Sample Complexity Bounds for Estimating the Wasserstein Distance under Invariances

Behrooz Tahmasebi [1] **Stefanie Jegelka** [1]

Abstract

Group-invariant probability distributions appear in many data-generative models in machine learning, such as graphs, point clouds, and images. In practice, one often needs to estimate divergences between such distributions. In this work, we study how the inherent invariances with respect to any smooth action of a Lie group on a manifold improve sample complexity when estimating the Wasserstein distance. Our result indicates a two-fold gain: (1) reducing the sample complexity by a multiplicative factor corresponding to the group size (for finite groups) or the normalized volume of the quotient space (for groups of positive dimension), (2) improving the exponent in the convergence rate (for groups of positive dimension). These results are completely new for groups of positive dimension and tighten recent bounds for finite group actions.

1. Introduction

Estimating the optimal transportation cost between probability measures is a fundamental problem in statistics, with many applications in machine learning, from Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) to domain adaptation (Flamary et al., 2016; Courty et al., 2014), geometric data processing (e.g., Wasserstein barycenters (Cuturi and Doucet, 2014)), and biomedical research (Zhang et al., 2021).

Estimating the Wasserstein distance is known to be a difficult task in general, and many algorithms suffer from the curse of dimensionality (Tsybakov, 2009). The slow convergence rate is unimprovable in general, as there exist difficult probability measures to estimate. However, those hard instances barely appear in practice when we study more structured probability measures. Indeed, in many applications (e.g., graphs, point clouds, molecules, spectral data) the underlying probability measures are *invariant* with respect to a group action on the input space. As observed in recent works (Birrell et al., 2022; Chen et al., 2023), considering the group invariances into the model can help improve the convergence rate of the Wasserstein distance, with applications in GANs for invariant data.

In this paper, we study the sample complexity of estimating the Wasserstein distance under group invariances, for any probability measures supported on a connected compact smooth manifold $\mathcal{M}$ being invariant with respect to a smooth action of a Lie group G on $\mathcal{M}$. Under this general setting, given any two (Borel) probability measures μ, ν supported on the manifold $\mathcal{M}$, we prove the following upper bound on the convergence rate of the Wasserstein distance $W_1(.,.)$ using the modified empirical measures $\hat{\mu}, \hat{\nu}$ (from n i.i.d. samples from each distribution):

$$\mathbb{E}[|W_1(\hat{\mu}, \hat{\nu}) - W_1(\mu, \nu)|] \leq 2C\Big(\frac{\mathrm{vol}(\mathcal{M}/G)}{n}\Big)^{\frac{1}{d}}, \quad (1)$$

where C is an absolute constant (does not depend on the group G or the number of samples n). Also, d and $\mathrm{vol}(\mathcal{M}/G)$ denote the dimension and the volume of the quotient space $\mathcal{M}/G$, respectively.

The new sample complexity bound shows two different aspects of gain of invariances. Compared to the general case (i.e., without invariances, $G = \{\mathrm{id}_G\}$), first, the exponent is improved from $1/\dim(\mathcal{M})$ to $1/d$, where d can be potentially as small as $\dim(\mathcal{M}) - \dim(G)$. Second, $\mathrm{vol}(\mathcal{M}/G)$ can be potentially much smaller than $\mathrm{vol}(\mathcal{M})$, as for finite groups, it can be $\mathrm{vol}(\mathcal{M})/|G|$. This shows that for finite groups (i.e., $\dim(G) = 0$), the gain of invariances for sample complexity (compared to the general case), is to replace n by $n \times |G|$ in the classical convergence rate of the Wasserstein distance estimation. This proves the intuitive belief about the gain of invariances for finite groups that each sample conveys the information of $|G|$ samples while comparing the invariant case to the general case.

The upper bound proved in this paper is completely new for groups of positive dimension, and for finite groups, extends the recent result on submanifolds (of full dimension) of

[1]MIT CSAIL. Correspondence to: Behrooz Tahmasebi <bzt@mit.edu>.

Presented at the 2^{nd} Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA.

$\mathbb{R}^d$ under 1-Lipschitz group actions (Chen et al., 2023) to arbitrary manifolds and arbitrary Lie groups.

