Statistical Guarantees of Group-Invariant GANs[*]

Ziyu Chen[†], Markos A. Katsoulakis[†], Luc Rey-Bellet[†], and Wei Zhu[†]

Abstract. This work presents the first statistical performance guarantees for group-invariant generative models. Many real-world datasets, such as images and molecules, are invariant to certain group symmetries, which can be taken advantage of to learn more efficiently, as we rigorously demonstrate in this work. Here we specifically study generative adversarial networks (GANs) and quantify the gains when incorporating symmetries into the model. Group-invariant GANs are a type in which the generators and discriminators are hardwired with group symmetries. Empirical studies have shown that these networks are capable of learning group-invariant distributions with significantly improved data efficiency. In this study, we aim to rigorously quantify this improvement by analyzing the reduction in sample complexity and in the discriminator approximation error for group-invariant GANs. Our findings indicate that when learning group-invariant distributions, the number of samples required for group-invariant GANs decreases proportionally by a factor of the group size, and the discriminator approximation error has a reduced lower bound. An important point is that the overall error reduction cannot be achieved merely through data augmentation on the training data. Numerical results substantiate our theory and highlight the stark contrast between learning with group-invariant GANs and using data augmentation. This work also sheds light on the study of other generative models with group symmetries, such as score-based generative models.

Key words. generative modeling, generative adversarial networks, group invariance, performance guarantees, generalization error

MSC codes. 62E10, 62E17, 60-08

DOI. 10.1137/24M1666628

1. Introduction. The machine learning community has shown a growing interest in generative models, which aim to understand the underlying distribution of data and generate new samples from it. Among the various generative models, generative adversarial networks (GANs) [20] have garnered significant attention due to their ability to learn and sample from complex data distributions. Empirical evidence shows that GANs achieve remarkable performance in diverse applications such as image synthesis and text generation [40, 51, 53]. Such success, however, hinges upon the availability of *abundant* training data.

Recent research has proposed to leverage the group-invariant structure of underlying distributions to improve the data efficiency of GANs [15, 7]. Additionally, the following related

[†]Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA, 01003 (ziyuchen@umass.edu, markos@umass.edu, luc@umass.edu, weizhu@umass.edu).

work has explored other generative models: reference [28] proposes equivariant flows, while [32] and [23] study equivariant diffusion models. Such approaches are motivated by the prevalence of *group symmetry* observed in various real-world distributions. For instance, in the case of medical images captured without orientation alignment, the distribution should be rotation-invariant, i.e., an image and its rotated copy are *equiprobable.* Furthermore, many physical, physicochemical, and biochemical systems possess intrinsic symmetries or equivariance structures, making structured probabilistic modeling crucial [39, 37, 25, 35, 30, 36, 38]. Empirical evidence from [28, 15, 7] shows that models designed to respect group symmetry can effectively learn a group-invariant distribution even with limited data. The key idea is to introduce structures into generative models, i.e., moving away from *generic* models to exploit essential structures in physicochemical models. This shift is expected to create *physics-informed* generative models that are more sample-efficient and reduce computational load and complexity. However, a clear theoretical understanding of these phenomena remains to be established.

In this paper, we provide statistical performance guarantees that explain why group-invariant generative models, in particular, group-invariant GANs [15, 7], can effectively learn group-invariant distributions with significantly fewer training data. Specifically, let Σ be a finite group, μ be a Σ-invariant target distribution supported on some compact domain $\mathcal{X} \subset \mathbb{R}^d$, and ρ be an easy-to-sample (noise) source distribution. Let $S^\Sigma[(g_{n,m}^*)_\sharp \rho]$ be the Σ-invariant generated distribution learned based on n independent and identically distributed (i.i.d.) training samples from the target μ and m random draws from the noise source ρ; see (3.15) for the exact definition. We show that, if m is sufficiently large and the network sizes are sufficiently large, then

$$(1.1) \qquad \mathbb{E}\left[\mathcal{W}_1(S^\Sigma[(g_{n,m}^*)_\sharp \rho], \mu)\right] \text{ is controlled by } \left(\frac{1}{|\Sigma|\, n}\right)^{1/d},$$

where $\mathcal{W}_1(P, Q)$ is the Wasserstein-1 distance between two distributions P and Q; see (3.2). This reduction by a factor of $|\Sigma|$ in the number of training samples needed to learn a Σ-invariant target μ partially explains the *enhanced* data efficiency and generalization guarantees of GANs compared to existing generalization errors, such as those reported in [24], by leveraging the group structure. This reduction in training sample complexity can be interpreted as follows: the performance of a group-invariant GAN utilizing n i.i.d. training samples is equivalent to a vanilla GAN with $|\Sigma|\, n$ i.i.d. training samples. This is crucial, especially when the datasets available are scarce. If we further assume that μ is supported on a smooth d^*-dimensional submanifold of $\mathbb{R}^d$, then

$$(1.2) \qquad \mathbb{E}\left[\mathcal{W}_1(S^\Sigma[(g_{n,m}^*)_\sharp \rho], \mu)\right] \text{ is controlled by } \left(\frac{1}{|\Sigma|\, n}\right)^{1/d^*}.$$

This estimation suggests that the improvement in the error bound does not suffer from the *curse of dimensionality,* as it depends only on the *intrinsic dimension d^** of the target's support, which could be much smaller than the ambient dimension d. More importantly, we demonstrate that group-invariant GANs function effectively as though the input training data are augmented by the group actions in an i.i.d. manner (in contrast, augmented data are conditionally independent but not i.i.d.) without an increase in the number of parameters and

with a reduced discriminator approximation error lower bound compared to vanilla GANs, with fixed network architecture (see Proposition 4.3). For a visual illustration of the significant difference between data augmentation and learning with group-invariant GANs, see the numerical experiments in section 6 (Figures 2, 3, and 4). We remark that we only present the dominating terms in (1.1) and (1.2), assuming m is sufficiently large and the network sizes are sufficiently large, and their precise statement can be found in Theorems 4.9 and 5.2.

To the best of our knowledge, our work presents the first step toward theoretically understanding the impact of harnessing group symmetry within group-invariant generative models. While our primary focus lies on the performance guarantees for group-invariant GANs, we hope the analysis developed herein can be generalized in future works to study performance guarantees for other group-invariant generative models. See, for instance, [11], discussed at the end of section 2, and Remark 4.4.

This paper is organized as follows. In section 2, we review some related work. Section 3 provides the background and goals of our study. Theoretical results for the case when the target distribution lies in the Euclidean space or a low-dimensional submanifold are presented in sections 4 and 5, respectively. A numerical example is provided in section 6. We conclude our work and discuss future directions in section 7. Some of the proofs are deferred to sections 8 and 9.

2. Related work. A burgeoning body of recent research on group-equivariant neural networks [12, 13, 48] has demonstrated remarkable empirical success achieved through leveraging group symmetries in various supervised machine learning tasks. These accomplishments have, in turn, fostered the development of group-invariant/equivariant generative models, enabling data-efficient unsupervised learning of group-invariant distributions. Notable endeavors in this area include equivariant normalizing flows [5, 8, 19, 28], group-equivariant GANs [15], structure-preserving GANs [7] that rely on structure-customized divergences [6], and discrete and continuous equivariant diffusion models [23, 32]. Despite the intuitive understanding, the extent to which and the underlying reasons why group symmetry can enhance the data efficiency of those group-invariant generative models remain largely unknown.

On the other hand, recent research [9, 24, 31] has made significant progress in the sample-complexity analysis of *vanilla* GANs (i.e., GANs without group-symmetric generators and discriminators) in distribution learning. These works quantify the disparity, often measured using probability divergences, between the distribution learned by GANs with finite training samples and the underlying data distribution. Compared to these works, our analysis follows a similar oracle inequality and bounds each term of the inequality separately. However, the novelty of our work lies in the following: (1) we propose a version of group-invariant generator and discriminator architectures, and (2) we treat each error term with incorporated group-invariant network architectures and utilize the fact that the target distribution is group-invariant. While recent work [10, 44] has explored the enhanced sample complexity of divergence estimation under group symmetry, it remains unclear how these findings can be extended to the statistical estimation of group-invariant GANs aimed at learning group-symmetric probability distributions. Moreover, the authors of [10] in particular did not consider distributions with low-dimensional support, and their technical assumptions on the group actions are notably restrictive. In our work, we extend the analysis of GANs to include

cases where distributions possess low-dimensional structure, and we adopt a more flexible approach toward the assumptions on finite group actions compared to those in [10].

While some other generative models, such as diffusion models, have gained traction recently; GANs as one-shot methods are still widely used in certain applications where computational efficiency and real-time generation are critical [33, 26, 41]. Moreover, the mathematical formulation of GANs is essentially the minimization of variational divergences. As the objective of general generative modeling can be viewed as minimizing the discrepancy between the generated and target distributions in terms of certain divergences, our work can shed light on the study of other generative models by considering the performance guarantees of invariant models evaluated by symmetry-adapted divergences. For example, in the recent work [11] we provide performance guarantees for score-based generative models with symmetries, inspired by the work in this paper.

3. Background and problem setup.

3.1. Integral probability metrics and GANs.
Let $\mathcal{X}$ be a measurable space, and let $\mathcal{P}(\mathcal{X})$ be the set of probability measures on $\mathcal{X}$. Given $\Gamma \subset \mathcal{M}_b(\mathcal{X})$, where $\mathcal{M}_b(\mathcal{X})$ is the space of bounded measurable functions on $\mathcal{X}$, the Γ-integral probability metric (Γ-IPM) [34, 43] between $\nu \in \mathcal{P}(\mathcal{X})$ and $\mu \in \mathcal{P}(\mathcal{X})$ is defined as

$$(3.1) \qquad d_\Gamma(\nu, \mu) := \sup_{\gamma \in \Gamma} \left\{ \mathbb{E}_\nu[\gamma] - \mathbb{E}_\mu[\gamma] \right\}.$$

For example, if $\Gamma = \mathrm{Lip}_H(\mathcal{X})$, the space of H-Lipschitz functions on $\mathcal{X}$, then by the Kantorovich–Rubinstein duality, we have

$$(3.2) \qquad \mathcal{W}_1(\nu, \mu) = H^{-1} d_\Gamma(\nu, \mu).$$

For simplicity, we omit the factor H^{-1} in the following presentation. When Γ is a class of neural networks, $d_\Gamma(\nu, \mu)$ defined by (3.1) is called the "neural network distance" [2].

GANs, originally proposed by Goodfellow et al. [20], learn to approximate a target data distribution μ by generating samples from a noise source distribution ρ (typically chosen as uniform or Gaussian) through a two-player game. More specifically, GANs aim to solve the minmax problem

$$(3.3) \qquad \min_{g \in \mathcal{G}} \max_{f \in \mathcal{D}} \mathbb{E}_{x \sim \mu}[f(x)] - \mathbb{E}_{z \sim \rho}[f(g(z))],$$

where $\mathcal{G}$ is a class of functions called *generators,* and $\mathcal{D}$ is a class of functions called *discriminators.* This minmax problem can be equivalently formulated as

$$\min_{g \in \mathcal{G}} d_\mathcal{D}(\mu, g_\sharp \rho),$$

where $g_\sharp \rho := \rho \circ g^{-1}$ is the push-forward measure of ρ under g. For example, if $\mathcal{D} = \mathrm{Lip}_1(\mathcal{X})$ is the class of 1-Lipschitz functions, this minmax problem is the formulation of Wasserstein GANs (W-GANs) [1].

Suppose we have only finite samples $\{x_i\}_{i=1}^n$ from the target μ and $\{z_i\}_{i=1}^m$ from the source ρ; we define the optimal generator as the minimizer of the optimization problem [9, 24]

$$(3.4) \qquad g_{n,m} = \arg\min_{g \in \mathcal{G}} d_{\mathcal{D}}(g_\sharp \widehat{\rho}_m, \widehat{\mu}_n),$$

where $\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ is the empirical distribution of the target μ, and $\widehat{\rho}_m = \frac{1}{m} \sum_{i=1}^m \delta_{z_i}$ is the empirical distribution of the source ρ.

3.2. Group invariance and symmetrization operators. A *group* is a set Σ equipped with a group product satisfying the properties of associativity, identity, and invertibility. Given a group Σ and a set $\mathcal{X}$, a map $\theta : \Sigma \times \mathcal{X} \to \mathcal{X}$ is called a *group action on $\mathcal{X}$* if $\theta_\sigma := \theta(\sigma, \cdot) : \mathcal{X} \to \mathcal{X}$ is a bijection on $\mathcal{X}$ for any $\sigma \in \Sigma$, and $\theta_{\sigma_2} \circ \theta_{\sigma_1} = \theta_{\sigma_2 \cdot \sigma_1} \; \forall \sigma_1, \sigma_2 \in \Sigma$. If the context is clear, we will abbreviate $\theta(\sigma, x)$ as σx.

