# Sample Complexity of Probability Divergences under Group Symmetry

**Ziyu Chen** [1] **Markos A. Katsoulakis** [1] **Luc Rey-Bellet** [1] **Wei Zhu** [1]

## Abstract

We rigorously quantify the improvement in the sample complexity of variational divergence estimations for group-invariant distributions. In the cases of the Wasserstein-1 metric and the Lipschitz-regularized $\alpha$-divergences, the reduction of sample complexity is proportional to an ambient-dimension-dependent power of the group size. For the maximum mean discrepancy (MMD), the improvement of sample complexity is more nuanced, as it depends on not only the group size but also the choice of kernel. Numerical simulations verify our theories.

## 1. Introduction

Probability divergences provide means to measure the discrepancy between two probability distributions. They have broad applications in a variety of inference tasks, such as independence testing (Zhang et al., 2018; Kinney & Atwal, 2014), independent component analysis (Hyvarinen et al., 2002), and generative modeling (Goodfellow et al., 2014; Nowozin et al., 2016; Arjovsky et al., 2017; Gulrajani et al., 2017; Tolstikhin et al., 2018; Nietert et al., 2021).

A key task within the above applications is the computation and estimation of the divergences from finite data, which is known to be a difficult problem (Paninski, 2003; Gao et al., 2015). Empirical estimators based on the variational representations for the probability divergences are generally favored and widely used due to their scalability to both the data size and the ambient space dimension (Belghazi et al., 2018; Birrell et al., 2022b; Nguyen et al., 2007; 2010; Ruderman et al., 2012; Sreekumar & Goldfeld, 2022; Birrell et al., 2021; 2022d; Sriperumbudur et al., 2012; Gretton et al., 2006; 2007; 2012; Genevay et al., 2019).

Empirical computation of the probability divergences and theoretical analysis on their sample complexity are typically studied without any a priori *structural assumption* on the probability measures. Many distributions in real life, however, are known to have intrinsic structures, such as *group symmetry*. For example, the distribution of the medical images collected without preferred orientation should be *rotation-invariant*, i.e., an image is supposed to have the same likelihood as its rotated copies; see Figure 1. Such structural information could be leveraged to improve the accuracy and/or sample-efficiency for divergence estimation.

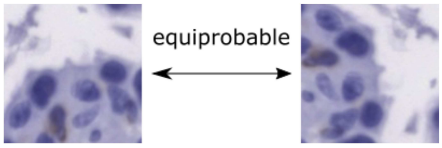


*Figure 1.* The distribution of the whole-slide prostate cancer images (LYSTO data set (Ciompi et al., 2019)) is *rotation-invariant*, i.e., an image and its rotated copies are equiprobable.

Indeed, the recent work by Birrell et al. (2022c) shows that one can develop an improved variational representation for divergences between group-invariant distributions. The key idea is to reduce the test function space in the variational formula to its subset of group-invariant functions, which effectively acts as an unbiased regularization. When used in a generative adversarial network (GAN) for group-invariant distribution learning, Birrell et al. (2022c) empirically show that divergence estimation/optimization based on their proposed variational representation under group symmetry leads to significantly improved sample generation, especially in the small data regime.

The purpose of this work is to rigorously quantify the performance gain of divergence estimation under group symmetry. More specifically, we analyze the reduction in sample complexity of divergence estimation in terms of the (finite) group size. We focus, in particular, on three types of probability divergences: the Wasserstein-1 metric, the maximum mean discrepancy (MMD), and the family of Lipschitz-regularized $\alpha$-divergences; see Section 3.1 for the exact definition. Our main results show that the reduction of samples needed for guaranteed fidelity in statistical estimation of divergences is proportional to a dimension-dependent power of the group size; see Theorem 4.1 for the Wasserstein-1 metric and Theorem 4.8 for the Lipschitz-regularized $\alpha$-

[1]Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA 01003, USA. Correspondence to: Ziyu Chen <ziyuchen@umass.edu>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. 

divergences respectively. In the case of MMD, the reduction in sample complexity due to the group invariance is more nuanced and depends on the properties of the kernel; see Theorem 4.10. As a byproduct, we also establish the consistency and sample complexity for the Lipschitz-regularized $\alpha$-divergences without group symmetry, which, to the best of our knowledge, is missing in the previous literature.

## 2. Related work

**Empirical estimation of probability divergences.** Probability divergences have been widely used, including in generative adversarial networks (GANs) (Arjovsky et al., 2017; Goodfellow et al., 2014; Nowozin et al., 2016; Birrell et al., 2022c; Gulrajani et al., 2017), uncertainty quantification (Chowdhary & Dupuis, 2013; Dupuis et al., 2016), independence determination through mutual information estimation (Belghazi et al., 2018), bounding risk in probably approximately correct (PAC) learning (Catoni et al., 2008; McAllester, 1999; Shawe-Taylor & Williamson, 1997), statistical mechanics and interacting particles (Kipnis & Landim, 1999), large deviations (Dupuis & Ellis, 2011), and parameter estimation (Broniatowski & Keziou, 2009).

A growing body of literature has been dedicated to the empirical estimation of divergences from finite data. Earlier works based on density estimation are known to work best for low dimensions (Kandasamy et al., 2015; Póczos et al., 2011). Recent research has shown that statistical estimators based on the variational representations of probability divergences scale better with dimensions; such studies include the KL-divergences (Belghazi et al., 2018), the more general $f$-divergences (Birrell et al., 2022b; Nguyen et al., 2007; 2010; Ruderman et al., 2012; Sreekumar & Goldfeld, 2022), Rényi divergences (Birrell et al., 2021; 2022d), integral probability metrics (IPMs) (Sriperumbudur et al., 2012; Gretton et al., 2006; 2007; 2012), and Sinkhorn divergences (Genevay et al., 2019). Such estimators are typically constructed to compare an *arbitrary* pair of probability measures without any a priori structural assumption, and are hence sub-optimal in estimating divergences between distributions with known structures, such as group symmetry.

**Group-invariant distributions.** Recent development in group-equivariant machine learning (Cohen & Welling, 2016; Cohen et al., 2019; Weiler & Cesa, 2019) has sparked a flurry of research in neural generative models for group-invariant distributions. Most of the works focus only on the guaranteed *generation*, through, e.g., an equivariant normalizing-flow, of the group-invariant distributions (Biloš & Günnemann, 2021; Boyda et al., 2021; Garcia Satorras et al., 2021; Köhler et al., 2019; Liu et al., 2019; Rezende et al., 2019); the divergence computation between the generated distribution and the ground-truth target, a crucial step in the optimization pipeline, however, does not leverage their group-invariant structure. Equivariant GANs for group-invariant distribution learning have also been proposed by modifying the inner loop of discriminator update through either data-augmentation (Zhao et al., 2020) or constrained optimization within a subspace of group-invariant discriminators (Dey et al., 2021); the theoretical justification of such procedures, as well as the resulting performance gain, have been explained by Birrell et al. (2022c) as an improved estimation of variational divergences under group symmetry via an unbiased regularization. The exact *quantification* of the improvement, in terms of reduction in sample complexity, is however still missing; this is the main focus of this work.

## 3. Background and motivation

### 3.1. Variational divergences and probability metrics

Let $\mathcal{X}$ be a measurable space, and $\mathcal{P}(\mathcal{X})$ be the set of probability measures on $\mathcal{X}$. A map $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, \infty]$ is called a *divergence* on $\mathcal{P}(\mathcal{X})$ if

$$D(P, Q) = 0 \iff P = Q \in \mathcal{P}(\mathcal{X}), \tag{1}$$

hence providing a notion of "distance" between probability measures. Many probability divergences of interest can be formulated using a variational representation

$$D(P, Q) = \sup_{\gamma \in \Gamma} H(\gamma; P, Q), \tag{2}$$

where $\Gamma \subset \mathcal{M}(\mathcal{X})$ is a space of test functions, $\mathcal{M}(\mathcal{X})$ is the set of measurable functions on $\mathcal{X}$, and $H : \mathcal{M}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [-\infty, \infty]$ is some objective functional. Through suitable choices of $H(\gamma; P, Q)$ and $\Gamma$, formula (2) includes many divergences and probability metrics. Below we list two specific classes of examples.

**(a) $\Gamma$-Integral Probability Metrics ($\Gamma$-IPMs).** Given $\Gamma \subset \mathcal{M}_b(\mathcal{X})$, the space of bounded measurable functions on $\mathcal{X}$, the $\Gamma$-IPM between $P$ and $Q$ is defined as

$$D^{\Gamma}(P, Q) \coloneqq \sup_{\gamma \in \Gamma} \{E_P[\gamma] - E_Q[\gamma]\}. \tag{3}$$

Some prominent examples of the $\Gamma$-IPMs include the Wasserstein-1 metric, the total variation metric, the Dudley metric, and the maximum mean discrepancy (MMD) (Müller, 1997; Sriperumbudur et al., 2012). Our work, in particular, focuses on the following two specific IPMs.

- The Wasserstein-1 metric, $W(P, Q) \coloneqq D^{\mathrm{Lip}_L(\mathcal{X})}(P, Q)$, i.e.,

$$W(P, Q) \coloneqq \sup_{\gamma \in \mathrm{Lip}_L(\mathcal{X})} \{E_P[\gamma] - E_Q[\gamma]\}, \tag{4}$$

  where $\mathrm{Lip}_L(\mathcal{X})$ is the space of $L$-Lipschitz functions on $\mathcal{X}$. We note that the normalizing factor $L^{-1}$ has been omitted from the formula.

- The *maximum mean discrepancy*, $\text{MMD}(P,Q) \coloneqq D^{B_\mathcal{H}}(P,Q)$, i.e.,

$$\text{MMD}(P,Q) \coloneqq \sup_{\gamma \in B_\mathcal{H}} \{E_P[\gamma] - E_Q[\gamma]\}, \quad (5)$$

  where $B_\mathcal{H}$ is the unit ball of some reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ on $\mathcal{X}$.

**(b) $(f,\Gamma)$-divergences.** Let $f : [0,\infty) \to \mathbb{R}$ be convex and lower semi-continuous, with $f(1) = 0$ and $f$ strictly convex at $x = 1$. Given $\Gamma \subset \mathcal{M}_b(\mathcal{X})$ that is closed under the shift transformations $\gamma \mapsto \gamma + \nu, \nu \in \mathbb{R}$, the $(f,\Gamma)$-divergence introduced by Birrell et al. (2022a) is defined as

$$D_f^\Gamma(P\|Q) = \sup_{\gamma\in\Gamma}\{E_P[\gamma] - E_Q[f^*(\gamma)]\}, \quad (6)$$

where $f^*$ denotes the Legendre transform of $f$. Formula (6) includes, as a special case when $\Gamma = \mathcal{M}_b(\mathcal{X})$, the widely known class of $f$-divergences, with notable examples such as the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), the total variation distance, the Jensen-Shannon divergence, the $\chi^2$-divergence, the Hellinger distance, and more generally the family of $\alpha$-divergences (Nowozin et al., 2016). Of particular interest to us is the class of the *Lipschitz-regularized* $\alpha$-divergences,

$$D_{f_\alpha}^\Gamma(P\|Q), \;\; \Gamma = \text{Lip}_L(\mathcal{X}), \; f_\alpha(x) = \frac{x^\alpha - 1}{\alpha(\alpha-1)}, \quad (7)$$

where $\alpha > 0$ and $\alpha \neq 1$ is a positive parameter.

An important observation that will be useful in one of our results, Theorem 4.8, is that $D_{f_\alpha}^\Gamma$ admits an equivalent representation, which writes

$$D_{f_\alpha}^\Gamma(P\|Q) = \sup_{\gamma\in\Gamma,\nu\in\mathbb{R}} \{E_P[\gamma+\nu] - E_Q[f_\alpha^*(\gamma+\nu)]\} \quad (8)$$

due to the invariance of $\Gamma = \text{Lip}_L(\mathcal{X})$ under the shift map $\gamma \mapsto \gamma + \nu$ for $\nu \in \mathbb{R}$.

### 3.2. Empirical estimation of variational divergences

Given i.i.d. samples $X = \{x_1, x_2, \cdots, x_m\}$ and $Y = \{y_1, y_2, \cdots, y_n\}$, respectively, from two unknown probability measures $P, Q \in \mathcal{P}(\mathcal{X})$, it is often of interest—in applications such as two-sample testing (Bickel, 1969; Gretton et al., 2006; 2012; Cheng & Xie, 2021) and independence testing (Gretton et al., 2007; 2012; Zhang et al., 2018; Kinney & Atwal, 2014)—to estimate the divergence between $P$ and $Q$ (Sriperumbudur et al., 2012; Birrell et al., 2021; Nguyen et al., 2007; 2010). For variational divergences $D^\Gamma(P,Q)$ and $D_f^\Gamma(P\|Q)$ in the form of (3) and (6), their empirical estimators can naturally be given by

$$D^\Gamma(P_m, Q_n) = \sup_{\gamma\in\Gamma}\left\{\sum_{i=1}^m \frac{\gamma(x_i)}{m} - \sum_{i=1}^n \frac{\gamma(y_i)}{n}\right\}, \quad (9)$$

$$D_f^\Gamma(P_m\|Q_n) = \sup_{\gamma\in\Gamma}\left\{\sum_{i=1}^m \frac{\gamma(x_i)}{m} - \sum_{i=1}^n \frac{f^*(\gamma(y_i))}{n}\right\} \quad (10)$$

where $P_m = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ and $Q_n = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$ represent the empirical distributions of $P$ and $Q$, respectively.

The consistency and sample complexity of the empirical estimators $W(P_m, Q_n)$ and $\text{MMD}(P_m, Q_n)$ in the form of (9) for, respectively, the Wasserstein-1 metric (4) and MMD (5) between two *general distributions* $P, Q \in \mathcal{P}(\mathcal{X})$ have been well studied (Sriperumbudur et al., 2012; Gretton et al., 2012). However, for probability measures with special structures, such as *group symmetry*, one can potentially obtain a divergence estimator with substantially improved sample complexity as empirically observed by Birrell et al. (2022c). We provide, in the following section, a brief review of group-invariant distributions and the improved variational representations for probability divergences under group symmetry, which serves as a motivation and foundation for our theoretical analysis in Section 4.

### 3.3. Variational divergences under group symmetry

A *group* is a set $\Sigma$ equipped with a group product satisfying the axioms of associativity, identity, and invertibility. Given a group $\Sigma$ and a set $\mathcal{X}$, a map $\theta : \Sigma \times \mathcal{X} \to \mathcal{X}$ is called a *group action on* $\mathcal{X}$ if $\theta_\sigma \coloneqq \theta(\sigma,\cdot) : \mathcal{X} \to \mathcal{X}$ is an automorphism on $\mathcal{X}$ for all $\sigma \in \Sigma$, and $\theta_{\sigma_2} \circ \theta_{\sigma_1} = \theta_{\sigma_2\cdot\sigma_1}$, $\forall \sigma_1, \sigma_2 \in \Sigma$. By convention, we will abbreviate $\theta(\sigma, x)$ as $\sigma x$ throughout the paper.

