W_2(\mu, \nu)$ for all $\sigma \geq 0$. These examples raise a natural question: how well does $W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ approximate $W_2(\mu, \nu)$ when σ is small, and how does the answer to this question depend on the measures μ and ν?

The main goal if this paper is to give a sharp answer to this question for *finitely supported* measures. We focus on the finite support case for two reasons. First, when μ and ν are finitely supported, $\mu * \mathcal{N}_\sigma$ and $\nu * \mathcal{N}_\sigma$ are each finite mixtures of Gaussians, and the behavior of Wasserstein distances for such measures is a topic of active research [DD20, CGT19]. Second, as our results indicate, the behavior of this quantity for finitely supported measures is unexpectedly rich, with a sharp dichotomy in rates depending on the structure of the optimal transport plan between μ and ν: we show that when the *unique* optimal transport plan between μ and ν is a *perfect matching*, then there exist positive σ_* and c such that

$$0 \leq W_2(\mu, \nu) - W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) \lesssim e^{-c/\sigma^2} \quad \forall \sigma \in (0, \sigma_*)\,.$$

In other words, for sufficiently small σ, the GOT distance approximates the standard W_2 distance exponentially well, substantially sharpening (1). More strikingly, we establish the existence of a phase transition: for $\sigma < \sigma_*$, the gap is exponentially small, whereas for $\sigma > \sigma_*$, the gap scales linearly. By contrast, if the optimal transport plan between μ and ν is not unique or is not a perfect matching, then no phase transition appears: the upper bound of (1) is tight even in a neighborhood of $\sigma = 0$.

To locate exactly where the phase transition happens, we introduce a notion of robustness of the optimal transport plan between μ and ν, which is motivated by the concept of *cyclical monotonicity* [Roc66, Roc70, Roc87]. (See definition in Section 2.) A fundamental result in the theory of optimal transport [Roc66] is that the support of the optimal transport plan in the definition of W_2^2 is cyclically monotone. We define a robust version of this property and show that it characterizes measures for which the gap between $W_2(\mu, \nu)$ and $W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ is exponentially small. We show that the critical σ_* can be described in terms of the strong convexity of the *potentials* appearing in the dual of the optimal transport problem. The strong convexity of these potentials has previously been explored in computational and statistical contexts [VV21, PdC20], but to our knowledge its connection to Gaussian smoothed optimal transport is new.

Our work provides a precise understanding on how GOT resembles vanilla OT in the vanishing noise ($\sigma \downarrow 0$) regime. These results complement those recently obtained by [CNW20] in the large noise regime, who show that if μ and ν have n matching moments, $n \geq 1$, then $W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) = O(\sigma^{-n})$ as $\sigma \to \infty$.

Along with results in [CNW20], our work completes the limiting picture of the Euclidean heat semigroup acting on atomic measures under the Wasserstein distance. All the relevant rates are presented in Table 1.

Regime	Condition	$\lim(W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma))$	Rate	Reference
$\sigma \downarrow 0$	Unique OT plan	$W_2(\mu, \nu)$	e^{-c/σ^2}	Theorem 3.1
$\sigma \downarrow 0$	Non-unique OT plan	$W_2(\mu, \nu)$	σ	Theorem 3.3
$\sigma \uparrow \infty$	μ and ν agree up to nth moment	0	σ^{-n}	[CNW20]

Table 1: Limiting behavior of $W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ for atomic measures μ and ν.

We note that our work leaves open the question of characterizing the rates for non-atomic measures. It is possible to show that, for general measures, there are measures exhibiting polynomial rates intermediate between σ and e^{-c/σ^2}; however, these rates appear to depend delicately on the geometry of the measures and their support. Giving a full characterization of the rate for general probability measures is an attractive open question.

2 Cyclical monotonicity and implementability

We are concerned with the optimal transport problem between discrete measures

$$\mu = \frac{1}{k}\sum_{i=1}^{k} \delta(x_i), \quad \nu = \frac{1}{k}\sum_{j=1}^{k} \delta(y_j)$$

in the space $\mathbb{R}^d$, equipped with the squared Euclidean cost function $c(x, y) = \|x - y\|^2/2$. We are mainly interested in transport plans in the form of perfect matchings between $\{x_i\}$ and $\{y_i\}$. By relabeling the points, we may assume without loss of generality that the optimal transport plan between μ and ν is the matching given by

$$\Gamma = \{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)\}.$$

Our techniques are based on a robust notion of optimality for Γ. We recall the following definition of cyclical monotonicity, which serves as an important certification of an optimal transport plan.

