Sample Complexity Bounds for Estimating Probability Divergences under Invariances

Behrooz Tahmasebi [1] **Stefanie Jegelka** [2]

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 1-Wasserstein distance, the Sobolev Integral Probability Metrics (Sobolev IPMs), the Maximum Mean Discrepancy (MMD), and also the complexity of the density estimation problem (in the L^2 and L^∞ distance). Our results indicate 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 extend recent bounds for finite group actions.

1. Introduction

Estimating the optimal transportation distance (Villani, 2021; Villani et al., 2009; Santambrogio, 2015) between probability measures is a fundamental problem in statistics, with many applications in machine learning, from Generative Adversarial Networks (GANs) (Goodfellow et al., 2020; Arjovsky et al., 2017; Salimans et al., 2018; Mallasto et al., 2019) to domain adaptation and generalization (Flamary et al., 2016; Courty et al., 2014; Chuang et al., 2021), geometric data processing (e.g., Wasserstein barycenters (Cuturi & Doucet, 2014) and intrinsic dimension estimation (Block et al., 2022)), biomedical research (Zhang et al., 2021), and

[1] MIT CSAIL [2] TU Munich and MIT CSAIL. Correspondence to: Behrooz Tahmasebi <bzt@mit.edu>.

control and dynamical systems (Bunne, 2023).

Estimating the 1-Wasserstein distance is known to be a difficult task in general, and it suffers from the curse of dimensionality (Tsybakov, 2009). The slow convergence rate is generally unimprovable, as there exist probability measures that are difficult to estimate. However, 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 in the mathematical model can help improve the convergence rate of the 1-Wasserstein distance and the Sobolev Integral Probability Metrics (Sobolev IPMs), with applications, e.g., to generative models for invariant data.

In this paper, we study the sample complexity gain of invariances for estimating probability measures in the 1-Wasserstein distance, the Sobolev IPMs, the Maximum Mean Discrepancy (MMD), and also for the density estimation problem (in L^2 and L^∞ distance). We consider probability measures supported on a given[1] connected compact smooth manifold $\mathcal{M}$ that are invariant with respect to a smooth action of a Lie group G on $\mathcal{M}$. In this general setting, given a (Borel) probability measure μ, supported on the manifold $\mathcal{M}$, we prove that there exists an estimator $\tilde{\mu}$ (as a function of n i.i.d. samples from μ), such that

$$\mathbb{E}[D(\mu, \tilde{\mu})] \lesssim \left(\frac{\delta_G}{n}\right)^{\kappa_G}, \tag{1}$$

where D can be replaced by the 1-Wasserstein distance, Sobolev IPMs, MMD, and also the $L^2(\mathcal{M})$ and $L^\infty(\mathcal{M})$ distance between distributions (the density estimation problem in $L^2(\mathcal{M})$ and $L^\infty(\mathcal{M})$). The exponent κ_G and the factor δ_G depend on the distribution's smoothness properties and the quotient space's dimension and volume.

The new sample complexity bound shows two different aspects of gain of invariances. First, the quantity κ_G is observed to be a non-increasing function of the quotient

---

[1] Throughout this paper, we always assume that the underlying space is a known manifold; this is distinguished from a related body of work on estimating distributions supported on unknown low-dimensional embedded manifolds (Divol, 2022).

Table 1. Summary of the main results.

Divergence Measure	δ_G	κ_G
1-Wasserstein Distance (Theorem 1)	$\mathrm{vol}(\mathcal{M}/G)$	$1/d$
1-Wasserstein Distance (s-smooth, Theorem 2)	$\mathrm{vol}(\mathcal{M}/G)$	$(s+1)/(2s+d)$
Sobolev IPMs ($\alpha < d/2$, Theorem 3)	$\mathrm{vol}(\mathcal{M}/G)$	$(s+\alpha)/(2s+d)$
Sobolev IPMs ($\alpha > d/2$, MMD regime, Theorem 4)	$\mathcal{Z}(\alpha; G)$	$1/2$
Density Estimation in L^2 Distance (Theorem 5)	$\mathrm{vol}(\mathcal{M}/G)$	$s/(2s+d)$
Density Estimation in L^∞ Distance ($s > d/2$, Theorem 6)	$(\mathrm{vol}(\mathcal{M}/G))^{\frac{2s}{s-d/2}}$	$(s-d/2)/(2s+d)$

*For any divergence measure D in the table, we prove that there exists an estimator $\tilde{\mu}$, as a function of n i.i.d. samples from μ, such that $\mathbb{E}[D(\mu, \tilde{\mu})] \lesssim \left(\frac{\delta_G}{n} \right)^{\kappa_G}$. Sobolev IPMs use the Sobolev space $\mathcal{H}^\alpha(\mathcal{M})$ as test functions, for some $\alpha \geq 0$. Moreover, $d\mu/dx \in \mathcal{H}^s(\mathcal{M})$ for some $s \geq 0$. Also, d denotes the dimension of the quotient space $\mathcal{M}/G$, and $\mathcal{Z}(\alpha; G)$ is the zeta function of the manifold when it is restricted to the invariant eigenfunction of the Laplace-Beltrami operator with respect to the group G.

space's dimension d, so compared to the general case (i.e., without invariances, $G = \{\mathrm{id}_G\}$), the exponent is improved, as d can be potentially as small as $\dim(\mathcal{M}) - \dim(G)$.

Second, the factor δ_G is a non-decreasing function of the quotient space's volume $\mathrm{vol}(\mathcal{M}/G)$, which can be potentially much smaller than $\mathrm{vol}(\mathcal{M})$. For instance, it can be $\mathrm{vol}(\mathcal{M})/|G|$ for finite groups. In particular, we have $\delta_G = \mathrm{vol}(\mathcal{M}/G)$ for estimating the 1-Wasserstein distance. Therefore, for finite groups (i.e., $\dim(G) = 0$), we may view the gain of invariances for sample complexity (compared to the general case) as replacing the number of samples n by $n \times |G|$ in the classical convergence rate of the 1-Wasserstein distance estimation (without invariances). This is intuitively reasonable: the gain of invariances for finite groups is that each sample effectively conveys the information of $|G|$ samples when comparing the invariant case to the general case. Our result proves this intuition formally.

The upper bound proved in this paper is completely new for groups of positive dimension. For finite groups, it extends a recent result for submanifolds (of full dimension) of $\mathbb{R}^d$ under 1-Lipschitz group actions (Chen et al., 2023) to arbitrary manifolds and arbitrary Lie groups.

We further study the convergence rate of estimating probability divergences for smooth distributions. Indeed, for probability measures having a density with respect to the uniform distribution on the manifold, with square-integrable derivatives up to order s (known as being in the Sobolev space $\mathcal{H}^s(\mathcal{M})$), we prove upper bounds on the convergence rate that exhibit the same two-fold gain for the sample complexity as discussed above, namely a factor in the *effective* number of samples, and an improvement in the exponent. Note that all the proven upper bounds in the paper reduce to the known tight bounds on estimating without invariances if we set $G = \{\mathrm{id}_G\}$ (i.e., the trivial group).

Our findings cannot be derived immediately from known results on estimating the 1-Wasserstein distance or Sobolev IPMs. Instead of the idea of using covering numbers, which

are used in a recent work (Chen et al., 2023), we use a Fourier approach to bounding the error. Specifically, 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, we prove the main result.

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

- We prove convergence rates for estimating the 1-Wasserstein distance for any invariant probability measure to a smooth Lie group action on a connected compact manifold (Theorem 1). Moreover, we extend this result to smooth distributions under invariances (Theorem 2).

- We prove convergence rates for estimating the Sobolev IPMs (Theorem 3) and the Maximum Mean Discrepancy (MMD, Theorem 4) under group invariances.

- We prove convergence rates for the density estimation problem in the $L^2(\mathcal{M})$ and $L^\infty(\mathcal{M})$ distance under invariances (Theorem 5 and Theorem 6).

2. Related Work

Optimal transportation has been an extensive area of research for the last few decades (Villani, 2021; Villani et al., 2009). The Kantorovich duality (Kantorovich, 2006) allows the dual formulation of the optimal transportation problem used in this paper. The study of the convergence rate of the 1-Wasserstein distance is a classical problem (Tsybakov, 2009; Fournier & Guillin, 2015; Weed & Bach, 2019; Boissard & Le Gouic, 2014; Singh et al., 2018; Lei, 2020; Butkovsky, 2014; Panaretos & Zemel, 2019; Arjovsky et al., 2017; Mroueh et al., 2017; Mena & Niles-Weed, 2019; Rigollet & Stromme, 2022; Genevay et al., 2019). Canas & Rosasco (2012) use the 1-Wasserstein distance for learning probability measures on manifolds. The computational

aspects of optimal transport in machine learning are also addressed in (Peyré et al., 2019).

Group-invariant probability measures have many applications in machine learning; for example, in addition to what was presented in the previous section, in group equivariant GANs (Dey et al., 2020; Birrell et al., 2022), normalizing flow (Biloš & Günnemann, 2021) and equivariant flows (Köhler et al., 2019).

In a recent paper, Chen et al. (2023) studied the gain of invariances for estimating the 1-Wasserstein distance, and they proved upper bounds on the convergence rate for finite groups with 1-Lipschitz actions of submanifolds (of full dimension) of $\mathbb{R}^d$, while our rates hold for any group (including groups of positive dimension), and any manifold, thus extending the previous results. In another closely related work, the gain of group invariances is investigated for the kernel ridge regression problem on manifolds (Tahmasebi & Jegelka, 2023). While we study completely different divergence estimation problems here, we follow the approach proposed in (Tahmasebi & Jegelka, 2023) to use the Laplace-Beltrami operator (i.e., Fourier basis) to study the benefits of invariances. For more on the benefits of invariances, see (Bietti et al., 2021).

It is known that the convergence rate of estimating the 1-Wasserstein distance can be improved if the densities are smooth (Singh & Póczos, 2018; Fournier & Guillin, 2015; Liang, 2021; Niles-Weed & Berthet, 2022). The study of convergence rates of the Sobolev IPMs for smooth distributions has applications, e.g., in Sobolev GANs (Mroueh et al., 2017). See (Fukumizu et al., 2007; Gretton et al., 2006; 2007; Smola et al., 2007; Gretton et al., 2012; Anderson et al., 2019; Tolstikhin et al., 2017; 2016; Sriperumbudur et al., 2011; 2010) for the study of achievable convergence rates and other properties of MMDs and (Hendriks, 1990) for the density estimation problem (without invariances).

3. Preliminaries

Let $\mathcal{M}$ denote an arbitrary compact, connected, and smooth manifold without boundary[2]. Let $\mathcal{P}(\mathcal{M})$ denote the set of Borel probability measures on $\mathcal{M}$, and also let $\text{Lip}(\mathcal{M})$ denote the set of all measurable functions $f : \mathcal{M} \to \mathbb{R}$ such that $|f(x) - f(y)| \leq \text{dist}(x, y)$, for all $x, y \in \mathcal{M}$, where $\text{dist}(.,.)$ denotes the geodesic distance between points on $\mathcal{M}$.

Definition 1. *The 1-Wasserstein distance between any two $\mu, \nu \in \mathcal{P}(\mathcal{M})$ is defined as follows:*

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

For any (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 $D(\mu, \nu)$, where D denotes a metric distance between distributions (such as the 1-Wasserstein distance or a Sobolev Integral Probability Metric), using the $2n$ independent samples. Note that given estimations of the (unknown) probability measures μ, ν as a function of their i.i.d. samples, such as $\tilde{\mu}, \tilde{\nu}$, one can estimate $D(\mu, \nu)$ by the triangle inequality[3]:

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

Thus, $D(\tilde{\mu}, \tilde{\nu})$ gives an estimation of the true distance $D(\mu, \nu)$. This means that to study the convergence of the probability divergences $D(\mu, \nu)$, one just needs to prove an upper bound on $D(\mu, \tilde{\mu})$ for any probability measure $\mu \in \mathcal{P}(\mathcal{M})$. In particular, we can focus on the problem of estimating probability measures from samples in the $D(.,.)$ distance, and study the effects of group invariances on its sample complexity.