A function $\gamma : \mathcal{X} \to \mathbb{R}$ is called $\Sigma$-*invariant* if $\gamma \circ \theta_\sigma = \gamma, \forall \sigma \in \Sigma$. Let $\Gamma$ be a set of measurable functions $\gamma : \mathcal{X} \to \mathbb{R}$; its subset, $\Gamma_\Sigma$, of $\Sigma$-invariant functions is defined as

$$\Gamma_\Sigma \coloneqq \{\gamma \in \Gamma : \gamma \circ \theta_\sigma = \gamma, \forall \sigma \in \Sigma\}. \quad (11)$$

On the other hand, a probability measure $P \in \mathcal{P}(\mathcal{X})$ is called $\Sigma$-*invariant* if $P = (\theta_\sigma)_* P, \forall \sigma \in \Sigma$, where $(\theta_\sigma)_* P \coloneqq P \circ (\theta_\sigma)^{-1}$ is the push-forward measure of $P$ under $\theta_\sigma$. We denote the set of all $\Sigma$-invariant distributions on $\mathcal{X}$ as $\mathcal{P}_\Sigma(\mathcal{X}) \coloneqq \{P \in \mathcal{P}(\mathcal{X}) : P \text{ is } \Sigma\text{-invariant}\}$.

Finally, for a compact Hausdorff topological group $\Sigma$ (Folland, 1999), we define two *symmetrization operators*, $S_\Sigma : \mathcal{M}_b(\mathcal{X}) \to \mathcal{M}_b(\mathcal{X})$ and $S^\Sigma : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$, on functions and probability measures, respectively, as follows

$$S_\Sigma[\gamma](x) \coloneqq \int_\Sigma \gamma(\sigma x)\mu_\Sigma(d\sigma), \; \forall \gamma \in \mathcal{M}_b(\mathcal{X}) \quad (12)$$

$$E_{S^\Sigma[P]}\gamma \coloneqq E_P S_\Sigma[\gamma], \; \forall P \in \mathcal{P}(\mathcal{X}), \forall \gamma \in \mathcal{M}_b(\mathcal{X}) \quad (13)$$

where $\mu_\Sigma$ is the unique Haar probability measure on $\Sigma$. The operators $S_\Sigma[\gamma]$ and $S^\Sigma[P]$ can be intuitively understood, respectively, as "averaging" the function $\gamma$ or "'spreading" the probability mass $P$ across the group orbits in $\mathcal{X}$; one

can easily verify that they are *projection operators* onto the corresponding invariant subsets $\Gamma_\Sigma \subset \Gamma$ and $\mathcal{P}_\Sigma(\mathcal{X}) \subset \mathcal{P}(\mathcal{X})$ (Birrell et al., 2022c).

The main result by Birrel et al. (2022c), which we summarize in Result 3.1, is that for $\Sigma$-invariant distributions, the function space $\Gamma$ in the variational formulae (3) and (6) can be reduced to its invariant subset $\Gamma_\Sigma \subset \Gamma$.

**Result 3.1** (paraphrased from (Birrell et al., 2022c))**.** *If $S_\Sigma[\Gamma] \subset \Gamma$ and $P, Q \in \mathcal{P}(X)$, then*

$$D^\Gamma(S^\Sigma[P], S^\Sigma[Q]) = D^{\Gamma_\Sigma}(P, Q), \tag{14}$$

$$D_f^\Gamma(S^\Sigma[P] \| S^\Sigma[Q]) = D_f^{\Gamma_\Sigma}(P \| Q), \tag{15}$$

*where $D^\Gamma(P, Q)$ and $D_f^\Gamma(P\|Q)$ are given by* (3) *and* (6)*. In particular, if $P, Q \in \mathcal{P}_\Sigma(\mathcal{X})$ are $\Sigma$-invariant, then*

$$D^\Gamma(P, Q) = D^{\Gamma_\Sigma}(P, Q), \quad D_f^\Gamma(P\|Q) = D_f^{\Gamma_\Sigma}(P\|Q).$$

Result 3.1 motivates a potentially more sample-efficient way to estimate the divergences $D^\Gamma(P, Q)$ and $D_f^\Gamma(P\|Q)$ between $\Sigma$-invariant distributions $P, Q \in \mathcal{P}(\mathcal{X})$ using

$$D^{\Gamma_\Sigma}(P_m, Q_n) = \sup_{\gamma \in \Gamma_\Sigma} \left\{ \sum_{i=1}^m \frac{\gamma(x_i)}{m} - \sum_{i=1}^n \frac{\gamma(y_i)}{n} \right\}, \tag{16}$$

$$D_f^{\Gamma_\Sigma}(P_m \| Q_n) = \sup_{\gamma \in \Gamma_\Sigma} \left\{ \sum_{i=1}^m \frac{\gamma(x_i)}{m} - \sum_{i=1}^n \frac{f^*(\gamma(y_i))}{n} \right\}. \tag{17}$$

Compared to Eq. (9) and (10), the estimators (16) and (17) have the benefit of optimizing over a reduced space $\Gamma_\Sigma \subset \Gamma$ of test functions, effectively acting as an *unbiased regularization*, and their efficacy has been empirically observed by Birrell et al. (2022c) in neural generation of group-invariant distributions with substantially improved data-efficiency. However, the theoretical understanding of the performance gain is still lacking.

The purpose of this work is to rigorously quantify the improvement in sample complexity of the divergence estimations (16) and (17) for group-invariant distributions. To contextualize the idea, we will focus our analysis on three specific types of probability divergences, the Wasserstein-1 metric (4), the MMD (5), and the Lipschitz-regularized $\alpha$ divergence (6)(7) between $\Sigma$-invariant $P, Q \in \mathcal{P}_\Sigma(\mathcal{X})$,

$$W(P, Q) = W^\Sigma(P, Q) \approx W^\Sigma(P_m, Q_n), \tag{18}$$

$$\text{MMD}(P, Q) = \text{MMD}^\Sigma(P, Q) \approx \text{MMD}^\Sigma(P_m, Q_n) \tag{19}$$

$$D_{f_\alpha}^\Gamma(P\|Q) = D_{f_\alpha}^{\Gamma_\Sigma}(P\|Q) \approx D_{f_\alpha}^{\Gamma_\Sigma}(P_m\|Q_n), \tag{20}$$

where

$$W^\Sigma(P, Q) \coloneqq D^{[\text{Lip}_L(\mathcal{X})]_\Sigma}(P, Q), \tag{21}$$

$$\text{MMD}^\Sigma(P, Q) \coloneqq D^{[B_\mathcal{H}]_\Sigma}(P, Q), \tag{22}$$

and the definition of $D_{f_\alpha}^{\Gamma_\Sigma}(P\|Q)$ is given by Equations (6), (7) and (11).

### 3.4. Further notations and assumptions

For the rest of the paper, we assume the measurable space $\mathcal{X} \subset \mathbb{R}^d$ is a bounded subset of $\mathbb{R}^d$ equipped with the Euclidean metric $\|\cdot\|_2$ and the group $\Sigma$ acting on $\mathcal{X}$ is assumed to be finite, i.e., $|\Sigma| < \infty$, where $|\Sigma|$ is the cardinality of $\Sigma$. The Haar measure $\mu_\Sigma$ is thus a uniform probability measure over $\Sigma$, and the symmetrization $S_\Sigma[\gamma]$ [Eq. (12)] is an average of $\gamma$ over the group orbit. We next introduce the concept of *fundamental domain* in the following definition.

**Definition 3.1.** A subset $\mathcal{X}_0 \subset \mathcal{X}$ is called a *fundamental domain* of $\mathcal{X}$ under the $\Sigma$-action if for each $x \in \mathcal{X}$, there exists a unique $x_0 \in \mathcal{X}_0$ such that $x = \sigma x_0$ for some $\sigma \in \Sigma$.

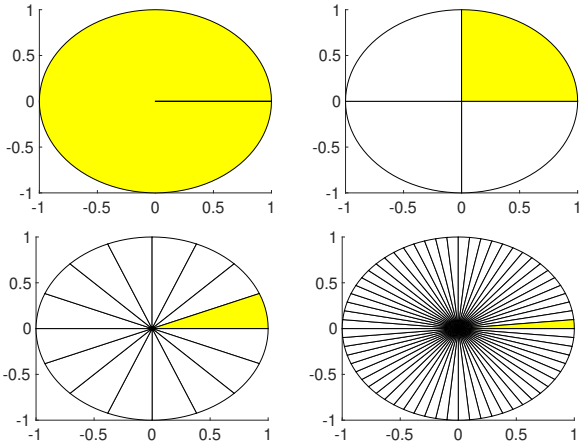

*Figure 2.* The unit disk $\mathcal{X} \subset \mathbb{R}^2$ with the action of the (discrete) rotation groups $\Sigma = C_n$, $n = 1, 4, 16, 64$. The fundamental domain $\mathcal{X}_0$ for each $C_n$ is filled with yellow color.

Figure 2 displays an example where $\mathcal{X}$ is the unit disk in $\mathbb{R}^2$, and $\Sigma = C_n, n = 1, 4, 16, 64$, are the discrete rotation groups acting on $\mathcal{X}$; the fundamental domain $\mathcal{X}_0$ for each $\Sigma = C_n$ is filled with yellow color. We note that the choice of the fundamental domain $\mathcal{X}_0$ is not unique. We will slightly abuse the notation $\mathcal{X} = \Sigma \times \mathcal{X}_0$ to denote $\mathcal{X}_0$ being a fundamental domain of $\mathcal{X}$ under the $\Sigma$-action. We define $T_0 : \mathcal{X} \to \mathcal{X}_0$

$$T_0(x) \coloneqq y \in \mathcal{X}_0, \text{ if } y = \sigma x \text{ for some } \sigma \in \Sigma, \tag{23}$$

i.e., $T_0$ maps $x \in \mathcal{X}$ to its unique orbit representative in $\mathcal{X}_0$. In addition, we denote by $P_{\mathcal{X}_0} \in \mathcal{P}(\mathcal{X}_0)$ the distribution on the fundamental domain $\mathcal{X}_0$ induced by a $\Sigma$-invariant distribution $P \in \mathcal{P}_\Sigma(\mathcal{X})$ on $\mathcal{X}$; that is,

$$P_{\mathcal{X}_0} = (T_0)_\sharp P. \tag{24}$$

The *diameter* of $\mathcal{X} \subset \mathbb{R}^d$ is defined as

$$\text{diam}(\mathcal{X}) = \sup_{x, y \in \mathcal{X}} \|x - y\|_2. \tag{25}$$

Finally, part of our results in Section 4 relies heavily on the concept of *covering numbers* which we define below.

**Definition 3.2** (Covering number)**.** Let $(\mathcal{X}, \rho)$ be a metric space. A subset $S \subset \mathcal{X}$ is called a $\delta$-cover of $\mathcal{X}$ if for any $x \in \mathcal{X}$ there is an $s \in S$ such that $\rho(s, x) \leq \delta$. The $\delta$-covering number of $\mathcal{X}$ is defined as

$$\mathcal{N}(\mathcal{X}, \delta, \rho) := \min\{|S| : S \text{ is a } \delta\text{-cover of } \mathcal{X}\}.$$

When $\rho(x, y) = \|x - y\|_2$ is the Euclidean metric in $\mathbb{R}^d$, we abbreviate $\mathcal{N}(\mathcal{X}, \delta, \rho)$ as $\mathcal{N}(\mathcal{X}, \delta)$.

## 4. Sample complexity under group invariance

In this section, we outline our main results for the sample complexity of divergence estimation under group invariance. In particular, we focus on three cases: the Wasserstein-1 metric (18), the MMD (19) and the $(f_\alpha, \Gamma)$-divergence (20). While the convergence rate in the bounds for the Wasserstein-1 metric and the $(f_\alpha, \Gamma)$-divergence depends on the dimension of the ambient space, that for the MMD case does not. In all the numerical experiments, for simplicity, we choose $X = \{x_1, x_2, \cdots, x_m\}$ and $Y = \{y_1, y_2, \cdots, y_n\}$ to sample from the same $\Sigma$-invariant distribution $P = Q$ for easy visualization and clear benchmark.

### 4.1. Wasserstein-1 metric, $W(P, Q)$

In this section, we set $\Gamma = \text{Lip}_L(\mathcal{X})$ to be the set of $L$-Lipschitz functions on $\mathcal{X}$; see Eq. (4). We further assume that the $\Sigma$-actions on $\mathcal{X}$ is 1-Lipschitz, i.e., $\|\sigma x - \sigma y\|_2 \leq \|x - y\|_2, \forall \sigma \in \Sigma, \forall x, y \in \mathcal{X}$, so that $S_\Sigma[\Gamma] \subset \Gamma$ (see Lemma A.3 for a proof). Due to Result 3.1, we have $W(P, Q) = W^\Sigma(P, Q)$ for $\Sigma$-invariant probability measures $P, Q \in \mathcal{P}_\Sigma(\mathcal{X})$.

To convey the main message, we provide a summary of our result in Theorem 4.1 for the sample complexity under group invariance for the Wasserstein-1 metric. The detailed statement and the technical assumption of the theorem as well as its proof are deferred to Appendix A.1. Readers are referred to Section 3 for the notations.

**Theorem 4.1.** *Let $\mathcal{X} = \Sigma \times \mathcal{X}_0$ be a subset of $\mathbb{R}^d$ equipped with the Euclidean distance. Suppose $P, Q \in \mathcal{P}_\Sigma(\mathcal{X})$ are $\Sigma$-invariant distributions on $\mathcal{X}$. If the number $m, n$ of samples drawn from $P$ and $Q$ are sufficiently large, then we have with high probability,*

*1) when $d \geq 2$, for any $s > 0$,*

$$\left|W(P, Q) - W^\Sigma(P_m, Q_n)\right| \leq C\left[\left(\frac{1}{|\Sigma|\, m}\right)^{\frac{1}{d+s}} + \left(\frac{1}{|\Sigma|\, n}\right)^{\frac{1}{d+s}}\right], \quad (26)$$

*where $C > 0$ depends only on $d, s$ and $\mathcal{X}$, and is independent of $m$ and $n$;*

*2) for $d = 1$, we have*

$$\left|W(P, Q) - W^\Sigma(P_m, Q_n)\right| \leq C \cdot \text{diam}(\mathcal{X}_0)\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right), \quad (27)$$

*where $C > 0$ is an absolute constant independent of $\mathcal{X}, \mathcal{X}_0, m$ and $n$.*

*Remark* 4.2. In the case for $d \geq 2$, the $s > 0$ in Theorem 4.1 means the rate can be arbitrarily close to $-\frac{1}{d}$. If we further assume that $\mathcal{X}_0$ is connected, then the bound can be improved to $\left|W(P, Q) - W^\Sigma(P_m, Q_n)\right| \leq C\left[\left(\frac{1}{|\Sigma|m}\right)^{\frac{1}{2}} \ln m + \left(\frac{1}{|\Sigma|n}\right)^{\frac{1}{2}} \ln n\right]$ for $d = 2$, and $\left|W(P, Q) - W^\Sigma(P_m, Q_n)\right| \leq C\left[\left(\frac{1}{|\Sigma|m}\right)^{\frac{1}{d}} + \left(\frac{1}{|\Sigma|n}\right)^{\frac{1}{d}}\right]$ for $d \geq 3$, without the dependence of $s$, which matches the rate in (Fournier & Guillin, 2015). See Remark A.7 after Lemma A.6 in the Appendix.