Definition 2.1 (See, e.g., Roc66)**.** *A set $S \subseteq \mathbb{R}^d \times \mathbb{R}^d$ is* cyclically monotone *if for any $(a_1, b_1), \ldots, (a_n, b_n) \in S$, we have*

$$\sum_{i=1}^{n} \|a_i - b_i\|^2 \leq \sum_{i=1}^{n} \|a_i - b_{i+1}\|^2,$$

where we set $b_{n+1} := b_1$.

The significance of this notion is the following fundamental result.

Theorem 2.2 (See Vil08, Theorem 5.10)**.** *If $\pi \in \Pi(\mu, \nu)$ has cyclically monotone support, then it is an optimal transport plan between μ and ν.*

We strengthen this notion by insisting that the inequalities in the definition of cyclical monotonicity be strict.

Definition 2.3. *We say $f : [k] \times [k] \to \mathbb{R}_{\geq 0}$ is a positive residual function on $[k]$, if $f(i, i) = 0$, $f(i, j) > 0$ for $i \neq j$, and $f(i, j) = f(j, i)$ for all $i, j \in [k]$.*

Definition 2.4 (Strong cyclical monotonicity)**.** *For a positive residual function f on $[k]$, we say that Γ is f-strongly cyclically monotone, if for any $n \in \mathbb{N}$ and $\sigma(1), \sigma(2), \ldots, \sigma(n) \in [k]$ with $\sigma(i) \neq \sigma(i+1)$ (the convention is $\sigma(n+1) = \sigma(1)$), we have*

$$\sum_{i=1}^{n} \|x_{\sigma(i)} - y_{\sigma(i)}\|^2 \leq \sum_{i=1}^{n} \|x_{\sigma(i)} - y_{\sigma(i+1)}\|^2 - \sum_{i=1}^{n} f(\sigma(i), \sigma(i+1)),$$

or equivalently,

$$\sum_{i=1}^{n} \langle x_{\sigma(i)}, y_{\sigma(i)} - y_{\sigma(i+1)} \rangle \geq \sum_{i=1}^{n} f(\sigma(i), \sigma(i+1)).$$

Strong cyclical monotonicity indicates that the optimal plan with support Γ is superior to any other plan by a positive margin in its transport cost. In [Roc87], the author introduced the notion *implementability* and established it as an equivalent condition of cyclical monotonicity. In parallel to the results in [Roc87], we also introduce the following stronger condition of implementability.

Definition 2.5 (Strong implementability)**.** *For a positive residual function f on $[k]$, we say that Γ is f-strongly implementable, if there exists a potential function φ, such that for any $i, j \in [k]$, we have*

$$\langle x_i, y_i - y_j \rangle \geq \varphi(y_i) - \varphi(y_j) + f(i, j).$$

Analogous to the equivalence result in [Roc87], we show that strong cyclical monotonicity and strong implementability are both equivalent to the uniqueness and optimality of Γ.

Proposition 2.6. *The following three statements are equivalent:*

(i) Γ *is f-strongly cyclically monotone for some f;*

(ii) Γ *is f-strongly implementable;*

(iii) Γ *is the unique optimal transport plan from $\{x_i\}$ to $\{y_i\}$.*

The positive payment function constructed in the equivalence between (iii) and (i) in Proposition 2.6 is of the form $f(i, j) = \frac{\lambda}{2}\|y_i - y_j\|^2$ for some $\lambda > 0$, in which case the implementability condition reads

$$\langle x_i, y_i - y_j \rangle \geq \varphi(y_i) - \varphi(y_j) + \frac{\lambda}{2}\|y_i - y_j\|^2.$$

This condition is equivalent to the existence of a λ-strongly convex potential φ satisfying $\nabla\varphi(y_i) = x_i$ for all $i \in [k]$ [THG17], or, equivalently, the existence of a Lipschitz *Brenier map* from μ to ν [Bre87]. More generally, we have the following theorem characterizing the properties of strongly implementable plans with residual functions of quadratic type.

Theorem 2.7. *The following conditions are equivalent:*

(i) For some positive numbers $\alpha < \beta$, there exists a potential function $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ which is α-strongly convex and β-smooth, such that $x_i = \nabla\varphi(y_i)$ for all $i \in [k]$.