A function $\gamma : \mathcal{X} \to \mathbb{R}$ is said to be Σ-*invariant* if it is constant over group orbits of $\mathcal{X}$, i.e., $\gamma \circ \theta_\sigma = \gamma \; \forall \; \sigma \in \Sigma$. Let $\Gamma \subset \mathcal{M}_b(\mathcal{X})$ be a set of bounded measurable functions $\gamma : \mathcal{X} \to \mathbb{R}$. We define its subset, $\Gamma_\Sigma \subset \Gamma$, of Σ-invariant functions as

$$(3.5) \qquad \Gamma_\Sigma := \{\gamma \in \Gamma : \gamma \circ \theta_\sigma = \gamma \; \forall \sigma \in \Sigma\}.$$

Likewise, a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ is said to be Σ-*invariant* if $\mu = (\theta_\sigma)_\sharp \mu$ for all $\sigma \in \Sigma$. We define the set of all Σ-invariant distributions on $\mathcal{X}$ as

$$(3.6) \qquad \mathcal{P}_\Sigma(\mathcal{X}) := \{\mu \in \mathcal{P}(\mathcal{X}) : \mu \text{ is } \Sigma\text{-invariant}\}.$$

Finally, following [7, 10], 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

$$(3.7) \qquad S_\Sigma[\gamma](x) := \int_\Sigma \gamma(\sigma x) \mu_\Sigma(d\sigma) \; \forall \gamma \in \mathcal{M}_b(\mathcal{X}),$$

$$(3.8) \qquad \mathbb{E}_{S^\Sigma[\mu]} \gamma := \mathbb{E}_\mu S_\Sigma[\gamma] \; \forall \mu \in \mathcal{P}(\mathcal{X}), \forall \gamma \in \mathcal{M}_b(\mathcal{X}),$$

where μ_Σ is the unique Haar probability measure on a compact Hausdorff topological group Σ [18]. One can easily verify that S_Σ and S^Σ are, respectively, *projection operators* onto their corresponding invariant subsets $\Gamma_\Sigma \subset \Gamma$ and $\mathcal{P}_\Sigma(\mathcal{X}) \subset \mathcal{P}(\mathcal{X})$ [7]. One of the main results of [7] that motivates the construction of group-invariant GANs is summarized in the following lemma.

Lemma 3.1 (paraphrased from [7]). *If $S_\Sigma[\Gamma] \subset \Gamma$ and $\nu, \mu \in \mathcal{P}(X)$, then*

$$d_\Gamma(S^\Sigma[\nu], S^\Sigma[\mu]) = d_{\Gamma_\Sigma}(\nu, \mu) = \sup_{\gamma \in \Gamma_\Sigma} \{\mathbb{E}_\nu[\gamma] - \mathbb{E}_\mu[\gamma]\}.$$

In particular, if $\nu, \mu \in \mathcal{P}_\Sigma(\mathcal{X})$ are Σ-invariant, then $d_\Gamma(\nu, \mu) = d_{\Gamma_\Sigma}(\nu, \mu)$.

In other words, to "tell the difference" between two Σ-invariant distributions $\nu, \mu \in \mathcal{P}_\Sigma(\mathcal{X})$, one need only optimize over all Σ-invariant discriminators $\gamma \in \Gamma_\Sigma$. *Group-invariant GANs* are thus a type where the discriminators are all Σ-invariant, and the generators are parameterized in such a way that the generated distributions are always Σ-invariant. We will further elaborate on the exact construction of group-invariant GANs in section 3.4.

3.3. Assumptions on the group actions. We assume $\mathcal{X} \subset \mathbb{R}^d$ is a bounded subset of $\mathbb{R}^d$ equipped with the Euclidean metric $\|\cdot\|_2$. In addition, we make the following assumptions on the group Σ and its action on $\mathcal{X}$.

Assumption 3.2. The group Σ and its action on $\mathcal{X} \subset \mathbb{R}^d$ satisfy the following:
1. Σ is finite, i.e., $|\Sigma| < \infty$;
2. Σ acts linearly on $\mathcal{X}$; that is, $\sigma(ax_1 + bx_2) = a\sigma x_1 + b\sigma x_2 \; \forall x_1, x_2 \in \mathcal{X}, \sigma \in \Sigma, a, b \in \mathbb{R}$;
3. the Σ-actions on $\mathcal{X}$ are isometric; that is,

$$\|\sigma x_1 - \sigma x_2\|_2 = \|x_1 - x_2\|_2 \; \forall x_1, x_2 \in \mathcal{X}, \sigma \in \Sigma.$$

Remark 3.3. In terms of the third condition, one can instead assume that the Σ-actions on $\mathcal{X}$ are 1-Lipschitz: that is, $\|\sigma x_1 - \sigma x_2\|_2 \le \|x_1 - x_2\|_2 \; \forall x_1, x_2 \in \mathcal{X}, \sigma \in \Sigma$. Weyl's unitary trick tells us that for every finite group action, there exists a basis in which the action is isometric, such that the third condition can always be satisfied with a simple change of basis.

The Σ action on $\mathcal{X}$ induces a *fundamental domain* $\mathcal{X}_0 \subset \mathcal{X}$, a terminology which we adopt from [10].

Definition 3.4 (fundamental domain). *A subset $\mathcal{X}_0 \subset \mathcal{X}$ is called a fundamental domain of $\mathcal{X}$ under the Σ-action if for any $x \in \mathcal{X}$, there exists a unique $x_0 \in \mathcal{X}_0$ such that $x = \sigma x_0$ for some $\sigma \in \Sigma$.*

We remark that the choice of the fundamental domain $\mathcal{X}_0$ is generally not unique. Following [10], we abuse the notation $\mathcal{X} = \Sigma \times \mathcal{X}_0$ to denote $\mathcal{X}_0$ as a fundamental domain of $\mathcal{X}$ under the Σ-action. Given a specific choice of the fundamental domain $\mathcal{X}_0$, we define the projection $T_0 : \mathcal{X} \to \mathcal{X}_0$ as

$$(3.9) \qquad T_0(x) := y \in \mathcal{X}_0 \text{ if } y = \sigma x \text{ for some } \sigma \in \Sigma.$$

In other words, T_0 maps any $x \in \mathcal{X}$ to its unique orbit representative in $\mathcal{X}_0$. Given a Σ-invariant distribution $\mu \in \mathcal{P}_\Sigma(\mathcal{X})$ on $\mathcal{X}$, the map T_0 induces a distribution $\mu_{\mathcal{X}_0} \in \mathcal{P}(\mathcal{X}_0)$ on the fundamental domain $\mathcal{X}_0$ defined by

$$(3.10) \qquad \mu_{\mathcal{X}_0} = (T_0)_\sharp \mu.$$

We next introduce an important concept, the *covering number,* which appears in many proofs throughout our paper.

Definition 3.5 (covering number). *Let $(\mathcal{X}, \tau)$ be a metric space. A subset $S \subset \mathcal{X}$ is called an ϵ-cover of $\mathcal{X}$ if for any $x \in \mathcal{X}$ there is an $s \in S$ such that $\tau(s, x) \le \epsilon$. Define the ϵ-covering number of $\mathcal{X}$ as*

$$\mathcal{N}(\mathcal{X}, \epsilon, \tau) := \min\{|S| : S \text{ is an } \epsilon\text{-cover of } \mathcal{X}\}.$$

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

With Definitions 3.4 and 3.5, we specify below our second technical assumption on the group action over $\mathcal{X}$, which is motivated by [42, 10].

Assumption 3.6. There exists some constant ϵ_Σ and a set-valued function $A_0(\epsilon)$, such that for any $\epsilon \in (0, \epsilon_\Sigma)$, we have $A_0(\epsilon) \subset \mathcal{X}_0$, and
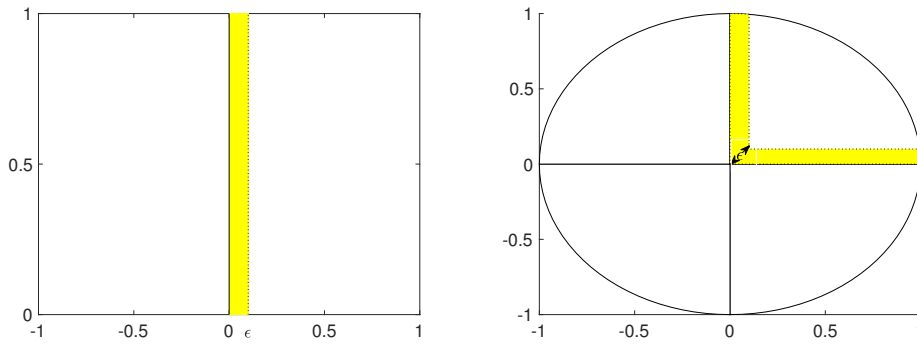
Figure 1. *Examples of Assumption* 3.6. *Left: Mirror reflection in* $\mathbb{R}^2$. $\mathcal{X}_0 = [0,1] \times [0,1]$. *The group actions are generated by* $\sigma(x,y) = (-x,y)$. *Right: Circular rotations within the unit disk in* $\mathbb{R}^2$. *The group actions are generated by the* $\pi/2$-*rotation with respect to the origin.* $\mathcal{X}_0 = (\rho\cos\theta, \rho\sin\theta)$, $\rho \in [0,1]$, $\theta \in [0,\pi/2)$. $A_0(\epsilon)$ *are colored yellow in both subfigures.*

1. $\|\sigma x - \sigma' x'\|_2 > 2\epsilon \ \forall x, x' \in \mathcal{X}_0 \backslash A_0(\epsilon), \sigma \neq \sigma' \in \Sigma$;
2. $\|\sigma x - \sigma x'\|_2 = \|x - x'\|_2 \ \forall x, x' \in \mathcal{X}_0 \backslash A_0(\epsilon), \sigma \in \Sigma$.

Moreover, for some $0 \leq r \leq d$, we have

$$(3.11) \qquad\qquad \limsup_{\epsilon \to 0^+} \frac{\mathcal{N}(A_0(\epsilon), \epsilon)}{\mathcal{N}(\mathcal{X}_0, \epsilon)\epsilon^r} < \infty.$$

Remark 3.7. Assumption 3.6 is notably less restrictive compared to the sufficient condition of Theorem 3 in [42], which can be viewed as a special case of Assumption 3.6 when $A(\epsilon) = \emptyset$ and $r = 0$. Such relaxation is crucial to accommodate symmetries such as mirror reflections and circular rotations in the following examples.

Example 3.8 (mirror reflection in $\mathbb{R}^2$). $\mathcal{X} = [-1,1] \times [0,1]$. The group actions are generated by $\sigma(x,y) = (-x,y)$. In this case, we have $r = 1$ in Assumption 3.6. See the left subfigure in Figure 1.

Example 3.9 (rotations in $\mathbb{R}^2$). $\mathcal{X} = (\rho\cos\theta, \rho\sin\theta)$, $\rho \in [0,1]$, $\theta \in [0,2\pi)$ and $\mathcal{X}_0 = (\rho\cos\theta, \rho\sin\theta)$, $\rho \in [0,1]$, $\theta \in [0,\pi/2)$. The group actions are generated by the $\pi/2$-rotation with respect to the origin. In this case, we also have $r = 1$ in Assumption 3.6. See the right subfigure in Figure 1.

Remark 3.10. The introduction of A_0 implies that group actions are *effective*: if $x_0 \in \mathcal{X}_0$, then there is an infinitesimal lower bound ϵ between σx_0 and $\mathcal{X}_0$ for any $\sigma \neq id$. For example, in Example 3.8, if $x_0 \in \mathcal{X}_0$ is arbitrarily close to the border of $\mathcal{X}_0$, i.e., $\{0\} \times [0,1]$, then $\sigma_1 x_0$ will be arbitrarily close to $\mathcal{X}_0$, where σ_1 is the reflection action; similarly for Example 3.9, when $x_0 \in \mathcal{X}_0$ is close to $\theta = 0$ and σ_1 is a counterclockwise rotation by $\pi/2$. Also note that the assumptions in [42, 10] both fail to satisfy the two cases in Figure 1.

3.4. Network architecture for group-invariant GANs. In GANs, both the generator class and the discriminator class are parameterized by neural networks. In this paper, we consider fully connected feed-forward Rectified Linear Unit (ReLU) networks as the architecture for both generators and discriminators.

A fully connected feed-forward ReLU neural network, $\phi : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{L+1}}$, can be formulated as

$$\phi(x) = W_L \cdot \text{ReLU}(W_{L-1} \cdots \text{ReLU}(W_0 x + b_0) \cdots + b_{L-1}) + b_L, \tag{3.12}$$

where $W_i \in \mathbb{R}^{d_{i+1} \times d_i}$ and $b_i \in \mathbb{R}^{d_{i+1}}$ are network weights, $i = 0, 1, \ldots, L$, and the activation function is given by $\text{ReLU}(x) := \max\{0, x\}$, $x \in \mathbb{R}$. We call the numbers $W = \max\{d_1, \ldots, d_L\}$ and L the width and the depth of the neural network ϕ, respectively. When the input and output dimensions are clear from the context, we denote by $\mathcal{NN}(W, L, N, K)$ a ReLU *network architecture* that consists of fully connected feed-forward ReLU neural networks with width at most W, depth at most L, number of weights at most N, and magnitude of entries of W_i and b_i no more than K. We also write $\mathcal{NN}(W, L)$ or $\mathcal{NN}(W, L, N)$ when either N or K is not specified.