*Sketch of the proof.* Using the group invariance and the map $T_0$ defined in (23), we can transform the i.i.d. samples on $\mathcal{X}$ to i.i.d. samples on $\mathcal{X}_0$, which are effectively sampled from $P_{\mathcal{X}_0}$ and $Q_{\mathcal{X}_0}$ [cf. Eq. (24)]. Hence the supremum after applying the triangle inequality to the error (26) can be taken over $L$-Lipschitz functions defined on the fundamental domain $\mathcal{X}_0$, i.e., $\text{Lip}_L(\mathcal{X}_0)$, instead of over the original space $\text{Lip}_L(\mathcal{X})$. We further demonstrate in Lemma A.4 that the supremum can be taken over an even smaller function space

$$\mathcal{F}_0 = \{\gamma \in \text{Lip}_L(\mathcal{X}_0) : \|\gamma\|_\infty \leq M\} \subset \text{Lip}_L(\mathcal{X}_0), \quad (28)$$

with some uniformly bounded $L^\infty$-norm $M$ due to the translation-invariance of $\gamma$ in definition (4). Using Dudley's entropy integral (Bartlett et al., 2017), the error can be bounded in terms of the metric entropy of $\mathcal{F}_0$ with $m$ i.i.d. samples,

$$\inf_{\alpha > 0}\left\{8\alpha + \frac{24}{\sqrt{m}} \int_\alpha^M \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta\right\}. \quad (29)$$

For $d \geq 2$, we establish the relations between the metric entropy, $\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)$, of $\mathcal{F}_0$ and the covering numbers of $\mathcal{X}_0$ and $\mathcal{X}$ via Lemma A.6 and Lemma A.8:

$$\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty) \leq \mathcal{N}(\mathcal{X}_0, \frac{c_2\delta}{L}) \ln(\frac{c_1 M}{\delta}), \quad (30)$$

$$\frac{\mathcal{N}(\mathcal{X}_0, \delta)}{\mathcal{N}(\mathcal{X}, \delta)} \leq \frac{1}{|\Sigma|}, \text{ for small enough } \delta, \quad (31)$$

which yields a factor in terms of the group size $|\Sigma|$ in Eq. (26). The dominant term of the bound based on the singularity of the entropy integral at $\alpha = 0$ is shown in Eq. (26). For $d = 1$, the entropy integral is not singular at the origin, and we bound the covering number of $\mathcal{F}_0$ by $\text{diam}(\mathcal{X}_0)$ instead. The probability bound is from the application of the McDiarmid's inequality.

*Remark* 4.3. Even though we present in Theorem 4.1 only the dominant terms showing the rate of convergence for

the estimator, our result for sample complexity is actually non-asymptotic. See Theorem A.2 in Appendix A.1 for a complete description of the result.

*Remark* 4.4. When $|\Sigma| = 1$, i.e., no group symmetry is leveraged in the divergence estimation, our result reduces to the case considered in, e.g., (Sriperumbudur et al., 2012), for general distributions $P, Q \in \mathcal{P}_\Sigma(\mathcal{X}) = \mathcal{P}(\mathcal{X})$.

*Remark* 4.5. The factor $\mathrm{diam}(\mathcal{X}_0)$ in the case for $d = 1$ is not necessarily directly related to the group size $|\Sigma|$. We refer to Example 4.6 below and its explanation in Remark A.10 for cases where we can achieve a factor of $|\Sigma|^{-1}$ in the rate.

**Example 4.6.** Let $\mathcal{X} = [0, 1)$ and $\Gamma = \mathrm{Lip}_L([0, 1))$, i.e., $d = 1$. We consider the $\Sigma$-actions on $\mathcal{X}$ generated by the translation $x \mapsto (x + \frac{1}{r}) \mod 1$, where $r = 1, 4, 16, 64, 256$, so that $|\Sigma| = r$ is the group size. We draw samples $x_i \sim P = Q \in \mathcal{P}_\Sigma(\mathcal{X})$ on $\mathcal{X}$ in the following way: $x_i = r^{-1}\xi_i^{1/3} + \eta_i$, where $\xi_i$ are i.i.d. uniformly distributed random variables on $[0, 1)$ and $\eta_i$ take values over $\{0, \frac{1}{r}, \ldots, \frac{r-1}{r}\}$ with equal probabilities. One can easily verify that $P = Q$ are indeed $\Sigma$-invariant. The numerical results for the empirical estimation of $W(P, Q) = 0$ using $W^\Sigma(P_n, Q_n)$ with different group size $|\Sigma| = r$, $r = 1, 4, 16, 64, 256$, are shown in the left panel of Figure 3. One can clearly observe a significant improvement of the estimator as the group size $|\Sigma|$ increases. Furthermore, the right panel of Figure 3 displays the ratios between the adjacent curves, all of which converge to 4, which is the ratio between the consecutive group size. This matches our calculation in Remark A.10; see also Remark 4.5.

**Example 4.7.** We let $\mathcal{X} = \mathbb{R}^2$, i.e., $d = 2$. The probability distributions $P = Q$ are the mixture of 8 Gaussians centered at $\big(\cos(\frac{2\pi r}{8}), \sin(\frac{2\pi r}{8})\big)$, $r = 0, 1, \ldots, 7$, with the same covariance. The distribution has $C_8$-rotation symmetry, but we pretend that it is only $C_1, C_2$ and $C_4$; that is, the $\Sigma$ used in the empirical estimation $W^\Sigma(P_m, Q_n)$ does not reflect the entire invariance structure. Even though in this case the domain $\mathcal{X}$ is unbounded, which is beyond our theoretical assumptions, we can still see in Figure 4 that as we increase the group size $|\Sigma|$ in the computation of $W^\Sigma(P_m, Q_n)$, fewer samples are needed to reach the same accuracy level in the approximation. The ratios between adjacent curves in this case are slightly above the predicted value $\sqrt{2} \approx 1.414$ according to our theory (see Remark 4.2), suggesting that the complexity bound could be further improved. For instance, in (Sriperumbudur et al., 2012), a logarithmic correction term can be revealed for $d = 2$ after a more thorough analysis.

### 4.2. Lipschitz-regularized $\alpha$-divergence, $D^\Gamma_{f_\alpha}(P\|Q)$

The Lipschitz-regularized $\alpha$-divergence is used in the symmetry-preserving GANs (Birrell et al., 2022c), where it

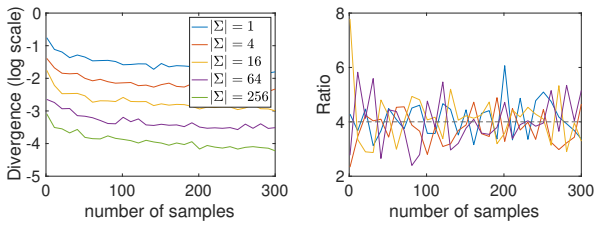


*Figure 3.* Left: the Wasserstein-1 distance with different group sizes on $[0, 1)$, averaged over 10 replicas. Right: the ratio of the average of the Wasserstein-1 distance between different group sizes: $|\Sigma| = 1$ over $|\Sigma| = 4$ (blue), $|\Sigma| = 4$ over $|\Sigma| = 16$ (red), $|\Sigma| = 16$ over $|\Sigma| = 64$ (orange), $|\Sigma| = 64$ over $|\Sigma| = 256$ (purple). The black horizontal dashed line refers to the ratio equal to 4, which is the value theoretically predicted in Theorem 4.1 for $d = 1$. See Example 4.6 and Remark A.10 for the detail.

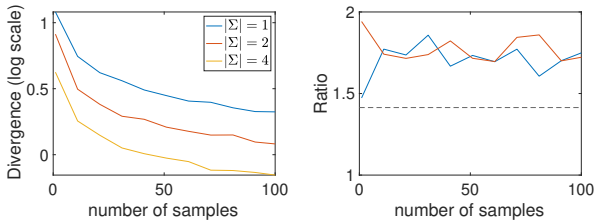


*Figure 4.* Left: The Wasserstein-1 distance assuming different group sizes in $\mathbb{R}^2$, averaged over 10 replicas. Right: the ratio of the average of the Wasserstein-1 distance between different group sizes: $|\Sigma| = 1$ over $|\Sigma| = 2$ (blue), $|\Sigma| = 2$ over $|\Sigma| = 4$ (red). The black horizontal dashed line refers to the ratio equal to $\sqrt{2}$, which is the value theoretically predicted in Theorem 4.1 for $d = 2$. The ratios are slightly above the reference line, suggesting that the complexity bound could be further improved. See Example 4.7 and Remark 4.2 for the detail.

allowed them to systematically include symmetries and gave a vastly improved performance on real data sets. The space $\Gamma$ in this section is always set to $\Gamma = \mathrm{Lip}_L(\mathcal{X})$; see Eq. (7). We only consider $\alpha > 1$, as the case when $0 < \alpha < 1$ can be derived in a similar manner. For $\alpha > 1$, the Legendre transform of $f_\alpha$, which is defined in (7), is

$$f^*_\alpha(y) = \left(\alpha^{-1}(\alpha - 1)^{\frac{\alpha}{\alpha-1}} y^{\frac{\alpha}{\alpha-1}} + \frac{1}{\alpha(\alpha - 1)}\right) \mathbf{1}_{y>0}.$$

We provide a theorem for the sample complexity for the $(f_\alpha, \Gamma)$-divergence under group invariance, whose detailed statement and proof can be found in Appendix A.2. We note that this is a new sample complexity result for the $(f_\alpha, \Gamma)$-divergence even without the group structure, which is still missing in the literature.

**Theorem 4.8.** *Let $\mathcal{X} = \Sigma \times \mathcal{X}_0$ be a subset of $\mathbb{R}^d$ equipped with the Euclidean distance. Let $f_\alpha(x) = \frac{x^\alpha - 1}{\alpha(\alpha-1)}$, $\alpha > 1$*

*and $\Gamma = Lip_L(\mathcal{X})$. Suppose $P$ and $Q$ are $\Sigma$-invariant distributions on $\mathcal{X}$. If the number of samples $m, n$ drawn from $P$ and $Q$ are sufficiently large, we have with high probability,*

*1) when $d \geq 2$, for any $s > 0$ ,*

$$\left|D_{f_\alpha}^{\Gamma}(P\|Q) - D_{f_\alpha}^{\Gamma_\Sigma}(P_m\|Q_n)\right| \leq C_1 \left(\frac{1}{|\Sigma|\, m}\right)^{\frac{1}{d+s}} + C_2 \left(\frac{1}{|\Sigma|\, n}\right)^{\frac{1}{d+s}}, \quad (32)$$

*where $C_1$ depends only on $d, s$ and $\mathcal{X}$; $C_2$ depends only on $d, s$, $\mathcal{X}$ and $\alpha$; both $C_1$ and $C_2$ are independent of $m$ and $n$;*

*2) for $d = 1$, we have*

$$\left|D_{f_\alpha}^{\Gamma}(P\|Q) - D_{f_\alpha}^{\Gamma_\Sigma}(P_m\|Q_n)\right| \leq \operatorname{diam}(\mathcal{X}_0)\left(\frac{C_1}{\sqrt{m}} + \frac{C_2}{\sqrt{n}}\right), \quad (33)$$

*where $C_1$ and $C_2$ are independent of $\mathcal{X}_0$, $m$ and $n$; $C_2$ depends on $\alpha$.*

*Sketch of the proof.* The idea is similar to the proof of Theorem 4.1. The only difference is that we need to tackle the $f_\alpha^*(\gamma)$ term separately, since it is not translation-invariant in $\gamma$. Using the equivalent form (8), we can obtain a different Lipschitz constant associated with $f_\alpha^*$, as well as a different $L^\infty$ bound $M$ than that in Eq. (28) by Lemma A.12. This results in the $\alpha$ dependence for $C_2$.

### 4.3. Maximum mean discrepancy, MMD$(P, Q)$

Though one can utilize the results on the covering number of the unit ball of a reproducing kernel Hilbert space, e.g. (Zhou, 2002; Kühn, 2011), to derive the sample complexity bounds that depend on the dimension $d$, we provide a dimension-independent bound as in (Gretton et al., 2012) without the use of the covering numbers. In the MMD case, we let $B_{\mathcal{H}}$ represent the unit ball in some reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ on $\mathcal{X}$; see Eq. (5). In addition, we make the following assumptions for the kernel $k(x, y)$.

**Assumption 4.9.** The kernel $k(x, y)$ for $\mathcal{H}$ satisfies

- $k(x, y) \geq 0$ and $k(\sigma(x), \sigma(y)) = k(x, y), \forall \sigma \in \Sigma, x, y \in \mathcal{X}$;
- Let $K := \max_{x,y\in\mathcal{X}} k(x, y)$, then $k(x, y) = K$ if and only if $x = y$;
- There exists $c_{\Sigma,k} \in (0, 1)$ such that for any $\sigma \in \Sigma$ and $\sigma$ is not the identity element and $x \in \mathcal{X}_0$, we have $k(\sigma x, x) \leq c_{\Sigma,k} K$.

Intuitively, the third condition in Assumption 4.9 suggests uniform decay of the kernel on the group orbits. See Remark 4.12 and Example 4.13 for more details and a related example.

From Lemma C.1 in (Birrell et al., 2022c), we know $S_\Sigma[\Gamma] \subset \Gamma$ by the first assumption. Below is an abbreviated result for the sample complexity for the MMD, whose detailed statement and proof can be found in Appendix A.3.

**Theorem 4.10.** *Let $\mathcal{X} = \Sigma \times \mathcal{X}_0$ be a subset of $\mathbb{R}^d$. $\mathcal{H}$ is a RKHS on $\mathcal{X}$ whose kernel satisfies Assumption 4.9. Suppose $P$ and $Q$ are $\Sigma$-invariant distributions on $\mathcal{X}$. Then for $m, n$ sufficiently large, we have with high probability,*

$$\left|\mathrm{MMD}(P, Q) - \mathrm{MMD}^\Sigma(P_m, Q_n)\right| < O\left(C_{\Sigma,k}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right)\right), \quad (34)$$

*where $C_{\Sigma,k} = \sqrt{\frac{1+c_{\Sigma,k}(|\Sigma|-1)}{|\Sigma|}}$, and $c_{\Sigma,k}$ is the constant in Assumption 4.9.*

*Sketch of the proof.* Based on Result 3.1, we use the equality $\mathrm{MMD}^\Sigma(P_m, Q_n) = \mathrm{MMD}(S^\Sigma[P_m], S^\Sigma[Q_n])$ to expand the divergence over all the orbit elements. The error bound is controlled in terms of the Rademacher average, whose supremum is attained at some known witness function due to the structure of the RKHS using Lemma A.14. Since the Rademacher average is estimated without covering numbers, the rate is independent of the dimension $d$. Then we use the decay of the kernel to obtain the bound.

*Remark* 4.11. When $|\Sigma| = 1$, the proof is reduced to that in (Sriperumbudur et al., 2012).

*Remark* 4.12. Unlike the cases for the Wasserstein metric and the Lipschitz-regularized $\alpha$-divergence in Theorem 4.1 and Theorem 4.8, the improvement of the sample complexity under group symmetry for MMD (measured by $C_{\Sigma,k}$ in Theorem 4.10) depends on not only the group size $|\Sigma|$ but also the kernel $k(x, y)$. For a fixed $\mathcal{X}$ and kernel $k(x, y)$, simply increasing the group size $|\Sigma|$ does not necessarily lead to a reduced sample complexity beyond a certain threshold; see the first four subfigures in Figure 5. However, we show in Example 4.13 below that, by adaptively picking a suitable kernel $k$ depending on the group size $|\Sigma|$, one can obtain an improvement in sample complexity by $C_{\Sigma,k} \approx \frac{1}{\sqrt{|\Sigma|}}$ for arbitrarily large $|\Sigma|$.