(ii) Γ *is strongly implementable for*

$$f(i, j) := \frac{1}{2(\beta - \alpha)} \left(\|x_i - x_j\|^2 + \alpha\beta\|y_i - y_j\|^2 - 2\alpha\langle y_i - y_j, x_i - x_j \rangle \right), \tag{2}$$

or equivalently, there exists $\{\tilde{\varphi}(y_i)\}_{i=1}^k \subset \mathbb{R}^d$, such that for all $i, j \in [k]$ $(i \neq j)$,

$$\begin{aligned} \langle x_i, y_i - y_j \rangle \geq\ & \tilde{\varphi}(y_i) - \tilde{\varphi}(y_j) \\ & + \frac{1}{2(\beta - \alpha)} \left(\|x_i - x_j\|^2 + \alpha\beta\|y_i - y_j\|^2 - 2\alpha\langle y_i - y_j, x_i - x_j \rangle \right) \end{aligned} \tag{3}$$

Proof. This is a direct application of Theorem 4 in [THG17]. □

Remark 2.8. *We should emphasize that the f defined in Theorem 2.7 is indeed a positive residual function given $\alpha < \beta$, since Cauchy-Schwartz gives*

$$2\alpha\langle y_i - y_j, x_i - x_j \rangle \leq \|x_i - x_j\|^2 + \alpha^2\|y_i - y_j\|^2 < \|x_i - x_j\|^2 + \alpha\beta\|y_i - y_j\|^2.$$

As a direct consequence of the direction (ii) to (i) in Theorem 2.7, if Γ is strongly implementable for a positive residual function f which is quadratic in $y_i - y_j$ and $x_i - x_j$, we will have guarantee on strong convexity and smoothness of the potential function.

Corollary 2.9. *Suppose Γ is strongly implementable for*

$$f(i, j) = \frac{1}{2} \left(\lambda_{xx}\|x_i - x_j\|^2 + \lambda_{yy}\|y_i - y_j\|^2 - 2\lambda_{xy}\langle y_i - y_j, x_i - x_j \rangle \right)$$

where $\lambda_{xx}, \lambda_{xy}$ and λ_{yy} are positive numbers which satisfy $\lambda_{xy}^2 + \lambda_{xy} = \lambda_{xx}\lambda_{yy}$. Then there exists a potential function $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ which is $\frac{\lambda_{xy}}{\lambda_{xx}}$-strongly convex and $\frac{\lambda_{yy}}{\lambda_{xy}}$-smooth, such that $\nabla x_i = \varphi(y_i)$ for all $i \in [k]$.

A crucial property of strongly cyclical monotone (or strongly implementable) transport plans is that they are robust to small perturbations in the sources and targets. We quantify the robustness of the map Γ in the following definition.

Definition 2.10. *For $\epsilon \geq 0$, we say Γ is ϵ-robust, if for any distinct $\sigma(1), \sigma(2), \ldots, \sigma(n) \in [k]$, and any $\alpha_{\sigma(1)}, \alpha_{\sigma(2)}, \ldots, \alpha_{\sigma(n)} \in \mathbb{R}^d$ such that*

$$\max_i \|\alpha_{\sigma(i)}\| \leq \epsilon,$$

there holds

$$\sum_{i=1}^n \|x_{\sigma(i)} - y_{\sigma(i)}\|^2 \leq \sum_{i=1}^n \|(x_{\sigma(i)} + \alpha_{\sigma(i)}) - (y_{\sigma(i+1)} + \alpha_{\sigma(i+1)})\|^2.$$

In the case that Γ is an optimal transport plan, also denote

$$R(\Gamma) := \sup \{\epsilon \geq 0 \ : \ \Gamma \textit{ is } \epsilon\textit{-robust}\}.$$

The quantity $R(\Gamma)$, which we call "robustness of optimality", is crucial to understanding the behavior of the optimal transport cost between μ and ν corrupted with noise.

Proposition 2.11. *Γ is strongly cyclically monotone if and only if $R(\Gamma) > 0$.*

The following proposition quantifies the relation between $R(\Gamma)$ and a positive residual f for which Γ is strongly implementable, and provides a certification of a lower bound of $R(\Gamma)$ in $O(k^2)$ time.

Proposition 2.12. *Suppose Γ is strongly implementable for a positive residual function f. Then T is ϵ-robust for*

$$\epsilon \leq \frac{1}{2} \inf_{i \neq j} \frac{f(i,j)}{\|x_i - x_j\| + \|y_i - y_j\|}. \tag{4}$$

This implies that

$$R(\Gamma) \geq \frac{1}{2} \inf_{i \neq j} \frac{f(i,j)}{\|x_i - x_j\| + \|y_i - y_j\|}.$$

A special case of Proposition 2.12 is when the optimal transport plan is strongly implementable with residual functions of quadratic type. In this case, we are able to derive a simple closed-form lower bound of $R(\Gamma)$.