Let $\mathcal{D}_{NN}$ and $\mathcal{G}_{NN}$, respectively, denote a class of discriminators and a class of generators realized by fully connected ReLU networks, which are generally not Σ-invariant. We explain next how to construct Σ-invariant discriminators and generators from given $\mathcal{D}_{NN}$ and $\mathcal{G}_{NN}$.

Since Σ acts linearly on $\mathcal{X}$ (Assumption 3.2), there is a matrix representation $W_\sigma \in \mathbb{R}^{d \times d}$ for any $\sigma \in \Sigma$ such that $\sigma x = W_\sigma x \forall x \in \mathbb{R}^d$. We can formulate Σ-invariant discriminators, denoted by $\mathcal{D}_{NN}^{\Sigma}$, using the function symmetrization operator S_Σ introduced in (3.7).

Definition 3.11 (Σ-invariant discriminators). *For any class of discriminators with ReLU network architecture $\mathcal{D}_{NN} = \mathcal{NN}(W, L, N, K)$, its corresponding class of Σ-invariant discriminators $\mathcal{D}_{NN}^{\Sigma}$ is defined as*

$$\mathcal{D}_{NN}^{\Sigma} := \left\{ \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} \phi(\sigma_i^{-1} x) : \phi(x) \in \mathcal{D}_{NN} \right\}. \tag{3.13}$$

That is, $\mathcal{D}_{NN}^{\Sigma}$ consists of neural networks of the form

$$\phi(x) = \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} W_L \cdot ReLU(W_{L-1} \cdots ReLU(W_0(W_{\sigma_i} x) + b_0) \cdots + b_{L-1}) + b_L, \tag{3.14}$$

where $\{W_i, b_i : i = 0, \ldots, L\}$ is the same set of weights as that of $\mathcal{D}_{NN}$.

By Definition 3.11, it is straightforward to verify that $S_\Sigma[\mathcal{D}_{NN}^{\Sigma}] = \mathcal{D}_{NN}^{\Sigma}$.

Remark 3.12. Definition 3.11 requires $|\Sigma|$ passes through the network; however, in numerous practical scenarios of generative modeling, particularly in medical imaging, the foremost limitation is typically the scarcity of sufficient data rather than computational resources. Consequently, while Definition 3.11 increases computational demands, it represents a necessary trade-off to effectively address this significant challenge in sample-constrained applications.

Similarly, the Σ-invariant generators can be defined using the distribution symmetrization operator S^Σ introduced in (3.8).

Definition 3.13 (Σ-invariant generators). *For any class of generators with ReLU network architecture* $\mathcal{G}_{NN} = \mathcal{NN}(W, L, N, K)$, *its corresponding class of Σ-invariant generators* $\mathcal{G}_{NN}^\Sigma$ *is defined as*

$$g_\Sigma \in \mathcal{G}_{NN}^\Sigma \iff \exists g(x) \in \mathcal{G}_{NN} \ s.t. \ g_\Sigma(\rho) = S^\Sigma[g_\sharp \rho] \ \forall \ noise \ source \ distribution \ \rho.$$

In other words, we add a Σ-symmetrization layer S^Σ that draws a random σ according to the Haar measure μ_Σ and transforms the output $g(x)$ of $\mathcal{G}_{NN}$ to $\sigma g(x)$ the output of $\mathcal{G}_{NN}^\Sigma$.

In this paper, we consider learning a Σ-invariant target data distribution $\mu \in \mathcal{P}_\Sigma(\mathcal{X})$ using Σ-invariant GANs whose generators and discriminators are both Σ-invariant and where the noise source probability measure ρ is absolutely continuous on $\mathbb{R}$ (with respect to the Lebesgue measure). More specifically, in contrast to (3.4), we denote

$$(3.15) \qquad\qquad g_{n,m}^* = \underset{g \in \mathcal{G}_{NN}}{\arg\min} \, d_{\mathcal{D}_{NN}^\Sigma}(S^\Sigma[g_\sharp \widehat{\rho}_m], \widehat{\mu}_n).$$

Note that $\{S^\Sigma[g_\sharp \rho] : g \in \mathcal{G}_{NN}\}$ is a set of Σ-invariant distributions in $\mathcal{X}$ for any easy-to-sample source distribution ρ on $\mathbb{R}$.

Remark 3.14. Our results also hold if ρ is an absolutely continuous probability measure on $\mathbb{R}^k$, since any such ρ can be projected to an absolutely continuous probability measure on $\mathbb{R}$ by a linear mapping, as was shown in Remark 7 of [24].

Remark 3.15. In [9, 24], the generator $g_n^* = \arg\min_{g \in \mathcal{G}_{NN}} d_{\mathcal{D}_{NN}^\Sigma}(S^\Sigma[g_\sharp \rho], \widehat{\mu}_n)$ is also considered. We will not provide the details for g_n^* since, in practice, GANs are trained with finite samples from ρ. Moreover, the bound in our main result, Theorem 4.9, directly applies to the case for g_n^*.

3.5. Notation. We write $A \lesssim B$ if $A \leq CB$ for some constant $C > 0$, and if both $A \lesssim B$ and $B \gtrsim A$ hold, then we write $A \simeq B$. We also write $A \lesssim_s B$ if the factor C depends on some quantity s. Given a subset $\mathcal{X} \subset \mathbb{R}^d$ of $\mathbb{R}^d$, its *diameter* is denoted as $\text{diam}(\mathcal{X}) := \sup_{x,y \in \mathcal{X}} \|x - y\|_2$. For two functions $f(n)$ and $g(n)$, we write $f(n) = o(g(n))$ when $\lim_{n \to \infty} f(n)/g(n) = 0$.

4. Learning distributions in Euclidean spaces. In this section, we quantify the error between the generated distribution $S^\Sigma[(g_{n,m}^*)_\sharp \rho]$ and the target data distribution $\mu \in \mathcal{P}_\Sigma(\mathcal{X})$ in terms of the Wasserstein-1 metric; that is, $\Gamma = \text{Lip}_H(\mathcal{X})$ and we estimate $d_\Gamma(S^\Sigma[(g_{n,m}^*)_\sharp \rho], \mu)$. All of the proofs can be found in the appendix. Since the value of the Γ-IPM does not change if we replace Γ by $\Gamma + c$ for any $c \in \mathbb{R}$, by Lemma 2 in [10], it is equivalent to replace Γ by $\Gamma = \{f : f \in \text{Lip}_H(\mathcal{X}), \|f\|_\infty \leq M\}$ if $\mathcal{X}$ is bounded, where $M = H \cdot \text{diam}(\mathcal{X}_0)$.

The proof of Theorem 4.9 hinges upon the following error decomposition based on the proof of Lemma 9 in [24] and Lemma 3.1. Compared to [24], our analysis makes use of the group-invariant architectures of the generator and discriminator as well as the fact that the target distribution has group-invariant structure so that each error term will reveal the gain or no extra cost of group invariance.

Lemma 4.1 (error decomposition). *Suppose $g_{n,m}^*$ is the optimal GAN generator from* (3.15); *then for any function class Γ defined on $\mathcal{X} \subset \mathbb{R}^d$, we have*

$$d_\Gamma(S^\Sigma[(g_{n,m}^*)_\sharp \rho], \mu) \leq \underbrace{2 \sup_{f \in \Gamma_\Sigma} \inf_{f_\omega \in \mathcal{D}_{NN}^\Sigma} \|f - f_\omega\|_\infty}_{\text{Invariant discriminator approximation error: } \Delta_1}$$

$$+ \underbrace{\inf_{g \in \mathcal{G}_{NN}} d_{\mathcal{D}_{NN}^\Sigma}(S^\Sigma[g_\sharp \rho], \widehat{\mu}_n)}_{\text{Invariant generator approximation error: } \Delta_2} + \underbrace{d_{\Gamma_\Sigma}(\widehat{\mu}_n, \mu)}_{\text{Statistical error from target: } \Delta_3}$$

$$+ \underbrace{2d_{\mathcal{D}_{NN}^\Sigma \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m)}_{\text{Statistical error from source: } \Delta_4} .$$

We provide lemmas that bound each of the Δ_i's.

Lemma 4.2 (Σ-invariant discriminator approximation error). *Let*

$$\Gamma = \{f : f \in Lip_H(\mathcal{X}), \|f\|_\infty \leq H \cdot diam(\mathcal{X}_0)\}.$$

For any $\epsilon \in (0,1)$, there exists a class of Σ-invariant discriminators

$$\mathcal{D}_{NN}^\Sigma = S_\Sigma[\mathcal{NN}(W_1, L_1, N_1, K_1)]$$

with $L_1 \lesssim \log(1/\epsilon)$, $N_1 \lesssim \epsilon^{-d} \log(1/\epsilon)$ and K_1 depends on $H, \sup_{x \in \mathcal{X}} \|x\|_\infty$ and $diam(\mathcal{X}_0)$, such that

$$\sup_{f \in \Gamma_\Sigma} \inf_{f_\omega \in \mathcal{D}_{NN}^\Sigma} \|f - f_\omega\|_\infty \leq \epsilon.$$

In addition to Lemma 4.2, we take a closer look at each individual approximation to illustrate the importance of using Σ-invariant discriminators in the next proposition, whose proof can be found in subsection 8.2.

Proposition 4.3. *Suppose f is a Σ-invariant function and f_ω is not necessarily Σ-invariant. Assuming Σ has unitary matrix representations, i.e., the Jacobian of each θ_σ is a unitary matrix, we have*

$$\|f - f_\omega\|_\infty \geq \sqrt{\frac{1}{vol(\mathcal{X})} \left(\|f - S_\Sigma[f_\omega]\|_{L^2(\mathcal{X})}^2 + \|f_\omega - S_\Sigma[f_\omega]\|_{L^2(\mathcal{X})}^2 \right)}$$

$$\geq \sqrt{\frac{1}{vol(\mathcal{X})}} \|f_\omega - S_\Sigma[f_\omega]\|_{L^2(\mathcal{X})}.$$

Remark 4.4 (deviation from invariance). The right-hand side of Proposition 4.3 can be viewed as a measurement of the noninvariance of f_ω and can be large if f_ω is "far from" being Σ-invariant. This lower bound vanishes whenever f_ω is Σ-invariant, which suggests that we use Σ-invariant discriminators $\mathcal{D}_{NN}^\Sigma$ instead of $\mathcal{D}_{NN}$ to achieve a smaller discriminator approximation error. See also Figure 4 for numerical illustrations. This measurement for the deviation from invariance seems general in generative models with symmetries. For example, [11] uses a similar notion to measure the nonequivariance of a vector field in the context of score-based generative models for score-matching approximations.

Lemma 4.5 (Σ-invariant generator approximation error). *Suppose $W_2 \geq 7d + 1, L_2 \geq 2$. Let ρ be an absolutely continuous probability measure on $\mathbb{R}$. If $n \leq \frac{W_2 - d - 1}{2} \lfloor \frac{W_2 - d - 1}{6d} \rfloor \lfloor \frac{L_2}{2} \rfloor + 2$, we have $\inf_{g \in \mathcal{G}_{NN}} d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[g_\sharp \rho], \widehat{\mu}_n) = 0$, where $\mathcal{G}_{NN} = \mathcal{NN}(W_2, L_2)$.*

Remark 4.6. We also refer the reader to Theorem 4.6 in [7] for the necessity of using Σ-invariant generators in the following case: Suppose the discriminator architecture is Σ-invariant. Since $d_{\mathcal{D}_{NN}^{\Sigma}}(g_\sharp \rho, \widehat{\mu}_n) = d_{\mathcal{D}_{NN}}(S^{\Sigma}[g_\sharp \rho], S^{\Sigma}[\widehat{\mu}_n])$, the generator may only learn to generate distributions whose symmetrization is the target distribution, thus causing "mode collapse." That is, $d_{\mathcal{D}_{NN}^{\Sigma}}$ is a probability metric on $\mathcal{P}_\Sigma(\mathcal{X})$ rather than $\mathcal{P}(\mathcal{X})$. We also refer the reader to the qualitative numerical examples of this phenomenon in [7].

Lemma 4.7 (statistical error from the target). *Let $\mathcal{X} = \Sigma \times \mathcal{X}_0$ be a subset of $\mathbb{R}^d$ satisfying the conditions in Assumptions 3.2 and 3.6. Let $\Gamma = \{f : f \in Lip_H(\mathcal{X}), \|f\|_\infty \leq M\}$. Suppose μ is a Σ-invariant probability measure on $\mathcal{X}$. Then we have the following:*

(1) If $d \geq 2$, then for any $s > 0$ and n sufficiently large, we have

$$\mathbb{E}[d_{\Gamma_\Sigma}(\widehat{\mu}_n, \mu)] \leq \quad C_{\mathcal{X}, H, d, s} \left(\frac{1}{|\Sigma| \, n} \right)^{\frac{1}{d+s}} + o \left(\left(\frac{1}{n} \right)^{\frac{1}{d+s}} \right),$$

where the coefficient $C_{\mathcal{X}, H, d, s}$ of the dominating term does not depend on Σ or n;

(2) If $d = 1$ and $\mathcal{X}_0$ is an interval, then for n sufficiently large, we have

$$\mathbb{E}[d_{\Gamma_\Sigma}(\widehat{\mu}_n, \mu)] \leq \frac{cH \cdot \text{diam}(\mathcal{X}_0)}{\sqrt{n}} + o \left(\frac{1}{n} \right),$$

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

Lemma 4.8 (statistical error from the source). *Suppose $L_1 \lesssim \log n$, $N_1 \lesssim n \log n$, $W_2^2 L_2 \lesssim n$; then we have $d_{\mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m) \lesssim \sqrt{\frac{n^2 \log^2 n \log m}{m}}$. In particular, if $m \gtrsim n^{2+2/d} \log^3 n$, we have $\mathbb{E}[d_{\mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m)] \lesssim n^{-1/d}$.*

With those lemmas, we now state and prove the main theorem.