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; see (Tahmasebi & Jegelka, 2023) for more details. 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. Without loss of generality and just for simplicity, in this paper, we assume that $\text{vol}(\mathcal{M}) = 1$, and we denote the volume element on the manifold by $dx := d\text{vol}_g(x)$.

4. Main Results

This section presents our main results on bounding $D(\mu, \tilde{\mu})$ under invariances. We start with the 1-Wasserstein distance for arbitrary probability measures (Theorem 1) and then state the results for smooth distributions (Theorem 2). Then, we study the sample complexity of Sobolev IPMs (Theorem 3) and MMDs (Theorem 4). Finally, we state our results for the density estimation problem in the $L^2(\mathcal{M})$ distance (Theorem 5) and in the $L^\infty(\mathcal{M})$ distance (Theorem 6).

4.1. 1-Wasserstein Distance

For any probability measure $\mu \in \mathcal{P}(\mathcal{M})$, assume that we are given samples $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mu$. Let $\hat{\mu} := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ denote the empirical measure given $X_1, X_2, \ldots, X_n$, where δ_x denotes the Dirac measure supported on $x \in \mathcal{M}$. The empirical measure $\hat{\mu}$ is probably the most straightforward way to generate an estimator of μ from

---

[2]The results can also be generalized to manifolds with boundaries. But we consider boundaryless manifolds here for simplicity.

[3]For more on tightness of this upper bound, see (Liang, 2019).

samples. However, for G-invariant μ, the empirical measure $\hat{\mu}$ is not necessarily G-invariant. To address this issue, we introduce and use a modified empirical measure, where the estimation is restricted to only the non-trivial parts of the Fourier transform of the measure (see the proof of Theorem 1), as for measures that exhibit invariance, parts of the spectrum that correspond to non-invariant functions are zero. Thus, the modified empirical measure is always G-invariant. The first result of this paper is the following theorem on the convergence of the 1-Wasserstein distance for arbitrary (Borel) probability measures using the modified empirical measure.

Theorem 1 (Convergence rate of the 1-Wasserstein distance under invariances)**.** *For any G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$, there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$, as a function of n i.i.d. samples $X_1, X_2 \ldots, X_n \sim \mu$, such that*

$$\mathbb{E}[W_1(\mu, \tilde{\mu})] \lesssim \left(\frac{\mathrm{vol}(\mathcal{M}/G)}{n}\right)^{\frac{1}{d}}, \qquad (4)$$

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

$$\mathbb{E}[|W_1(\tilde{\mu}, \tilde{\nu}) - W_1(\mu, \nu)|] \lesssim \left(\frac{\mathrm{vol}(\mathcal{M}/G)}{n}\right)^{\frac{1}{d}}, \quad (5)$$

for any G-invariant probability measures $\mu, \nu \in \mathcal{P}(\mathcal{M})$.

In proving this theorem, there are some complications to overcome. For instance, 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; this can make the Fourier approach inapplicable to this problem (Tahmasebi & Jegelka, 2023). 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 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 $\mathbb{E}[W_1(\mu, \hat{\mu})] \lesssim \left(\frac{\mathrm{vol}(\mathcal{M})}{n}\right)^{\frac{1}{\dim(\mathcal{M})}}$, as expected from the standard results for arbitrary probability measures (Fournier & Guillin, 2015).

This shows that the sample complexity of estimating the 1-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 greater than $\frac{1}{\dim(\mathcal{M})}$, (2) the number of samples n is multiplied by $\mathrm{vol}(\mathcal{M})/\mathrm{vol}(\mathcal{M}/G)$. For finite groups, if the action on $\mathcal{M}$ is effective[4], then Theorem 1 shows that

$$\mathbb{E}[W_1(\mu, \hat{\mu})] \lesssim \left(\frac{\mathrm{vol}(\mathcal{M})}{n|G|}\right)^{\frac{1}{\dim(\mathcal{M})}}. \qquad (6)$$

This could be interpreted as each sample being worth the same as $|G|$ samples under invariances compared to the general (non-invariant) case. This improves a recent result on the convergence of the 1-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, and hyperbolic spaces, and also for arbitrary groups, not necessarily finite groups. Indeed, to the best of our knowledge, the improvement in the exponent is new for the convergence of the 1-Wasserstein distance under invariances.

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

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

$$\{p_1, p_2, \ldots, p_m\} \cong \{Ap_1, Ap_2, \ldots, Ap_m\}, \qquad (7)$$

for any orthogonal matrix A. Also, permuting the points will not change the point clouds. Let G denote the group of invariances for point clouds defined on $(\mathbb{R}/\mathbb{Z})^{3m}$ as above. Then, by Theorem 1, the gain of invariances (i.e., estimating the 1-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 $m!$.

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

To prove the theorem, we focus on an approach for upper bounding the 1-Wasserstein distance using the orthonormal basis $\phi_\ell \in L^2(\mathcal{M})$, $\ell = 0, 1, \ldots$, of eigenfunctions of the Laplace-Beltrami operator on $\mathcal{M}$ in $L^2(\mathcal{M})$ (see Appendix

---

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

A for more details). This allows us to conclude that

$$W_1^2(\mu,\nu) \le \sum_{\ell=1}^{\infty} \frac{(\mu_\ell - \nu_\ell)^2}{\lambda_\ell}, \tag{8}$$

where $\mu_\ell = \int_{\mathcal{M}} \phi_\ell d\mu$ for each ℓ (defined similarly for ν), and λ_ℓ, $\ell = 0, 1, \ldots$, are the eigenvalues of the Laplace-Beltrami operator on $\mathcal{M}$. This approach shows that to upper bound the 1-Wasserstein distance, all we need is to know how sparse the sequence μ_ℓ, $\ell = 0, 1, \ldots$, 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 $(\mathcal{M}, g)$ (Tahmasebi & Jegelka, 2023).

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 Laplace-Beltrami basis) and use the theory of heat kernels on manifolds to achieve the final result. Further detailed explanations are provided in Appendix B, where the complete proof is presented.

4.2. 1-Wasserstein Distance for Smooth Densities

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

It turns out that in this special case, the convergence rate of estimating the 1-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 1-Wasserstein distance for smooth distributions under invariances)**.** *For any G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$ with $\frac{d\mu}{dx} \in \mathcal{H}^s(\mathcal{M})$ for some $s \ge 0$, there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$, as a function of n i.i.d. samples $X_1, X_2 \ldots, X_n \sim \mu$, such that*

$$\mathbb{E}[W_1(\mu,\tilde{\mu})] \lesssim \left(\frac{\text{vol}(\mathcal{M}/G)}{n}\right)^{\frac{s+1}{2s+d}} \left\|\frac{d\mu}{dx}\right\|_{\mathcal{H}^s(\mathcal{M})}^{\frac{d-2}{2s+d}},$$

where $\text{vol}(\mathcal{M}/G)$ is the volume of the quotient space $\mathcal{M}/G$ and $d := \dim(\mathcal{M}/G) \ge 3$.

Theorem 2 shows that the gain of invariances for estimating the 1-Wasserstein distance for smooth distributions under invariances follows the same behavior as before, for any

$s \ge 0$. The two-fold gain is observed in 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 grows, the exponent converges to $1/2$, as expected. Moreover, if $G = \{\text{id}_G\}$, the bound reduces to the known convergence rate of 1-Wasserstein distance estimation under smoothness (without invariances) (Liang, 2021; Niles-Weed & Berthet, 2022).

Proof sketch for Theorem 2. 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 $\phi_\ell^{\text{inv}} \in L^2(\mathcal{M})$, $\ell = 0, 1, \ldots$ that consists of the eigenfunctions of the Laplace-Beltrami operator on $\mathcal{M}$ (see Appendix A for more details). For example, for the circle $\mathcal{M} = \mathbb{S}^1$, these functions correspond to the sinusoidal waves that are invariant under the group action. Given n samples $X_1, X_2, \ldots, X_n$, 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:

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

where T is a fixed positive integer (to be set). For any T, $\tilde{\mu}$ is a Borel measure, but generally, it can be a signed measure. We can then take the closest probability measure to $\tilde{\mu}$ in the 1-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.

The parameter T indicates when the sum is terminated. Note that estimating the higher-order coefficients in the Fourier basis requires many samples, and so if the Fourier coefficients decay quickly, one can neglect them and truncate the sum at a finite T with a small error. Indeed, choosing a higher number T of eigenfunctions reduces the bias of the estimator, but it also increases the variance due to the randomness of sampling and the difficulty of estimating the higher-order terms. Therefore, optimizing T to balance the bias and variance terms, according to the problem's parameters, allows us to achieve the best algorithm of this type (in terms of the convergence rate). Note that the decay of the Fourier coefficients is provided using the Sobolev space assumption (see Appendix C more details). We follow this approach to prove Theorem 2.

4.3. Sobolev Integral Probability Metrics (Sobolev IPMs)

Sobolev Integral Probability Metrics (IPMs) are a family of integral probability metrics (Müller, 1997; Sriperumbudur et al., 2012) that use the Sobolev space $\mathcal{H}^\alpha(\mathcal{M})$ as the set of test functions to define a divergence between probability measures.

Definition 2 (Sobolev IPMs). *For any $\alpha \geq 0$, the Sobolev IPM with parameter α is defined as*

$$D_\alpha(\mu, \nu) := \sup_{\substack{f \in \mathcal{H}^\alpha(\mathcal{M}) \\ \|f\|_{\mathcal{H}^\alpha(\mathcal{M})} \leq 1}} \left\{ \mathbb{E}_{x \sim \mu}[f(x)] - \mathbb{E}_{x \sim \nu}[f(x)] \right\},$$

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

Here, we study the sample complexity of estimating Sobolev IPMs under invariances, where we assume that we have samples from a smooth probability measure μ; more precisely, $\frac{d\mu}{dx} \in \mathcal{H}^s(\mathcal{M})$ for some $s \geq 0$. Note that for $\alpha > d/2$, the Sobolev IPMs are special cases of the Maximum Mean Discrepancies (MMDs) with the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}^\alpha(\mathcal{M})$. Since we cover the convergence rates of MMDs in the next section, we focus on the non-MMD regime where $\alpha \leq d/2$.

Theorem 3 (Convergence rate of Sobolev IPMs under invariances). *Consider a G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$ with $\frac{d\mu}{dx} \in \mathcal{H}^s(\mathcal{M})$ for some $s \geq 0$.*

- *If $\alpha < d/2$, then there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$, as a function of n i.i.d. samples $X_1, X_2 \ldots, X_n \sim \mu$, such that*

$$\mathbb{E}[D_\alpha(\mu, \tilde{\mu})] \lesssim \left(\frac{\text{vol}(\mathcal{M}/G)}{n} \right)^{\frac{s+\alpha}{2s+d}} \left\| \frac{d\mu}{dx} \right\|_{\mathcal{H}^s(\mathcal{M})}^{\frac{d-2\alpha}{2s+d}},$$

  where the constant only depends on the manifold and $\text{vol}(\mathcal{M}/G)$ is the volume of the quotient space $\mathcal{M}/G$ with $d := \dim(\mathcal{M}/G)$.

- *If $\alpha = d/2$, then there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$ such that*

$$\mathbb{E}[D_\alpha(\mu, \tilde{\mu})] \lesssim \sqrt{\frac{\text{vol}(\mathcal{M}/G) \log(n)}{n}}.$$

We note that the same two-fold gain is observed in the convergence of the Sobolev IPMs when $\alpha \leq d/2$.

The convergence results behave differently in the three regimes. First, if $\alpha < d/2$, it extends the 1-Wasserstein distance convergence rate for smooth distributions to the Sobolev IPMs. Second, if $\alpha > d/2$, as we will see in the next section, the rate saturates at $O(n^{-1/2})$. Finally, in the third convergence regime where $\alpha = d/2$, we get a convergence rate of the order $O(\sqrt{\log(n)/n})$.

Remark 1. *While Theorem 3 holds when (at least) $\frac{d\mu}{dx} \in L^2(\mathcal{M})$, we can extend it to all (Borel) probability measures (for which the rate corresponds to $s = 0$), by the same approach as the proof of Theorem 1.*

Proof sketch for Theorem 3. We follow the same approach as Theorem 2. Here, the coefficients of the test functions

$f \in \mathcal{H}^\alpha(\mathcal{M})$ in the Laplace-Beltrami basis can be upper bounded using the parameter α (by the definition of Sobolev spaces). Then, we use an appropriate frequency T, as a function of α, to simultaneously minimize the bias and variance terms.