Proposition 2.13. *When the equivalence in Theorem 2.7 holds, T is ϵ-robust for*

$$\epsilon \leq \frac{1}{2} \inf_{i \neq j} \frac{\max\left\{\frac{1}{\beta}\|x_i - x_j\|^2, \alpha\|y_i - y_j\|^2\right\}}{\|x_i - x_j\| + \|y_i - y_j\|}. \tag{5}$$

This implies that

$$R(\Gamma) \geq \frac{1}{2} \inf_{i \neq j} \frac{\max\left\{\frac{1}{\beta}\|x_i - x_j\|^2, \alpha\|y_i - y_j\|^2\right\}}{\|x_i - x_j\| + \|y_i - y_j\|}.$$

Remark 2.14. *When condition (i) in Theorem 2.7 holds, α-strong convexity and β-smoothness implies*

$$\frac{1}{\beta}\|x_i - x_j\| \leq \|y_i - y_j\| \leq \frac{1}{\alpha}\|x_i - x_j\|.$$

Thus the condition (5) *may be replaced by the bound*

$$\epsilon \leq \frac{1}{2} \inf_{i \neq j} \max\left\{\frac{\alpha}{1+\beta}\|x_i - x_j\|, \frac{\alpha}{\beta(1+\alpha)}\|y_i - y_j\|\right\}, \tag{6}$$

which is easier to verify in practice.

3 Case I: perfect matching

Our main results show that the robustness of optimality $R(\Gamma)$ controls the gap between $W_2(\mu,\nu)$ and $W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$.

Theorem 3.1. *If $\sigma_* = R(\Gamma) > 0$, then for $\sigma \in (0, \sigma_*)$,*

$$W_2(\mu,\nu) - W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) \lesssim \sqrt{\sigma_* \sigma} e^{-\sigma_*^2/4\sigma^2}.$$

In the regime where σ does not exceed $R(\Gamma)$, the above theorem tells that the GOT distance is an excellent approximation of the OT distance. Our second main result is a converse to that statement, showing that if σ goes beyond $R(\Gamma)$, we show that the loss $W_2(\mu,\nu) - W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ is linear in σ. We start with the following proposition, which quantifies a "violation of cyclical monotonicity" under possibly large perturbations in the sources and targets.

Proposition 3.2. *If Γ is an optimal transport plan, for any $M \geq 0$, denote*

$$G(M) := \sup \left\{ \sum_{i=1}^n \|x_{\sigma(i)} - y_{\sigma(i)}\|^2 - \sum_{i=1}^n \|(x_{\sigma(i)} + \alpha_{\sigma(i)}) - (y_{\sigma(i+1)} + \alpha_{\sigma(i+1)})\|^2 \ : \|\alpha_{\sigma(i)}\| \leq M \right\}$$

Then $G(M)$ is a concave function of M for $M \in [0, +\infty)$.

Note that $G(M)$ vanishes for $M < \sigma_*$. The next theorem shows that as long as $G(M)$ is not negligible for $M \gtrsim \sigma_*$, the approximation loss for $\sigma \geq \sigma_*$ is linear in σ.

Theorem 3.3. *If $\sigma_* = R(\Gamma) > 0$, then*

$$W_2^2(\mu,\nu) - W_2^2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) \gtrsim \sup_{M > \sigma_*} e^{-cM^2/\sigma^2} G(M).$$

Here $G(M)$ is defined as in Proposition 3.2. In particular, if $G(3\sigma_*) \gtrsim \sigma_*$, then for $\sigma \in (0, 2\sigma_*)$,

$$W_2^2(\mu,\nu) - W_2^2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) \gtrsim \sigma e^{-c\sigma_*^2/\sigma^2}.$$

To prove Theorem 3.1, we need the following lemma, which tells that no loss in W_2 is incurred by a local perturbation on μ and ν. The proof of Lemma 3.4 can be found in the supplementary material.