Theorem 4.9 (main theorem). *Let $\mathcal{X} = \Sigma \times \mathcal{X}_0$ be a compact subset of $\mathbb{R}^d (d \geq 2)$ satisfying Assumptions 3.2 and 3.6, and let $\Gamma = Lip_H(\mathcal{X})$. Suppose the target distribution μ is Σ-invariant on $\mathcal{X}$, and the noise source distribution ρ is absolutely continuous on $\mathbb{R}$. Then there exists Σ-invariant discriminator architecture $\mathcal{D}_{NN}^{\Sigma} = S_\Sigma[\mathcal{D}_{NN}]$, where $\mathcal{D}_{NN} = \mathcal{NN}(W_1, L_1, N_1)$ as defined in (3.14) with $N_1 \lesssim n \log n$ and $L_1 \lesssim \log n$, and Σ-invariant generator architecture $\mathcal{G}_{NN}^{\Sigma}$, where $\mathcal{G}_{NN} = \mathcal{NN}(W_2, L_2)$, with $W_2^2 L_2 \lesssim n$, such that if $m \gtrsim n^{2+2/d} \log^3 n$, we have*

$$(4.1) \qquad \mathbb{E}\left[d_\Gamma(S^{\Sigma}[(g_{n,m}^*)_\sharp \rho], \mu) \right] \leq C_{\mathcal{X}, H, d, s} \left(\frac{1}{|\Sigma| \, n} \right)^{\frac{1}{d+s}} + o \left(\left(\frac{1}{n} \right)^{\frac{1}{d+s}} \right),$$

for any $s > 0$, where the coefficient $C_{\mathcal{X}, H, d, s}$ of the dominating term does not depend on Σ or n. See (8.5) in the appendix for details.

Proof. We can choose $\epsilon \simeq n^{-1/d}$ in Lemma 4.2 with $L_1 \lesssim \log n^{1/d} \simeq \log n$ and $N_1 \lesssim n \log n$, so that $\sup_{f \in \Gamma_\Sigma} \inf_{f_\omega \in \mathcal{D}_{NN}^\Sigma} \|f - f_\omega\|_\infty \lesssim n^{-1/d}$. On the other hand, we can make $\inf_{g \in \mathcal{G}_{NN}} d_{\mathcal{D}_{NN}^\Sigma}(S^\Sigma[g_\sharp \rho], \widehat{\mu}_n) = 0$ with $W_2^2 L_2 \lesssim n$ by Lemma 4.5. Finally, $\mathbb{E}[d_{\Gamma_\Sigma}(\widehat{\mu}_n, \mu)] \lesssim_s (|\Sigma| n)^{-1/(d+s)}$ and $\mathbb{E}[d_{\mathcal{D}_{NN}^\Sigma \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m)] \lesssim n^{-1/d}$ by Lemmas 4.7 and 4.8 respectively. ∎

The statement of the theorem in the case of $d = 1$ can be found in the appendix.

Remark 4.10 (data efficiency). This improved training sample complexity can be interpreted as follows: For vanilla GANs without built-in group symmetry, the group size is 1, representing the baseline sample complexity. However, for a group-invariant GAN with nontrivial group $|\Sigma| > 1$, its performance when utilizing n i.i.d. training samples is equivalent to a vanilla GAN with $|\Sigma| n$ i.i.d. training samples.

Remark 4.11 (optimality of the rate). The rate obtained in Theorem 4.9 is less sharp than $\lesssim n^{-1/d} \vee n^{-1/2} \log n$ provided in [24] since in [24] it is assumed that $\mathcal{X} = \mathcal{X}_0 = [0, 1]^d$, but here we do not assume the fundamental domain is connected. If we assume that $\mathcal{X}_0$ is connected, then the rate can be improved to $\lesssim (|\Sigma| n)^{-1/d}$ for $d \geq 3$ and $\lesssim_s (|\Sigma| n)^{-1/2} \log n$ for $d = 2$ since Lemma 8.5 in the proof can be replaced by a sharper covering number bound in [29].

5. Target distributions with low-dimensional structure.
While Theorem 4.9 explains the improved data efficiency of group-invariant GANs, the error bound increases as the ambient dimension d increases. However, empirical evidence suggests that numerous probability distributions found in nature are concentrated on submanifolds of low *intrinsic dimensions*. In light of this, we present an enhanced analysis of sample complexity for group-invariant GANs, specifically when learning distributions of low intrinsic dimension under group symmetry. Suppose the target distribution μ is supported on some compact d^*-dimensional smooth submanifold of $\mathbb{R}^d$. We first prove the following bound for the covering number of a compact smooth submanifold of $\mathbb{R}^d$, which reflects the intrinsic dimension of the submanifold.

Lemma 5.1. *Let $\mathcal{M}$ be a compact d^*-dimensional smooth submanifold of $\mathbb{R}^d$; then for small $\epsilon > 0$, we have*

$$\mathcal{N}(\mathcal{M}, \epsilon) \leq C_\mathcal{M} \left(\frac{1}{\epsilon}\right)^{d^*},$$

where the metric in the covering number is the Euclidean metric on $\mathbb{R}^d$, and the constant $C_\mathcal{M}$ depends on $\mathcal{M}$.

Proof. For any point $x \in \mathcal{M}$, there exists a chart (U_x, φ_x), where $U_x \subset \mathbb{R}^d$ is an open set containing x, and $\varphi_x : U_x \cap \mathcal{M} \to \varphi_x(U_x \cap \mathcal{M}) \subset \mathbb{R}^{d^*}$ is a diffeomorphism such that $\varphi_x(U_x \cap \mathcal{M})$ is an open set of $\mathbb{R}^{d^*}$. Without loss of generality, we can assume $\varphi_x(x) = 0 \in \mathbb{R}^{d^*}$. In addition, we can assume that $\varphi_x(U_x \cap \mathcal{M}) = B_{d^*}(0, R_x)$ for some $R_x > 0$, and $\|D\varphi_x^{-1}\|_2 \leq c_x$ on $B_{d^*}(0, R_x)$ for some constant c_x, where $D\varphi_x^{-1}$ is the Jacobian matrix of φ_x^{-1}. (This is possible since $\varphi_x(U_x)$ is open; we can pick some $r > 0$ such that $B_{d^*}(0, r) \subset \varphi_x(U_x)$, so that $\overline{B_{d^*}(0, r/2)}$ is compact, and $\|D\varphi_x^{-1}\|_2$ is bounded on $\overline{B_{d^*}(0, r/2)}$. Then we can set $U_x \cap \mathcal{M} = \varphi_x^{-1}(B_{d^*}(0, r/4))$.)

Let $\epsilon_x \leq \frac{R_x}{3}$; then any open ball in $\mathbb{R}^{d^*}$ that intersects $B_{d^*}(0, \epsilon_x)$ must lie within $B_{d^*}(0, R_x)$. For any small $\epsilon \in (0, 1)$, we know that $B_{d^*}(0, \epsilon_x)$ can be covered by $\frac{(\epsilon_x + \epsilon/2)^{d^*}}{(\epsilon/2)^{d^*}}$ balls with radius ϵ in $\mathbb{R}^{d^*}$ (cf. Proposition 4.2.12 in [47]). Moreover, since $\|D\varphi_x^{-1}\|_2 \leq c_x$, φ_x^{-1} maps each of these

ϵ-balls to some open subset of U_x whose diameter is no more than $c_x\epsilon$ and hence contained in some open ball of radius $2c_x\epsilon$ in $\mathbb{R}^d$. Therefore, $\varphi_x^{-1}(B_{d^*}(0,\epsilon_x))$ can be covered by $\frac{(\epsilon_x+\epsilon/2)^{d^*}}{(\epsilon/2)^{d^*}}$ balls with radius $2c_x\epsilon$ in $\mathbb{R}^d$. By a rescaling, $\varphi_x^{-1}(B_{d^*}(0,\epsilon_x))$ can be covered by $\frac{(\epsilon_x+\epsilon/4c_x)^{d^*}}{(\epsilon/4c_x)^{d^*}}$ balls with radius ϵ in $\mathbb{R}^d$. Also note that $\cup_{x\in\mathcal{M}}[\varphi_x^{-1}(B_{d^*}(0,\epsilon_x))]$ forms an open cover of $\mathcal{M}$, so there exist finitely many $x_1,\ldots,x_n$ such that $\mathcal{M}=\cup_{i=1}^n[\varphi_{x_i}^{-1}(B_{d^*}(0,\epsilon_{x_i}))]$, since $\mathcal{M}$ is compact. Thus $\mathcal{M}$ can be covered by $\sum_{i=1}^n \frac{(\epsilon_{x_i}+\epsilon/4c_{x_i})^{d^*}}{(\epsilon/4c_{x_i})^{d^*}} = \sum_{i=1}^n(\frac{4c_{x_i}\epsilon_{x_i}}{\epsilon}+1)^{d^*} \leq \sum_{i=1}^n(\frac{4c_{x_i}\epsilon_{x_i}+1}{\epsilon})^{d^*} \leq \frac{C_\mathcal{M}}{\epsilon^{d^*}}$ balls with radius ϵ in $\mathbb{R}^d$, where $C_\mathcal{M}=n\cdot\max_i\{(4c_{x_i}\epsilon_{x_i}+1)^{d^*}\}$. That is, $\mathcal{N}(\mathcal{M},\epsilon)\leq C_\mathcal{M}(\frac{1}{\epsilon})^{d^*}$. $\blacksquare$

Theorem 5.2. *Let $\mathcal{X}=\Sigma\times\mathcal{X}_0\subset\mathbb{R}^d$, and let $\mathcal{X}_0$ be a compact d^*-dimensional smooth submanifold of $\mathbb{R}^d(d^*\geq 2)$ satisfying Assumption 3.6 with some $0\leq r\leq d^*$ and $\Gamma=Lip_H(\mathcal{X})$. Suppose the target distribution μ is Σ-invariant on $\mathcal{X}$ and the noise source distribution ρ is absolutely continuous on $\mathbb{R}$. Then there exists Σ-invariant discriminator architecture $\mathcal{D}_{NN}^\Sigma = S_\Sigma[\mathcal{D}_{NN}]$, where $\mathcal{D}_{NN}=\mathcal{NN}(W_1,L_1,N_1)$ as defined in (3.14) with $N_1\lesssim n\log n$ and $L_1\lesssim\log n$, and Σ-invariant generator architecture $\mathcal{G}_{NN}^\Sigma$, where $\mathcal{G}_{NN}=\mathcal{NN}(W_2,L_2)$, with $W_2^2 L_2\lesssim n$, such that if $m\gtrsim n^{2+2/d^*}\log^3 n$, we have*

$$\mathbb{E}\left[d_\Gamma(S^\Sigma[(g_{n,m}^*)_\sharp\rho],\mu)\right]\leq C_{\mathcal{X},H,d^*,s}\left(\frac{1}{|\Sigma|n}\right)^{\frac{1}{d^*+s}}+o\left(\left(\frac{1}{n}\right)^{\frac{1}{d^*+s}}\right),$$

for any $s>0$, where the coefficient $C_{\mathcal{X},H,d^*,s}$ of the dominating term does not depend on Σ or n.