We also study the convergence rate of estimating the Wasserstein distance for smooth distributions. Indeed, for probability measures having a density with respect to the uniform distribution on the manifold, with s times square-integrable derivatives (known as being in the Sobolev space $H^s(\mathcal{M})$), we prove an upper bound on the convergence rate which also exhibits the same two-fold gain for the sample complexity of estimating the Wasserstein distance under invariances. Note that all the proven upper bounds in the paper reduce to the known tight bounds on estimating the Wasserstein distance if we set $G = \{\mathrm{id}_G\}$ (i.e., the trivial group).

We note that our findings cannot be derived immediately from the known results on estimating the Wasserstein distance under invariances. Instead of the idea of using covering numbers, which is used in a recent work (Chen et al., 2023), we use a Fourier approach to bounding the Wasserstein distance. We use the theory of the Laplace-Beltrami operator on manifolds, and via a new version of Weyl's law which captures the sparsity of the Fourier series on manifolds, as well as ideas from differential geometry and Fourier analysis (such as mollifiers), we prove the main result.

In short, in this paper, we make the following contributions:

- We prove an upper bound on the sample complexity of estimating the Wasserstein distance under group invariances, for any smooth Lie group action on a connected compact manifold (Theorem 1).
- We also study the convergence rate of the Wasserstein distance for smooth distributions and prove an upper bound for the same setup as above (Theorem 2).

2. Problem Statement

Let $\mathcal{M}$ denote an arbitrary compact, connected, and smooth manifold without boundary[1]. Let $\mathcal{P}(\mathcal{M})$ denote the set of Borel probability measures on $\mathcal{M}$, and also let $\mathrm{Lip}(\mathcal{M})$ denote the set of all measurable functions $f : \mathcal{M} \to \mathbb{R}$ such that $|f(x) - f(y)| \leq \mathrm{dist}(x, y)$, for all $x, y \in \mathcal{M}$, where $\mathrm{dist}(.,.)$ denotes the geodesic distance between points on $\mathcal{M}$. The Wasserstein distance between any two $\mu, \nu \in \mathcal{P}(\mathcal{M})$ is defined as follows:

$$W_1(\mu, \nu) := \sup_{f \in \mathrm{Lip}(\mathcal{M})} \Big\{ \int_{\mathcal{M}} f d\mu - \int_{\mathcal{M}} f d\nu \Big\}. \quad (2)$$

For two arbitrary (unknown) probability measures $\mu, \nu \in \mathcal{P}(\mathcal{M})$, assume that we are given independent samples $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mu$ and $Y_1, Y_2, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \nu$. The goal is to estimate $W_1(\mu, \nu)$ using the given $2n$ independent samples. Let $\hat{\mu} := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ denote the empirical measure of μ given $X_1, X_2, \ldots, X_n$, where δ_x denotes the Dirac measure supported on $x \in \mathcal{M}$. The modified empirical measure modifies the empirical measure by restricting it to the invariant eigenfunctions in the Fourier basis (see the proof of Theorem 1 in the full version). We denote it (with a slight abuse of notation) by $\hat{\mu}$, too. Define $\hat{\nu}$ similarly, and consider the modified empirical estimation $W_1(\hat{\mu}, \hat{\nu})$, as a candidate to estimate $W_1(\mu, \nu)$ with samples. Note that by the triangle inequality:

$$|W_1(\hat{\mu}, \hat{\nu}) - W_1(\mu, \nu)| \leq W_1(\mu, \hat{\mu}) + W_1(\nu, \hat{\nu}). \quad (3)$$

Thus, to study the convergence of the empirical estimator $W_1(\hat{\mu}, \hat{\nu})$, one just needs to prove an upper bound on $W_1(\mu, \hat{\mu})$ for any arbitrary measure $\mu \in \mathcal{P}(\mathcal{M})$.

Let G be an arbitrary Lie group acting smoothly on $\mathcal{M}$. Without loss of generality, we assume that $\mathcal{M}$ is equipped with a Riemannian metric g such that the action of G is *isometric* on $\mathcal{M}$ with respect to g. A probability measure $\mu \in \mathcal{P}(\mathcal{M})$ is called G-invariant if for any Borel set $A \subseteq \mathcal{M}$ and all $\tau \in G$, one has $\mu(A) = \mu(\tau A)$. For example, the uniform distribution on $(\mathcal{M}, g)$ is invariant with respect to any isometric group action.