**Example 4.13.** Let $\mathcal{X} = \{(r\cos\theta, r\sin\theta) \in \mathbb{R}^2 : r \in [0, 1], \theta \in [0, 2\pi)\}$ be the unit disk centered at the origin, and let $k_s(x, y) = e^{-\frac{\|x-y\|_2^2}{2s^2}}$, $x, y \in \mathcal{X}$, be the Gaussian kernel. Consider the group actions generated by a rotation (with respect to the origin) of $\frac{2\pi}{l}$, $l = 1, 4, 16, 64, 256$, so that $|\Sigma| = l$ is the group size. The fundamental domain $\mathcal{X}_0$ under the $\Sigma$-action is $\mathcal{X}_0 = [0, 1] \times [0, \frac{2\pi}{l})$ (see Figure 2

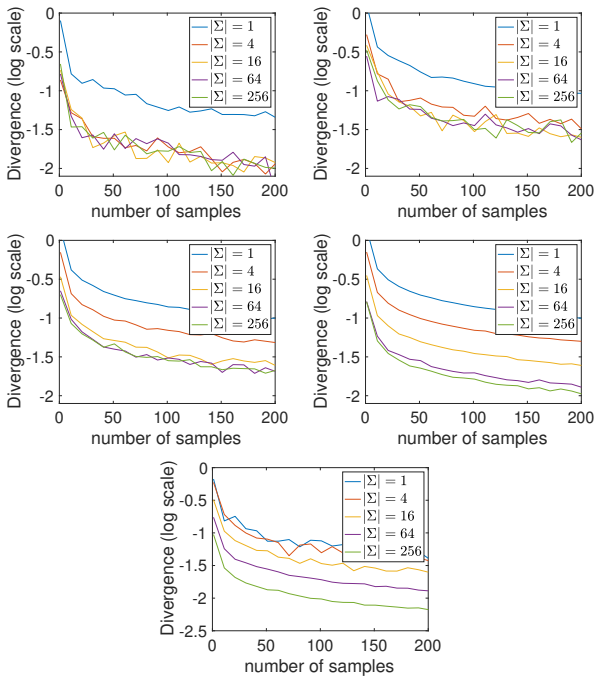


*Figure 5.* MMD simulations with Gaussian kernels $k_s(x,y) = e^{-\frac{\|x-y\|_2^2}{2s^2}}$. From left to right, top to bottom: $s = \frac{2\pi}{1\times 6}$, $s = \frac{2\pi}{4\times 6}$, $s = \frac{2\pi}{16\times 6}$, $s = \frac{2\pi}{64\times 6}$, $s = \frac{2\pi}{6|\Sigma|}$. The first four subfigures (top two rows) show that the Gaussian kernel with a fixed bandwidth $s > 0$ satisfies the third condition in Assumption 4.9 up to a group size of $|\Sigma| = l, l = 1, 4, 16, 64$, and thus an improvement of sample complexity of order $C_{\Sigma,k} \approx |\Sigma|^{-1/2}$ persists till $|\Sigma| = l$; when $|\Sigma| > l$, no further reduction in sample complexity can be observed. The last subfigure demonstrates that with an adaptive bandwidth $s$ inversely scaled with $|\Sigma|$, nonstop improvement of the sample complexity can be achieved as the group size $|\Sigma|$ increases. See Example 4.13 for the detail and explanations.

for a visual illustration). We draw samples $x_i \sim P = Q \in \mathcal{P}_\Sigma(\mathcal{X})$ in the following way,

$$x_i = \sqrt{\xi_i}\left(\cos\left[\frac{2\pi}{l}\theta_i^{1/3} + \eta_i\frac{2\pi}{l}\right], \sin\left[\frac{2\pi}{l}\theta_i^{1/3} + \eta_i\frac{2\pi}{l}\right]\right)$$

where $\xi_i$ and $\theta_i$ are i.i.d. uniformly distributed random variables on $[0, 1)$ and $\eta_i$ take values over $\{0, \frac{1}{l}, \ldots, \frac{l-1}{l}\}$ with equal probabilities. We select the kernel bandwidth $s > 0$ in different ways:

- Fixed $s$ with changing group size $|\Sigma| = l$. We intuitively follow the "three-sigma rule" in the argument direction to pick different $s$. Since the angle of each sector is $\frac{2\pi}{l}$, we select $s = \frac{2\pi}{6l}$, $l = 1, 4, 16, 64$. Smaller bandwidth $s$ corresponds to faster decay of the kernel $k_s(x, y)$, such that for a fixed bandwidth $s = \frac{2\pi}{6l}$, the third condition in 4.9 is satisfied with a small $c_k$ for any group $\Sigma$ such that $|\Sigma| \le l$, i.e., $C_{\Sigma,k} \approx |\Sigma|^{-1/2}$. On the other hand, it is difficult to observe the improvement by further increasing the group size $|\Sigma|$ beyond $|\Sigma| > l$, since the third condition in 4.9 is not satisfied with any uniformly small $c$. See the top two rows in Figure 5 for the results for $s = \frac{2\pi}{l\times 6}$, $l = 1, 4, 16, 64$. Notice that the sample complexity improvement stops right at $|\Sigma| = l$, perfectly matching our theoretical result Theorem 4.10.
- $s$ inversely scales with $|\Sigma|$, i.e., $s = \frac{2\pi}{6|\Sigma|}$. Unlike the fixed $s$ discusses previously, with these adaptive selections of kernels, we can observe nonstop improvement of the sample complexity as the group size $|\Sigma|$ increases; see the last row of Figure 5. This numerical result is explained by the third condition in Assumption 4.9; that is, in order to continuously observe the benefit from the increasing group size $|\Sigma|$, we need to have a faster decay in the kernel $k_s$ (i.e., smaller $s$) so that $c_{\Sigma,k_s}$ is uniformly small for all $|\Sigma|$.

*Remark* 4.14. The bound provided in Theorem 4.10 for the MMD case is almost sharp in the sense that, by a direct calculation, one can obtain that

$$\frac{E_{\mathcal{X}}\mathrm{MMD}^\Sigma(P, P_n)^2}{E_{\mathcal{X}}\mathrm{MMD}(P, P_n)^2} \approx C_{\Sigma,k}^2,$$

if the kernel bandwidth $s \propto \frac{\sqrt{2c_{\Sigma,k}\pi}}{|\Sigma|}$.

## 5. Conclusion and future work

We provide rigorous analysis to quantify the reduction in sample complexity for variational divergence estimations between group-invariant distributions. We obtain a reduction on the error bound by a power of the group size. The exponent on the group size depends on the ambient dimension for the Wasserstein-1 metric and the Lipschitz-regularized $\alpha$-divergence; that exponent, however, is independent of the ambient dimension for the MMD with a proper choice of the kernel.

This work also motivates some possible future directions. For the Wasserstein-1 metric in $\mathbb{R}^2$, one could potentially derive a sharper bound in terms of the group size. For the MMD with Gaussian kernels, it is worth investigating how to choose the bandwidth to make as much use of the group structure as possible. Further applications of the theories on machine learning, such as neural generative models or neural estimations of divergence under symmetry, are also expected.

## Acknowledgements

The research of Z.C., M.K. and L.R.-B. was partially supported by the Air Force Office of Scientific Research

(AFOSR) under the grant FA9550-21-1-0354. The research of M. K. and L.R.-B. was partially supported by the National Science Foundation (NSF) under the grants DMS-2008970 and TRIPODS CISE-1934846. The research of Z.C and W.Z. was partially supported by NSF under DMS-2052525 and DMS-2140982. We thank Yulong Lu for the insightful discussions.

## References

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

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/belghazi18a.html.

Bickel, P. J. A distribution free version of the smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, 40(1):1–23, 1969.

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., Dupuis, P., Katsoulakis, M. A., Rey-Bellet, L., and Wang, J. Variational representations and neural network estimation of Rényi divergences. *SIAM Journal on Mathematics of Data Science*, 3(4):1093–1116, 2021.

Birrell, J., Dupuis, P., Katsoulakis, M. A., Pantazis, Y., and Rey-Bellet, L. $(f, \Gamma)$-Divergences: Interpolating between $f$-Divergences and Integral Probability Metrics. *Journal of Machine Learning Research*, 23(39):1–70, 2022a.

Birrell, J., Katsoulakis, M. A., and Pantazis, Y. Optimizing variational representations of divergences and accelerating their statistical estimation. *IEEE Transactions on Information Theory*, 2022b.

Birrell, J., Katsoulakis, M. A., Rey-Bellet, L., and Zhu, W. Structure-preserving GANs. *Proceedings of the 39th International Conference on Machine Learning, PMLR 162*, pp. 1982–2020, 2022c.

Birrell, J., Pantazis, Y., Dupuis, P., Katsoulakis, M. A., and Rey-Bellet, L. Function-space regularized Rényi divergences. *arXiv preprint arXiv:2210.04974*, 2022d.

Boyda, D., Kanwar, G., Racanière, S., Rezende, D. J., Albergo, M. S., Cranmer, K., Hackett, D. C., and Shanahan, P. E. Sampling using su (n) gauge equivariant flows. *Physical Review D*, 103(7):074504, 2021.

Broniatowski, M. and Keziou, A. Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100(1):16 – 36, 2009. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2008.03.011. URL http://www.sciencedirect.com/science/article/pii/S0047259X08001036.

Catoni, O., Euclid, P., Library, C. U., and Press, D. U. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Lecture notes-monograph series. Cornell University Library, 2008. URL https://books.google.gr/books?id=-EtrnQAACAAJ.

Cheng, X. and Xie, Y. Kernel two-sample tests for manifold data. *arXiv preprint arXiv:2105.03425*, 2021.

Chowdhary, K. and Dupuis, P. Distinguishing and integrating aleatoric and epistemic variation in uncertainty quantification. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47(3):635–662, 2013. doi: 10.1051/m2an/2012038.

Ciompi, F., Jiao, Y., and van der Laak, J. Lymphocyte assessment hackathon (LYSTO), October 2019. URL https://doi.org/10.5281/zenodo.3513571.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/cohenc16.html.

Cohen, T. S., Geiger, M., and Weiler, M. A general theory of equivariant CNNs on homogeneous spaces. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/b9cfe8b6042cf759dc4c0cccb27a6737-Paper.pdf.

Dey, N., Chen, A., and Ghafurian, S. Group equivariant generative adversarial networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rgFNuJHHXv.

Dupuis, P. and Ellis, R. S. *A weak convergence approach to the theory of large deviations*, volume 902. John Wiley & Sons, 2011.

Dupuis, P., Katsoulakis, M. A., Pantazis, Y., and Plechac, P. Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1): 80–111, 2016. doi: 10.1137/15M1025645.

Folland, G. B. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

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.

Gao, S., Ver Steeg, G., and Galstyan, A. Efficient estimation of mutual information for strongly dependent variables. In *Artificial intelligence and statistics*, pp. 277–286. PMLR, 2015.

Garcia Satorras, V., Hoogeboom, E., Fuchs, F., Posner, I., and Welling, M. E (n) equivariant normalizing flows. *Advances in Neural Information Processing Systems*, 34, 2021.

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 nets. *Advances in neural information processing systems*, 27, 2014.

Gottlieb, L.-A., Kontorovich, A., and Krauthgamer, R. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Transactions on Information Theory*, 63(8):4838–4849, 2017.

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.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of Wasserstein GANs. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

Hyvarinen, A., Karhunen, J., and Oja, E. Independent component analysis. *Studies in informatics and control*, 11(2):205–207, 2002.

Kandasamy, K., Krishnamurthy, A., Poczos, B., Wasserman, L., et al. Nonparametric von mises estimators for entropies, divergences and mutual informations. *Advances in Neural Information Processing Systems*, 28, 2015.

Kinney, J. B. and Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9): 3354–3359, 2014.

Kipnis, C. and Landim, C. *Scaling Limits of Interacting Particle Systems*. Springer-Verlag, 1999.

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

Kolmogorov, A. N. $\varepsilon$-entropy and $\varepsilon$-capacity of sets in function spaces. *American Mathematical Society Translations*, 17(2):277–364, 1961.

Kühn, T. Covering numbers of gaussian reproducing kernel hilbert spaces. *Journal of Complexity*, 27(5):489–499, 2011.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

Liu, J., Kumar, A., Ba, J., Kiros, J., and Swersky, K. Graph normalizing flows. *Advances in Neural Information Processing Systems*, 32, 2019.

McAllester, D. A. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, pp. 164–170, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581131674. doi: 10.1145/307400.307435. URL https://doi.org/10.1145/307400.307435.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

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

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Nonparametric estimation of the likelihood ratio and divergence functionals. In *2007 IEEE International Symposium on Information Theory*, pp. 2016–2020. IEEE, 2007.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Nietert, S., Goldfeld, Z., and Kato, K. Smooth $p$-wasserstein distance: Structure, empirical approximation, and statistical applications. In *International Conference on Machine Learning*, pp. 8172–8183. PMLR, 2021.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

Paninski, L. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Póczos, B., Xiong, L., and Schneider, J. Nonparametric divergence estimation with applications to machine learning on distributions. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 599–608, 2011.

Rezende, D. J., Racanière, S., Higgins, I., and Toth, P. Equivariant hamiltonian flows. *arXiv preprint arXiv:1909.13739*, 2019.

Ruderman, A., Reid, M. D., García-García, D., and Petterson, J. Tighter variational representations of f-divergences via restriction to probability measures. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1155–1162, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Shawe-Taylor, J. and Williamson, R. C. A PAC analysis of a Bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, COLT '97, pp. 2–9, New York, NY, USA, 1997. Association for Computing Machinery. ISBN 0897918916. doi: 10.1145/267460.267466. URL https://doi.org/10.1145/267460.267466.

Sokolic, J., Giryes, R., Sapiro, G., and Rodrigues, M. Generalization error of invariant classifiers. In *Artificial Intelligence and Statistics*, pp. 1094–1103. PMLR, 2017.

Sreekumar, S. and Goldfeld, Z. Neural estimation of statistical divergences. *Journal of Machine Learning Research*, 23(126):1–75, 2022.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkL7n1-0b.

Van Handel, R. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Weiler, M. and Cesa, G. General E(2)-equivariant steerable CNNs. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/45d6637b718d0f24a237069fe41b0db4-Paper.pdf.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.

Zhao, S., Liu, Z., Lin, J., Zhu, J.-Y., and Han, S. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33: 7559–7570, 2020.

Zhou, D.-X. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

# A. Theorems and Proofs

In this section, we provide detailed statements of the theorems introduced in the main text as well as their proofs.

## A.1. Wasserstein-1 metric

**Assumption A.1.** Let $\mathcal{X} = \Sigma \times \mathcal{X}_0 \subset \mathbb{R}^d$. Assume that there exists some $\delta_0 > 0$ such that

1) $\|\sigma(x) - \sigma'(x')\|_2 > 2\delta_0, \forall x, x' \in \mathcal{X}_0, \sigma \neq \sigma' \in \Sigma$; and
2) $\|\sigma(x) - \sigma(x')\|_2 \geq \|x - x'\|_2, \forall x, x' \in \mathcal{X}_0, \sigma \in \Sigma$,

where $\|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^d$.

Example 4.6 provides a simple example when this assumption holds.