Proof. Following the notation in the proof of Lemma 4.7, we have, by Lemma 8.5,

$$\log\mathcal{N}(\mathcal{F}_0,\epsilon,\|\cdot\|_\infty)\leq\mathcal{N}\left(\mathcal{X}_0,\frac{c_2\epsilon}{H}\right)\log\left(\frac{c_1 M}{\epsilon}\right),$$

where $\mathcal{F}_0=\{\gamma\in Lip_H(\mathcal{X}_0):\|\gamma\|_\infty\leq M\}$. By Assumption 3.6 and Lemma 5.1, we have

$$\mathcal{N}\left(\mathcal{X}_0,\frac{c_2\epsilon}{H}\right)\log\left(\frac{c_1 M}{\epsilon}\right)\leq\mathcal{N}\left(\mathcal{X}_0\backslash A_0\left(\frac{c_2\epsilon}{H}\right),\frac{c_2\epsilon}{H}\right)\log\left(\frac{c_1 M}{\epsilon}\right)$$
$$+\mathcal{N}\left(A_0\left(\frac{c_2\epsilon}{H}\right),\frac{c_2\epsilon}{H}\right)\log\left(\frac{c_1 M}{\epsilon}\right)$$
$$\leq\frac{\mathcal{N}\left(\mathcal{X},\frac{c_2\epsilon}{H}\right)}{|\Sigma|}\log\left(\frac{c_1 M}{\epsilon}\right)+\bar{c}\epsilon^r\mathcal{N}\left(\mathcal{X},\frac{c_2\epsilon}{H}\right)\log\left(\frac{c_1 M}{\epsilon}\right)$$
$$\leq\frac{C_\mathcal{X}H^{d^*}}{|\Sigma|c_2^{d^*}\epsilon^{d^*}}\log\left(\frac{c_1 M}{\epsilon}\right)+\frac{\bar{c}C_\mathcal{X}H^{d^*}}{c_2^{d^*}\epsilon^{d^*-r}}\log\left(\frac{c_1 M}{\epsilon}\right)$$

when $\frac{c_2\epsilon^*}{H}\leq\epsilon_\Sigma$. Therefore, we have the statistical error from the target Δ_3 as

$$\mathbb{E}[d_{\Gamma_\Sigma}(\widehat{\mu}_n,\mu)]\leq C_{\mathcal{X},H,d^*,s}\left(\frac{1}{|\Sigma|n}\right)^{\frac{1}{d^*+s}}+o\left(\left(\frac{1}{n}\right)^{\frac{1}{d^*+s}}\right).$$

The rest of the prooffollows that of Theorem 4.9. $\blacksquare$

Remark 5.3. The statement of the theorem for 1-dimensional submanifolds can be found in the appendix.

Remark 5.4. The condition in Theorem 5.2 can be relaxed to be C^1-submanifolds, as the proof of Lemma 5.1 only requires a C^1 condition.

Remark 5.5. If the target distribution does not lie on a manifold, such as on fractals, we can introduce the Minkowski dimension, whose definition itself involves the covering number, so that we can derive the bound similarly as in Theorem 5.2.

6. Numerical experiments. We provide a synthetic example to illustrate efficient learning by incorporating group symmetry into the generator and discriminator networks. Specifically, the target is a mixture of four 2-dimensional Gaussian distributions centered at $[\pm 10, \pm 10]$, each with a standard deviation of 5, resulting in a target distribution with C_4 symmetry.

We train three different GANs based on the vanilla GAN that has three hidden fully connected layers with 64, 32, and 16 nodes in sequence with ReLU activations for both the generator and the discriminator. The discriminators and generators of C_2 and C_4 GANs are defined as in Definitions 3.11 and 3.13 with C_2 and C_4 groups, respectively. We applied a gradient penalty as described in [22] with a (soft) Lipschitz constant of 1 and penalty of 10 (see Theorem 31 in [6] for a theoretical justification). The noise source is selected to be a 10-dimensional Gaussian.

Figure 2 displays 5000 samples generated by the trained generator of each GAN, where we use 100 training samples, incorporating C_1 (no invariance, or vanilla), C_2, and C_4 symmetries. The GAN with C_4 symmetry (third row) achieves the best result, and the one with no invariance (top row) performs the worst. Additionally, the vanilla GAN without symmetry, even with full data augmentation, i.e., augmenting all the training data using group actions in C_4 and thus giving us 400 non-i.i.d. training samples (bottom row), performed only slightly better than the GAN without symmetry. This highlights the stark contrast between structured distribution learning with group-invariant GANs and using data augmentation.
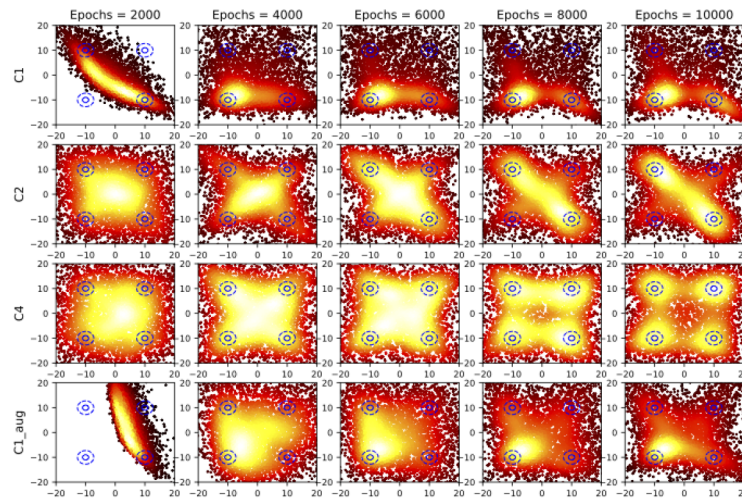

Figure 2. *Heat map of* 5000 *GAN generated samples learned from a* 2D *Gaussian mixture. Top row: GAN with no symmetry; second row: GAN with* C_2 *(partial) symmetry; third row: GAN with* C_4 *(full) symmetry; bottom row: GAN with no symmetry but with* C_4 *augmentation on the training data. Yellow parts refer to higher density.*
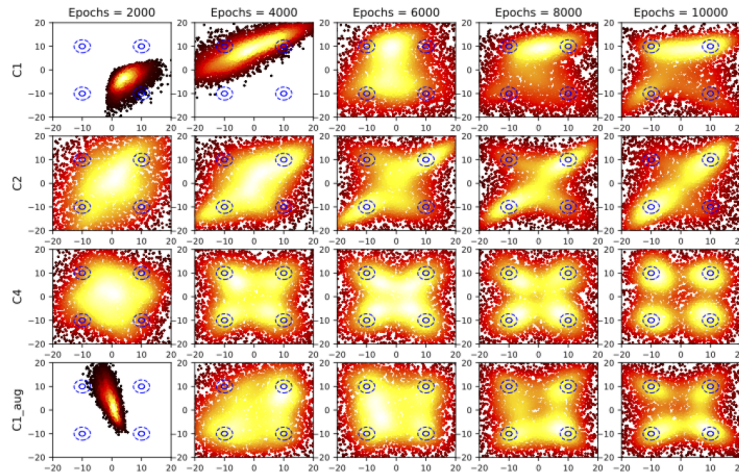
Figure 3. *Heat map of* 5000 *GAN generated samples learned from a* 2D *Gaussian mixture embedded in a higher-dimensional ambient space* $\mathbb{R}^{12}$. *The figure shows the* 2D *projection of the generated samples onto the intrinsic support plane* $(d^* = 2)$ *of the distribution. Top row: GAN with no symmetry; second row: GAN with* C_2 *(partial) symmetry; third row: GAN with* C_4 *(full) symmetry; bottom row: GAN with no symmetry but with* C_4 *augmentation on the training data. Compared to Figure* 2, *this result qualitatively suggests that the convergence depends only on the intrinsic dimension, as discussed in Theorem* 5.2.
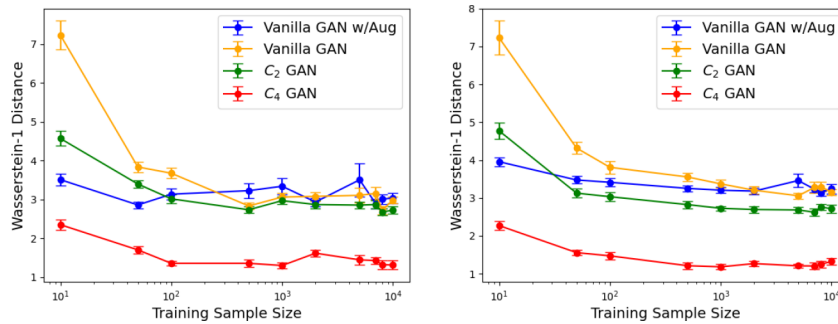

Figure 4. *Wasserstein-*1 *distance between* 10,000 *samples drawn from the generated and target distributions with different GAN implementations over* 20 *runs. Left:* 2D *example. Right:* 12D *example.* C_4 *GAN achieves the best performance in both cases.*

Moreover, Figure 3 presents similar results with the same (hyper)parameters, except that the 2D Gaussian mixture is linearly embedded into a higher-dimensional ambient space $\mathbb{R}^{12}$.

In Figure 4, we quantitatively evaluate the performance of methods by calculating the Wasserstein-1 distance between 10,000 samples drawn from the generated distributions by the trained GANs for 10,000 epochs and the target distribution of the Gaussian mixture, respectively. We apply the linear program in [16, 17] for the calculation of the Wasserstein-1 distance, and we set the maximal number of iterations to 10^6, and no cases reaching this limit were reported. For each training data size, we perform 20 independent runs for each method. We observe that the C_4 GAN consistently achieves the best performance in both examples. This is because the C_4 GAN not only has a smaller statistical error Δ_3 but also provides a smaller invariant discriminator approximation error Δ_1, which becomes the dominant error in

Lemma 4.1 when the number of training samples is large. (See Proposition 4.3 and Remark 4.4 for the reduced Δ_1 error by the invariant architecture.) We keep the same network architecture and hyperparameters in our numerical experiments to ensure consistency of performance across different methods.

7. Conclusion and future work. In this work, we quantitatively established the improved generalization guarantees of group-invariant GANs. Our results indicate that when learning group-invariant target distributions, potentially supported on a manifold with small intrinsic dimension, the required number of samples for group-invariant GANs is reduced by a factor of the group size. In essence, group-invariant GANs function effectively as though the input data were augmented in the statistical sense, without an increase in the number of parameters, compared to vanilla GANs. This becomes crucial when we face data scarcity. On the other hand, however, the overall improved bound cannot simply be derived from data augmentation since invariant GANs have a reduced discriminator approximation error, and our numerical results further support this distinction. Our findings present several potential avenues for further exploration. First, it is essential to explore how the dimension of the support of the noise source ρ influences the performance of GANs. A recent experimental study [54] suggests that the dimension of the source should not be significantly smaller than the intrinsic dimension of the target distribution's support. Second, exploring the case of unbounded support for the target distribution, particularly when heavy-tailed distributions are involved, holds promise for further investigation. To the best of our knowledge, the theoretical study of how the dimension of the support and the distribution characteristics of the noise source ρ influence the performance of GANs is still open. Third, it will also be worthwhile to study the statistical estimation and expressive power of group-invariant GANs constructed from equivariant convolutional neural networks (CNNs), based on the universality result of CNNs (cf. [52]). Lastly, our work assumes the group is finite, which is crucial for defining our group-invariant generators and discriminators. Extending the analysis to compact Lie groups should also be explored. While the statistical error from the target Δ_3 has been studied in [10, 44] for continuous groups, it remains an open problem for a novel design of invariant discriminator architecture and its universality for infinite groups.

8. Proofs of results related to Theorem 4.9.

8.1. Proof of Lemma 4.1.

Proof. Based on the definition of d_Γ in (3.1), we have

$$
\begin{aligned}
d_\Gamma(S^\Sigma[(g_{n,m}^*)_\sharp \rho], \mu) &= d_{\Gamma_\Sigma}(S^\Sigma[(g_{n,m}^*)_\sharp \rho], \mu) \\
&\leq d_{\Gamma_\Sigma}(S^\Sigma[(g_{n,m}^*)_\sharp \rho], \widehat{\mu}_n) + d_{\Gamma_\Sigma}(\widehat{\mu}_n, \mu) \\
&\leq d_{\mathcal{D}_{NN}^\Sigma}(S^\Sigma[(g_{n,m}^*)_\sharp \rho], \widehat{\mu}_n) + 2 \sup_{f \in \Gamma_\Sigma} \inf_{f_\omega \in \mathcal{D}_{NN}^\Sigma} \|f - f_\omega\|_\infty + d_{\Gamma_\Sigma}(\widehat{\mu}_n, \mu) \\
&\leq d_{\mathcal{D}_{NN}^\Sigma}(S^\Sigma[(g_{n,m}^*)_\sharp \rho], S^\Sigma[(g_{n,m}^*)_\sharp \widehat{\rho}_m]) + d_{\mathcal{D}_{NN}^\Sigma}(S^\Sigma[(g_{n,m}^*)_\sharp \widehat{\rho}_m], \widehat{\mu}_n) \\
&\quad + 2 \sup_{f \in \Gamma_\Sigma} \inf_{f_\omega \in \mathcal{D}_{NN}^\Sigma} \|f - f_\omega\|_\infty + d_{\Gamma_\Sigma}(\widehat{\mu}_n, \mu),
\end{aligned}
$$

where the first equality is due to Lemma 3.1, the first inequality is given by the triangle inequality, the second inequality is due to Lemma 24 in [24], and the last inequality is also given by the triangle inequality.