4.4. Maximum Mean Discrepancy (MMD)

The Maximum Mean Discrepancy (MMD) is an integral probability metric associated with a Positive Definite Symmetric (PDS) kernel, where the set of test functions is a Reproducing Kernel Hilbert Space (RKHS) (Berlinet & Thomas-Agnan, 2011).

Definition 3. *For any PDS kernel $K : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ with an RKHS denoted by $\mathcal{H}$, define*

$$D_\mathcal{H}(\mu, \nu) := \sup_{\substack{f \in \mathcal{H} \\ \|f\|_\mathcal{H} \leq 1}} \left\{ \mathbb{E}_{x \sim \mu}[f(x)] - \mathbb{E}_{x \sim \nu}[f(x)] \right\},$$

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

Note that when $\alpha > d/2$, the Sobolev IPMs are MMDs, since in that case, $\mathcal{H}^\alpha(\mathcal{M})$ is an RKHS (Sobolev Inequality). For simplicity, we assume that the PDS kernel K is diagonalizable in the Laplace-Beltrami basis:

$$K(x, y) = \sum_{\ell=0}^\infty \xi_\ell \phi_\ell(x) \phi_\ell(y), \tag{10}$$

for some $\xi_\ell \in \mathbb{R}$. This holds, for example, for the dot-product kernels on the sphere $\mathbb{S}^{d-1}$ as well as all Sobolev spaces $\mathcal{H}^\alpha(\mathcal{M})$ with $\alpha > d/2$.

To study the gain of invariances for estimating MMDs, we need to define a quantity that measures how many eigen-functions can take non-zero coefficients when restricted to the G-invariant functions. This weighted quantity must also depend on the kernel decomposition in the Laplace-Beltrami operator. To this end, define

$$\text{Tr}(K; G) := \sum_{\lambda \neq 0}^\infty m(\lambda; G) |\xi_\ell(\lambda)|, \tag{11}$$

where the sum is over all the eigenvalues of the Laplace-Beltrami operator on the manifold $(\mathcal{M}, g)$, and $m(\lambda; G)$ denotes the multiplicity of the eigenvalue λ when restricted to invariant eigenfunctions (see Appendix A for more details).

The following theorem related the quantity $\text{Tr}(K; G)$ to the sample complexity gain of invariances for the MMD estimation problem.

Theorem 4. *Consider a G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$ with $\frac{d\mu}{dx} \in \mathcal{H}$. Then, there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$ such that*

$$\mathbb{E}[D_\mathcal{H}(\mu, \hat{\mu})] \leq \sqrt{\frac{\text{Tr}(K; G)}{n}}. \tag{12}$$

This allows us to immediately conclude the following result for the Sobolev IPMs with $\alpha > d/2$ (i.e., the MMD regime).

Corollary 1 (Gain of invariances for the Sobolev IPMs in the MMD regime). *For any $\alpha > d/2$, there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$ such that*

$$\mathbb{E}[D_\alpha(\mu, \tilde{\mu})] \leq \sqrt{\frac{\mathcal{Z}(\alpha; G)}{n}},$$

where $\mathcal{Z}(\alpha; G)$ is the zeta function associated with the Lie group action G on the manifold $\mathcal{M}$:

$$\mathcal{Z}(\alpha; G) := \sum_{\lambda \neq 0} m(\lambda; G)\lambda^{-\alpha}. \quad (13)$$

Note that if $\alpha > d/2$ (i.e., the MMD regime), the exponent is always $\frac{1}{2}$ and it does not depend on the group of invariances. Moreover, the gain of invariances in this case can intuitively be understood as having *effectively*

$$n \times \frac{\mathcal{Z}(\alpha)}{\mathcal{Z}(\alpha; G)} \quad (14)$$

samples instead of n samples, where $\mathcal{Z}(\alpha)$ is the original zeta function of the manifold (achieved via the trivial group $G = \{\mathrm{id}_G\}$). Therefore, the quantity $\mathcal{Z}(\alpha)/\mathcal{Z}(\alpha; G)$ represents the sample complexity gain of invariances in the kernel regime.

We provide an example to see how the behavior of the quantity $\mathcal{Z}(\alpha)/\mathcal{Z}(\alpha; G)$ can depend on more delicate information about the group action (i.e., more than the dimension of the group and the volume of the quotient space).

Example 2. *Consider the unit circle $\mathbb{S}^1$ with the round metric. Note that if we parameterize functions defined on $\mathbb{S}^1$ as functions of the angle $\theta \in [0, 2\pi)$, then the eigenfunctions of the Laplace-Beltrami operator on $\mathbb{S}^1$ are the sinusoidal functions:*

$$\phi_k(\theta) = \exp(ik\theta) \quad \text{with} \quad \lambda = k^2; \quad k \in \mathbb{Z}. \quad (15)$$

This allows us to compute the zeta function of the unit circle as

$$\mathcal{Z}(\alpha) = \sum_{\lambda \neq 0} m(\lambda)\lambda^{-\alpha} = 2\sum_{k=1}^\infty k^{-2\alpha} = 2\zeta(2\alpha), \quad (16)$$

where ζ denotes the Riemann zeta function.

Now consider the finite group of rotations around the center: $G := \left\{ \frac{2\pi j}{|G|} : j \in \mathbb{Z} \right\}$. An eigenfunction ϕ_k is invariant with respect to the action of G if and only if we have

$$\phi_k(\theta) = \phi_k\left(\theta + \frac{2\pi j}{|G|}\right), \quad (17)$$

for any $\theta \in [0, 2\pi)$ and $j \in \mathbb{Z}$. This is equivalent to having k divide $|G|$. Thus, we get

$$\mathcal{Z}(\alpha; G) = \sum_{\lambda \neq 0}^\infty m(\lambda; G)\lambda^{-\alpha} = 2|G|^{-2\alpha}\sum_{k=1}^\infty k^{-2\alpha} \quad (18)$$

$$= 2|G|^{-2\alpha}\zeta(2\alpha). \quad (19)$$

This shows that the gain of invariances in sample complexity, in this case, is to have effectively

$$n \times |G|^{2\alpha} \quad (20)$$

many samples instead of n samples. Note that $\alpha > d/2$ in the MMD regime. Therefore, this example shows how the multiplicative gain can be improved from $|G|$ to $|G|^{2\alpha}$ in the MMD regime, which is in contrast to the problem of estimating Sobolev IPMs in the non-MMD regime (Theorem 3).

Example 3. *Let $\mathbb{T}^d = [0, 1]^d$ denote the d-dimensional flat torus, and consider the isometric action of the group $G = \mathbb{T}^{\dim(G)}$ with the circular shift on the first $\dim(G)$ coordinates of $\mathbb{T}^d$; for any $\tau \in G$ and $x \in \mathbb{T}^d$:*

$$\tau x := \left(\tau_1 + x_1, \tau_2 + x_2, \ldots, \tau_{\dim(G)} + x_{\dim(G)}, \quad (21)\right.$$

$$\left. x_{\dim(G)+1}, \ldots, x_d\right), \quad (22)$$

where the coordinates are summed modulo one. Consider the heat kernel on $\mathbb{T}^d$, defined as:

$$K_\beta(x, y) = \sum_{0 \neq v \in \mathbb{Z}^d} \exp(-\beta\|v\|_2^2 + 2\pi i\langle v, x+y\rangle).$$

Note that

$$\mathrm{Tr}(K_\beta; G) = \sum_{0 \neq v \in \mathbb{Z}^{d-\dim(G)}} \exp(-\beta\|v\|_2^2) \quad (23)$$

$$= \Theta_\beta^{(d-\dim(G))} - 1, \quad (24)$$

where

$$\Theta_\beta := \sum_{v \in \mathbb{Z}} \exp(-\beta v^2) > 1. \quad (25)$$

This shows that the gain of having invariances with respect to the group G for estimating the MMD with the heat kernel is to have effectively

$$n \times \frac{\mathrm{Tr}(K; \{\mathrm{id}_G\})}{\mathrm{Tr}(K; G)} = n \times \frac{\Theta_\beta^d - 1}{\Theta_\beta^{d-\dim(G)} - 1}$$

$$\approx n \times \Theta_\beta^{\dim(G)},$$

many samples, instead of n. This shows that the gain of invariances can exponentially depend on the Lie group dimension $\dim(G)$ while estimating MMDs.

Proof sketch for Theorem 4. The spectrum of the test functions $f \in \mathcal{H}$, in the Laplace-Beltrami basis, decays quickly in the MMD case, and it can be shown that we can take $T \to \infty$, and use a similar approach to Theorem 3 to prove the convergence result.

4.5. Density Estimation in $L^2(\mathcal{M})$

In this section, we present convergence results for estimating probability density functions $\frac{d\mu}{dx}$ in $L^2(\mathcal{M})$ distance while having n i.i.d. samples from $\mu \in \mathcal{P}(\mathcal{M})$.

Theorem 5 (Convergence rate for density estimation in $L^2(\mathcal{M})$ distance under invariances). *Consider a G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$ with $\frac{d\mu}{dx} \in \mathcal{H}^s(\mathcal{M})$. Then, there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$ such that*

$$\mathbb{E}\Big[\Big\|\frac{d\tilde{\mu}}{dx} - \frac{d\mu}{dx}\Big\|_{L^2(\mathcal{M})}\Big] \lesssim \Big(\frac{\mathrm{vol}(\mathcal{M}/G)}{n}\Big)^{\frac{s}{2s+d}} \Big\|\frac{d\mu}{dx}\Big\|_{L^2(\mathcal{M})}^{\frac{d}{2s+d}},$$

where the constant only depends on the manifold and $\mathrm{vol}(\mathcal{M}/G)$ denotes the volume of the quotient space $\mathcal{M}/G$ with $d := \dim(\mathcal{M}/G) \geq 3$.

We observe that the density estimation problem in the $L^2(\mathcal{M})$ distance exploits the same two-fold gain as previous cases.

4.6. Density Estimation in $L^\infty(\mathcal{M})$

In this part, we study the gain of group invariances for the density estimation problem in the $L^\infty(\mathcal{M})$ distance.

Theorem 6 (Convergence rate for density estimation in $L^\infty(\mathcal{M})$ under invariances). *Consider a G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$ with $\frac{d\mu}{dx} \in \mathcal{H}^s(\mathcal{M})$ for some $s > d/2$. Then, there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$, such that*

$$\mathbb{E}\Big[\Big\|\frac{d\tilde{\mu}}{dx} - \frac{d\mu}{dx}\Big\|_{L^\infty(\mathcal{M})}\Big]$$
$$\lesssim \Big(\frac{(\mathrm{vol}(\mathcal{M}/G))^{\frac{2s}{s-d/2}}}{n}\Big)^{\frac{s-d/2}{2s+d}} \Big\|\frac{d\mu}{dx}\Big\|_{\mathcal{H}^s(\mathcal{M})}^{\frac{d}{s+d/2}},$$

where the constant only depends on the manifold.

Note that here we only get a convergence result if $s > d/2$; otherwise, the density $\frac{d\mu}{dx}$ is not necessarily bounded, since $\mathcal{H}^s(\mathcal{M})$ contains unbounded functions in that case (Sobolev inequality).