**Theorem A.2.** *Let $\mathcal{X} = \Sigma \times \mathcal{X}_0$ be a subset of $\mathbb{R}^d$ satisfying the conditions in Assumption A.1. Suppose $P$ and $Q$ are $\Sigma$-invariant probability measures on $\mathcal{X}$.*

*1) If $d \geq 2$, then for any $s > 0, \epsilon > 0$ and $m, n$ sufficiently large, we have with probability at least $1 - \epsilon$,*

$$\left|W(P,Q) - W^{\Sigma}(P_m, Q_n)\right| \leq \left(8 + \frac{24}{(\frac{d+s}{2} - 1)}\right)\left[\left(\frac{9D^2_{\mathcal{X},L}}{|\Sigma|\, m}\right)^{\frac{1}{d+s}} + \left(\frac{9D^2_{\mathcal{X},L}}{|\Sigma|\, n}\right)^{\frac{1}{d+s}}\right] + \bar{D}_{\mathcal{X}_0,L}\left(\frac{24}{\sqrt{m}} + \frac{24}{\sqrt{n}}\right) + L \cdot \mathrm{diam}(\mathcal{X}_0)\sqrt{\frac{2(m+n)}{mn}\ln\frac{1}{\epsilon}},$$

*where $D_{\mathcal{X},L}$ depends only on $\mathcal{X}$ and $L$; $\bar{D}_{\mathcal{X}_0,L}$ depends only on $\mathcal{X}_0$ and $L$, and is increasing in $\mathcal{X}_0$, i.e., $\bar{D}_{A_1,L} \leq \bar{D}_{A_2,L}$ for $A_1 \subset A_2$;*
*2) If $d = 1$, then for any $\epsilon > 0$ and $m, n$ sufficiently large, we have with probability at least $1 - \epsilon$,*

$$\left|W(P,Q) - W^{\Sigma}(P_m, Q_n)\right| \leq cL \cdot \mathrm{diam}(\mathcal{X}_0)\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) + L \cdot \mathrm{diam}(\mathcal{X}_0)\sqrt{\frac{2(m+n)}{mn}\ln\frac{1}{\epsilon}},$$

*where $c > 0$ is an absolute constant independent of $\mathcal{X}$ and $\mathcal{X}_0$.*

Before proving this theorem, we have the following lemmas.

**Lemma A.3.** *Suppose the $\Sigma$-actions on $\mathcal{X}$ are 1-Lipschitz, i.e., $\|\sigma x - \sigma y\|_2 \leq \|x - y\|_2$ for any $x, y \in \mathcal{X}$ and $\sigma \in \Sigma$, then we have $S_\Sigma[\Gamma] \subset \Gamma$, where $\Gamma = Lip_L(\mathcal{X})$.*

*Proof.* For any $x, y \in \mathcal{X}$ and $f \in \Gamma$, we have

$$\begin{aligned}
|S_\Sigma(f)(x) - S_\Sigma(f)(y)| &= \left|\frac{1}{|\Sigma|}\sum_{\sigma\in\Sigma} f(\sigma x) - \frac{1}{|\Sigma|}\sum_{\sigma\in\Sigma} f(\sigma y)\right| \\
&\leq \frac{1}{|\Sigma|}\sum_{\sigma\in\Sigma} |f(\sigma x) - f(\sigma y)| \\
&\leq \frac{1}{|\Sigma|}\sum_{\sigma\in\Sigma} L\,\|\sigma x - \sigma y\|_2 \\
&\leq \frac{1}{|\Sigma|}\sum_{\sigma\in\Sigma} L\,\|x - y\|_2 \\
&= L\,\|x - y\|_2 .
\end{aligned}$$

Therefore, we have $S_\Sigma(f) \in \Gamma$. □

**Lemma A.4.** *For any $\gamma \in Lip_L(\mathcal{X}_0)$, there exists $\nu \in \mathbb{R}$, such that $\|\gamma + \nu\|_\infty \leq L \cdot \text{diam}(\mathcal{X}_0)$.*

*Proof.* Suppose $\gamma \in \text{Lip}_L(\mathcal{X}_0)$ and $\|\gamma(x)\|_\infty > L \cdot \text{diam}(\mathcal{X}_0)$. Without loss of generality, we can assume $\sup_{x\in\mathcal{X}_0} \gamma(x) > L \cdot \text{diam}(\mathcal{X}_0)$. Since $\gamma$ is $L$-Lipschitz on $\mathcal{X}_0$, we have $\sup_{x\in\mathcal{X}_0} \gamma(x) - \inf_{x\in\mathcal{X}_0} \gamma(x) \leq L \cdot \text{diam}(\mathcal{X}_0)$, so that

$$\inf_{x\in\mathcal{X}_0} \gamma(x) \geq \sup_{x\in\mathcal{X}_0} \gamma(x) - L \cdot \text{diam}(\mathcal{X}_0) > 0.$$

Hence we can select $\nu = -\frac{\inf_{x\in\mathcal{X}_0} \gamma(x)}{2}$, so that $\|\gamma + \nu\|_\infty < \|\gamma\|_\infty$. □

We provide a variant of the Dudley's entropy integral as well as its proof for completeness.

**Lemma A.5.** *Suppose $\mathcal{F}$ is a family of functions mapping the metric space $(\mathcal{X}, \rho)$ to $[-M, M]$ for some $M > 0$. Also assume that $0 \in \mathcal{F}$ and $\mathcal{F} = -\mathcal{F}$. Let $\xi = \{\xi_1, \ldots, \xi_m\}$ be a set of independent random variables that take values on $\{-1, 1\}$ with equal probabilities, $i = 1, \ldots, m$. $x_1, x_2, \ldots, x_m \in \mathcal{X}$. Then we have*

$$E_\xi \sup_{f\in\mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^m \xi_i f(x_i) \right| \leq \inf_{\alpha>0} 4\alpha + \frac{12}{\sqrt{m}} \int_\alpha^M \sqrt{\ln \mathcal{N}(\mathcal{F}, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta.$$

The proof of Lemma A.5 is standard using the dyadic path., e.g. see the proof of Lemma A.5. in (Bartlett et al., 2017).

*Proof.* Let $N$ be an arbitrary positive integer and $\delta_k = M2^{-(k-1)}$, $k = 1, \ldots, N$. Let $V_k$ be the cover achieving $\mathcal{N}(\mathcal{F}, \delta_k, \|\cdot\|_\infty)$ and denote $|V_k| = \mathcal{N}(\mathcal{F}, \delta_k, \|\cdot\|_\infty)$. For any $f \in \mathcal{F}$, let $\pi_k(f) \in V_k$, such that $\|f - \pi_k(f)\|_\infty \leq \delta_k$. We have

$$\begin{aligned}
&E_\xi \sup_{f\in\mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^m \xi_i f(x_i) \right| \\
&\leq E_\xi \sup_{f\in\mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^m \xi_i \left( f(x_i) - \pi_N(f)(x_i) \right) \right| + \sum_{j=1}^{N-1} E_\xi \sup_{f\in\mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^m \xi_i \left( \pi_{j+1}(f)(x_i) - \pi_j(f)(x_i) \right) \right| \\
&\quad + E_\xi \sup_{f\in\mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^m \xi_i \pi_1(f)(x_i) \right|.
\end{aligned}$$

The first on the right hand side is bounded by $\delta_N$. Note that we can choose $V_1 = \{0\}$, so that $\pi_1(f)$ is the zero function. For each $j$, let $W_j = \{\pi_{j+1}(f) - \pi_j(f) : f \in \mathcal{F}\}$. We have $|W_j| \leq |V_{j+1}|\,|V_j| \leq |V_{j+1}|^2$. Then we have

$$\sum_{j=1}^{N-1} E_\xi \sup_{f\in\mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^m \xi_i \left( \pi_{j+1}(f)(x_i) - \pi_j(f)(x_i) \right) \right| = \sum_{j=1}^{N-1} E_\xi \sup_{w\in W_j} \left| \frac{1}{m} \sum_{i=1}^m \xi_i w(x_i) \right|.$$

In addition, we have

$$\begin{aligned}
&\sup_{w\in W_j} \sqrt{\sum_{i=1}^m w(x_i)^2} \\
&= \sup_{f\in\mathcal{F}} \sqrt{\sum_{i=1}^m \left( \pi_{j+1}(f)(x_i) - \pi_j(f)(x_i) \right)^2} \\
&\leq \sup_{f\in\mathcal{F}} \sqrt{\sum_{i=1}^m \left( \pi_{j+1}(f)(x_i) - f(x_i) \right)^2} + \sup_{f\in\mathcal{F}} \sqrt{\sum_{i=1}^m \left( f(x_i) - \pi_j(f)(x_i) \right)^2} \\
&\leq \sqrt{m \cdot \delta_{j+1}^2} + \sqrt{m \cdot \delta_j^2} \\
&= \sqrt{m}(\delta_{j+1} + \delta_j)
\end{aligned}$$

$$= 3\sqrt{m}\delta_{j+1}.$$

By the Massart finite class lemma (see, e.g. (Mohri et al., 2018)), we have

$$E_\xi \sup_{w\in W_j}\left|\frac{1}{m}\sum_{i=1}^m \xi_i w(x_i)\right| \le \frac{3\sqrt{m}\delta_{j+1}\sqrt{2\ln|W_j|}}{m} \le \frac{6\delta_{j+1}\sqrt{\ln|V_{j+1}|}}{\sqrt{m}}.$$

Therefore,

$$\begin{aligned} E_\xi \sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \xi_i f(x_i)\right| &\le \delta_N + \frac{6}{\sqrt{m}}\sum_{j=1}^{N-1}\delta_{j+1}\sqrt{\ln\mathcal{N}(\mathcal{F},\delta_{j+1},\|\cdot\|_\infty)} \\ &\le \delta_N + \frac{12}{\sqrt{m}}\sum_{j=1}^{N}(\delta_j-\delta_{j+1})\sqrt{\ln\mathcal{N}(\mathcal{F},\delta_j,\|\cdot\|_\infty)} \\ &\le \delta_N + \frac{12}{\sqrt{m}}\int_{\delta_{N+1}}^M \sqrt{\ln\mathcal{N}(\mathcal{F},\delta,\|\cdot\|_\infty)}\,\mathrm{d}\delta. \end{aligned}$$

Finally, select any $\alpha\in(0,M)$ and let $N$ be the largest integer with $\delta_{N+1}>\alpha$, (implying $\delta_{N+2}\le\alpha$ and $\delta_N = 4\delta_{N+2}\le 4\alpha$), so that

$$\delta_N + \frac{12}{\sqrt{m}}\int_{\delta_{N+1}}^M \sqrt{\ln\mathcal{N}(\mathcal{F},\delta_j,\|\cdot\|_\infty)}\,\mathrm{d}\delta \le 4\alpha + \frac{12}{\sqrt{m}}\int_\alpha^M \sqrt{\ln\mathcal{N}(\mathcal{F},\delta,\|\cdot\|_\infty)}\,\mathrm{d}\delta.$$

□

We can easily extend Lemma 6 in (Gottlieb et al., 2017) to the following lemma by meshing on the range $[-M,M]$ rather than $[0,1]$.

**Lemma A.6.** *Let $\mathcal{F}$ be the family of L-Lipschitz functions mapping the metric space $(\mathcal{X},\|\cdot\|_2)$ to $[-M,M]$ for some $M>0$. Then we have*

$$\mathcal{N}(\mathcal{F},\delta,\|\cdot\|_\infty) \le \left(\frac{c_1 M}{\delta}\right)^{\mathcal{N}(\mathcal{X},\frac{c_2\delta}{L})},$$

*where $c_1\ge 1$ and $c_2\le 1$ are some absolute constants not depending on $\mathcal{X}$, $M$, and $\delta$.*

*Remark* A.7. If $\mathcal{X}$ is connected, then the bound can be improved to $\mathcal{N}(\mathcal{F},\delta,\|\cdot\|_\infty)\le e^{\mathcal{N}(\mathcal{X},\frac{c_2\delta}{L})}$ by the result in (Kolmogorov, 1961).

**Lemma A.8** (Theorem 3 in (Sokolic et al., 2017))**.** *Assume that $\mathcal{X}=\Sigma\times\mathcal{X}_0$. If for some $\delta>0$ we have*
*1) $\|\sigma(x)-\sigma'(x')\|_2 > 2\delta$, $\forall x,x'\in\mathcal{X}_0, \sigma\ne\sigma'\in\Sigma$; and*
*2) $\|\sigma(x)-\sigma(x')\|_2 \ge \|x-x'\|_2$, $\forall x,x'\in\mathcal{X}_0,\sigma\in\Sigma$,*
*then we have*

$$\frac{\mathcal{N}(\mathcal{X}_0,\delta)}{\mathcal{N}(\mathcal{X},\delta)} \le \frac{1}{|\Sigma|}.$$

In addition, we provide the following lemma for the scaling of covering numbers.

**Lemma A.9.** *Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ and $\bar\delta>0$. Then there exists a constant $C_{d,\bar\delta}$ that depends on $d$ and $\bar\delta$ such that for $\delta\in(0,1)$ we have*

$$\mathcal{N}(\mathcal{X},\delta) \le C_{d,\bar\delta}\cdot\frac{\mathcal{N}(\mathcal{X},\bar\delta)}{\delta^d}.$$

*Proof.* Let $N:=\mathcal{N}(\mathcal{X},\bar\delta)$. Then $\mathcal{X}$ can be covered by $N$ balls with radius $\bar\delta$. From Proposition 4.2.12 in (Vershynin, 2018), we know that each ball with radius $\bar\delta$ can be covered by $\frac{(\bar\delta+\delta/2)^d}{(\delta/2)^d}$ balls with radius $\delta$. This implies that $\mathcal{X}$ can be covered by $N\cdot\frac{(\bar\delta+\delta/2)^d}{(\delta/2)^d}$ balls with radius $\delta$, so that $\mathcal{N}^{\text{ext}}(\mathcal{X},\delta)\le N\cdot\frac{(\bar\delta+\delta/2)^d}{(\delta/2)^d}$, where $\mathcal{N}^{\text{ext}}(\mathcal{X},\delta)$ is the exterior covering number of $\mathcal{X}$ with radius $\delta$. Therefore, $\mathcal{N}(\mathcal{X},\delta)\le\mathcal{N}^{\text{ext}}(\mathcal{X},\delta/2)\le N\cdot\frac{(\bar\delta+\delta/4)^d}{(\delta/4)^d} = N\cdot(\frac{4\bar\delta}{\delta}+1)^d \le N\cdot\frac{(4\bar\delta+1)^d}{\delta^d}$. □

*Proof of Theorem A.2.*

$$
\begin{aligned}
&\left|W(P,Q) - W^{\Sigma}(P_m, Q_n)\right| \\
&= \left| \sup_{\gamma\in\Gamma_\Sigma} \{E_P[\gamma] - E_Q[\gamma]\} - \sup_{\gamma\in\Gamma_\Sigma} \{E_{P_m}[\gamma] - E_{Q_n}[\gamma]\} \right| \\
&\le \sup_{\gamma\in\Gamma_\Sigma} \left| E_P[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma(x_i) - \left( E_Q[\gamma] - \frac{1}{n}\sum_{i=1}^n \gamma(y_i)\right)\right| \\
&= \sup_{\gamma\in\Gamma_\Sigma} \left| E_P[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma\left(T_0(x_i)\right) - \left( E_Q[\gamma] - \frac{1}{n}\sum_{i=1}^n \gamma\left(T_0(y_i)\right)\right)\right| \\
&\overset{(a)}{\le} \sup_{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)} \left| E_{P_{\mathcal{X}_0}}[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma\left(T_0(x_i)\right) - \left( E_{Q_{\mathcal{X}_0}}[\gamma] - \frac{1}{n}\sum_{i=1}^n \gamma\left(T_0(y_i)\right)\right)\right| \qquad (35) \\
&:= f(x_1,\dots,x_m,y_1,\dots,y_n),
\end{aligned}
$$

where inequality $(a)$ is due to the fact that $E_P[\gamma] = E_{P_{\mathcal{X}_0}}[\gamma|_{\mathcal{X}_0}]$ and $E_Q[\gamma] = E_{Q_{\mathcal{X}_0}}[\gamma|_{\mathcal{X}_0}]$ since $P$ and $Q$ are both $\Sigma$-invariant and $\gamma \in \Gamma_\Sigma$, and the fact that if $\gamma \in \Gamma_\Sigma$, then $\gamma|_{\mathcal{X}_0} \in \mathrm{Lip}_L(\mathcal{X}_0)$, where $\gamma|_{\mathcal{X}_0}$ is the restriction of $\gamma$ on $\mathcal{X}_0$.