Therefore we have two different terms to bound: $d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[(g_{n,m}^*)_{\sharp}\rho], S^{\Sigma}[(g_{n,m}^*)_{\sharp}\widehat{\rho}_m])$ and $d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[(g_{n,m}^*)_{\sharp}\widehat{\rho}_m], \widehat{\mu}_n)$. For the first term, we have

$$d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[(g_{n,m}^*)_{\sharp}\rho], S^{\Sigma}[(g_{n,m}^*)_{\sharp}\widehat{\rho}_m]) = d_{\mathcal{D}_{NN}^{\Sigma}}((g_{n,m}^*)_{\sharp}\rho, (g_{n,m}^*)_{\sharp}\widehat{\rho}_m)$$

due to Lemma 3.1 and the fact that $S_{\Sigma}[\mathcal{D}_{NN}^{\Sigma}] = \mathcal{D}_{NN}^{\Sigma}$. Hence we have

$$\begin{aligned} d_{\mathcal{D}_{NN}^{\Sigma}}((g_{n,m}^*)_{\sharp}\rho, (g_{n,m}^*)_{\sharp}\widehat{\rho}_m) &= \sup_{h \in \mathcal{D}_{NN}^{\Sigma}} \left\{ E_{(g_{n,m}^*)_{\sharp}\rho}[h] - E_{(g_{n,m}^*)_{\sharp}\widehat{\rho}_m}[h] \right\} \\ &= \sup_{h \in \mathcal{D}_{NN}^{\Sigma}} \left\{ E_{\rho}[h \circ g_{n,m}^*] - E_{\widehat{\rho}_m}[h \circ g_{n,m}^*] \right\} \\ &\leq d_{\mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m). \end{aligned}$$

For the second term, by the definition in (3.15), we have, for any $g \in \mathcal{G}_{NN}$,

$$\begin{aligned} d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[(g_{n,m}^*)_{\sharp}\widehat{\rho}_m], \widehat{\mu}_n) &\leq d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[g_{\sharp}\widehat{\rho}_m], \widehat{\mu}_n) \\ &\leq d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[g_{\sharp}\widehat{\rho}_m], S^{\Sigma}[g_{\sharp}\rho]) + d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[g_{\sharp}\rho], \widehat{\mu}_n) \\ &\leq d_{\mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m) + d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[g_{\sharp}\rho], \widehat{\mu}_n). \end{aligned}$$

Taking the infimum over $g \in \mathcal{G}_{NN}$, we have

$$d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[(g_{n,m}^*)_{\sharp}\widehat{\rho}_m], \widehat{\mu}_n) \leq d_{\mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m) + \inf_{g \in \mathcal{G}_{NN}} d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[g_{\sharp}\rho], \widehat{\mu}_n).$$

Therefore,

$$\begin{aligned} d_{\Gamma}(S^{\Sigma}[(g_{n,m}^*)_{\sharp}\rho], \mu) &\leq \inf_{g \in \mathcal{G}_{NN}} d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[g_{\sharp}\rho], \widehat{\mu}_n) + 2 \sup_{f \in \Gamma_{\Sigma}} \inf_{f_{\omega} \in \mathcal{D}_{NN}^{\Sigma}} \|f - f_{\omega}\|_{\infty} \\ &\quad + d_{\Gamma_{\Sigma}}(\widehat{\mu}_n, \mu) + 2 d_{\mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m). \end{aligned}$$

This completes the proof of Lemma 4.1. ∎

8.2. Bound for Δ_1. We first cite the following network universal approximation result from [50].

Lemma 8.1 (universal approximation of Lipschitz functions; Theorem 1 in [50]). *Let $\mathcal{X}$ be a compact domain in $\mathbb{R}^d$, and let $\sup_{x \in \mathcal{X}} \|x\|_{\infty} \leq F$. Given any $\epsilon \in (0,1)$, there exists a ReLU network architecture $\mathcal{NN}(W, L, N)$ such that, for any $f \in Lip_H(\mathcal{X})$ and $\|f\|_{\infty} \leq M$, with appropriately chosen network weights, the network provides a function $\hat{f}$ such that $\left\|f - \hat{f}\right\|_{\infty} \leq \epsilon$. Such a network has no more than $c_1(\log \frac{1}{\epsilon} + 1)$ layers and at most $c_2 \epsilon^{-d}(\log \frac{1}{\epsilon} + 1)$ neurons and weight parameters, where the constants c_1 and c_2 depend on d, H, F, and M.*

Remark 8.2. Though the original Theorem 1 in [49] assumes $\mathcal{X}$ is a cube in $\mathbb{R}^d$, it also holds when $\mathcal{X}$ is some compact domain, since we can extend any $f \in \text{Lip}_H(\mathcal{X})$ to some cube containing $\mathcal{X}$ while preserving the Lipschitz constant, by the Kirszbraun theorem [27].

Proof of Lemma 4.2. By Lemma 8.1, for any $\epsilon \in (0,1)$, there exists a ReLU network architecture $\mathcal{NN}(W_1, L_1, N_1, K_1)$ such that

$$\sup_{f \in \Gamma_\Sigma} \inf_{\phi \in \mathcal{NN}(W_1, L_1, N_1, K_1)} \|f(x) - \phi(x)\|_\infty \leq \epsilon,$$

where W_1, L_1, N_1 are given by Lemma 8.1, and K_1 depends on H, $\sup_{x \in \mathcal{X}} \|x\|_\infty$ and M. Hence for any $f \in \Gamma_\Sigma$ and $\delta > 0$, there exists $\phi \in \mathcal{NN}(W_1, L_1, N_1, K_1)$ such that

$$\begin{aligned}
\left\| f(x) - \phi(\sigma^{-1}x) \right\|_\infty &= \left\| f(\sigma^{-1}x) - \phi(\sigma^{-1}x) \right\|_\infty \\
&= \|f(x) - \phi(x)\|_\infty \\
&\leq \epsilon + \delta
\end{aligned}$$

for any $\sigma \in \Sigma$. Therefore,

$$\begin{aligned}
\left\| f(x) - \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} \phi(\sigma_i^{-1}x) \right\|_\infty &= \left\| \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} f(x) - \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} \phi(\sigma_i^{-1}x) \right\|_\infty \\
&\leq \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} \left\| f(x) - \phi(\sigma_i^{-1}x) \right\|_\infty \\
&\leq \epsilon + \delta.
\end{aligned}$$

Note that $\frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} \phi(\sigma_i^{-1}x) \in \mathcal{D}_{NN}^\Sigma$, and δ can be arbitrarily small, and hence we have

$$\inf_{f_\omega \in \mathcal{D}_{NN}^\Sigma} \|f(x) - f_\omega(x)\|_\infty \leq \epsilon.$$

Since f can be any function in Γ_Σ, we have

$$\sup_{f \in \Gamma_\Sigma} \inf_{f_\omega \in \mathcal{D}_{NN}^\Sigma} \|f - f_\omega\|_\infty \leq \epsilon.$$

This completes the proof of Lemma 4.2. ∎

Since the weight of the Σ-symmetrization layer W_Σ is determined by Σ, and the ReLU function is 1-Lipschitz, combined with the uniform bound for the weights by K_1, the Lipschitz constant of each layer has a uniform bound. Therefore, we have $\sup_{f_\omega \in \mathcal{D}_{NN}^\Sigma} \|f\|_{\mathrm{Lip}} \leq \tilde{H}$ for some $\tilde{H} > 0$ that depends on Σ, W_1, L_1, and K_1, where $\|f\|_{\mathrm{Lip}}$ stands for the Lipschitz constant of the function f. This observation is useful to prove the bound for Δ_2 in subsection 8.3 following.

Proof of Proposition 4.3. We have

$$\int_\mathcal{X} |f - f_\omega|^2 \, dx \leq \mathrm{vol}(\mathcal{X}) \|f - f_\omega\|_\infty^2.$$

For the left-hand side, we claim that

$$\int_\mathcal{X} |f - f_\omega|^2 \, dx = \int_\mathcal{X} |f - S_\Sigma[f_\omega]|^2 \, dx + \int_\mathcal{X} |f_\omega - S_\Sigma[f_\omega]|^2 \, dx.$$

To prove the above equality, it is sufficient to show that

$$(8.1) \qquad \int_{\mathcal{X}} f f_\omega \, \mathrm{d}x = \int_{\mathcal{X}} \left(S_\Sigma[f_\omega] f - (S_\Sigma[f_\omega])^2 + S_\Sigma[f_\omega] f_\omega \right) \mathrm{d}x.$$

We show $\int (S_\Sigma[f_\omega])^2 \, \mathrm{d}x = \int S_\Sigma[f_\omega] f_\omega \, \mathrm{d}x$ and $\int f f_\omega \, \mathrm{d}x = \int S_\Sigma[f_\omega] f \, \mathrm{d}x$. First, we have

$$
\begin{aligned}
\int_{\mathcal{X}} (S_\Sigma[f_\omega])^2 \, \mathrm{d}x &= \int_{\mathcal{X}} \frac{1}{|\Sigma|} \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} \sum_{j=1}^{|\Sigma|} f_\omega(\sigma_i^{-1} x) f_\omega(\sigma_j^{-1} x) \, \mathrm{d}x \\
&= \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} \int_{\mathcal{X}} \frac{1}{|\Sigma|} \sum_{j=1}^{|\Sigma|} f_\omega(x) f_\omega(\sigma_j^{-1} \sigma_i x) \, \mathrm{d}x \\
&= \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} \int_{\mathcal{X}} \frac{1}{|\Sigma|} \sum_{k=1}^{|\Sigma|} f_\omega(x) f_\omega(\sigma_k^{-1} x) \, \mathrm{d}x \\
&= \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} \int_{\mathcal{X}} f_\omega(x) S_\Sigma[f_\omega(x)] \, \mathrm{d}x \\
&= \int_{\mathcal{X}} S_\Sigma[f_\omega(x)] f_\omega(x) \, \mathrm{d}x.
\end{aligned}
$$

On the other hand, we have

$$
\begin{aligned}
\int_{\mathcal{X}} S_\Sigma[f_\omega(x)] f \, \mathrm{d}x &= \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} \int_{\mathcal{X}} f_\omega(\sigma_i^{-1} x) f(x) \, \mathrm{d}x \\
&= \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} \int_{\mathcal{X}} f_\omega(x) f(\sigma_i x) \, \mathrm{d}x \\
&= \frac{1}{|\Sigma|} \sum_{i=1}^{|\Sigma|} \int_{\mathcal{X}} f_\omega(x) f(x) \, \mathrm{d}x \\
&= \int_{\mathcal{X}} f_\omega(x) f(x) \, \mathrm{d}x,
\end{aligned}
$$

where the second-to-last equality is due to f being Σ-invariant. ■

8.3. Bound for Δ_2. Before we bound the generator approximation error, we cite a few useful results. We denote by $\mathcal{S}^d(z_0, \dots, z_{N+1})$ the set of all continuous piecewise linear functions $f : \mathbb{R} \to \mathbb{R}^d$ which have breakpoints at $z_0 < z_1 < \cdots < z_{N+1}$ and are constant on $(-\infty, z_0)$ and (z_{N+1}, ∞). The following lemma from [49] extends the 1-dimensional approximation result in [14] to higher dimensions.

Lemma 8.3 (Lemma 3.1 in [49]). *Suppose $W \geq 7d + 1, L \geq 2$, and $n \leq (W - d - 1)\lfloor \frac{W-d-1}{6d} \rfloor \lfloor \frac{L}{2} \rfloor$. Then for any $z_0 < z_1 < \cdots < z_{n+1}$, we have $\mathcal{S}^d(z_0, \dots, z_{n+1}) \subset \mathcal{NN}(W, L)$.*

We next provide the proof of the generator approximation error bound from Lemma 4.5. The construction in the first step of the proof follows that of Lemma 3.2 in [49], and we add the group symmetrization in the second step of the proof.

Proof of Lemma 4.5. First, by Lemma 3.1, we have

$$d_{\mathcal{D}_{NN}^\Sigma}(S^\Sigma[g_\sharp\rho], \widehat{\mu}_n) = d_{\mathcal{D}_{NN}^\Sigma}(S^\Sigma[g_\sharp\rho], S^\Sigma[\widehat{\mu}_n])$$

due to $S_\Sigma[\mathcal{D}_{NN}^\Sigma] = \mathcal{D}_{NN}^\Sigma$.

Step 1. Without loss of generality, we assume that $m := n - 1 \geq 1$ and $\widehat{\mu}_n = \frac{1}{n}\sum_{i=0}^m \delta_{x_i}$. Let ϵ be sufficiently small such that $0 < \epsilon < \frac{m}{n}\|x_i - x_{i-1}\|_2$ for all $i = 1, \ldots, m$. By the absolute continuity of ρ, we can choose $2m$ points on $\mathbb{R}, z_{1/2} < z_1 < z_{3/2} < \cdots < z_{m-1/2} < z_m$, such that

$$\rho((-\infty, z_{1/2})) = \frac{1}{n},$$

$$\rho((z_{i-1/2}, z_i)) = \frac{\epsilon}{m\|x_i - x_{i-1}\|_2}, \qquad\qquad 1 \leq i \leq m,$$

$$\rho((z_i, z_{i+1/2})) = \frac{1}{n} - \frac{\epsilon}{m\|x_i - x_{i-1}\|_2}, \qquad 1 \leq i \leq m-1,$$

$$\rho((z_m, \infty)) = \frac{1}{n} - \frac{\epsilon}{m\|x_m - x_{m-1}\|_2}.$$

We define the continuous piecewise linear function $\phi : \mathbb{R} \to \mathbb{R}^d$ by

$$\phi(z) := \begin{cases} x_0, & z \in (-\infty, z_{1/2}), \\ \dfrac{z_i - z}{z_i - z_{i-1/2}}x_{i-1} + \dfrac{z - z_{i-1/2}}{z_i - z_{i-1/2}}x_i, & z \in [z_{i-1/2}, z_i), \\ x_i, & z \in [z_i, z_{i+1/2}), \\ x_m, & z \in [z_m, \infty). \end{cases}$$

Since $\phi \in \mathcal{S}^d(z_{1/2}, \ldots, z_m)$ has $2m = 2n - 2 \leq (W_2 - d - 1)\lfloor\frac{W_2-d-1}{6d}\rfloor\lfloor\frac{L_2}{2}\rfloor + 2$ breakpoints, by Lemma 8.3, $\phi \in \mathcal{NN}(W_2, L_2)$. We denote the line segment joining x_{i-1} and x_i by $\mathcal{L}_i := \{(1-t)x_{i-1} + tx_i \in \mathbb{R}^d : 0 < t \leq 1\}$.