Remark 2. *Note that in the density estimation in $L^2(\mathcal{M})$ distance, for finite groups, the effective number of samples is $n \times |G|$, as intuitively expected (for all s, d). However, for the $L^\infty(\mathcal{M})$ distance, better rates are achievable according to Theorem 6. For instance, let us take $d = 4$ and $s = 6$; then, for finite groups, the convergence rate becomes*

$$\mathbb{E}\Big[\Big\|\frac{d\tilde{\mu}}{dx} - \frac{d\mu}{dx}\Big\|_{L^\infty(\mathcal{M})}\Big] \lesssim \Big(\frac{1}{n|G|^3}\Big)^{1/4}. \tag{26}$$

This shows that the effective number of samples can potentially be more than the linear gain, i.e., in this example, we have an effective number of samples $n \times |G|^3$.
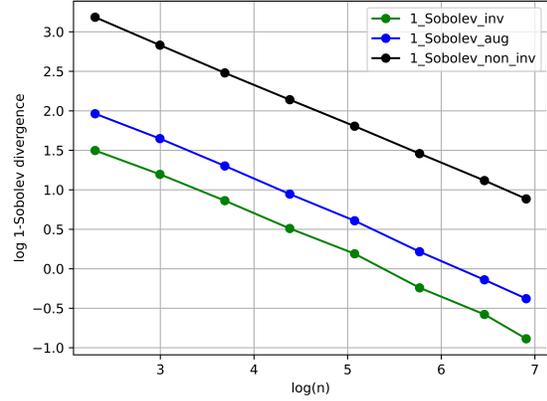

Figure 1. The sample complexity gain of estimation the Sobolev IPM with $\alpha = 1$ for invariant distributions.

We also notice that in this paper we used a specific simple algorithm to achieve the bound for the density estimation in $L^\infty(\mathcal{M})$, and it can be improved when there is no invariances (Uppal et al., 2019). We leave the problem of proving tighter bounds for this case under invariances to future works.

5. Experiments

To observe the gain of invariances, we conduct a simple experiment on the following synthetic dataset. The input space is the six-dimensional flat torus $\mathbb{T}^6 = [0, 1]^6$, and the group of invariances acts as the circular shifts on the last two coordinates; see Example 3 for more details about this action. We consider a non-invariant distribution μ_{non} and an invariant distribution μ_{inv} in this setting as follows. Samples from the non-invariant distribution are generated based on the sum of three i.i.d. random variables $X = X_1 + X_2 + X_3$, each chosen uniformly from $[0, 1/3]^6$. Moreover, samples from the invariant distribution such as (X, X_5, X_6) are generated based on the sum of four i.i.d. random variables, denoted by $X = X_1 + X_2 + X_3 + X_4$, each chosen uniformly from $[0, 0.25]^4$, and for the last two coordinates, X_5, X_6 are chosen uniformly from $[0, 1]$. With this specific choice, one can show that the distributions of μ_{non} and μ_{inv} both lie in the Sobolev space with $s = 6$.

The convergence rates for estimating the Sobolev IPM with $\alpha = 1$ are shown in Figure 5, where we used the proposed estimators in the paper to achieve the results. Indeed, we didn't optimize the parameter T, and just used a fixed regularization parameter, which leads to having parallel plots in the logarithmic scale. Still, the gain of invariances is evident. We also report the convergence rates for the full data augmentation setting in the plot. Note that even though the group of invariances here has infinitely many elements,

8

we can report the full data augmentation results using the Fourier approach. The results show that full data augmentation is not enough to achieve the performance of the invariant estimator that is used in the paper. This shows the crucial role of regularization (i.e., balancing the bias and variance terms) in the problem.

6. Conclusion

In this paper, we studied the sample complexity gain of group invariances for estimating various divergence measures on manifolds, given a number of i.i.d. samples. We extended the recent theoretical results for finite group actions on submanifolds (of full dimension) of $\mathbb{R}^d$ to arbitrary manifolds (including spheres, tori, hyperbolic spaces, etc.), and arbitrary groups (including groups of positive dimension). The gain of invariances in sample complexity is two-fold: improving the exponent and multiplying the number of samples by a factor depending on the group of invariances. We also studied faster convergence rates for smooth distributions and showed a similar gain in that case.

Acknowledgments

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

Anderson, B., Hy, T.-S., and Kondor, R. Cormorant: Covariant molecular neural networks. *Advances in Neural Information Processing Systems*, 2019.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Bietti, A., Venturi, L., and Bruna, J. On the sample complexity of learning under geometric stability. *Advances in Neural Information Processing Systems*, 34:18673–18684, 2021.

Biloš, M. and Günnemann, S. Scalable normalizing flows for permutation invariant densities. In *International Conference on Machine Learning*, pp. 957–967. PMLR, 2021.

Birrell, J., Katsoulakis, M. A., Rey-Bellet, L., and Zhu, W. Structure-preserving gans. *arXiv preprint arXiv:2202.01129*, 2022.

Block, A., Jia, Z., Polyanskiy, Y., and Rakhlin, A. Intrinsic dimension estimation using wasserstein distance. *The Journal of Machine Learning Research*, 23(1):14124–14160, 2022.

Boissard, E. and Le Gouic, T. On the mean speed of convergence of empirical and occupation measures in wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pp. 539–563, 2014.

Bunne, C. Optimal transport in learning, control, and dynamical systems. *ICML Tutorial*, 2023.

Butkovsky, O. Subgeometric rates of convergence of markov processes in the wasserstein metric. *Annals of Applied Probability*, 2014.

Canas, G. and Rosasco, L. Learning probability measures with respect to optimal transport metrics. *Advances in Neural Information Processing Systems*, 25, 2012.

Chen, Z., Katsoulakis, M. A., Rey-Bellet, L., and Zhu, W. Sample complexity of probability divergences under group symmetry. *arXiv preprint arXiv:2302.01915*, 2023.

Chuang, C.-Y., Mroueh, Y., Greenewald, K., Torralba, A., and Jegelka, S. Measuring generalization with optimal transport. *Advances in Neural Information Processing Systems*, 34:8294–8306, 2021.

Courty, N., Flamary, R., and Tuia, D. 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*, pp. 274–289. Springer, 2014.

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

Dey, N., Chen, A., and Ghafurian, S. Group equivariant generative adversarial networks. *arXiv preprint arXiv:2005.01683*, 2020.

Divol, V. Measure estimation on manifolds: an optimal transport approach. *Probability Theory and Related Fields*, 183(1-2):581–647, 2022.

Evans, L. C. and Gariepy, R. F. *Measure theory and fine properties of functions*. CRC press, 2015.

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

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.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. *Advances in neural information processing systems*, 20, 2007.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1574–1583. PMLR, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Hendriks, H. Nonparametric estimation of a probability density on a Riemannian manifold using fourier expansions. *The Annals of Statistics*, pp. 832–849, 1990.

Kantorovich, L. V. On the translocation of masses. *Journal of mathematical sciences*, 133(4):1381–1382, 2006.

Köhler, J., Klein, L., and Noé, F. Equivariant flows: sampling configurations for multi-body systems with symmetric energies. *arXiv preprint arXiv:1910.00753*, 2019.

Lei, J. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 2020.

Liang, T. Estimating certain integral probability metric (ipm) is as hard as estimating under the ipm. *arXiv preprint arXiv:1911.00730*, 2019.

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

Mallasto, A., Frellsen, J., Boomsma, W., and Feragen, A. (q, p)-wasserstein gans: Comparing ground metrics for wasserstein gans. *arXiv preprint arXiv:1902.03642*, 2019.

Mena, G. and Niles-Weed, J. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in neural information processing systems*, 32, 2019.

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., and Cheng, Y. Sobolev gan. *arXiv preprint arXiv:1711.04894*, 2017.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29 (2):429–443, 1997.

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

Panaretos, V. M. and Zemel, Y. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Rigollet, P. and Stromme, A. J. On the sample complexity of entropic optimal transport. *arXiv preprint arXiv:2206.13472*, 2022.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Singh, S. and Póczos, B. Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M., and Póczos, B. Nonparametric density estimation under adversarial losses. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pp. 13–31. Springer, 2007.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 (7), 2011.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. On the empirical estimation of integral probability metrics. 2012.

Tahmasebi, B. and Jegelka, S. The exact sample complexity gain from invariances for kernel regression on manifolds. *arXiv preprint arXiv:2303.14269*, 2023.

Tolstikhin, I., Sriperumbudur, B. K., and Muandet, K. Minimax estimation of kernel mean embeddings. *The Journal of Machine Learning Research*, 18(1):3002–3048, 2017.

Tolstikhin, I. O., Sriperumbudur, B. K., and Schölkopf, B. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems*, 29, 2016.

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

Uppal, A., Singh, S., and Póczos, B. Nonparametric density estimation & convergence rates for gans under besov ipm losses. In *Advances in neural information processing systems*, volume 32, 2019.

Villani, C. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 2019.

Zhang, J., Zhong, W., and Ma, P. A review on modern computational optimal transport methods with applications in biomedical research. *Modern Statistical Methods for Health Research*, pp. 279–300, 2021.

A. Preliminaries for Proofs

This section reviews some essential preliminaries on group actions on manifolds (Tahmasebi & Jegelka, 2023). We consider a compact connected boundaryless smooth Riemannian manifold $(\mathcal{M}, g)$ with dimension $\dim(\mathcal{M})$. The Riemannian metric g allows us to define the volume element $d\operatorname{vol}_g(x)$ on $\mathcal{M}$, and without loss of generality, we assume that it is normalized such that $\operatorname{vol}(\mathcal{M}) = \int_{\mathcal{M}} d\operatorname{vol}_g(x) = 1$. The volume element is also denoted by dx whenever the choice of the Riemannian metric is evident from the context.

The Laplace-Beltrami operator Δ_g is the unique continuous operator satisfying the integration by parts formula:

$$\int_{\mathcal{M}} \langle \nabla_g \phi(x), \nabla_g \psi(x) \rangle_{L^2(\mathcal{M})} d\operatorname{vol}_g(x) = \int_{\mathcal{M}} \Delta_g \phi(x) \psi(x) d\operatorname{vol}_g(x), \tag{27}$$

for any smooth function $\phi, \psi : \mathcal{M} \to \mathbb{R}$. One can see that the Euclidean Laplacian $\Delta = \sum_i \partial_i^2$ satisfies this definition by Green's identities.

The Laplace-Beltrami operator eigenfunctions, denoted by ϕ_ℓ, $\ell = 0, 1, \ldots$, are the the sequence of functions satisfying the equation $\Delta_g \phi + \lambda_\ell \phi = 0$ on the manifold $\mathcal{M}$, where $0 = \lambda_0 < \lambda_1 \le \lambda_2 \le \ldots$ denote the eigenvalues of $(-\Delta_g)$. Let us assume that $\|\phi_\ell\|_{L^2(\mathcal{M})} = \int_{\mathcal{M}} \phi_\ell^2(x) d\operatorname{vol}_g(x) = 1$, i.e., the eigenfunctions are normalized in $L^2(\mathcal{M})$. One important property of the Laplace-Beltrami eigenfunctions it that they constitute a basis for the function space $L^2(\mathcal{M})$.

It can be shown that the sequence $\lambda_\ell \to \infty$ as $\ell \to \infty$, thus it does not accumulate around any finite number. Moreover, according to the celebrated Weyl's law, the density of eigenvalues is given by the following asymptotic formula:

$$N(\lambda) := \#\{\lambda_\ell \le \lambda\} = \frac{\omega_{\dim(\mathcal{M})}}{(2\pi)^{\dim(\mathcal{M})}} \operatorname{vol}(\mathcal{M}) \lambda^{\dim(\mathcal{M})/2} + O(\lambda^{(\dim(\mathcal{M})-1)/2}), \tag{28}$$

where $\omega_{\dim(\mathcal{M})}$ is the volume of the unit $\dim(\mathcal{M})$-dimensional ball in the Euclidean space $\mathbb{R}^{\dim(\mathcal{M})}$ (with the usual metric). Even more, one can show that Weyl's law holds locally around each point on the manifold:

$$N_x(\lambda) := \sum_{\lambda_\ell \le \lambda} \phi_\ell^2(x) = \frac{\omega_{\dim(\mathcal{M})}}{(2\pi)^{\dim(\mathcal{M})}} \operatorname{vol}(\mathcal{M}) \lambda^{\dim(\mathcal{M})/2} + O(\lambda^{(\dim(\mathcal{M})-1)/2}), \tag{29}$$

where the error term is uniformly bounded for all $x \in \mathcal{M}$. Note that integrating the above formula on the manifolds gives the original version of Weyl's law.

A Lie group G acts on $\mathcal{M}$ isometrically if for all $\tau \in G$, the corresponding map $x \mapsto \tau x$ on the manifold is an isometry. A function $f : \mathcal{M} \to \mathcal{M}$ is called $G-$invariant, if and only if one has $f(x) = f(\tau x)$, for all $\tau \in G$, in $L^2(\mathcal{M})$.