Lemma 3.4. *If $\sigma_* = R(\Gamma) > 0$, then for any measure Q in $\mathbb{R}^d$ supported on $B(0, \sigma_*)$,*

$$W_2(\mu,\nu) = W_2(\mu * Q, \nu * Q).$$

Proof of Theorem 3.3. For $M > \sigma_*$, pick $\sigma(1), \sigma(2), \ldots, \sigma(n) \in [k]$ and $\{\alpha_{\sigma(i)}\}_{i=1}^n \subset \mathbb{R}^d$ such that $\|\alpha_{\sigma(i)}\| \leq M$ and

$$\sum_{i=1}^n \|x_{\sigma(i)} - y_{\sigma(i)}\|^2 - \sum_{i=1}^n \|(x_{\sigma(i)} + \alpha_{\sigma(i)}) - (y_{\sigma(i+1)} + \alpha_{\sigma(i+1)})\|^2 = G(M).$$

For every $i \in [k]$, denote $B_{\sigma(i)}$ the ball centered at $x_{\sigma(i)} + \alpha_{\sigma(i)}$ with radius σ, and $\hat{B}_{\sigma(i)}$ the ball centered at $y_{\sigma(i)} + \alpha_{\sigma(i)}$ with radius σ. Also denote

- $\gamma \in \Pi(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ the law of $(X + Z, Y + Z)$, where $(X, Y) \sim \frac{1}{k} \sum_{i=1}^k \delta(x_i, y_i)$ and $Z \sim \mathcal{N}_\sigma$ are independent.
- $\gamma_{\sigma(i)} \in \Pi(\mathsf{Unif}(B_{\sigma(i)}), \mathsf{Unif}(\hat{B}_{\sigma(i)}))$ the coupling associated with the transport map
$$x \mapsto x + y_{\sigma(i)} - x_{\sigma(i)};$$
- $\tilde{\gamma}_{\sigma(i)} \in \Pi(\mathsf{Unif}(B_{\sigma(i)}), \mathsf{Unif}(\hat{B}_{\sigma(i+1)}))$ the coupling associated with the transport map
$$x \mapsto x + y_{\sigma(i+1)} - x_{\sigma(i)};$$
- A constant $m = c_d \exp\left(-\frac{(M+\sigma)^2}{2\sigma^2}\right)$, where c_d is a constant only dependent on the dimension d.

Consider the following measure in $\mathbb{R}^d \times \mathbb{R}^d$:

$$\tilde{\gamma} := \gamma - m \sum_{i=1}^{n} \gamma_{\sigma(i)} + m \sum_{i=1}^{n} \tilde{\gamma}_{\sigma(i)}.$$

We shall show that $\tilde{\gamma} \in \Pi(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$. We first verify that $\tilde{\gamma}$ is a positive measure on $\mathbb{R}^d \times \mathbb{R}^d$. In fact, for $x, y \in \mathbb{R}^d$,

$$\gamma(dx, dy) = \frac{1}{k} \sum_{i=1}^{k} \left(\frac{1}{(\sqrt{2\pi}\sigma)^d} e^{-\frac{\|x - x_i\|^2}{2\sigma^2}} dx \cdot \delta_{x - x_i + y_i}(dy) \right).$$

Meanwhile,

$$\left(m \sum_{i=1}^{n} \gamma_{\sigma(i)} \right)(dx, dy) = m \sum_{i=1}^{n} \left(\frac{\mathbf{1}\{x \in B_{\sigma(i)}\}}{\mathsf{Vol}(B_{\sigma(i)})} dx \cdot \delta_{x - x_{\sigma(i)} + y_{\sigma(i)}}(dy) \right).$$

For every $\sigma(i)$ such that $x \in B_{\sigma(i)}$, note that

$$\|x - x_{\sigma(i)}\| \le \|x - (x_{\sigma(i)} + \alpha_{\sigma(i)})\| + \|\alpha_{\sigma(i)}\| \le \sigma + M,$$

hence (with a proper choice of c_d)

$$\frac{1}{k} \frac{1}{(\sqrt{2\pi}\sigma)^d} e^{-\frac{\|x - x_{\sigma(i)}\|^2}{2\sigma^2}} \ge \frac{1}{k} \frac{1}{(\sqrt{2\pi}\sigma)^d} e^{-\frac{(M+\sigma)^2}{2\sigma^2}} \ge \frac{m}{\mathsf{Vol}(B_{\sigma(i)})}.$$