Note that the quantity inside the absolute value in (35) will not change if we replace $\gamma$ by $\gamma + \nu$ and we still have $\gamma + \nu \in \mathrm{Lip}_L(\mathcal{X}_0)$ for any $\nu \in \mathbb{R}$. Therefore, by Lemma A.4, the supremum in (35) can be taken over $\gamma \in \mathrm{Lip}_L(\mathcal{X}_0)$, where $\|\gamma\|_\infty \le L \cdot \mathrm{diam}(\mathcal{X}_0)$. The denominator in the exponent when applying the McDiarmid's inequality is thus equal to

$$
m\left(\frac{2L\cdot\mathrm{diam}(\mathcal{X}_0)}{m}\right)^2 + n\left(\frac{2L\cdot\mathrm{diam}(\mathcal{X}_0)}{n}\right)^2 = 4L^2\cdot\mathrm{diam}(\mathcal{X}_0)^2\frac{m+n}{mn}. \qquad (36)
$$

Denoting by $X' = \{x'_1, x'_2, \dots, x'_m\}$ and $Y' = \{y'_1, y'_2, \dots, y'_n\}$ the i.i.d. samples drawn from $P_{\mathcal{X}_0}$ and $Q_{\mathcal{X}_0}$. Also note that $T_0(x_1), \dots, T_0(x_m)$ and $T_0(y_1), \dots, T_0(y_n)$ can be viewed as i.i.d. samples on $\mathcal{X}_0$ drawn from $P_{\mathcal{X}_0}$ and $Q_{\mathcal{X}_0}$ respectively, such that the expectation

$$
\begin{aligned}
&E_{X,Y} f(x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_n) \\
&= E_{X,Y} \sup_{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)} \left| E_{P_{\mathcal{X}_0}}[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma(T_0(x_i)) - \left( E_{Q_{\mathcal{X}_0}}[\gamma] - \frac{1}{n}\sum_{i=1}^n \gamma(T_0(y_i))\right)\right|
\end{aligned}
$$

can be replaced by the equivalent quantity

$$
E_{X,Y} \sup_{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)} \left| E_{P_{\mathcal{X}_0}}[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma(x_i) - \left( E_{Q_{\mathcal{X}_0}}[\gamma] - \frac{1}{n}\sum_{i=1}^n \gamma(y_i)\right)\right|,
$$

where $X = \{x_1, x_2, \dots, x_m\}$ and $Y = \{y_1, y_2, \dots, y_n\}$ are are i.i.d. samples on $\mathcal{X}_0$ drawn from $P_{\mathcal{X}_0}$ and $Q_{\mathcal{X}_0}$ respectively. Then we have

$$
\begin{aligned}
&E_{X,Y} \sup_{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)} \left| E_{P_{\mathcal{X}_0}}[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma(x_i) - \left( E_{Q_{\mathcal{X}_0}}[\gamma] - \frac{1}{n}\sum_{i=1}^n \gamma(y_i)\right)\right| \\
&= E_{X,Y} \sup_{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)} \left| E_{X'}\left(\frac{1}{m}\sum_{i=1}^m \gamma(x'_i)\right) - \frac{1}{m}\sum_{i=1}^m \gamma(x_i) - E_{Y'}\left(\frac{1}{n}\sum_{i=1}^n \gamma(y'_i)\right) + \frac{1}{n}\sum_{i=1}^n \gamma(y_i)\right| \\
&\le E_{X,Y,X',Y'} \sup_{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)} \left| \frac{1}{m}\sum_{i=1}^m \gamma(x'_i) - \frac{1}{m}\sum_{i=1}^m \gamma(x_i) - \frac{1}{n}\sum_{i=1}^n \gamma(y'_i) + \frac{1}{n}\sum_{i=1}^n \gamma(y_i)\right| \\
&= E_{X,Y,X',Y',\xi,\xi'} \sup_{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)} \left| \frac{1}{m}\sum_{i=1}^m \xi_i\left(\gamma(x'_i) - \gamma(x_i)\right) - \frac{1}{n}\sum_{i=1}^n \xi'_i\left(\gamma(y'_i) - \gamma(y_i)\right)\right| \\
&\le E_{X,X',\xi} \sup_{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)} \left| \frac{1}{m}\sum_{i=1}^m \xi_i\left(\gamma(x'_i) - \gamma(x_i)\right)\right| + E_{Y,Y',\xi'} \sup_{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)} \left| \frac{1}{n}\sum_{i=1}^n \xi'_i\left(\gamma(y'_i) - \gamma(y_i)\right)\right|
\end{aligned}
$$

$$\leq \inf_{\alpha>0} 8\alpha + \frac{24}{\sqrt{m}} \int_{\alpha}^{M} \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta + \inf_{\alpha>0} 8\alpha + \frac{24}{\sqrt{n}} \int_{\alpha}^{M} \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta,$$

where $\mathcal{F}_0 = \{\gamma \in \mathrm{Lip}_L(\mathcal{X}_0) : \|\gamma\|_\infty \leq M\}$ and $M = L \cdot \mathrm{diam}(\mathcal{X}_0)$ by Lemma A.4.

For $d \geq 2$, from Lemma A.6, we have $\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty) \leq \mathcal{N}(\mathcal{X}_0, \frac{c_2\delta}{L}) \ln(\frac{c_1 M}{\delta})$. We fix a $\bar{\delta} > 0$ such that $\mathcal{N}(\mathcal{X}, \frac{c_2\bar{\delta}}{L}) = 1$, and select $\delta^*$ such that $\frac{c_2\delta^*}{L} \leq 1$ and $\frac{c_2\delta^*}{L} \leq \delta_0$; that is, $\delta^* \leq \min\left(\frac{L}{c_2}, \frac{L\delta_0}{c_2}\right)$, so that by Lemma A.8 and A.9, we have

$$\mathcal{N}(\mathcal{X}_0, \frac{c_2\delta}{L}) \ln(\frac{c_1 M}{\delta}) \leq \frac{\mathcal{N}(\mathcal{X}, \frac{c_2\delta}{L})}{|\Sigma|} \ln(\frac{c_1 M}{\delta}) \leq \frac{C_{d,\bar{\delta}} L^d}{|\Sigma|\, c_2^d \delta^d} \ln(\frac{c_1 M}{\delta}),$$

when $\delta < \delta^*$. Therefore, for sufficiently small $\alpha$, we have

$$\begin{aligned}
&\int_{\alpha}^{M} \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta \\
&= \int_{\alpha}^{\delta^*} \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta + \int_{\delta^*}^{M} \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta \\
&\leq \int_{\alpha}^{\delta^*} \sqrt{\frac{C_{d,\bar{\delta}} L^d}{|\Sigma|\, c_2^d \delta^d} \ln(\frac{c_1 M}{\delta})}\, \mathrm{d}\delta + \int_{\delta^*}^{M} \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta. \qquad (37)
\end{aligned}$$

For any $s > 0$, we can choose $\delta^*$ to be sufficiently small, such that we have $\ln(\frac{c_1 M}{\delta}) \leq \frac{1}{\delta^s}$ when $\delta \leq \delta^*$. Therefore, if we let $D_{\mathcal{X},L} = \sqrt{\frac{C_{d,\bar{\delta}} L^d}{c_2^d}}$, we will have

$$\begin{aligned}
\int_{\alpha}^{\delta^*} \sqrt{\frac{C_{d,\bar{\delta}} L^d}{|\Sigma|\, c_2^d \delta^d} \ln(\frac{c_1 M}{\delta})}\, \mathrm{d}\delta &\leq D_{\mathcal{X},L} \int_{\alpha}^{\delta^*} \sqrt{\frac{1}{|\Sigma|\, \delta^{d+s}}}\, \mathrm{d}\delta \\
&\leq D_{\mathcal{X},L} \int_{\alpha}^{\infty} \sqrt{\frac{1}{|\Sigma|\, \delta^{d+s}}}\, \mathrm{d}\delta \\
&= \frac{D_{\mathcal{X},L}}{\sqrt{|\Sigma|}} \cdot \frac{\alpha^{1-\frac{d+s}{2}}}{\frac{d+s}{2} - 1}.
\end{aligned}$$

Notice that the second integral in (37) is bounded while the first integral diverges as $\alpha$ tends to zero, so we can optimize the majorizing terms

$$8\alpha + \frac{24}{\sqrt{m}} \cdot \frac{D_{\mathcal{X},L}}{\sqrt{|\Sigma|}} \cdot \frac{\alpha^{1-\frac{d+s}{2}}}{\frac{d+s}{2} - 1}$$

with respect to $\alpha$, to obtain

$$\alpha = (\frac{9}{m})^{\frac{1}{d+s}} \cdot (\frac{D_{\mathcal{X},L}^2}{|\Sigma|})^{\frac{1}{d+s}},$$

so that

$$\begin{aligned}
&\inf_{\alpha>0} 8\alpha + \frac{24}{\sqrt{m}} \int_{\alpha}^{M} \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta \\
&\leq 8(\frac{9}{m})^{\frac{1}{d+s}} \cdot (\frac{D_{\mathcal{X},L}^2}{|\Sigma|})^{\frac{1}{d+s}} + \frac{24}{(\frac{d+s}{2} - 1)} (\frac{9}{m})^{\frac{1}{d+s}} \cdot (\frac{D_{\mathcal{X},L}^2}{|\Sigma|})^{\frac{1}{d+s}} + \frac{24}{\sqrt{m}} \int_{\delta^*}^{M} \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta.
\end{aligned}$$

Therefore, for sufficiently large $m$ and $n$, we have

$$E_{X,Y} \sup_{\gamma \in \mathrm{Lip}_L(\mathcal{X}_0)} \left| E_{P_{\mathcal{X}_0}}[\gamma] - \frac{1}{m} \sum_{i=1}^{m} \gamma(x_i) - \left( E_{Q_{\mathcal{X}_0}}[\gamma] - \frac{1}{n} \sum_{i=1}^{n} \gamma(y_i) \right) \right|$$

$$\leq \left(8 + \frac{24}{(\frac{d+s}{2} - 1)}\right) \left[\left(\frac{9D_{\mathcal{X},L}^2}{|\Sigma|\, m}\right)^{\frac{1}{d+s}} + \left(\frac{9D_{\mathcal{X},L}^2}{|\Sigma|\, n}\right)^{\frac{1}{d+s}}\right] + \left(\frac{24}{\sqrt{m}} + \frac{24}{\sqrt{n}}\right) \int_{\delta^*}^{M} \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta.$$

For $d = 1$, the first integral in (37) in the one-dimensional case does not have a singularity at $\alpha = 0$. On the other hand, replacing the interval $[0, 1]$ by an interval of length $\mathrm{diam}(\mathcal{X}_0)$ in Lemma 5.16 in (Van Handel, 2014), there exists a constant $c > 0$ such that

$$\mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty) \leq e^{\frac{cL \cdot \mathrm{diam}(\mathcal{X}_0)}{\delta}} \text{ for } \delta < M = L \cdot \mathrm{diam}(\mathcal{X}_0).$$

Therefore, we have

$$8\alpha + \frac{24}{\sqrt{m}} \int_\alpha^M \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta \leq 8\alpha + \frac{24}{\sqrt{m}} \int_\alpha^M \sqrt{\frac{cL \cdot \mathrm{diam}(\mathcal{X}_0)}{\delta}}\, \mathrm{d}\delta,$$

whose minimum is achieved at $\alpha = \frac{9cL \cdot \mathrm{diam}(\mathcal{X}_0)}{m}$. This implies that

$$\begin{aligned}\inf_{\alpha > 0} 8\alpha + \frac{24}{\sqrt{m}} \int_\alpha^M \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta &\leq \frac{72cL \cdot \mathrm{diam}(\mathcal{X}_0)}{m} + \frac{48L\sqrt{c} \cdot \mathrm{diam}(\mathcal{X}_0)}{\sqrt{m}} - \frac{144cL \cdot \mathrm{diam}(\mathcal{X}_0)}{m} \\ &= \frac{48L\sqrt{c} \cdot \mathrm{diam}(\mathcal{X}_0)}{\sqrt{m}} - \frac{72cL \cdot \mathrm{diam}(\mathcal{X}_0)}{m}.\end{aligned}$$

Hence, we have

$$\begin{aligned}&E_{X,Y} \sup_{\gamma \in \mathrm{Lip}_L(\mathcal{X}_0)} \left| E_{P_{\mathcal{X}_0}}[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma(x_i) - \left(E_{Q_{\mathcal{X}_0}}[\gamma] - \frac{1}{n}\sum_{i=1}^n \gamma(y_i)\right)\right| \\ &\leq \frac{48L\sqrt{c} \cdot \mathrm{diam}(\mathcal{X}_0)}{\sqrt{m}} - \frac{72cL \cdot \mathrm{diam}(\mathcal{X}_0)}{m} + \frac{48L\sqrt{c} \cdot \mathrm{diam}(\mathcal{X}_0)}{\sqrt{n}} - \frac{72cL \cdot \mathrm{diam}(\mathcal{X}_0)}{n}.\end{aligned}$$

Finally, by a simple change of variable for the probability provided in (36), we prove the theorem. □

*Remark* A.10. Though we do not directly observe the effect under the group invariance in the case when $d = 1$ in Theorem A.2, the upper bound can be improved in some special cases. Here we analyze Example 4.6 as an example. Replacing the interval $[0, 1]$ by $\mathcal{X}_0 = [0, \frac{1}{|\Sigma|})$ in Lemma 5.16 in (Van Handel, 2014), there exists a constant $c > 0$ such that

$$\mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty) \leq e^{\frac{cL}{|\Sigma|\delta}} \text{ for } \delta < M = L \cdot \mathrm{diam}(\mathcal{X}_0).$$

Therefore, we have

$$8\alpha + \frac{24}{\sqrt{m}} \int_\alpha^M \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta = 8\alpha + \frac{24}{\sqrt{m}} \int_\alpha^M \sqrt{\frac{cL}{|\Sigma|\, \delta}}\, \mathrm{d}\delta,$$

whose minimum is achieved at $\alpha = \frac{9cL}{m|\Sigma|}$. This implies that

$$\inf_{\alpha > 0} 8\alpha + \frac{24}{\sqrt{m}} \int_\alpha^M \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\, \mathrm{d}\delta = \frac{72cL}{|\Sigma|\, m} + \frac{48L\sqrt{c}}{|\Sigma|\, \sqrt{m}} - \frac{144cL}{|\Sigma|\, m} = \frac{48L\sqrt{c}}{|\Sigma|\, \sqrt{m}} - \frac{72cL}{|\Sigma|\, m}.$$

Hence, we have

$$E_{X,Y} \sup_{\gamma \in \mathrm{Lip}_L(\mathcal{X}_0)} \left| E_{P_{\mathcal{X}_0}}[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma(x_i) - \left(E_{Q_{\mathcal{X}_0}}[\gamma] - \frac{1}{n}\sum_{i=1}^n \gamma(y_i)\right)\right| \leq \frac{48L\sqrt{c}}{|\Sigma|\, \sqrt{m}} - \frac{72cL}{|\Sigma|\, m} + \frac{48L\sqrt{c}}{|\Sigma|\, \sqrt{n}} - \frac{72cL}{|\Sigma|\, n}.$$

This matches the numerical result in Figure 3 where the ratio curves are around 4, since our group sizes are $|\Sigma| = 1, 4, 16, 64, 256$, increasing by a factor of 4,

### A.2. $(f_\alpha, \Gamma)$-divergence

We assume Assumption A.1 also holds in this case.