For any function $f \in L^2(\mathcal{M})$, one has $f = \sum_{\ell=0}^{\infty} \langle f, \phi_\ell \rangle_{L^2(\mathcal{M})} \phi_\ell$. However, for any G-invariant function, one can show that many coefficients in the above series are a priory known to be zero. Indeed, for any eigenspace of the Laplace-Beltrami operator, such as V_λ, with eigenvalue λ, there exists a subspace $V_{\lambda,G} \subseteq V_\lambda$ such that all G-invariant functions $f \in V_\lambda$ can be written as a linear combination of functions in $V_{\lambda,G}$ (Tahmasebi & Jegelka, 2023). Without loss of generality, we assume that the eigenfunctions ϕ_ℓ are chosen such that for any λ there exists a subset of $\{\phi_\ell : \ell = 0, 1, \ldots\}$ that forms an orthonormal basis for $V_{\lambda,G}$. Let $\mathbb{1}_G(\ell)$ indicate whether $\phi_\ell \in V_{\lambda,G}$ (with one) or not (with zero). Let $N_x(\lambda; G) := \sum_{\ell:\lambda_\ell \le \lambda} \mathbb{1}_G(\ell) \phi_\ell^2(x)$.

We are interested in asymptotic estimation formulae for $N_x(\lambda; G)$ to study the gain of invariances. Note that if $G = \{\operatorname{id}_G\}$, then Weyl's law provides such estimation. The following theorem shows how to extend it to non-trivial groups.

Theorem 7 (Local Weyl's law for G-invariant functions, (Tahmasebi & Jegelka, 2023)). *For any Lie group G acting smoothly on a connected compact boundaryless manifold $\mathcal{M}$, one has for all $x \in \mathcal{M}$,*

$$N_x(\lambda; G) = \sum_{\ell:\lambda_\ell \le \lambda} \mathbb{1}_G(\ell) \phi_\ell^2(x) = \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G) \lambda^{d/2} + O(\lambda^{(d-1)/2}), \tag{30}$$

where $d := \dim(\mathcal{M}/G)$ and $\operatorname{vol}(\mathcal{M}/G)$ denote the dimension and the volume of the principal part of the quotient space, respectively. Note that one always has $N_x(\lambda; G) \le N_x(\lambda)$.

A Hilbert space $\mathcal{H} \subseteq L^2(\mathcal{M})$ is called a Reproducing Kernel Hilbert Spaces (RKHS) if and only if for any $x \in \mathcal{M}$ and $f \in \mathcal{H}$, one has $|f(x)| \le C_x \|f\|_{\mathcal{H}}$ for some constant $C_x < \infty$. Associated with $\mathcal{H}$, there exists a PDS kernel

$K : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$, and according to Mercer's theorem, it can be diagonalized in some basis in $L^2(\mathcal{M})$. In this paper, for the sake of simplicity, we consider kernels that can be diagonalized in the Laplace-Beltrami basis as follows:

$$K(x, y) = \sum_{\ell=0}^{\infty} \xi_\ell \phi_\ell(x) \phi_\ell(y), \tag{31}$$

for some coefficients $\xi_\ell \in \mathbb{R}$. This includes dot-product kernels over spheres. We further assume that the coefficients, for each ℓ, only depend on λ_ℓ, thus we can denote them as $\xi(\lambda)$, too. A set of sufficient conditions for kernels on manifolds to satisfy such types of diagonalization is given in (Tahmasebi & Jegelka, 2023).

We also use the following definition of Sobolev spaces $\mathcal{H}^s(\mathcal{M})$ for any $s \geq 0$:

$$\mathcal{H}^s(\mathcal{M}) := \Big\{ f \in L^2(\mathcal{M}) : \|f\|_{\mathcal{H}^s(\mathcal{M})}^2 := \langle f, \phi_0 \rangle_{L^2(\mathcal{M})}^2 + \sum_{\ell=1}^{\infty} \lambda_\ell^s \langle f, \phi_\ell \rangle_{L^2(\mathcal{M})}^2 < \infty \Big\}. \tag{32}$$

Sobolev spaces are indeed Hilbert spaces under the following inner product:

$$\langle f_1, f_2 \rangle_{\mathcal{H}^s(\mathcal{M})}^2 := \langle f_1, \phi_0 \rangle_{L^2(\mathcal{M})} \langle f_2, \phi_0 \rangle_{L^2(\mathcal{M})} + \sum_{\ell=1}^{\infty} \lambda_\ell^s \langle f_1, \phi_\ell \rangle_{L^2(\mathcal{M})} \langle f_2, \phi_\ell \rangle_{L^2(\mathcal{M})}. \tag{33}$$

Note that Sobolev spaces are Reproducing Kernel Hilbert Spaces if and only if $s > d/2$ (Tahmasebi & Jegelka, 2023). For any $0 \leq \alpha \leq \beta < \infty$, one has $\mathcal{H}^\beta(\mathcal{M}) \subseteq \mathcal{H}^\alpha(\mathcal{M}) \subseteq L^2(\mathcal{M})$. Moreover, $H^0(\mathcal{M}) = L^2(\mathcal{M})$.

B. Proof of Theorem 1

Theorem 1 (Convergence rate of the 1-Wasserstein distance under invariances). *For any G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$, there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$, as a function of n i.i.d. samples $X_1, X_2 \ldots, X_n \sim \mu$, such that*

$$\mathbb{E}[W_1(\mu, \tilde{\mu})] \lesssim \Big(\frac{\text{vol}(\mathcal{M}/G)}{n} \Big)^{\frac{1}{d}}, \tag{4}$$

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

$$\mathbb{E}[|W_1(\tilde{\mu}, \tilde{\nu}) - W_1(\mu, \nu)|] \lesssim \Big(\frac{\text{vol}(\mathcal{M}/G)}{n} \Big)^{\frac{1}{d}}, \tag{5}$$

for any G-invariant probability measures $\mu, \nu \in \mathcal{P}(\mathcal{M})$.

Proof. Let ϕ_ℓ, $\ell = 0, 1, \ldots$, denote the eigenfunctions of the Laplace-Beltrami operator, corresponding to the eigenvalues $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \ldots$. Each eigenspace is denoted by V_{λ_ℓ}, and its subspace of G-invariant eigenfunctions is denoted by $V_{\lambda_\ell, G}$ (see Appendix A for more details). For any probability measure $\mu \in \mathcal{P}(\mathcal{M})$, define $\mu_\ell = \int_{\mathcal{M}} \phi_\ell d\mu$ for each $\ell = 0, 1, \ldots$, and note that $\mu_0 = 1$. Let $dx = \frac{1}{\text{vol}(\mathcal{M})} d \, \text{vol}_g(x)$ denote the uniform measure on the manifold and define the inner-product in $L^2(\mathcal{M})$ with respect to the measure $d \, \text{vol}_g(x)$. If a probability measure μ is absolutely continuous with respect to the uniform measure, then $\frac{d\mu}{dx}$ denotes its Radon-Nikodym derivative.

Lemma 1. *Consider two Borel measures $\mu, \nu \in \mathcal{P}(\mathcal{M})$ with $\int_{\mathcal{M}} d\mu = \int_{\mathcal{M}} d\nu = 1$. Also, assume that $\frac{d\mu}{dx}, \frac{d\nu}{dx} \in L^2(\mathcal{M})$ with $\text{vol}(\mathcal{M}) = 1$. Then,*

$$W_1(\mu, \nu) \leq D_1(\mu, \nu), \tag{34}$$

where $D_1(\mu, \nu)$ is the Sobolev IPM with parameter $\alpha = 1$.

Proof. Note that $\text{Lip}(\mathcal{M}) \subset L^2(\mathcal{M})$. By Rademacher's theorem (Evans & Gariepy, 2015), any $f \in \text{Lip}(\mathcal{M})$ is differentiable almost everywhere, and thus $|\nabla_g f(x)|_g \leq 1$ for almost every $x \in \mathcal{M}$. Therefore,

$$\|\nabla_g f\|_{L^2(\mathcal{M})}^2 \leq \int_{\mathcal{M}} |\nabla_g f(x)|_g^2 d\text{vol}_g(x) \leq \int_{\mathcal{M}} d\text{vol}_g(x) = 1. \tag{35}$$

13

For any $f \in L^2(\mathcal{M})$, one can write $f = \sum_{\ell=0}^{\infty} f_\ell \phi_\ell$ with $f_\ell := \langle f, \phi_\ell \rangle_{L^2(\mathcal{M})}$. Thus, we have

$$W_1(\mu, \nu) = \sup_{f \in \text{Lip}(\mathcal{M})} \left\{ \int_{\mathcal{M}} f d\mu - \int_{\mathcal{M}} f d\nu \right\} \tag{36}$$

$$\leq \sup_{\substack{f \in L^2(\mathcal{M}) \\ \|\nabla_g f\|_{L^2(\mathcal{M})}^2 \leq 1}} \left\{ \int_{\mathcal{M}} f d\mu - \int_{\mathcal{M}} f d\nu \right\}. \tag{37}$$

Now note that since $\mu_0 = \nu_0$, one can restrict the above optimization problem to functions with $f_0 = 0$. For those functions, we have $\|f\|_{\mathcal{H}^1(\mathcal{M})} = \|\nabla_g f\|_{L^2(\mathcal{M})} = 1$, and thus we have

$$W_1(\mu, \nu) \leq \sup_{\substack{f \in \mathcal{H}^1(\mathcal{M}) \\ \|f\|_{\mathcal{H}^1(\mathcal{M})} \leq 1}} \left\{ \int_{\mathcal{M}} f d\mu - \int_{\mathcal{M}} f d\nu \right\} = D_1(\mu, \nu). \tag{38}$$

$\square$

To prove Theorem 1, we use Lemma 1. First, note that it is impossible to immediately apply Lemma 1 since the unknown measure μ is not necessarily absolutely continuous with respect to the uniform measure on the manifold. Nevertheless, let us define the sequence

$$\tilde{\mu}_\ell := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_G(\ell) \phi_\ell(X_i), \tag{39}$$

for each $\ell = 0, 1, \ldots$.

Given the probability measure $\mu \in \mathcal{P}(\mathcal{M})$, define the twisted measure $\mu_\star$ as follows:

$$\frac{d\mu_\star}{dx} = \sum_{\ell=0}^{\infty} \exp(-\sigma \lambda_\ell) \mu_\ell \phi_\ell, \tag{40}$$

where σ is a fixed positive constant (to be set later). Note that for a fixed σ, the new measure $\mu_\star$ is absolutely continuous with respect to the uniform measure. Indeed, its density is even smooth (i.e., having smooth derivatives with arbitrary orders), since the tail of the above summation goes to zero exponentially fast as $\ell \to \infty$. Also, note that $\int_{\mathcal{M}} d\mu_\star = 1$, and with a bit more consideration, we can observe that $\mu_\star$ is a probability measure. Let us also define $\tilde{\mu}_\star$ similarly, where μ_ℓ is replaced with its empirical estimation as above. Thus, the measure $\tilde{\mu}_\star$ is computable from the given samples.

There is also another interpretation for the probability measure $\mu_\star$. Let $X \sim \mu$ denote a sample from μ, and consider the heat diffusion corresponding to the Brownian motion on the manifold, started from X and stopped at time $t = \sigma$. The Brownian motion is assumed to be independent of the sample X, and it can be shown that the law of the process at time $t = \sigma$ is the probability measure $\mu_\star$. Therefore, this gives a natural coupling between the two measures μ and $\mu_\star$.

Hence, we can write

$$W_1(\mu, \tilde{\mu}_\star) \leq W_1(\mu, \mu_\star) + W_1(\mu_\star, \tilde{\mu}_\star) \tag{41}$$

$$\leq C\sqrt{\sigma} + W_1(\mu_\star, \tilde{\mu}_\star), \tag{42}$$

where we use the fact that for any probability measure $\nu \in \mathcal{P}(\mathcal{M})$, one has $W_1(\nu, \nu_\star) \leq W_2(\nu, \nu_\star) \leq C\sqrt{\sigma}$ for some constant C depending on the manifold. The last inequality is due to the fact that the law of the Brownian motion on the manifold at time $t = \sigma$ behaves (locally) similarly to the Gaussian distribution on Euclidean spaces.