As a result, $\gamma - m \sum_{i=1}^{n} \gamma_{\sigma(i)} \ge 0$, and $\tilde{\gamma}$ is a positive measure. Also note that its first marginal (i.e. the marginal on the first d dimensions) and second marginal (i.e. the marginal on the last d dimensions) agree with the respective marginals of γ. Thus we conclude that $\tilde{\gamma} \in \Pi(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$. Now note that

$$\begin{aligned}
&\int c(x, y) d\gamma(x, y) - \int c(x, y) d\tilde{\gamma}(x, y) \\
&= m \left(\sum_{i=1}^{n} \|x_{\sigma(i)} - y_{\sigma(i)}\|^2 - \sum_{i=1}^{n} \|(x_{\sigma(i)} + \alpha_{\sigma(i)}) - (y_{\sigma(i+1)} + \alpha_{\sigma(i+1)})\|^2 \right) \\
&= m \cdot G(M).
\end{aligned}$$

In the meantime,

$$\int c(x, y) d\gamma(x, y) = \frac{1}{2k} \sum_{i=1}^{k} \|x_i - y_i\|^2 = W_2^2(\mu, \nu),$$

therefore,

$$\begin{aligned}
&W_2^2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) \\
&\le \int c(x, y) d\tilde{\gamma}(x, y) \\
&\le W_2^2(\mu, \nu) - G(M) \cdot c_d \exp\left(-\frac{(M + \sigma)^2}{2\sigma^2} \right).
\end{aligned}$$

In particular, choosing $M = \sigma + \sigma_*$ yields

$$W_2^2(\mu, \nu) - W_2^2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) \gtrsim G(\sigma + \sigma_*) \exp\left(-c \frac{\sigma_*^2}{\sigma^2} \right).$$

The rest follows from the observation that, for $\sigma \in (0, 2\sigma_*)$,

$$G(\sigma + \sigma_*) = G(\sigma + \sigma_*) - G(\sigma_*) \ge \frac{G(3\sigma_*) - G(\sigma_*)}{2\sigma_*} \cdot \sigma$$

since G is concave by Proposition 3.2. □

4 Case II: no perfect matching

In the case that $R(\Gamma) = 0$, or equivalently by Proposition 2.6 and Proposition 2.11 that the optimal transport map between μ and ν is not a perfect matching, Theorems 3.1 and 3.3 are not applicable. In this situation, we are able to show that the approximation error is linear, even in a neighborhood of zero. In fact, this holds whenever there exists an optimal transport plan between μ and ν which is not a perfect matching.

To analyze this case, we generalize our setting to optimal transport problems between two discrete measures that do not necessarily have the same number of atoms, and whose mass may not be evenly distributed:

$$\mu = \sum_{i=1}^{m} \alpha_i \delta(x_i), \quad \nu = \sum_{j=1}^{n} \beta_j \delta(y_j)\,. \tag{7}$$

Here $\{\alpha_i\}_{i=1}^m$ and $\{\beta_j\}_{j=1}^n$ are positive numbers such that $\sum_{i=1}^m \alpha_i = \sum_{j=1}^n \beta_j = 1$. For the sake of notational convenience and without loss of generality, we also assume that $\{x_i\}$ and $\{y_j\}$ are all different. We prove the following result.

Theorem 4.1. *For μ, ν defined per* (7)*, unless the optimal transport plan T between μ and ν is unique and a perfect matching, i.e. $m = n$ and there exists a permutation p on $[m]$ such that $\alpha_i = \beta_{p(i)}$ and $T^{-1}(y_{p(i)}) = \{x_i\}$ for all $i \in [m]$, there exists $c_0 > 0$ such that for $\sigma \in (0, c_0)$,*

$$W_2^2(\mu, \nu) - W_2^2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) \gtrsim \sigma.$$

Theorem 4.1 tells that, unless the optimal transport plan between μ and ν is unique and a perfect matching, the loss from approximating the OT distance with the GOT distance is at least linear in σ. To proceed with the proof, we need the following lemma. Its proof can be found in the supplementary material.

Lemma 4.2. *Let x, y_1 and y_2 be different points in $\mathbb{R}^d$. For $\mu_0 := \delta(x)$ and $\nu_0 := \frac{1}{2}\delta(y_1) + \frac{1}{2}\delta(y_2)$, there exists $c_0 > 0$, such that for $\sigma \in (0, c_0)$, we have*

$$W_2^2(\mu_0, \nu_0) - W_2^2(\mu_0 * \mathcal{N}_\sigma, \nu_0 * \mathcal{N}_\sigma) \gtrsim \sigma. \tag{8}$$