Step 2. Note that $\sigma\mathcal{L}_i$ is the line segment joining σx_{i-1} and σx_i for any $\sigma \in \Sigma$ since the group action is linear. Then $S^\Sigma[\phi_\sharp\rho]$ is supported on $\bigcup_{j=1}^{|\Sigma|}(\cup_{i=1}^m \sigma_j\mathcal{L}_i \cup \{\sigma_j x_0\})$, and $S^\Sigma[\phi_\sharp\rho](\{\sigma_j x_0\}) = \frac{1}{n|\Sigma|}$, $S^\Sigma[\phi_\sharp\rho](\{\sigma_j x_i\}) = \frac{1}{n|\Sigma|} - \frac{\epsilon}{m|\Sigma|\|x_i-x_{i-1}\|_2}$, $S^\Sigma[\phi_\sharp\rho](\sigma_j\mathcal{L}_i) = \frac{1}{n|\Sigma|}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, |\Sigma|$. We define the sum of product measures

$$\gamma = \sum_{j=1}^{|\Sigma|} S^\Sigma[\phi_\sharp\rho]|_{\{\sigma_j x_0\}} \times \delta_{\sigma_j x_0} + \sum_{j=1}^{|\Sigma|}\sum_{i=1}^m S^\Sigma[\phi_\sharp\rho]|_{\sigma_j\mathcal{L}_i} \times \delta_{\sigma_j x_i}.$$

It is easy to verify γ is a coupling of $S^\Sigma[\phi_\sharp\rho]$ and $S^\Sigma[\widehat{\mu}_n]$. Thus we have

$$\mathcal{W}_1(S^\Sigma[g_\sharp\rho], S^\Sigma[\widehat{\mu}_n]) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2\, \mathrm{d}\gamma(x, y)$$

$$= \sum_{j=1}^{|\Sigma|}\sum_{i=1}^m \int_{\sigma_j\mathcal{L}_i \setminus \{\sigma_j x_i\}} \|\sigma_j x_i - y\|_2\, \mathrm{d}S^\Sigma[\phi_\sharp\rho](y)$$

$$\leq \sum_{j=1}^{|\Sigma|} \sum_{i=1}^{m} \|\sigma_j x_i - \sigma_j x_{i-1}\|_2 \, S^{\Sigma}[\phi_\sharp \rho](\sigma_j \mathcal{L}_i \backslash \{\sigma_j x_i\})$$

$$\leq \sum_{j=1}^{|\Sigma|} \sum_{i=1}^{m} \|x_i - x_{i-1}\|_2 \, \frac{\epsilon}{m \, |\Sigma| \, \|x_i - x_{i-1}\|_2}$$

$$= \epsilon,$$

where the last inequality is due to Σ-actions being 1-Lipschitz. Since functions in $\mathcal{D}_{NN}^{\Sigma}$ are $\tilde{H}$-Lipschitz, we have

$$d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[g_\sharp \rho], S^{\Sigma}[\widehat{\mu}_n]) \leq \tilde{H} \cdot \mathcal{W}_1(S^{\Sigma}[g_\sharp \rho], S^{\Sigma}[\widehat{\mu}_n]) \leq \tilde{H}\epsilon.$$

Since ϵ can be arbitrarily small, we have $d_{\mathcal{D}_{NN}^{\Sigma}}(S^{\Sigma}[g_\sharp \rho], S^{\Sigma}[\widehat{\mu}_n]) = 0$. ∎

8.4. Bound for Δ_3. We first cite some useful lemmas before establishing the bound for Δ_3. First, we need the following bound of the Rademacher complexity by Dudley's entropy integral.

Lemma 8.4. *Suppose $\mathcal{F}$ is a family of functions mapping $\mathcal{X}$ to $[-M, M]$ for some $M > 0$. Also assume that $0 \in \mathcal{F}$ and $\mathcal{F} = -\mathcal{F}$. Let $\xi = \{\xi_1, \ldots, \xi_n\}$ be a set of independent random variables that take values on $\{-1, 1\}$ with equal probabilities, $i = 1, \ldots, n$. $x_1, x_2, \ldots, x_n \in \mathcal{X}$. Then we have*

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

and therefore, if μ is a distribution on $\mathcal{X}$ and $\widehat{\mu}_n$ is its empirical distribution, we have

$$\mathbb{E}[d_{\mathcal{F}}(\widehat{\mu}_n, \mu)] \leq \inf_{\alpha > 0} 8\alpha + \frac{24}{\sqrt{n}} \int_\alpha^M \sqrt{\log \mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_\infty)} \, \mathrm{d}\epsilon.$$

The proof of Lemma 8.4 is standard using the dyadic path., e.g., see the proof of Lemma A.5. in [3], which is modified in Lemma 3 in [10].

The following lemma is a direct consequence of Lemma 6 in [21].

Lemma 8.5. *Let $\mathcal{F}$ be the family of H-Lipschitz functions mapping the metric space $(\mathcal{X}, \|\cdot\|_2)$ to $[-M, M]$ for some $M > 0$. Then we have*

$$\mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_\infty) \leq \left(\frac{c_1 M}{\epsilon} \right)^{\mathcal{N}(\mathcal{X}, \frac{c_2 \epsilon}{H})},$$

where $c_1 \geq 1$ and $c_2 \leq 1$ are some absolute constants independent of $\mathcal{X}$, M, and ϵ.

Lemma 8.6 (Theorem 3 in [42]). *Assume that $\mathcal{X} = \Sigma \times \mathcal{X}_0$. If for some $\epsilon > 0$ we have*
(1) $\|\sigma(x) - \sigma'(x')\|_2 > 2\epsilon, \; \forall x, x' \in \mathcal{X}_0, \sigma \neq \sigma' \in \Sigma$; and
(2) $\|\sigma(x) - \sigma(x')\|_2 \geq \|x - x'\|_2, \; \forall x, x' \in \mathcal{X}_0, \sigma \in \Sigma$,

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

then we have

Lemma 8.7 (Lemma 6 in [10]). *Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$, and let $\bar{\epsilon} > 0$. Then there exists a constant $C_{d,\bar{\epsilon}}$ that depends on d and $\bar{\epsilon}$ such that for $\epsilon \in (0,1)$ we have*

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

Following the statistical analysis in [10], we can bound the statistical error $d_{\Gamma_\Sigma}(\widehat{\mu}_n, \mu)$ as follows.

Proof of Lemma 4.7. By the definition of the probability metric d_Γ (3.1), we have

$$
\begin{aligned}
d_{\Gamma_\Sigma}(\widehat{\mu}_n, \mu) &= |d_{\Gamma_\Sigma}(\widehat{\mu}_n, \mu)| \\
&= \left| \sup_{\gamma \in \Gamma_\Sigma} \left\{ \frac{1}{n} \sum_{i=1}^{n} \gamma(x_i) - \mathbb{E}_\mu[\gamma] \right\} \right| \\
&\le \sup_{\gamma \in \Gamma_\Sigma} \left| \mathbb{E}_\mu[\gamma] - \frac{1}{n} \sum_{i=1}^{n} \gamma(x_i) \right| \\
&= \sup_{\gamma \in \Gamma_\Sigma} \left| \mathbb{E}_\mu[\gamma] - \frac{1}{n} \sum_{i=1}^{n} \gamma(T_0(x_i)) \right| \\
&\overset{(a)}{\le} \sup_{\gamma \in \mathrm{Lip}_H(\mathcal{X}_0)} \left| \mathbb{E}_{\mu_{\mathcal{X}_0}}[\gamma] - \frac{1}{n} \sum_{i=1}^{n} \gamma(T_0(x_i)) \right|,
\end{aligned}
$$

(8.2)

where inequality (a) is due to the fact that $\mathbb{E}_\mu[\gamma] = \mathbb{E}_{\mu_{\mathcal{X}_0}}[\gamma|_{\mathcal{X}_0}]$, since μ is Σ-invariant and $\gamma \in \Gamma_\Sigma$, and the fact that if $\gamma \in \Gamma_\Sigma$, then $\gamma|_{\mathcal{X}_0} \in \mathrm{Lip}_H(\mathcal{X}_0)$, where $\gamma|_{\mathcal{X}_0}$ is the restriction of γ on $\mathcal{X}_0$.

Denote by $X' = \{x'_1, x'_2, \ldots, x'_n\}$ i.i.d. samples drawn from $\mu_{\mathcal{X}_0}$. Also note that

$$T_0(x_1), \ldots, T_0(x_n)$$

can be viewed as i.i.d. samples on $\mathcal{X}_0$ drawn from $\mu_{\mathcal{X}_0}$. Therefore, the expectation

$$
\mathbb{E}_X \sup_{\gamma \in \mathrm{Lip}_H(\mathcal{X}_0)} \left| E_{\mu_{\mathcal{X}_0}}[\gamma] - \frac{1}{n} \sum_{i=1}^{n} \gamma(T_0(x_i)) \right|
$$

can be replaced by the equivalent quantity

$$
\mathbb{E}_X \sup_{\gamma \in \mathrm{Lip}_H(\mathcal{X}_0)} \left| E_{\mu_{\mathcal{X}_0}}[\gamma] - \frac{1}{n} \sum_{i=1}^{n} \gamma(x_i) \right|,
$$

where $X = \{x_1, x_2, \ldots, x_m\}$ are i.i.d. samples on $\mathcal{X}_0$ drawn from $\mu_{\mathcal{X}_0}$. Then we have

$$
\begin{aligned}
\mathbb{E}_X \sup_{\gamma \in \mathrm{Lip}_H(\mathcal{X}_0)} \left| E_{\mu_{\mathcal{X}_0}}[\gamma] - \frac{1}{n} \sum_{i=1}^{n} \gamma(x_i) \right| &= \mathbb{E}_X \sup_{\gamma \in \mathrm{Lip}_H(\mathcal{X}_0)} \left| E_{X'} \left(\frac{1}{n} \sum_{i=1}^{n} \gamma(x'_i) \right) - \frac{1}{n} \sum_{i=1}^{n} \gamma(x_i) \right| \\
&\le \mathbb{E}_{X,X'} \sup_{\gamma \in \mathrm{Lip}_H(\mathcal{X}_0)} \left| \frac{1}{n} \sum_{i=1}^{n} \gamma(x'_i) - \frac{1}{n} \sum_{i=1}^{n} \gamma(x_i) \right|
\end{aligned}
$$

$$= \mathbb{E}_{X,X',\xi} \sup_{\gamma \in \mathrm{Lip}_H(\mathcal{X}_0)} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \left(\gamma(x_i') - \gamma(x_i) \right) \right|$$

$$\leq 2\mathbb{E}_{X,\xi} \sup_{\gamma \in \mathrm{Lip}_H(\mathcal{X}_0)} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \gamma(x_i') \right|$$

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

where $\mathcal{F}_0 = \{\gamma \in \mathrm{Lip}_H(\mathcal{X}_0) : \|\gamma\|_{\infty} \leq M\}$.