Let us now bound $\mathbb{E}[W_1(\mu_\star, \tilde{\mu}_\star)]$. Since the two measures $\mu_\star, \tilde{\mu}_\star$ are absolutely continuous with respect to the uniform measure on the manifold, we can use Lemma 1 and Lemma 2; thus,

$$\mathbb{E}[W_1(\mu_\star, \tilde{\mu}_\star)] \leq \sqrt{\mathbb{E}\left[\sum_{\ell=1}^{\infty} \exp(-2\sigma \lambda_\ell) \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell}\right]}. \tag{43}$$

Let us define

$$R(\lambda) := \mathbb{E}\Big[\sum_{\ell:\lambda_\ell \leq \lambda} (\mu_\ell - \tilde{\mu}_\ell)^2 \Big], \tag{44}$$

and note that we can rewrite Equation (43) as follows:

$$\mathbb{E}[W_1(\mu_\star, \tilde{\mu}_\star)] \leq \sqrt{\int_{0^+}^{\infty} \exp(-2\sigma\lambda)\frac{dR(\lambda)}{\lambda}}, \tag{45}$$

where the above formula must be understood as a Riemann-Stieltjes integral.

Using integration by parts and Lemma 3, we have

$$\int_{0^+}^{\infty} \exp(-2\sigma\lambda)\frac{dR(\lambda)}{\lambda} = \underbrace{\exp(-2\sigma\lambda)\frac{R(\lambda)}{\lambda}\Big|_{0^+}^{\infty}}_{=0} + \int_{0^+}^{\infty} R(\lambda)\frac{2\sigma\lambda+1}{\lambda^2}\exp(-2\sigma\lambda)d\lambda. \tag{46}$$

Now we use Lemma 3 and Theorem 7 to find an upper bound on the above integral:

$$\int_{0^+}^{\infty} R(\lambda)\frac{2\sigma\lambda+1}{\lambda^2}\exp(-2\sigma\lambda)d\lambda \lesssim \frac{1}{n}\int_{0^+}^{\infty} \frac{\omega_d}{(2\pi)^d}\mathrm{vol}(\mathcal{M}/G)\lambda^{d/2}\frac{2\sigma\lambda+1}{\lambda^2}\exp(-2\sigma\lambda)d\lambda \tag{47}$$

$$= \frac{1}{n}\frac{\omega_d}{(2\pi)^d}\mathrm{vol}(\mathcal{M}/G)\int_{0^+}^{\infty} (2\sigma\lambda^{d/2-1} + \lambda^{d/2-2})\exp(-2\sigma\lambda)d\lambda. \tag{48}$$

By a change of variables in the above integral, we conclude that

$$\int_{0^+}^{\infty} R(\lambda)\frac{2\sigma\lambda+1}{\lambda^2}\exp(-2\sigma\lambda)d\lambda \lesssim \frac{\mathrm{vol}(\mathcal{M}/G)}{n}\sigma^{1-d/2}. \tag{49}$$

Therefore,

$$\mathbb{E}[W_1(\mu, \tilde{\mu}_\star)] \lesssim \sqrt{\sigma} + \sqrt{\frac{\mathrm{vol}(\mathcal{M}/G)}{n}}\sigma^{(2-d)/4}. \tag{50}$$

To minimize the above upper bound, we consider the function $p(\sigma) = \sqrt{\sigma} + \sqrt{\frac{\mathrm{vol}(\mathcal{M}/G)}{n}}\sigma^{(2-d)/4}$ for $\sigma \in (0,\infty)$ that achieves its minimum at

$$\sigma = \sigma_{\mathrm{opt}} = \Big((1-d/2)^2\frac{\mathrm{vol}(\mathcal{M}/G)}{n}\Big)^{2/d}, \tag{51}$$

and thus we have

$$\mathbb{E}[W_1(\mu, \tilde{\mu}_\star)] \lesssim \Big(\frac{\mathrm{vol}(\mathcal{M}/G)}{n}\Big)^{1/d}, \tag{52}$$

which completes the proof. $\qquad\square$

C. Proof of Theorem 2

Theorem 2 (Convergence rate of the 1-Wasserstein distance for smooth distributions under invariances). *For any G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$ with $\frac{d\mu}{dx} \in \mathcal{H}^s(\mathcal{M})$ 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 $X_1, X_2 \ldots, X_n \sim \mu$, such that*

$$\mathbb{E}[W_1(\mu,\tilde{\mu})] \lesssim \Big(\frac{\mathrm{vol}(\mathcal{M}/G)}{n}\Big)^{\frac{s+1}{2s+d}} \Big\|\frac{d\mu}{dx}\Big\|_{\mathcal{H}^s(\mathcal{M})}^{\frac{d-2}{2s+d}},$$

where $\mathrm{vol}(\mathcal{M}/G)$ is the volume of the quotient space $\mathcal{M}/G$ and $d := \dim(\mathcal{M}/G) \geq 3$.

Proof. According to Lemma 1, we have the upper bound

$$W_1(\mu, \nu) \leq D_1(\mu, \nu), \tag{53}$$

for any measures $\frac{d\mu}{dx}, \frac{d\nu}{dx} \in L^2(\mathcal{M})$. Now consider the estimator $\mu^\star$ for μ, for which we are given n i.i.d samples, defined in the proof of Theorem 3. From the assumption, we have $\frac{d\mu}{dx} \in \mathcal{H}^s(\mathcal{M})$ and clearly, $\frac{d\tilde{\mu}}{dx} \in \mathcal{H}^s(\mathcal{M})$ (from its own definition). Thus, we can use the proof of Theorem 3 specified to $\alpha = 1$ to achieve the result for the 1-Wasserstein distance of smooth distributions.

$\square$

D. Proof of Theorem 3

Theorem 3 (Convergence rate of Sobolev IPMs under invariances). *Consider a G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$ with $\frac{d\mu}{dx} \in \mathcal{H}^s(\mathcal{M})$ for some $s \geq 0$.*

- *If $\alpha < d/2$, then there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$, as a function of n i.i.d. samples $X_1, X_2 \ldots, X_n \sim \mu$, such that*

$$\mathbb{E}[D_\alpha(\mu, \tilde{\mu})] \lesssim \left(\frac{\mathrm{vol}(\mathcal{M}/G)}{n} \right)^{\frac{s+\alpha}{2s+d}} \left\| \frac{d\mu}{dx} \right\|_{\mathcal{H}^s(\mathcal{M})}^{\frac{d-2\alpha}{2s+d}},$$

  where the constant only depends on the manifold and $\mathrm{vol}(\mathcal{M}/G)$ is the volume of the quotient space $\mathcal{M}/G$ with $d := \dim(\mathcal{M}/G)$.

- *If $\alpha = d/2$, then there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$ such that*

$$\mathbb{E}[D_\alpha(\mu, \tilde{\mu})] \lesssim \sqrt{\frac{\mathrm{vol}(\mathcal{M}/G) \log(n)}{n}}.$$

Proof. We first need to prove the following upper bound on the Sobolev IPMs.

Lemma 2. *For Borel measures μ, ν with $\int_\mathcal{M} d\mu = \int_\mathcal{M} d\mu = 1$, such that $\frac{d\mu}{dx}, \frac{d\nu}{dx} \in L^2(\mathcal{M})$ and assuming $\mathrm{vol}(\mathcal{M}) = 1$, one has*

$$D_\alpha^2(\mu, \nu) = \sum_{\ell=1}^\infty \frac{(\mu_\ell - \nu_\ell)^2}{\lambda_\ell^\alpha}, \tag{54}$$

for any $\alpha \geq 0$, where $\mu_\ell = \mathbb{E}_{x \sim \mu}[\phi_\ell(x)]$ and $\nu_\ell = \mathbb{E}_{x \sim \nu}[\phi_\ell(x)]$ for any ℓ.

Proof. Any function $f \in \mathcal{H}^\alpha(\mathcal{M})$ can be written as $f = \sum_{\ell=0}^\infty f_\ell \phi_\ell$ with $\|f\|_{\mathcal{H}^\alpha(\mathcal{M})}^2 = f_0^2 + \sum_{\ell=1}^\infty \lambda_\ell^\alpha f_\ell^2 < \infty$. From the definition,

$$D_\alpha(\mu, \nu) = \sup_{\substack{f \in \mathcal{H}^\alpha(\mathcal{M}) \\ \|f\|_{\mathcal{H}^\alpha(\mathcal{M})} \leq 1}} \left\{ \int_\mathcal{M} f d\mu - \int_\mathcal{M} f d\nu \right\} \tag{55}$$

$$= \sup_{\substack{f \in \mathcal{H}^\alpha(\mathcal{M}) \\ \|f\|_{\mathcal{H}^\alpha(\mathcal{M})} \leq 1}} \left\{ \int_\mathcal{M} f \frac{d\mu}{dx} dx - \int_\mathcal{M} f \frac{d\nu}{dx} dx \right\} \tag{56}$$

$$= \sup_{\substack{f \in \mathcal{H}^\alpha(\mathcal{M}) \\ \|f\|_{\mathcal{H}^\alpha(\mathcal{M})} \leq 1}} \left\langle f, \frac{d\mu}{dx} - \frac{d\nu}{dx} \right\rangle_{L^2(\mathcal{M})} \tag{57}$$

$$= \sup_{\substack{f \in \mathcal{H}^\alpha(\mathcal{M}) \\ \|f\|_{\mathcal{H}^\alpha(\mathcal{M})} \leq 1}} \sum_{\ell=0}^\infty f_\ell(\mu_\ell - \nu_\ell) \tag{58}$$

$$\overset{(a)}{=} \sup_{\substack{f \in \mathcal{H}^\alpha(\mathcal{M}) \\ \|f\|_{\mathcal{H}^\alpha(\mathcal{M})} \leq 1}} \sum_{\ell=1}^\infty \lambda_\ell^{\alpha/2} f_\ell \times \frac{(\mu_\ell - \nu_\ell)}{\lambda_\ell^{\alpha/2}} \tag{59}$$

$$\overset{(b)}{=} \sup_{\substack{f \in \mathcal{H}^\alpha(\mathcal{M}) \\ \|f\|_{\mathcal{H}^\alpha(\mathcal{M})} \leq 1}} \underbrace{\sqrt{\sum_{\ell=1}^\infty \lambda_\ell^\alpha f_\ell^2}}_{\leq 1} \times \sqrt{\sum_{\ell=1}^\infty \frac{(\mu_\ell - \nu_\ell)^2}{\lambda_\ell^\alpha}} \tag{60}$$

$$= \sqrt{\sum_{\ell=1}^\infty \frac{(\mu_\ell - \nu_\ell)^2}{\lambda_\ell^\alpha}}, \tag{61}$$

where (a) is due to the fact that $\mu_0 = \nu_0 = 1$, and (b) follows from the Cauchy–Schwarz inequality. $\qquad \square$

Proof of Theorem 3. Given i.i.d. samples from a probability measure μ with $\|\frac{d\mu}{dx}\|_{\mathcal{H}^s(\mathcal{M})} < \infty$, we propose the following estimator. First, fix a parameter T (to be set later) and let $\mathbb{1}_G(\ell; T) = \mathbb{1}_G(\ell)\mathbb{1}\{\ell < T\}$ for any $\ell = 0, 1, \dots$. Then, define the sequence

$$\tilde{\mu}_\ell := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_G(\ell; T)\phi_\ell(X_i), \tag{62}$$

for each ℓ and let $\frac{d\tilde{\mu}}{dx} := \sum_{\ell=0}^\infty \tilde{\mu}_\ell \phi_\ell$. The measure $\tilde{\mu}$ is thus well-defined, and $\int_\mathcal{M} d\tilde{\mu} = 1$, but it is not necessarily a probability measure since it can be a signed measure. However, due to the truncation of the sum at T, we have $\frac{d\tilde{\mu}}{dx} \in \mathcal{H}^s(\mathcal{M})$. Define a new probability measure $\mu^\star$ as

$$\mu^\star = \arg\min_{\nu \in \mathcal{P}(\mathcal{M})} D_\alpha(\nu, \tilde{\mu}). \tag{63}$$

This shows that by the triangle inequality

$$D_\alpha(\mu, \mu^\star) \leq D_\alpha(\mu, \tilde{\mu}) + D_\alpha(\tilde{\mu}, \mu^\star) \leq 2D_\alpha(\mu, \tilde{\mu}), \tag{64}$$

where in above, we used the definition of the probability measure $\mu^\star$.