**Theorem A.11.** *Let $\mathcal{X} = \Sigma \times \mathcal{X}_0$ be a subset of $\mathbb{R}^d$ equipped with the Euclidean distance, $f(x) = f_\alpha(x) = \frac{x^\alpha - 1}{\alpha(\alpha-1)}$, $\alpha > 1$ and $\Gamma = Lip_L(\mathcal{X})$. Suppose $P$ and $Q$ are $\Sigma$-invariant distributions on $\mathcal{X}$. We have*

*1) if $d \geq 2$, then for any $s > 0$ and $m, n$ sufficiently large, we have with probability at least $1 - \epsilon$,*

$$\left| D_{f_\alpha}^{\Gamma}(P\|Q) - D_{f_\alpha}^{\Gamma_\Sigma}(P_m\|Q_n) \right| \leq \left(8 + \frac{24}{(\frac{d+s}{2} - 1)}\right) \left[ \left( \frac{9 D_{\mathcal{X},L}^2}{|\Sigma|\, m} \right)^{\frac{1}{d+s}} + \left( \frac{9 D_{\mathcal{X},L'}^2}{|\Sigma|\, n} \right)^{\frac{1}{d+s}} \right] + \frac{24 \bar{D}_{\mathcal{X}_0,L}}{\sqrt{m}} + \frac{24 \bar{D}_{\mathcal{X}_0,L'}}{\sqrt{n}} + \sqrt{\frac{2(M_1^2 m + M_0^2 n)}{mn} \ln \frac{1}{\epsilon}},$$

*where $D_{\mathcal{X},L}$ depends only on $\mathcal{X}$ and $L$, and $D_{\mathcal{X},L'}$ depends only on $\mathcal{X}$, $L$ and $\alpha$; $\bar{D}_{\mathcal{X}_0,L}$ depends only on $\mathcal{X}_0$ and $L$, and $\bar{D}_{\mathcal{X}_0,L'}$ depends only on $\mathcal{X}_0$ and $L$ and $\alpha$, and both are increasing in $\mathcal{X}_0$; $M_0$ and $M_1$ both only depend on $\mathcal{X}$, $L$ and $\alpha$;*

*2) if $d = 1$, for any $\epsilon > 0$ and $m, n$ sufficiently large, we have with probability at least $1 - \epsilon$,*

$$\left| D_{f_\alpha}^{\Gamma}(P\|Q) - D_{f_\alpha}^{\Gamma_\Sigma}(P_m\|Q_n) \right| \leq \frac{48L\sqrt{c} \cdot \mathrm{diam}(\mathcal{X}_0)}{\sqrt{m}} - \frac{72cL \cdot \mathrm{diam}(\mathcal{X}_0)}{m} + \frac{48L'\sqrt{c} \cdot \mathrm{diam}(\mathcal{X}_0)}{\sqrt{n}} - \frac{72cL' \cdot \mathrm{diam}(\mathcal{X}_0)}{n} + \sqrt{\frac{2(M_1^2 m + M_0^2 n)}{mn} \ln \frac{1}{\epsilon}},$$

*where $c > 0$ is an absolute constant independent of $\mathcal{X}_0$; $L'$ depends only on $\mathcal{X}$, $L$ and $\alpha$; $M_0$ and $M_1$ both only depend on $\mathcal{X}$, $L$ and $\alpha$.*

Before proving this theorem, we first provide the following lemma.

**Lemma A.12.** $D_{f_\alpha}^{\Gamma}(P\|Q) = D_{f_\alpha}^{\mathcal{F}}(P\|Q)$, *where*

$$\mathcal{F} = \left\{ \gamma \in Lip_L(\mathcal{X}) : \|\gamma\|_\infty \leq (\alpha - 1)^{-1} + L \cdot diam(\mathcal{X}) \right\},$$

*and $P$ and $Q$ are probability distributions on $\mathcal{X}$ that are not necessarily $\Sigma$-invariant.*

*Proof.* For any fixed $\gamma \in \Gamma$, let $h(\nu) = E_P[\gamma + \nu] - E_Q[f_\alpha^*(\gamma + \nu)]$. We know that $\sup_{x \in \mathcal{X}} \gamma(x) - \inf_{x \in \mathcal{X}} \gamma(x) \leq L \cdot \mathrm{diam}(\mathcal{X})$, so interchanging the integration with differentiation is allowed by the dominated convergence theorem: $h'(\nu) = 1 - E_Q[f_\alpha^{*\prime}(\gamma + \nu)]$, where

$$f_\alpha^{*\prime}(y) = (\alpha - 1)^{\frac{1}{\alpha-1}} y^{\frac{1}{\alpha-1}} \mathbf{1}_{y>0}.$$

If $\inf_{x \in \mathcal{X}} \gamma(x) > (\alpha - 1)^{-1}$, then $h'(0) < 0$. So there exists some $\nu_0 < 0$ such that $E_P[\gamma + \nu_0] - E_Q[f_\alpha^*(\gamma + \nu_0)] = h(\nu_0) > h(0) = E_P[\gamma] - E_Q[f_\alpha^*(\gamma)]$. This indicates the supremum in $D_f^\Gamma(P\|Q)$ is attained only if $\sup_{x \in \mathcal{X}} \gamma(x) \leq (\alpha-1)^{-1} + L \cdot \mathrm{diam}(\mathcal{X})$. On the other hand, if $\sup_{x \in \mathcal{X}} \gamma(x) < 0$, then there exists $\nu_0 > 0$ that satisfies $\sup_{x \in \mathcal{X}} \gamma(x) + \nu_0 < 0$ such that $E_P[\gamma + \nu_0] - E_Q[f_\alpha^*(\gamma + \nu_0)] = E_P[\gamma] + \nu_0 > E_P[\gamma] = E_P[\gamma] - E_Q[f_\alpha^*(\gamma)]$. This indicates that the supremum in $D_f^\Gamma(P\|Q)$ is attained only if $\inf_{x \in \mathcal{X}} \gamma(x) \geq -L \cdot \mathrm{diam}(\mathcal{X})$. Therefore, we have that the supremum in $D_f^\Gamma(P\|Q)$ is attained only if $\|\gamma\|_\infty \leq (\alpha - 1)^{-1} + L \cdot \mathrm{diam}(\mathcal{X})$. □

*Proof of Theorem A.11.* Similar to the beginning of the proof of Theorem A.2, we have by Lemma A.12 that

$$\begin{aligned} &\left| D_{f_\alpha}^{\Gamma}(P\|Q) - D_{f_\alpha}^{\Gamma_\Sigma}(P_m, Q_n) \right| \\ &= \left| \sup_{\substack{\gamma \in \Gamma_\Sigma \\ \|\gamma\|_\infty \leq M_0}} \{E_P[\gamma] - E_Q[f_\alpha^*(\gamma)]\} - \sup_{\substack{\gamma \in \Gamma_\Sigma \\ \|\gamma\|_\infty \leq M_0}} \{E_{P_m}[\gamma] - E_{Q_n}[f_\alpha^*(\gamma)]\} \right| \end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{\substack{\gamma\in\Gamma_\Sigma\\ \|\gamma\|_\infty\leq M_0}} \left| E_P[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma(x_i) - \left( E_Q[f_\alpha^*(\gamma)] - \frac{1}{n}\sum_{i=1}^n f_\alpha^*\left(\gamma(y_i)\right)\right)\right| \\
&= \sup_{\substack{\gamma\in\Gamma_\Sigma\\ \|\gamma\|_\infty\leq M_0}} \left| E_P[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma\left(T_0(x_i)\right) - \left( E_Q[f_\alpha^*(\gamma)] - \frac{1}{n}\sum_{i=1}^n f_\alpha^*\left(\gamma(T_0(y_i))\right)\right)\right| \\
&\leq \sup_{\substack{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)\\ \|\gamma\|_\infty\leq M_0}} \left| E_{P_{\mathcal{X}_0}}[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma(T_0(x_i)) - \left( E_{Q_{\mathcal{X}_0}}[f_\alpha^*(\gamma)] - \frac{1}{n}\sum_{i=1}^n f_\alpha^*\left(\gamma(T_0(y_i))\right)\right)\right| \\
&:= g(x_1,\ldots,x_m,y_1,\ldots,y_n),
\end{aligned}
$$

where $T_0$ is the same as defined in (23). The denominator in the exponent when applying the McDiarmid's inequality is thus equal to

$$
m\left(\frac{2M_0}{m}\right)^2 + n\left(\frac{2M_1}{n}\right)^2 = \frac{4M_0^2}{m} + \frac{4M_1^2}{n},
$$

where $M_0 = (\alpha-1)^{-1} + L\cdot\mathrm{diam}(\mathcal{X})$, $M_1 = f_\alpha^*(M_0)$, since for any $\gamma$ such that $\|\gamma\|_\infty \leq M_0$, we have $\|f_\alpha^*\circ\gamma\|_\infty \leq M_1$. Denoting by $X' = \{x_1', x_2', \ldots, x_m'\}$ and $Y' = \{y_1', y_2', \ldots, y_n'\}$ the i.i.d. samples drawn from $P_{\mathcal{X}_0}$ and $Q_{\mathcal{X}_0}$. Also note that $T_0(x_1),\ldots,T_0(x_m)$ and $T_0(y_1),\ldots,T_0(y_n)$ can be viewed as i.i.d. samples on $\mathcal{X}_0$ drawn from $P_{\mathcal{X}_0}$ and $Q_{\mathcal{X}_0}$ respectively, such that the expectation

$$
\begin{aligned}
&E_{X,Y}\, g(x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n) \\
&= E_{X,Y} \sup_{\substack{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)\\ \|\gamma\|_\infty\leq M_0}} \left| E_{P_{\mathcal{X}_0}}[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma(T_0(x_i)) - \left( E_{Q_{\mathcal{X}_0}}[f_\alpha^*(\gamma)] - \frac{1}{n}\sum_{i=1}^n f_\alpha^*\left(\gamma(T_0(y_i))\right)\right)\right|
\end{aligned}
$$

can be replaced by the equivalent quantity

$$
E_{X,Y} \sup_{\substack{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)\\ \|\gamma\|_\infty\leq M_0}} \left| E_{P_{\mathcal{X}_0}}[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma(x_i) - \left( E_{Q_{\mathcal{X}_0}}[f_\alpha^*(\gamma)] - \frac{1}{n}\sum_{i=1}^n f_\alpha^*\left(\gamma(y_i)\right)\right)\right|,
$$

where $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ are are i.i.d. samples on $\mathcal{X}_0$ drawn from $P_{\mathcal{X}_0}$ and $Q_{\mathcal{X}_0}$ respectively. Then we have

$$
\begin{aligned}
&E_{X,Y} \sup_{\substack{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)\\ \|\gamma\|_\infty\leq M_0}} \left| E_{P_{\mathcal{X}_0}}[\gamma] - \frac{1}{m}\sum_{i=1}^m \gamma(x_i) - \left( E_{Q_{\mathcal{X}_0}}[f_\alpha^*(\gamma)] - \frac{1}{n}\sum_{i=1}^n f_\alpha^*\left(\gamma(y_i)\right)\right)\right| \\
&= E_{X,Y} \sup_{\substack{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)\\ \|\gamma\|_\infty\leq M_0}} \left| E_{X'}\left(\frac{1}{m}\sum_{i=1}^m \gamma(x_i')\right) - \frac{1}{m}\sum_{i=1}^m \gamma(x_i) - E_{Y'}\left(\frac{1}{n}\sum_{i=1}^n f_\alpha^*\left(\gamma(y_i')\right)\right) + \frac{1}{n}\sum_{i=1}^n f_\alpha^*\left(\gamma(y_i)\right)\right| \\
&\leq E_{X,Y,X',Y'} \sup_{\substack{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)\\ \|\gamma\|_\infty\leq M_0}} \left| \frac{1}{m}\sum_{i=1}^m \gamma(x_i') - \frac{1}{m}\sum_{i=1}^m \gamma(x_i) - \frac{1}{n}\sum_{i=1}^n f_\alpha^*\left(\gamma(y_i')\right) + \frac{1}{n}\sum_{i=1}^n f_\alpha^*\left(\gamma(y_i)\right)\right| \\
&= E_{X,Y,X',Y',\xi,\xi'} \sup_{\substack{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)\\ \|\gamma\|_\infty\leq M_0}} \left| \frac{1}{m}\sum_{i=1}^m \xi_i\left(\gamma(x_i') - \gamma(x_i)\right) - \frac{1}{n}\sum_{i=1}^n \xi_i'\left(f_\alpha^*\left(\gamma(y_i')\right) - f_\alpha^*\left(\gamma(y_i)\right)\right)\right| \\
&\leq E_{X,X',\xi} \sup_{\substack{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)\\ \|\gamma\|_\infty\leq M_0}} \left| \frac{1}{m}\sum_{i=1}^m \xi_i\left(\gamma(x_i') - \gamma(x_i)\right)\right| + E_{Y,Y',\xi'} \sup_{\substack{\gamma\in\mathrm{Lip}_L(\mathcal{X}_0)\\ \|\gamma\|_\infty\leq M_0}} \left| \frac{1}{n}\sum_{i=1}^n \xi_i'\left(f_\alpha^*\left(\gamma(y_i')\right) - f_\alpha^*\left(\gamma(y_i)\right)\right)\right| \\
&\leq \inf_{\alpha>0} 8\alpha + \frac{24}{\sqrt{m}}\int_\alpha^{M_0} \sqrt{\ln \mathcal{N}(\mathcal{F}_0, \delta, \|\cdot\|_\infty)}\,\mathrm{d}\delta + \inf_{\alpha>0} 8\alpha + \frac{24}{\sqrt{n}}\int_\alpha^{M_1} \sqrt{\ln \mathcal{N}(\mathcal{F}_1, \delta, \|\cdot\|_\infty)}\,\mathrm{d}\delta,
\end{aligned}
$$

where $\mathcal{F}_0 = \{\gamma \in \text{Lip}_L(\mathcal{X}_0) : \|\gamma\|_\infty \le M_0\}$ and $\mathcal{F}_1 = \{\gamma \in \text{Lip}_{L'}(\mathcal{X}_0) : \|\gamma\|_\infty \le M_1\}$, since for any $\gamma \in \mathcal{F}_0$, $\|f_\alpha^* \circ \gamma\|_\infty \le M_1$ and $\left|\frac{\mathrm{d}}{\mathrm{d}y} f_\alpha^*(y)\right| \le (\alpha-1)^{\frac{1}{\alpha-1}}(M_0)^{\frac{1}{\alpha-1}}$ for $|y| \le M_0$ such that $f_\alpha^* \circ \gamma$ is $L'$-Lipschitz, where $M_1 = f_\alpha^*(M_0)$ and $L' = L(\alpha-1)^{\frac{1}{\alpha-1}}(M_0)^{\frac{1}{\alpha-1}}$. Then the rest of the proof follows from the proof of Theorem A.2. □

### A.3. MMD

We assume the kernel $k(x,y)$ satisfies Assumption 4.9. Furthermore, let $\phi(x)$ be the evaluation functional at $x$ in $\mathcal{H}$: $\langle \phi(x), \phi(y)\rangle_{\mathcal{H}} = k(x,y), \forall x, y \in \mathcal{H}$.