Proof of Theorem 4.1. Suppose that there exists a transport plan T between μ and ν which achieves the optimal cost and is not a perfect matching, without loss of generality we assume that $x_1 \in T^{-1}(y_1) \cap T^{-1}(y_2)$. Denote π the associated coupling. Let $\lambda = \min\{\pi(x_1, y_1), \pi(x_1, y_2)\}$. We decompose μ and ν as

$$\begin{aligned} \hat{\mu} = \mu - 2\lambda\delta(x_1), \quad \tilde{\mu} = 2\lambda\delta(x_1), \\ \hat{\nu} = \nu - \lambda\left(\delta(y_1) + \delta(y_2)\right), \quad \tilde{\nu} = \lambda\left(\delta(y_1) + \delta(y_2)\right). \end{aligned}$$

By Lemma 4.2, there exists $c_0 > 0$ such that for $\sigma \in (0, c_0)$,

$$W_2^2(\tilde{\mu}, \tilde{\nu}) - W_2^2(\tilde{\mu} * \mathcal{N}_\sigma, \tilde{\nu} * \mathcal{N}_\sigma) \gtrsim \sigma.$$

Therefore, for $\sigma \in (0, c_0)$, we also have

$$\begin{aligned} & W_2^2(\mu, \nu) - W_2^2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) \\ \geq\ & W_2^2(\hat{\mu}, \hat{\nu}) - W_2^2(\hat{\mu} * \mathcal{N}_\sigma, \hat{\nu} * \mathcal{N}_\sigma) + W_2^2(\tilde{\mu}, \tilde{\nu}) - W_2^2(\tilde{\mu} * \mathcal{N}_\sigma, \tilde{\nu} * \mathcal{N}_\sigma) \\ \geq\ & W_2^2(\tilde{\mu}, \tilde{\nu}) - W_2^2(\tilde{\mu} * \mathcal{N}_\sigma, \tilde{\nu} * \mathcal{N}_\sigma) \\ \gtrsim\ & \sigma. \end{aligned}$$

□

We conclude that, for general discrete measures defined per (7), a phase transition in $W_2(\mu, \nu) - W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ only happens when the optimal transport plan between μ and ν is unique and a perfect matching with a positive $R(\Gamma)$. Otherwise, one would always suffer a linear loss in approximating the OT distance with the GOT distance.

5 Numerical example

In this section, we present a numerical example to demonstrate different regimes of the rate $W_2(\mu,\nu) - W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$, in respect of Theorem 3.1 and Theorem 3.3. For the sake of clarity, we consider atomic measures μ and ν both defined on $\mathbb{R}^2$. One of the simplest cases where a coupling Γ has $R(\Gamma) = 0$ is

$$\mu = \frac{1}{2}\left[\delta((-1,-1)) + \delta((1,1))\right],$$
$$\nu = \frac{1}{2}\left[\delta((-1,1)) + \delta((1,-1))\right]$$

It is easy to see that the optimal transport plan from μ to ν is not unique, which is also a consequence of Proposition 2.6, Proposition 2.11 and the fact that $R(\Gamma) = 0$ for the map

$$\Gamma = \{((-1,-1),(-1,1)),((1,1),(1,-1))\}$$

that achieves the optimal cost. We also consider the family

$$\mu_k = \frac{1}{2}\left[\delta((-1,-1+\frac{k}{10})) + \delta((1,1-\frac{k}{10}))\right],\ k \in [4]$$

The source and target distributions are demonstrated in Figure 2. For each k, the unique optimal transport plan from μ_k to ν is given by

$$\Gamma_k = \left\{((-1,-1+\frac{k}{10}),(-1,1)),((1,1-\frac{k}{10}),(1,-1))\right\}.$$

For each of these GOT tasks, we draw 500 samples from the source distribution $\mu_k * \mathcal{N}_\sigma$ and target distribution $\nu * \mathcal{N}_\sigma$, and use the empirical W_2 distance as an estimate of the true $W_2(\mu_k * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$. We repeat the process 20 times and report the mean, as shown in the following figure.

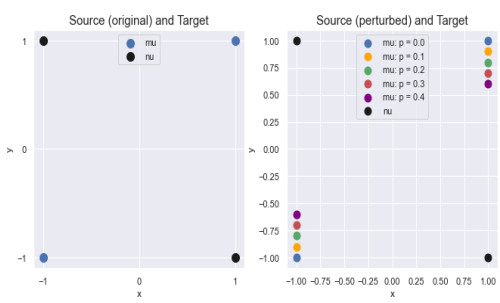

Figure 1: Source and Target distributions

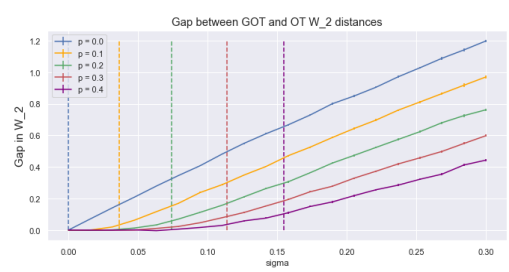

Figure 2: Rate of $W_2(\mu,\nu) - W_2(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ in the vanishing σ regime.