We claim that the estimated probability measure achieves the convergence rate claimed in the theorem. To this end, we need to prove upper bounds on $\mathbb{E}[D_\alpha(\mu, \tilde{\mu})]$. According to Lemma 2, one has

$$\mathbb{E}[D_\alpha(\mu, \mu^\star)] \leq 2\mathbb{E}[D_\alpha(\mu, \tilde{\mu})] \tag{65}$$

$$\leq 2\sqrt{\mathbb{E}[D_\alpha^2(\mu, \tilde{\mu})]} \tag{66}$$

$$= 2\sqrt{\mathbb{E}\Big[\sum_{\ell=1}^\infty \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha}\Big]}. \tag{67}$$

Note that

$$\mathbb{E}\Big[\sum_{\ell=1}^\infty \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha}\Big] = \mathbb{E}\Big[\sum_{\ell=1}^{T-1} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha}\Big] + \mathbb{E}\Big[\sum_{\ell \geq T} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha}\Big] \tag{68}$$

$$= \underbrace{\mathbb{E}\Big[\sum_{\ell=1}^{T-1} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha}\Big]}_{(\mathrm{I})} + \underbrace{\sum_{\ell \geq T} \frac{\mu_\ell^2}{\lambda_\ell^\alpha}}_{(\mathrm{II})}, \tag{69}$$

since $\tilde{\mu}_\ell = 0$ for $\ell > T$. $\qquad \square$

To upper bound the second term, we write

$$\sum_{\ell \geq T} \frac{\mu_\ell^2}{\lambda_\ell^\alpha} = \sum_{\ell > T} \mu_\ell^2 \lambda_\ell^s \lambda_\ell^{-(s+\alpha)} \leq \lambda_T^{-(s+\alpha)} \sum_{\ell \geq T} \mu_\ell^2 \lambda_\ell^s \leq \lambda_T^{-(s+\alpha)} \Big\|\frac{d\mu}{dx}\Big\|_{\mathcal{H}^s(\mathcal{M})}^2. \tag{70}$$

To bound the first term, we need to use the following lemma.

17

Lemma 3. *One has*

$$R(\lambda) := \mathbb{E}\Big[\sum_{\ell:\lambda_\ell \leq \lambda} (\mu_\ell - \tilde{\mu}_\ell)^2\Big] \leq \frac{1}{n}\mathbb{E}_{x\sim\mu}\Big[N_x(\lambda;G)\Big], \tag{71}$$

where $N_x(\lambda;G)$ is defined in Theorem 7.

Proof. Note that the coefficients $\tilde{\mu}_\ell$ are empirical estimates of μ_ℓ given n i.i.d. samples; thus

$$R(\lambda_T) = \mathbb{E}\Big[\sum_{\ell:\lambda_\ell \leq \lambda_T} (\mu_\ell - \tilde{\mu}_\ell)^2\Big] = \frac{1}{n}\Big\{\sum_{\ell:\lambda_\ell \leq \lambda_T} \mathbb{1}_G(\ell)\Big(\mathbb{E}_{x\sim\mu}[\phi_\ell^2(x)] - (\mathbb{E}_{x\sim\mu}[\phi_\ell(x)])^2\Big)\Big\} \tag{72}$$

$$\leq \frac{1}{n}\mathbb{E}_{x\sim\mu}\Big[\sum_{\ell:\lambda_\ell \leq \lambda_T} \mathbb{1}_G(\ell)\phi_\ell^2(x)\Big] \tag{73}$$

$$\leq \frac{1}{n}\mathbb{E}_{x\sim\mu}\Big[N_x(\lambda_T;G)\Big]. \tag{74}$$

$\square$

We use Lemma 3 to complete the proof. First, note that

$$\mathbb{E}\Big[\sum_{\ell=1}^{T-1} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha}\Big] \leq \int_{0^+}^{\lambda_T} \frac{dR(\lambda)}{\lambda^\alpha}, \tag{75}$$

where $R(\lambda)$ is defined in Lemma 3 and $dR(\lambda)$ denotes the Stieltjes measure corresponding to $R(\lambda)$. By integration by parts, one has

$$\mathbb{E}\Big[\sum_{\ell=1}^{T-1} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha}\Big] \leq \int_{0^+}^{\lambda_T} \frac{dR(\lambda)}{\lambda^\alpha} = \frac{R(\lambda)}{\lambda^\alpha}\Big|_{0^+}^{\lambda_T} + \alpha\int_{0^+}^{\lambda_T} \frac{R(\lambda)}{\lambda^{\alpha+1}}d\lambda. \tag{76}$$

Now assume $\alpha < d/2$. By Theorem 7 and Lemma 3, we have

$$\mathbb{E}\Big[\sum_{\ell=1}^{T-1} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha}\Big] \leq \frac{R(\lambda_T)}{\lambda_T^\alpha} + \alpha\int_{0^+}^{\lambda_T} \frac{R(\lambda)}{\lambda^{\alpha+1}}d\lambda \tag{77}$$

$$\leq \frac{1}{n}\Big\{\frac{\omega_d}{(2\pi)^d}\,\mathrm{vol}(\mathcal{M}/G)\lambda_T^{d/2-\alpha} + O(\lambda_T^{(d-1)/2-\alpha}) \tag{78}$$

$$+ \alpha\int_{0^+}^{\lambda_T}\Big\{\frac{\omega_d}{(2\pi)^d}\,\mathrm{vol}(\mathcal{M}/G)\lambda_T^{d/2-\alpha-1} + O(\lambda_T^{(d-1)/2-\alpha-1})\Big\}d\lambda\Big\} \tag{79}$$

$$\lesssim \frac{1}{n}\Big(1 + \frac{\alpha}{d/2-\alpha}\Big)\frac{\omega_d}{(2\pi)^d}\,\mathrm{vol}(\mathcal{M}/G)\lambda_T^{d/2-\alpha}, \tag{80}$$

where in $\lesssim$ the constant is absolute. Plugging in this estimation into Equation (69) results in

$$\mathbb{E}\Big[\sum_{\ell=1}^{\infty} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha}\Big] \lesssim \frac{1}{n}\frac{d/2}{d/2-\alpha}\frac{\omega_d}{(2\pi)^d}\,\mathrm{vol}(\mathcal{M}/G)\lambda_T^{d/2-\alpha} + \lambda_T^{-(s+\alpha)}\Big\|\frac{d\mu}{dx}\Big\|_{\mathcal{H}^s(\mathcal{M})}^2. \tag{81}$$

We can choose the parameter $\lambda_T \in (0,\infty)$ to minimize the above upper bound.

Note that the function $p(\lambda) = c_a\lambda^{-a} + c_b\lambda^b$ with $a,b,c_a,c_b > 0$ is minimized for $\lambda \in (0,\infty)$ when

$$\lambda = \lambda_{\mathrm{opt}} := \Big(\frac{ac_a}{bc_b}\Big)^{1/(a+b)}. \tag{82}$$

Taking $a = s+\alpha$, $b = d/2-\alpha$, and

$$c_a = \Big\|\frac{d\mu}{dx}\Big\|_{\mathcal{H}^s(\mathcal{M})}^2, \quad c_b = \frac{1}{n}\frac{d/2}{d/2-\alpha}\frac{\omega_d}{(2\pi)^d}\,\mathrm{vol}(\mathcal{M}/G), \tag{83}$$

suggests to take

$$\lambda_T = \left\{ \left((s+\alpha)\|d\mu/dx\|^2_{\mathcal{H}^s(\mathcal{M})} \right) \Big/ \left(\frac{d}{2n} \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G) \right) \right\}^{1/(s+d/2)}. \tag{84}$$

Therefore, we have the following upper bound

$$\mathbb{E}\left[\sum_{\ell=1}^{\infty} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha} \right] \tag{85}$$

$$\lesssim \frac{1}{n} \frac{d/2}{d/2-\alpha} \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G) \left\{ \left((s+\alpha)\|d\mu/dx\|^2_{\mathcal{H}^s(\mathcal{M})} \right) \Big/ \left(\frac{d}{2n} \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G) \right) \right\}^{(d/2-\alpha)/(s+d/2)} \tag{86}$$

$$+ \left\| \frac{d\mu}{dx} \right\|^2_{\mathcal{H}^s(\mathcal{M})} \left\{ \left((s+\alpha)\|d\mu/dx\|^2_{\mathcal{H}^\alpha(\mathcal{M})} \right) \Big/ \left(\frac{d}{2n} \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G) \right) \right\}^{-(s+\alpha)/(s+d/2)}, \tag{87}$$

which shows that

$$\mathbb{E}\left[\sum_{\ell=1}^{\infty} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha} \right] \lesssim \left(\frac{\operatorname{vol}(\mathcal{M}/G)}{n} \right)^{\frac{s+\alpha}{s+d/2}} \left\| \frac{d\mu}{dx} \right\|^{\frac{d-2\alpha}{s+d/2}}_{\mathcal{H}^s(\mathcal{M})}. \tag{88}$$

This completes the proof for $\alpha < d/2$, since

$$\mathbb{E}[D_\alpha(\mu, \mu^\star)] \leq 2\sqrt{ \mathbb{E}\left[\sum_{\ell=1}^{\infty} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha} \right] } \lesssim \left(\frac{\operatorname{vol}(\mathcal{M}/G)}{n} \right)^{\frac{s+\alpha}{2s+d}} \left\| \frac{d\mu}{dx} \right\|^{\frac{d-2\alpha}{2s+d}}_{\mathcal{H}^s(\mathcal{M})}. \tag{89}$$

Now, we study the case where $\alpha = d/2$. By Theorem 7 and Lemma 3, we have

$$\mathbb{E}\left[\sum_{\ell=1}^{T-1} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha} \right] \leq \frac{R(\lambda_T)}{\lambda_T^\alpha} + \alpha \int_{0^+}^{\lambda_T} \frac{R(\lambda)}{\lambda^{\alpha+1}} d\lambda \tag{90}$$

$$\leq \frac{\alpha}{n} \int_{\lambda_1}^{\lambda_T} \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G) \frac{d\lambda}{\lambda} + O(1/n) \tag{91}$$

$$= \frac{\alpha}{n} \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G) \ln(\lambda_T) + O(1/n). \tag{92}$$

Thus, from Equation (69) we have

$$\mathbb{E}\left[\sum_{\ell=1}^{\infty} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha} \right] \lesssim \frac{\alpha}{n} \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G) \ln(\lambda_T) + \lambda_T^{-(s+\alpha)} \left\| \frac{d\mu}{dx} \right\|^2_{\mathcal{H}^s(\mathcal{M})} + O(1/n). \tag{93}$$

To minimize the above upper bound as a function of λ_T, we consider the function $p(\lambda) = c_a \lambda^{-a} + c_b \ln(\lambda)$ and assume that c_b is small enough. Note that the minimum of $p(\lambda)$ for $\lambda \in (0, \infty)$ is achieved when

$$\lambda = \lambda_{\text{opt}} := \left(\frac{ac_a}{c_b} \right)^{1/a}. \tag{94}$$

This means that for $a = s + \alpha$, $c_a = \|d\mu/dx\|^2_{\mathcal{H}^s(\mathcal{M})}$, and $c_b = \frac{\alpha}{n} \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G)$, we can take

$$\lambda_T = \left\{ \left((s+\alpha)\|d\mu/dx\|^2_{\mathcal{H}^s(\mathcal{M})} \right) \Big/ \left(\frac{\alpha}{n} \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G) \right) \right\}^{1/(s+d/2)}, \tag{95}$$

to achieve

$$\mathbb{E}\left[\sum_{\ell=1}^{\infty} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha} \right] \lesssim \frac{d \log(n)}{(2s+d)n} \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G) + O(1/n). \tag{96}$$