**Theorem A.13.** *Let $\mathcal{X} = \Sigma \times \mathcal{X}_0$ be a subset of $\mathbb{R}^d (d \ge 1)$ and $\mathcal{H}$ be a RKHS on $\mathcal{X}$ whose kernel satisfies Assumption 4.9. Suppose $P$ and $Q$ are $\Sigma$-invariant distributions on $\mathcal{X}$. Then for $m, n$ sufficiently large and any $\epsilon > 0$ we have with probability at least $1-\epsilon$,*

$$\left|\mathit{MMD}(P,Q) - \mathit{MMD}^{\Sigma}(P_m, Q_n)\right| < 2K^{\frac{1}{2}}\left[1 + c(|\Sigma|-1)\right]^{\frac{1}{2}}\left(\frac{1}{\sqrt{|\Sigma|\, m}} + \frac{1}{\sqrt{|\Sigma|\, n}}\right) + \sqrt{\frac{2K(1+c(|\Sigma|-1))\ln(\frac{1}{\epsilon})}{|\Sigma|}}\sqrt{\frac{1}{m}+\frac{1}{n}},$$

*where $K$ and $c$ are the constants in Assumption 4.9.*

Before proving the theorem, we provide the following lemma.

**Lemma A.14.** *Suppose the kernel in an RKHS satisfies Assumption 4.9, and $\xi = \{\xi_1, \ldots, \xi_m\}$ is a set of independent random variables, each of which takes values on $\{-1, 1\}$ with equal probabilities. Then we have*

$$E_\xi \sup_{\|\gamma\|_{\mathcal{H}} \le 1}\left|\frac{1}{m\,|\Sigma|}\sum_{i=1}^{m}\xi_i\sum_{j=1}^{|\Sigma|}\gamma(\sigma_j x_i)\right| \le \frac{(1+c(|\Sigma|-1))\,K^{\frac{1}{2}}}{\sqrt{|\Sigma|\, m}}.$$

*Proof.* Since the witness function to attain the supremum is explicit, we can write

$$\begin{aligned}
E_\xi \sup_{\|\gamma\|_{\mathcal{H}} \le 1}\left|\frac{1}{m\,|\Sigma|}\sum_{i=1}^{m}\xi_i\sum_{j=1}^{|\Sigma|}\gamma(\sigma_j x_i)\right| &= E_\xi\left\|\frac{1}{m\,|\Sigma|}\sum_{i=1}^{m}\xi_i\sum_{j=1}^{|\Sigma|}\phi(\sigma_j x_i)\right\|_{\mathcal{H}} \\
&= \frac{1}{m\,|\Sigma|}E_\xi\left[\sum_{i,i'=1}^{m}\xi_i\xi_{i'}\sum_{j,j'=1}^{|\Sigma|}k(\sigma_j x_i, \sigma_{j'}x_{i'})\right]^{\frac{1}{2}} \\
&\le \frac{1}{m\,|\Sigma|}\left[E_\xi\sum_{i,i'=1}^{m}\xi_i\xi_{i'}\sum_{j,j'=1}^{|\Sigma|}k(\sigma_j x_i, \sigma_{j'}x_{i'})\right]^{\frac{1}{2}} \\
&= \frac{1}{m\,|\Sigma|}\left[E_\xi\sum_{i=1}^{m}(\xi_i)^2\sum_{j,j'=1}^{|\Sigma|}k(\sigma_j x_i, \sigma_{j'}x_i)\right]^{\frac{1}{2}} \\
&\le \frac{1}{m\,|\Sigma|}\left[m\cdot\Big(|\Sigma|\,K + c(|\Sigma|^2-|\Sigma|)K\Big)\right]^{\frac{1}{2}} \\
&= \frac{K^{\frac{1}{2}}\left[1+c(|\Sigma|-1)\right]^{\frac{1}{2}}}{\sqrt{|\Sigma|\, m}}.
\end{aligned}$$

□

*Proof of Theorem A.13.* The proof below is a generalization of the proof of Theorem 7 in (Gretton et al., 2012), which does not need the notion of covering numbers due to the structure of RKHS.

$$\left|\text{MMD}(P,Q) - \text{MMD}^{\Sigma}(P_m, Q_n)\right|$$

$$
\begin{aligned}
&= \left|\text{MMD}(P,Q) - \text{MMD}(S^{\Sigma}[P_m], S^{\Sigma}[Q_n])\right| \\
&= \left| \sup_{\|\gamma\|_{\mathcal{H}}\leq 1} \{E_P[\gamma] - E_Q[\gamma]\} - \sup_{\|\gamma\|_{\mathcal{H}}\leq 1} \{E_{S^{\Sigma}[P_m]}[\gamma] - E_{S^{\Sigma}[Q_n]}[\gamma]\} \right| \\
&= \left| \sup_{\|\gamma\|_{\mathcal{H}}\leq 1} \{E_P[\gamma] - E_Q[\gamma]\} - \sup_{\|\gamma\|_{\mathcal{H}}\leq 1} \{\frac{1}{m\,|\Sigma|}\sum_{i=1}^{m}\sum_{j=1}^{|\Sigma|}\gamma(\sigma_j x_i) - \frac{1}{n\,|\Sigma|}\sum_{i=1}^{n}\sum_{j=1}^{|\Sigma|}\gamma(\sigma_j y_i)\} \right| \\
&\leq \sup_{\|\gamma\|_{\mathcal{H}}\leq 1} \left| E_P[\gamma] - E_Q[\gamma] - \frac{1}{m\,|\Sigma|}\sum_{i=1}^{m}\sum_{j=1}^{|\Sigma|}\gamma(\sigma_j x_i) + \frac{1}{n\,|\Sigma|}\sum_{i=1}^{n}\sum_{j=1}^{|\Sigma|}\gamma(\sigma_j y_i) \right| \\
&:= f(x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n).
\end{aligned}
$$

Now we estimate the upper bound of the difference of $f$ if we change one of $x_i$'s.

$$
\begin{aligned}
&|f(x_1, \ldots, x_i, \ldots, y_1, \ldots, y_n) - f(x_1, \ldots, \tilde{x}_i, \ldots, y_1, \ldots, y_n)| \\
&\leq \sup_{\|\gamma\|_{\mathcal{H}}\leq 1} \left| \frac{1}{m\,|\Sigma|}\sum_{j=1}^{|\Sigma|}\gamma(\sigma_j x_i) - \gamma(\sigma_j \tilde{x}_i) \right| \\
&= \frac{1}{m\,|\Sigma|} \left\| \sum_{j=1}^{|\Sigma|}\phi(\sigma_j x_i) - \phi(\sigma_j \tilde{x}_i) \right\|_{\mathcal{H}} \\
&\leq \frac{1}{m\,|\Sigma|} \left( \left\| \sum_{j=1}^{|\Sigma|}\phi(\sigma_j x_i) \right\|_{\mathcal{H}} + \left\| \sum_{j=1}^{|\Sigma|}\phi(\sigma_j \tilde{x}_i) \right\|_{\mathcal{H}} \right).
\end{aligned} \tag{38}
$$

To bound $\left\|\sum_{j=1}^{|\Sigma|}\phi(\sigma_j x_i)\right\|_{\mathcal{H}}$, we have

$$
\begin{aligned}
\left\| \sum_{j=1}^{|\Sigma|}\phi(\sigma_j x_i) \right\|_{\mathcal{H}} &= \left[ \sum_{j=1}^{|\Sigma|} k(\sigma_j x_i, \sigma_j x_i) + \sum_{j\neq l} k(\sigma_j x_i, \sigma_l x_i) \right]^{\frac{1}{2}} \\
&= \left[ \sum_{j=1}^{|\Sigma|} k(\sigma_j x_i, \sigma_j x_i) + \sum_{\sigma_j \neq id} k(\sigma_j x_i, x_i) \right]^{\frac{1}{2}} \\
&\leq \left[ |\Sigma| \cdot K + \left( |\Sigma|^2 - |\Sigma| \right) \cdot cK \right]^{\frac{1}{2}}.
\end{aligned}
$$

The upper bound of the difference of $f$ if we change one of $y_i$'s can be derived in the same way. To apply the McDiarmid's inequality, the denominator in the exponent is thus

$$
\begin{aligned}
&m \cdot \frac{4\left[ |\Sigma| \cdot K + \left( |\Sigma|^2 - |\Sigma| \right) \cdot cK \right]}{m^2\,|\Sigma|^2} + n \cdot \frac{4\left[ |\Sigma| \cdot K + \left( |\Sigma|^2 - |\Sigma| \right) \cdot cK \right]}{n^2\,|\Sigma|^2} \\
&\leq 4K(\frac{1}{m} + \frac{1}{n}) \cdot \frac{1 + c(|\Sigma| - 1)}{|\Sigma|}.
\end{aligned}
$$

Moreover, we can extend inequality (16) in (Gretton et al., 2012) to take into account the group invariance. Denoting by $X' = \{x'_1, x'_2, \ldots, x'_m\}$ and $Y' = \{y'_1, y'_2, \ldots, y'_n\}$ the i.i.d. samples drawn from $P$ and $Q$, and $\xi = \{\xi_1, \ldots, \xi_m\}$, $\xi' = \{\xi'_1, \ldots, \xi'_n\}$ sets of independent random variables, each of which takes values on $\{-1, 1\}$ with equal probabilities, we have

$$E_{X,Y} f(x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n)$$

$$
\begin{aligned}
&= E_{X,Y} \sup_{\|\gamma\|_{\mathcal{H}} \le 1} \left| E_P[\gamma] - E_Q[\gamma] - \frac{1}{m\,|\Sigma|} \sum_{i=1}^{m} \sum_{j=1}^{|\Sigma|} \gamma(\sigma_j x_i) + \frac{1}{n\,|\Sigma|} \sum_{i=1}^{n} \sum_{j=1}^{|\Sigma|} \gamma(\sigma_j y_i) \right| \\
&= E_{X,Y} \sup_{\|\gamma\|_{\mathcal{H}} \le 1} \left| E_{X'} \left( \frac{1}{m\,|\Sigma|} \sum_{i=1}^{m} \sum_{j=1}^{|\Sigma|} \gamma(\sigma_j x_i') \right) - E_{Y'} \left( \frac{1}{n\,|\Sigma|} \sum_{i=1}^{n} \sum_{j=1}^{|\Sigma|} \gamma(\sigma_j y_i') \right) - \frac{1}{m\,|\Sigma|} \sum_{i=1}^{m} \sum_{j=1}^{|\Sigma|} \gamma(\sigma_j x_i) \right. \\
&\left. + \frac{1}{n\,|\Sigma|} \sum_{i=1}^{n} \sum_{j=1}^{|\Sigma|} \gamma(\sigma_j y_i) \right| \\
&\le E_{X,Y,X',Y'} \sup_{\|\gamma\|_{\mathcal{H}} \le 1} \left| \frac{1}{m\,|\Sigma|} \sum_{i=1}^{m} \sum_{j=1}^{|\Sigma|} (\gamma(\sigma_j x_i') - \gamma(\sigma_j x_i)) - \frac{1}{n\,|\Sigma|} \sum_{i=1}^{n} \sum_{j=1}^{|\Sigma|} (\gamma(\sigma_j y_i') - \gamma(\sigma_j y_i)) \right| \\
&= E_{X,Y,X',Y',\xi,\xi'} \sup_{\|\gamma\|_{\mathcal{H}} \le 1} \left| \frac{1}{m\,|\Sigma|} \sum_{i=1}^{m} \xi_i \sum_{j=1}^{|\Sigma|} (\gamma(\sigma_j x_i') - \gamma(\sigma_j x_i)) - \frac{1}{n\,|\Sigma|} \sum_{i=1}^{n} \xi_i' \sum_{j=1}^{|\Sigma|} (\gamma(\sigma_j y_i') - \gamma(\sigma_j y_i)) \right| \\
&\le E_{X,X',\xi} \sup_{\|\gamma\|_{\mathcal{H}} \le 1} \left| \frac{1}{m\,|\Sigma|} \sum_{i=1}^{m} \xi_i \sum_{j=1}^{|\Sigma|} (\gamma(\sigma_j x_i') - \gamma(\sigma_j x_i)) \right| + E_{Y,Y',\xi'} \sup_{\|\gamma\|_{\mathcal{H}} \le 1} \left| \frac{1}{n\,|\Sigma|} \sum_{i=1}^{n} \xi_i' \sum_{j=1}^{|\Sigma|} (\gamma(\sigma_j y_i') - \gamma(\sigma_j y_i)) \right| \\
&\le 2 \left[ \frac{K^{\frac{1}{2}} \left[1 + c(|\Sigma| - 1)\right]^{\frac{1}{2}}}{\sqrt{|\Sigma|\,m}} + \frac{K^{\frac{1}{2}} \left[1 + c(|\Sigma| - 1)\right]^{\frac{1}{2}}}{\sqrt{|\Sigma|\,n}} \right],
\end{aligned}
$$

where the last inequality is due to Lemma A.14. Therefore, by the McDiarmid's theorem, we have

$$
\begin{aligned}
&\mathbb{P}\left( \left| \mathrm{MMD}(P, Q) - \mathrm{MMD}^{\Sigma}(P_m, Q_n) \right| - 2K^{\frac{1}{2}} \left[1 + c(|\Sigma| - 1)\right]^{\frac{1}{2}} \left( \frac{1}{\sqrt{|\Sigma|\,m}} + \frac{1}{\sqrt{|\Sigma|\,n}} \right) > \epsilon \right) \\
&\qquad \le \exp\left( - \frac{\epsilon^2 mn\,|\Sigma|}{2K(m+n)(1 + c(|\Sigma| - 1))} \right).
\end{aligned}
$$

By a change of variable, we have with probability at least $1 - \epsilon$,

$$
\begin{aligned}
\left| \mathrm{MMD}(P, Q) - \mathrm{MMD}^{\Sigma}(P_m, Q_n) \right| &< 2K^{\frac{1}{2}} \left[1 + c(|\Sigma| - 1)\right]^{\frac{1}{2}} \left( \frac{1}{\sqrt{|\Sigma|\,m}} + \frac{1}{\sqrt{|\Sigma|\,n}} \right) \\
&\quad + \sqrt{\frac{2K(1 + c(|\Sigma| - 1)) \ln(\frac{1}{\epsilon})}{|\Sigma|}} \sqrt{\frac{1}{m} + \frac{1}{n}}.
\end{aligned}
$$

□