3. Main Results

Theorem 1 (Convergence rate of the Wasserstein distance under invariances)**.** *For any G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$,*

$$\mathbb{E}[W_1(\mu, \hat{\mu})] \leq C\Big(\frac{\mathrm{vol}(\mathcal{M}/G)}{n}\Big)^{\frac{1}{d}}, \quad (4)$$

where C can only depend on the manifold $\mathcal{M}$. Also, $\mathrm{vol}(\mathcal{M}/G)$ *is the volume of the quotient space $\mathcal{M}/G$ and* $d := \dim(\mathcal{M}/G) \geq 3$. *Consequently, one has*

$$\mathbb{E}[|W_1(\hat{\mu}, \hat{\nu}) - W_1(\mu, \nu)|] \leq 2C\Big(\frac{\mathrm{vol}(\mathcal{M}/G)}{n}\Big)^{\frac{1}{d}}, \quad (5)$$

for any $\mu, \nu \in \mathcal{P}(\mathcal{M})$.

Even though we used the notation $\dim(\mathcal{M}/G)$ and $\mathrm{vol}(\mathcal{M}/G)$, we notice that the quotient space $\mathcal{M}/G$ is *not* necessarily a manifold. Also, it may exhibit boundary, even though $\mathcal{M}$ is assumed to be boundaryless (see the full version for examples). To address this issue, we define $\dim(\mathcal{M}/G)$ (or $\mathrm{vol}(\mathcal{M}/G)$) as the dimension (or the volume) of the *principal part* of the quotient space. The principal part, denoted as $\mathcal{M}_0/G$, is a connected dense subset of $\mathcal{M}/G$, such that it has a manifold structure inherited from $\mathcal{M}$. Since it is a manifold, one can define

[1]The results can be generalized to manifolds with boundaries, too. But we consider boundaryless manifolds here for simplicity.

its dimension/volume in a natural way. It is guaranteed that the principal part exists, and is unique, under the assumptions of this paper. Besides the principal part, $\mathcal{M}/G$ is only a disjoint union of *finitely* many other manifolds, all of lower dimension than the principal part. Note that $\mathrm{vol}(\mathcal{M}/G)$ is defined with respect to the dimension of the quotient space $\dim(\mathcal{M}/G)$, so it is nonzero even if $\dim(\mathcal{M}/G) < \dim(\mathcal{M})$.

To compare the convergence rate with the general case (i.e., not necessarily G-invariant probability measures), note that if $G = \{\mathrm{id}_G\}$, then the convergence rate is $\mathbb{E}[W_1(\mu, \hat{\mu})] \leq C\Big(\frac{\mathrm{vol}(\mathcal{M}/G)}{n}\Big)^{\frac{1}{\dim(\mathcal{M})}}$, as expected from the standard results for arbitrary probability measures (Fournier and Guillin, 2015). This shows that the sample complexity of estimating the Wasserstein distance is improved under invariances; (1) the new exponent is $\frac{1}{d}$ with $d = \dim(\mathcal{M}/G)$, which can be potentially much better than $\frac{1}{\dim(\mathcal{M})}$, (2) the number of samples is multiplied by $\mathrm{vol}(\mathcal{M})/\mathrm{vol}(\mathcal{M}/G)$. For finite groups, if its action on $\mathcal{M}$ is effective[2], then Theorem 1 shows that

$$\mathbb{E}[W_1(\mu, \hat{\mu})] \leq C\Big(\frac{\mathrm{vol}(\mathcal{M})}{n|G|}\Big)^{\frac{1}{\dim(\mathcal{M})}}. \tag{6}$$

This means that under invariances, each sample is worth the same as $|G|$ samples compared to the general (non-invariant) case. This improves a recent result on the convergence of the Wasserstein metric under invariances (Chen et al., 2023). Chen et al. (2023) prove that this rate is achievable for finite group actions on a compact submanifold (of full dimension) of the Euclidean space $\mathbb{R}^d$. However, our result is more general, holding for arbitrary smooth compact manifolds, including spheres, tori, hyperbolic spaces, and also for arbitrary groups, not only finite groups. Indeed, to the best of our knowledge, the improvement in the exponent is new for the convergence of the Wasserstein distance under invariances.

Let us observe the result of Theorem 1 in the following example.