This completes the proof for $\alpha = d/2$, since

$$\mathbb{E}[D_\alpha(\mu, \mu^\star)] \leq 2\sqrt{ \mathbb{E}\left[\sum_{\ell=1}^{\infty} \frac{(\mu_\ell - \tilde{\mu}_\ell)^2}{\lambda_\ell^\alpha} \right] } \lesssim \sqrt{ \frac{\operatorname{vol}(\mathcal{M}/G) \log(n)}{n} }. \tag{97}$$

$\square$

19

E. Proof of Theorem 4

Theorem 4. *Consider a G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$ with $\frac{d\mu}{dx} \in \mathcal{H}$. Then, there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$ such that*

$$\mathbb{E}[D_{\mathcal{H}}(\mu, \hat{\mu})] \leq \sqrt{\frac{\mathrm{Tr}(K; G)}{n}}. \tag{12}$$

Proof. Note that a function $f \in \mathcal{H}$ can be written as $f = \sum_{\ell=0}^{\infty} f_\ell \phi_\ell$ such that $\|f\|_{\mathcal{H}}^2 = \sum_{\ell=0}^{\infty} \frac{f_\ell^2}{\xi_\ell} < \infty$. Thus, for any probability measures μ, ν such that $\frac{d\mu}{dx}, \frac{d\nu}{dx} \in \mathcal{H}$, we have

$$D_{\mathcal{H}}(\mu, \nu) := \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \left\{ \mathbb{E}_{x \sim \mu}[f(x)] - \mathbb{E}_{x \sim \nu}[f(x)] \right\} \tag{98}$$

$$= \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \left\{ \sum_{\ell=1}^{\infty} f_\ell(\mu_\ell - \nu_\ell) \right\} = \sqrt{\sum_{\ell=1}^{\infty} \xi_\ell(\mu_\ell - \nu_\ell)^2}. \tag{99}$$

Now consider the estimator $\tilde{\mu}$ that is defined similar to the proof of Theorem 1; we have

$$\tilde{\mu}_\ell := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_G(\ell) \phi_\ell(X_i), \tag{100}$$

for each $\ell = 0, 1, \ldots$, and $\tilde{\mu}$ denotes the measure[5] for which $\mathbb{E}_{x \sim \tilde{\mu}}[\phi_\ell(x)] = \tilde{\mu}_\ell$. Assume that ξ_ℓ depends only on λ_ℓ, and thus denote it by $\xi(\lambda)$. Using Lemma 3 and integration by parts, we have

$$\mathbb{E}[D_{\mathcal{H}}(\mu, \tilde{\mu})] \leq \sqrt{\mathbb{E}[D_{\mathcal{H}}(\mu, \tilde{\mu})^2]} = \sqrt{\mathbb{E}\left[\sum_{\ell=1}^{\infty} \xi_\ell(\mu_\ell - \tilde{\mu}_\ell)^2 \right]} \tag{101}$$

$$= \sqrt{\int_{0^+}^{\infty} \xi(\lambda) dR(\lambda)} \tag{102}$$

$$\leq \sqrt{\frac{1}{n} \int_{0^+}^{\infty} \xi(\lambda) dN_x(\lambda; G)}, \tag{103}$$

which holds when $\xi(\lambda)$ is non-increasing. Now, note that

$$\int_{0^+}^{\infty} \xi(\lambda) dN_x(\lambda; G) = \sum_{\ell=1}^{\infty} \mathbb{1}_G(\ell) \xi(\lambda_\ell), \tag{104}$$

and this completes the proof.

$\square$

F. Proof of Theorem 5

Theorem 5 (Convergence rate for density estimation in $L^2(\mathcal{M})$ distance under invariances)**.** *Consider a G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$ with $\frac{d\mu}{dx} \in \mathcal{H}^s(\mathcal{M})$. Then, there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$ such that*

$$\mathbb{E}\left[\left\| \frac{d\tilde{\mu}}{dx} - \frac{d\mu}{dx} \right\|_{L^2(\mathcal{M})} \right] \lesssim \left(\frac{\mathrm{vol}(\mathcal{M}/G)}{n} \right)^{\frac{s}{2s+d}} \left\| \frac{d\mu}{dx} \right\|_{L^2(\mathcal{M})}^{\frac{d}{2s+d}},$$

where the constant only depends on the manifold and $\mathrm{vol}(\mathcal{M}/G)$ denotes the volume of the quotient space $\mathcal{M}/G$ with $d := \dim(\mathcal{M}/G) \geq 3$.

---

[5]Indeed, one can also achieve the convergence rate by truncating the sum over $\lambda_\ell < T$ with appropriate T to make the estimation computable.

Proof. For any two probability measures μ, ν such that $\frac{d\mu}{dx}, \frac{d\nu}{dx} \in L^2(\mathcal{M})$, we have

$$D_0(\mu, \nu) = \sup_{\substack{f \in L^2(\mathcal{M}) \\ \|h\|_{f^2(\mathcal{M})}=1}} \mathbb{E}_\mu[f] - \mathbb{E}_\nu[f] \tag{105}$$

$$= \sup_{\substack{f \in L^2(\mathcal{M}) \\ \|f\|_{L^2(\mathcal{M})}=1}} \langle f, \frac{d\mu}{dx} - \frac{d\nu}{dx} \rangle_{L^2(\mathcal{M})} \tag{106}$$

$$= \left\| \frac{d\mu}{dx} - \frac{d\nu}{dx} \right\|_{L^2(\mathcal{M})}. \tag{107}$$

Therefore, we can use the estimator in the proof of Theorem 3 for $\alpha = 0$ to achieve the desired result. $\qquad\square$

G. Proof of Theorem 6

Theorem 6 (Convergence rate for density estimation in $L^\infty(\mathcal{M})$ under invariances). *Consider a G-invariant probability measure $\mu \in \mathcal{P}(\mathcal{M})$ with $\frac{d\mu}{dx} \in \mathcal{H}^s(\mathcal{M})$ for some $s > d/2$. Then, there exists an estimator $\tilde{\mu} \in \mathcal{P}(\mathcal{M})$, such that*

$$\mathbb{E}\left[\left\| \frac{d\tilde{\mu}}{dx} - \frac{d\mu}{dx} \right\|_{L^\infty(\mathcal{M})} \right]$$
$$\lesssim \left(\frac{(\mathrm{vol}(\mathcal{M}/G))^{\frac{2s}{s-d/2}}}{n} \right)^{\frac{s-d/2}{2s+d}} \left\| \frac{d\mu}{dx} \right\|_{\mathcal{H}^s(\mathcal{M})}^{\frac{d}{s+d/2}},$$

where the constant only depends on the manifold.

Proof. Consider the estimator $\tilde{\mu}$ proposed in the proof of Theorem 3 with a parameter T (to be set). Let μ_T denote the measure obtained by truncating the sum corresponding to μ at T. Note that for the estimator $\tilde{\mu}$ we have

$$\mathbb{E}\left[\left\| \frac{d\tilde{\mu}}{dx} - \frac{d\mu}{dx} \right\|_{L^\infty(\mathcal{M})} \right] \leq \underbrace{\mathbb{E}\left[\left\| \frac{d\tilde{\mu}}{dx} - \frac{d\mu_T}{dx} \right\|_{L^\infty(\mathcal{M})} \right]}_{\text{(I)}} + \underbrace{\left\| \frac{d\mu_T}{dx} - \frac{d\mu}{dx} \right\|_{L^\infty(\mathcal{M})}}_{\text{(II)}}. \tag{108}$$

To bound the first term, we use the Cauchy–Schwarz inequality for low degree functions; see the following lemma.

Lemma 4. *Assume that $f \in L^2(\mathcal{M})$ and $f = \sum_{\ell=0}^{T} f_\ell \mathbb{1}_G(\ell) \phi_\ell$. Then,*

$$\|f\|_{L^\infty(\mathcal{M})} \leq \|f\|_{L^2(\mathcal{M})} \sqrt{\sup_{x \in \mathcal{M}} N_x(\lambda; G)}. \tag{109}$$

Proof. By Theorem 7, we have

$$\|f\|_{L^\infty(\mathcal{M})} = \sup_{x \in \mathcal{M}} |f(x)| = \sup_{x \in \mathcal{M}} \left| \sum_{\ell=0}^{T} f_\ell \mathbb{1}_G(\ell) \phi_\ell \right| \tag{110}$$

$$\leq \sqrt{\sum_{\ell=0}^{T} f_\ell^2} \sqrt{\sup_{x \in \mathcal{M}} \sum_{\ell=0}^{T} \mathbb{1}_G(\ell) \phi_\ell(x)^2} \tag{111}$$

$$\leq \|f\|_{L^2(\mathcal{M})} \sqrt{\sup_{x \in \mathcal{M}} N_x(\lambda; G)}. \tag{112}$$

$$\square$$

Now we use the above lemma and Lemma 3 to write

$$\mathbb{E}\left[\left\| \frac{d\tilde{\mu}}{dx} - \frac{d\mu_T}{dx} \right\|_{L^\infty(\mathcal{M})} \right] \leq \sqrt{\sup_{x \in \mathcal{M}} N_x(\lambda_T; G)} \mathbb{E}\left[\left\| \frac{d\tilde{\mu}}{dx} - \frac{d\mu_T}{dx} \right\|_{L^2(\mathcal{M})} \right] \tag{113}$$

$$\leq \sqrt{\frac{1}{n} \sup_{x \in \mathcal{M}} N_x(\lambda_T; G) \times \mathbb{E}_{x \sim \mu}[N_x(\lambda_T; G)]}. \tag{114}$$

To bound the second term, we have

$$\left\| \frac{d\mu_T}{dx} - \frac{d\mu}{dx} \right\|_{L^\infty(\mathcal{M})} = \sup_{x \in \mathcal{M}} \left| \sum_{\ell > T} \mu_\ell \mathbb{1}_G(\ell) \phi_\ell \right| \tag{115}$$

$$= \sup_{x \in \mathcal{M}} \left| \sum_{\ell > T} \mu_\ell \mathbb{1}_G(\ell) \lambda_\ell^{s/2} \phi_\ell \lambda_\ell^{-s/2} \right| \tag{116}$$

$$\leq \sqrt{\sum_{\ell > T} \mu_\ell^2 \lambda_\ell^s} \sqrt{\sup_{x \in \mathcal{M}} \sum_{\ell > T} \mathbb{1}_G(\ell) \phi_\ell^2 \lambda_\ell^{-s}} \tag{117}$$

$$\leq \|f\|_{\mathcal{H}^s(\mathcal{M})} \sqrt{\sup_{x \in \mathcal{M}} \sum_{\ell > T} \mathbb{1}_G(\ell) \phi_\ell^2 \lambda_\ell^{-s}}. \tag{118}$$

Note that

$$\sup_{x \in \mathcal{M}} \sum_{\ell > T} \mathbb{1}_G(\ell) \phi_\ell^2 \lambda_\ell^{-s} = \sup_{x \in \mathcal{M}} \int_{\lambda_T}^\infty \lambda_\ell^{-s} dN_x(\lambda; G) \tag{119}$$

$$\leq \frac{d}{d - 2s} \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G) \lambda_T^{d/2-s}. \tag{120}$$

Thus, we combine the two terms and obtain

$$\mathbb{E}\left[\left\| \frac{d\tilde{\mu}}{dx} - \frac{d\mu}{dx} \right\|_{L^\infty(\mathcal{M})} \right]$$

$$\leq \sqrt{\frac{1}{n} \sup_{x \in \mathcal{M}} N_x(\lambda; G) \times \mathbb{E}_{x \sim \mu}[N_x(\lambda; G)]} + \|f\|_{\mathcal{H}^s(\mathcal{M})} \sqrt{\frac{d}{d - 2s} \frac{\omega_d}{(2\pi)^d} \operatorname{vol}(\mathcal{M}/G) \lambda_T^{d/2-s}}$$

$$\lesssim \frac{1}{\sqrt{n}} \operatorname{vol}(\mathcal{M}/G) \lambda_T^{d/2} + \|f\|_{\mathcal{H}^s(\mathcal{M})} \sqrt{\operatorname{vol}(\mathcal{M}/G)} \lambda_T^{d/4-s/2}.$$

By minimizing the above upper bound as a function of λ_T, we obtain the desired result.

$\square$