By Theorem 3.1 and Theorem 3.3, we expect $W_2^2(\mu_k,\nu) - W_2^2(\mu_k * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ to be of scale e^{-c/σ^2} for $\sigma \in (0, R(\Gamma_k))$, and $W_2^2(\mu_k,\nu) - W_2^2(\mu_k * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) \gtrsim \sigma$ for $\sigma \geq R(\Gamma_k)$. The phase transition happening at the vertical dashed line (here $R(\Gamma_k)$ is evaluated per Proposition 2.13) is indeed observed in the above experiment.

References

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[AWR17] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1961–1971, 2017.

[BLG14] Emmanuel Boissard and Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563, 2014.

[Bre87] Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *C. R. Acad. Sci. Paris Sér. I Math.*, 305(19):805–808, 1987.

[CFT14] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *ECML PKDD*, pages 274–289, 2014.

[CFTR17] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1853–1865, 2017.

[CGT19] Yongxin Chen, Tryphon T. Georgiou, and Allen R. Tannenbaum. Optimal transport for gaussian mixture models. *IEEE Access*, 7:6269–6278, 2019.

[CNW20] Hong-Bin Chen and Jonathan Niles-Weed. Asymptotics of smoothed wasserstein distances. *arXiv preprint arXiv:2005.00738*, 2020.

[Cut13] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *NIPS*, volume 2, page 4, 2013.

[DD20] Julie Delon and Agnès Desolneux. A wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

[Dud69] Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[DY95] V Dobrić and Joseph E Yukich. Asymptotics for transportation cost in high dimensions. *Journal of Theoretical Probability*, 8(1):97–118, 1995.

[FCVP17] Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017 - 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I*, pages 291–299, 2017.

[FG15] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

[GBG+18] Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. *arXiv preprint arXiv:1810.05728*, 2018.

[GCB+19] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

[GG20] Ziv Goldfeld and Kristjan Greenewald. Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. In *International Conference on Artificial Intelligence and Statistics*, pages 3327–3337. PMLR, 2020.

[GGK20] Ziv Goldfeld, Kristjan Greenewald, and Kengo Kato. Asymptotic guarantees for generative modeling based on the smooth wasserstein distance. *arXiv preprint arXiv:2002.01012*, 2020.

[GGNWP20] Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, 2020.

[GPC18] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1608–1617, 2018.

[MNW19] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

[MNW21] Tudor Manole and Jonathan Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs, 2021.

[NWB19] Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in wasserstein distance. *arXiv e-prints*, pages arXiv–1902, 2019.

[NWR19] Jonathan Niles-Weed and Philippe Rigollet. Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019.

[PdC20] François-Pierre Paty, Alexandre d'Aspremont, and Marco Cuturi. Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1222–1232. PMLR, 2020.

[PKD07] François Pitié, Anil C. Kokaram, and Rozenn Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1-2):123–137, 2007.

[Roc66] Ralph Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 17(3):497–510, 1966.

[Roc70] Ralph Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics*, 33(1):209–216, 1970.

[Roc87] Jean-Charles Rochet. A necessary and sufficient condition for rationalizability in a quasi-linear context. *Journal of mathematical Economics*, 16(2):191–200, 1987.

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[RW18] Philippe Rigollet and Jonathan Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11-12):1228–1235, 2018.

[SP18] Shashank Singh and Barnabás Póczos. Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

[SUL+18] Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos. Nonparametric density estimation under adversarial losses. *arXiv preprint arXiv:1805.08836*, 2018.

[THG17] Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Math. Program.*, 161(1-2, Ser. A):307–345, 2017.

[Vil08] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[VV21] Adrien Vacher and François-Xavier Vialard. Convex transport potential selection with semi-dual criterion. *arXiv preprint arXiv:2112.07275*, 2021.

[ZCR21] Yixing Zhang, Xiuyuan Cheng, and Galen Reeves. Convergence of gaussian-smoothed optimal transport distance with sub-gamma distributions and dependent samples. In *International Conference on Artificial Intelligence and Statistics*, pages 2422–2430. PMLR, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]