Example 1 (Point clouds)**.** *Consider a point cloud as a set $\{p_1, p_2, \dots, p_m\} \subseteq (\mathbb{R}/\mathbb{Z})^3$ of m points. For fixed m, we can think of each point cloud as a point on the manifold $(\mathbb{R}/\mathbb{Z})^{3m}$. Point clouds are typically assumed to be unchanged under a change of coordinates for all the points:*

$$\{p_1, p_2, \dots, p_m\} \cong \{Ap_1, Ap_2, \dots, Ap_m\}, \tag{7}$$

for any orthogonal matrix A. Also, permuting the points won't change the point clouds. Let G denote the group of invariances for point clouds defined on $(\mathbb{R}/\mathbb{Z})^{3m}$ as above. Then, after doing the calculations, the gain of invariances (i.e., estimating the Wasserstein distance on point clouds by considering the invariances of the problem) is (1) improving the exponent from $3m$ to $3m-6$, and (2) multiplying the number of samples n by $m!$.

[2] The action of a group G on a manifold $\mathcal{M}$ is called effective, if any $\tau \neq \mathrm{id}_G$ corresponds to a non-trivial bijection on $\mathcal{M}$.

Proof sketch for Theorem 1. In this part, we give a quick proof sketch for Theorem 1. The complete proof is available in the full version.

To prove the theorem, we focus on an approach for upper bounding the Wasserstein distance using the orthonormal basis $\phi_\ell \in L^2(\mathcal{M})$, $\ell = 0, 1, \dots$, of eigenfunctions of Laplacian on $\mathcal{M}$ in $L^2(\mathcal{M})$ (see (Bobkov and Ledoux, 2021) for more details). This allows us to conclude that

$$W_1^2(\mu, \nu) \leq \sum_{\ell=1}^{\infty} \frac{(\mu_\ell - \nu_\ell)^2}{\lambda_\ell} \times \mathrm{vol}(\mathcal{M}), \tag{8}$$

where $\mu_\ell = \int_{\mathcal{M}} \phi_\ell d\mu$ for each ℓ (defined similarly for ν), and λ_ℓ, $\ell = 0, 1, \dots$, are the eigenvalues of the Laplacian operator on $\mathcal{M}$. This approach shows that to upper bound the Wasserstein distance, all we need is to know how sparse the sequence μ_ℓ, $\ell = 0, 1, \dots$, is for a G-invariant probability measure μ. To this end, we use recent results on quantifying the sparsity of the series for G-invariant functions defined on a connected compact smooth manifold $(\mathcal{M}, g)$.

However, it turns out that using this method cannot guarantee a finite convergence rate since high-frequency components in the sum accumulate a lot of noise for empirical measures. To solve this issue, we use a mollifier function with exponential tail decay (in the Laplacian basis) and use the theory of heat kernel on manifolds to achieve the final result. We provide more explanations in the full version.

3.1. Convergence Rate for Smooth Distributions

Assume that $\mu \in \mathcal{P}(\mathcal{M})$ is absolutely continuous with respect to the uniform probability measure $dx = \frac{1}{\mathrm{vol}(\mathcal{M})} d\,\mathrm{vol}(x)$ on $(\mathcal{M}, g)$. Assume that $\frac{d\mu}{dx} \in H^s(\mathcal{M})$, for some $s \geq 0$, where $H^s(\mathcal{M})$ denotes the Sobolev space of real-valued measurable functions on $(\mathcal{M}, g)$ having s times square-integrable derivatives. In this spacial case, the probability measure is *smoother* as s grows.

It turns out that in this special case, the convergence rate of estimating the Wasserstein distance as a function of the number of samples can be improved using a new estimator $\tilde{\mu}$ (which is different from the modified empirical estimator $\hat{\mu}$). The following theorem states the main result for smooth distributions.

Theorem 2 (Convergence rate of the Wasserstein distance for smooth distributions under invariances)**.** *For any G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$ with $\frac{d\mu}{dx} \in H^s(\mathcal{M})$ for some $s \geq 0$, there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$, as a function of n i.i.d. samples*

$X_1, X_2 \dots, X_n \sim \mu$, such that

$$\mathbb{E}[W_1(\mu,\tilde{\mu})] \le \sqrt{\mathrm{vol}(\mathcal{M})} \times \Big(\frac{1}{4\kappa^2}\frac{\omega_d}{(2\pi)^d}\frac{\mathrm{vol}(\mathcal{M}/G)}{n}\Big)^{\frac{s+1}{2s+d}} \Big\|\frac{d\mu}{dx}\Big\|_{H^s(\mathcal{M})}^{\frac{s+1}{s+d/2}},$$

where $\kappa = \frac{s+1}{d-2}$, and ω_d is the volume of the unit ball in $\mathbb{R}^d$. Also, $\mathrm{vol}(\mathcal{M}/G)$ is the volume of the quotient space $\mathcal{M}/G$ and $d := \dim(\mathcal{M}/G) \ge 3$.

To define the estimator $\tilde{\mu}$, we need to review some facts about manifolds. The set of square-integrable G-invariant functions on a connected, compact, smooth manifold $\mathcal{M}$ has an orthonormal basis $\phi_\ell^{\mathrm{inv}} \in L^2(\mathcal{M})$, $\ell = 0, 1, \dots$ of eigenfunctions of Laplacian on $\mathcal{M}$ (see the full version for more details). For the case of the circle $\mathcal{M} = \mathbb{S}^1$, these functions correspond to the sinusoidal waves/Fourier basis that are invariant under the group action. Given n samples $X_1, X_2, \dots, X_n$, the Borel measure $\tilde{\mu}$ is defined using its Radon–Nikodym derivative with respect to the uniform probability measure on $(\mathcal{M}, g)$ as follows:

$$\frac{d\tilde{\mu}}{dx} := \frac{\mathrm{vol}(\mathcal{M})}{n}\sum_{\ell=0}^{T-1}\sum_{i=1}^{n}\phi_\ell^{\mathrm{inv}}(X_i)\phi_\ell, \tag{9}$$

where T is a fixed positive integer (to be set). For any T, $\tilde{\mu}$ is a Borel measure, but in general, it can be a signed measure with $\int_{\mathcal{M}} d\tilde{\mu} \ne 1$. We can then take the closest probability measure to $\tilde{\mu}$ in Wasserstein distance as the final estimation for μ. With a slight abuse of notation, we denote the final output of the algorithm by $\tilde{\mu}$ again.

Choosing larger T corresponds to higher *variance* due to the randomness of sampling while it reduces the *bias* of the estimator. Therefore, optimizing T to balance the bias and variance terms, according to the problem's parameters, allows to achieve the best algorithm of this type (in terms of the convergence rate). We follow this approach to prove Theorem 2.

Theorem 2 shows that the gain of invariances for estimating the Wasserstein distance for smooth distributions under invariances follows the same behavior as before, for any $s \ge 0$. The two-fold gain is observed on the exponent and the multiplicative factor. The new upper bound's exponent interpolates between the worst-case exponent $1/\dim(\mathcal{M})$ and $1/2$. As s enlarges, the exponent converges to $1/2$, as expected. If $G = \{\mathrm{id}_G\}$, the bound reduces to the known convergence rate of the Wasserstein distance estimation under smoothness (without invariances) (Liang, 2021; Niles-Weed and Berthet, 2022).

Acknowledgements

This research was supported by the NSF award CCF-2112665, the NSF Award 2134108, and the ONR grant N00014-20-1-2023 (MURI ML-SCOPE).

References

Jeremiah Birrell, Markos A Katsoulakis, Luc Rey-Bellet, and Wei Zhu. Structure-preserving gans. *arXiv preprint arXiv:2202.01129*, 2022.

Sergey G Bobkov and Michel Ledoux. A simple fourier analytic proof of the akt optimal matching theorem. *The Annals of Applied Probability*, 31(6):2567–2584, 2021.

Ziyu Chen, Markos A Katsoulakis, Luc Rey-Bellet, and Wei Zhu. Sample complexity of probability divergences under group symmetry. *arXiv preprint arXiv:2302.01915*, 2023.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, pages 274–289. Springer, 2014.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

R Flamary, N Courty, D Tuia, and A Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 1, 2016.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Tengyuan Liang. How well generative adversarial networks learn distributions. *The Journal of Machine Learning Research*, 22(1):10366–10406, 2021.

Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in wasserstein distance. *The Annals of Statistics*, 50(3):1519–1540, 2022.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Jingyi Zhang, Wenxuan Zhong, and Ping Ma. A review on modern computational optimal transport methods with applications in biomedical research. *Modern Statistical Methods for Health Research*, pages 279–300, 2021.