For $d \geq 2$, from Lemma 8.5, we have $\log \mathcal{N}(\mathcal{F}_0, \epsilon, \|\cdot\|_{\infty}) \leq \mathcal{N}(\mathcal{X}_0, \frac{c_2\epsilon}{H}) \log(\frac{c_1 M}{\epsilon})$. We fix an $\bar{\epsilon} > 0$ such that $\mathcal{N}(\mathcal{X}, \frac{c_2\bar{\epsilon}}{H}) = 1$ and select ϵ^* such that $\frac{c_2\epsilon^*}{H} \leq 1$ and $\frac{c_2\epsilon^*}{H} \leq \epsilon_\Sigma$; that is, $\epsilon^* \leq \min\left(\frac{H}{c_2}, \frac{H\epsilon_\Sigma}{c_2}\right)$, so that by Assumption 3.6, Lemma 8.6, and Lemma 8.7, we have

$$\mathcal{N}\left(\mathcal{X}_0, \frac{c_2\epsilon}{H}\right) \log\left(\frac{c_1 M}{\epsilon}\right) \leq \mathcal{N}\left(\mathcal{X}_0 \backslash A_0\left(\frac{c_2\epsilon}{H}\right), \frac{c_2\epsilon}{H}\right) \log\left(\frac{c_1 M}{\epsilon}\right)$$

$$+ \mathcal{N}\left(A_0\left(\frac{c_2\epsilon}{H}\right), \frac{c_2\epsilon}{H}\right) \log\left(\frac{c_1 M}{\epsilon}\right)$$

$$\leq \frac{\mathcal{N}\left(\mathcal{X}, \frac{c_2\epsilon}{H}\right)}{|\Sigma|} \log\left(\frac{c_1 M}{\epsilon}\right) + \bar{c}\epsilon^r \mathcal{N}\left(\mathcal{X}, \frac{c_2\epsilon}{H}\right) \log\left(\frac{c_1 M}{\epsilon}\right)$$

$$\leq \frac{C_{d,\bar{\epsilon}} H^d}{|\Sigma| c_2^d \epsilon^d} \log\left(\frac{c_1 M}{\epsilon}\right) + \frac{\bar{c} C_{d,\bar{\epsilon}} H^d}{c_2^d \epsilon^{d-r}} \log\left(\frac{c_1 M}{\epsilon}\right)$$

when $\epsilon < \epsilon^*$, where $\bar{c} > 0$ is some constant implied by Assumption 3.6. Therefore, for sufficiently small α, we have

$$\int_{\alpha}^{M} \sqrt{\log \mathcal{N}(\mathcal{F}_0, \epsilon, \|\cdot\|_{\infty})} \, \mathrm{d}\epsilon$$

$$= \int_{\alpha}^{\epsilon^*} \sqrt{\log \mathcal{N}(\mathcal{F}_0, \epsilon, \|\cdot\|_{\infty})} \, \mathrm{d}\epsilon + \int_{\epsilon^*}^{M} \sqrt{\log \mathcal{N}(\mathcal{F}_0, \epsilon, \|\cdot\|_{\infty})} \, \mathrm{d}\epsilon$$

(8.3)
$$\leq \int_{\alpha}^{\epsilon^*} \sqrt{\frac{C_{d,\bar{\epsilon}} H^d}{|\Sigma| c_2^d \epsilon^d} \log\left(\frac{c_1 M}{\epsilon}\right)} \, \mathrm{d}\epsilon + \int_{\alpha}^{\epsilon^*} \sqrt{\frac{\bar{c} C_{d,\bar{\epsilon}} H^d}{|\Sigma| c_2^d \epsilon^{d-r}} \log\left(\frac{c_1 M}{\epsilon}\right)} \, \mathrm{d}\epsilon$$

$$+ \int_{\epsilon^*}^{M} \sqrt{\log \mathcal{N}(\mathcal{F}_0, \epsilon, \|\cdot\|_{\infty})} \, \mathrm{d}\epsilon.$$

For any $s > 0$, we can choose ϵ^* to be sufficiently small, such that $\log(\frac{c_1 M}{\epsilon}) \leq \frac{1}{\epsilon^s}$ when $\epsilon \leq \epsilon^*$. Therefore, if we let $D_{\mathcal{X},H} = \sqrt{\frac{C_{d,\bar{\epsilon}} H^d}{c_2^d}}$, then for the first term in (8.3), we have

$$\int_{\alpha}^{\epsilon^*} \sqrt{\frac{C_{d,\bar{\epsilon}} H^d}{|\Sigma| c_2^d \epsilon^d} \log\left(\frac{c_1 M}{\epsilon}\right)} \, \mathrm{d}\epsilon \leq D_{\mathcal{X},H} \int_{\alpha}^{\epsilon^*} \sqrt{\frac{1}{|\Sigma| \epsilon^{d+s}}} \, \mathrm{d}\epsilon$$

$$\leq D_{\mathcal{X},H} \int_{\epsilon}^{\infty} \sqrt{\frac{1}{|\Sigma| \epsilon^{d+s}}} \, \mathrm{d}\epsilon$$

$$= \frac{D_{\mathcal{X},H}}{\sqrt{|\Sigma|}} \cdot \frac{\alpha^{1 - \frac{d+s}{2}}}{\frac{d+s}{2} - 1}.$$

Notice that the third integral in (8.3) is bounded, while the first integral diverges as α tends to zero; the second integral is either bounded (if $d - r < 2$ and s is small) or of order $\simeq \alpha^{1 - \frac{d-r+s}{2}}$ (if $d - r \geq 2$), so we can optimize the majorizing terms

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

with respect to α, to obtain

$$\alpha = \left(\frac{9}{n}\right)^{\frac{1}{d+s}} \cdot \left(\frac{D_{\mathcal{X},H}^2}{|\Sigma|}\right)^{\frac{1}{d+s}},$$

so that

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

(8.4)
$$\leq 8\left(\frac{9}{n}\right)^{\frac{1}{d+s}} \cdot \left(\frac{D_{\mathcal{X},H}^2}{|\Sigma|}\right)^{\frac{1}{d+s}} + \frac{24}{(\frac{d+s}{2} - 1)} \left(\frac{9}{n}\right)^{\frac{1}{d+s}} \left(\frac{D_{\mathcal{X},H}^2}{|\Sigma|}\right)^{\frac{1}{d+s}} + o\left(\frac{1}{n^{\frac{1}{d+s}}}\right).$$

Therefore, for sufficiently large n, we have

$$\mathbb{E}_X \sup_{\gamma \in \mathrm{Lip}_H(\mathcal{X}_0)} \left| E_{\mu_{\mathcal{X}_0}}[\gamma] - \frac{1}{n} \sum_{i=1}^n \gamma(x_i) \right| \leq \left(8 + \frac{24}{(\frac{d+s}{2} - 1)}\right) \left(\frac{9 D_{\mathcal{X},H}^2}{|\Sigma|\, n}\right)^{\frac{1}{d+s}} + o\left(\frac{1}{n^{\frac{1}{d+s}}}\right)$$

(8.5)
$$:= C_{\mathcal{X},H,d,s} \left(\frac{1}{|\Sigma|\, n}\right)^{\frac{1}{d+s}} + o\left(\left(\frac{1}{n}\right)^{\frac{1}{d+s}}\right).$$

For $d = 1$, the first integral in (8.3) does not have a singularity at $\alpha = 0$ if $1 + s < 2$. On the other hand, replacing the interval $[0, 1]$ by an interval of length $\mathrm{diam}(\mathcal{X}_0)$ in Lemma 5.16 in [46], there exists a constant $c > 0$ such that

$$\mathcal{N}(\mathcal{F}_0, \epsilon, \|\cdot\|_\infty) \leq e^{\frac{cH \cdot \mathrm{diam}(\mathcal{X}_0)}{\epsilon}} \quad \text{for } \epsilon < M.$$

Therefore, we have

$$8\alpha + \frac{24}{\sqrt{n}} \int_\alpha^M \sqrt{\log \mathcal{N}(\mathcal{F}_0, \epsilon, \|\cdot\|_\infty)}\, d\epsilon \leq 8\alpha + \frac{24}{\sqrt{n}} \int_\alpha^M \sqrt{\frac{cH \cdot \mathrm{diam}(\mathcal{X}_0)}{\epsilon}}\, d\epsilon,$$

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

$$\inf_{\alpha > 0} 8\alpha + \frac{24}{\sqrt{n}} \int_\alpha^M \sqrt{\log \mathcal{N}(\mathcal{F}_0, \epsilon, \|\cdot\|_\infty)}\, d\epsilon$$
$$\leq \frac{72cH \cdot \mathrm{diam}(\mathcal{X}_0)}{n} + \frac{48H\sqrt{c} \cdot \mathrm{diam}(\mathcal{X}_0)}{\sqrt{n}} - \frac{144cH \cdot \mathrm{diam}(\mathcal{X}_0)}{n}$$
$$= \frac{48H\sqrt{c} \cdot \mathrm{diam}(\mathcal{X}_0)}{\sqrt{n}} - \frac{72cH \cdot \mathrm{diam}(\mathcal{X}_0)}{n}.$$

Hence we have

$$\mathbb{E}_X \sup_{\gamma \in \mathrm{Lip}_H(\mathcal{X}_0)} \left| \mathbb{E}_{\mu_{\mathcal{X}_0}}[\gamma] - \frac{1}{n} \sum_{i=1}^n \gamma(x_i) \right| \le \frac{48 H \sqrt{c} \cdot \mathrm{diam}(\mathcal{X}_0)}{\sqrt{n}} - \frac{72 c H \cdot \mathrm{diam}(\mathcal{X}_0)}{n}.$$

This completes the proof. ∎

8.5. Bound for Δ_4. To prove the bound for $\mathbb{E}[d_{\mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m)]$, we introduce the notion of pseudodimension from [4], which is another measure of complexity for a class of functions.

Definition 8.8 (pseudodimension). Let $\mathcal{F}$ be a class of functions that map $\mathcal{X}$ to $\mathbb{R}$. The pseudodimension of $\mathcal{F}$, denoted by $Pdim(\mathcal{F})$, is the largest integer n for which there exists $(x_1, \dots, x_n, y_1, \dots, y_n) \in \mathcal{X}^n \times \mathbb{R}^n$ such that for any $(b_1, \dots, b_n) \in \{0,1\}^n$, there exists $f \in \mathcal{F}$ such that

$$\forall i : f(x_i) > y_i \quad iff \quad b_i = 1.$$

Proof of Lemma 4.8. First, we show that $\sup_{f \in \mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}} \|f\|_\infty$ is bounded. This is straightforward since we can add an additional clipping layer to the output of $\mathcal{D}_{NN}^{\Sigma}$ so that its output lies within, for example, $[-2M, 2M]$. Note that such a clipping does not impact the invariant discriminator approximation error if we require $\epsilon \le \frac{M}{2}$ in Lemma 4.2. By Corollary 35 in [24], we have

$$d_{\mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m) \lesssim \sqrt{\frac{\mathrm{Pdim}(\mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}) \log m}{m}}.$$

By Theorem 7 in [4], we have $\mathrm{Pdim}(\mathcal{NN}(W, L, N)) \lesssim NL \log N$. Note that we can rewrite $\mathcal{D}_{NN}^{\Sigma}$ defined in (3.14) as a ReLU network with a larger width (increased by a factor of the group size) and the same depth L_1 as $\mathcal{D}_{NN}$, despite the first layer, where we multiply input x by the W_{σ_i}'s and the last averaging layer. Importantly, the number of free parameters N_1 remains exactly the same as $\mathcal{D}_{NN}$. Hence, we have

$$\mathbb{E}[d_{\mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m)] \lesssim \sqrt{\frac{(N_1 + W_2^2 L_2)(L_1 + L_2) \log(N_1 + W_2^2 L_2) \log m}{m}},$$

where we use the trivial bound $N_2 \simeq W_2^2 L_2$. By Lemmas 4.2 and 4.5, we have $N_1 \lesssim n \log n$ and $W_2^2 L_2 \lesssim n$, so we have

$$\mathbb{E}[d_{\mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m)] \lesssim \sqrt{\frac{(n \log n + n)(\log n + n) \log(n \log n + n) \log m}{m}}$$

$$\lesssim \sqrt{\frac{n^2 \log^2 n \log m}{m}}.$$

Thus if $m \gtrsim n^{2 + 2/d} \log^3 n$, we have $\mathbb{E}[d_{\mathcal{D}_{NN}^{\Sigma} \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m)] \lesssim n^{-1/d}$. ∎

8.6. Theorem 4.9 for $d = 1$.

Theorem 8.9 (main theorem ($d = 1$)). *Let $\mathcal{X} = \Sigma \times \mathcal{X}_0$ be a subset of $\mathbb{R}$, let $\Gamma = Lip_H(\mathcal{X})$, and let $\mathcal{X}_0$ be an interval of finite length. Suppose the target distribution μ is Σ-invariant on $\mathcal{X}$, and the noise source distribution ρ is absolutely continuous on $\mathbb{R}$. Then there exists Σ-invariant discriminator architecture $\mathcal{D}_{NN}^{\Sigma} = S_\Sigma[\mathcal{D}_{NN}]$, where $\mathcal{D}_{NN} = \mathcal{NN}(W_1, L_1, N_1)$ as defined in (3.14) with $N_1 \lesssim n \log n$ and $L_1 \lesssim \log n$, and Σ-invariant generator architecture $\mathcal{G}_{NN}^{\Sigma}$, where $\mathcal{G}_{NN} = \mathcal{NN}(W_2, L_2)$, with $W_2^2 L_2 \lesssim n$, such that if $m \gtrsim n^4 \log^3 n$, we have*

$$\mathbb{E}\left[d_\Gamma(S^\Sigma[(g_{n,m}^*)_\sharp \rho], \mu) \right] \lesssim \frac{diam(\mathcal{X}_0)}{\sqrt{n}}.$$

Proof. It suffices to choose $\epsilon \simeq n^{-1}$ in Lemma 4.2 with $L_1 \lesssim \log n$ and $N_1 \lesssim n \log n$, so that $\sup_{f \in \Gamma_\Sigma} \inf_{f_\omega \in \mathcal{D}_{NN}^\Sigma} \|f - f_\omega\|_\infty \lesssim n^{-1}$. On the other hand, we can construct

$$\inf_{g \in \mathcal{G}_{NN}} d_{\mathcal{D}_{NN}^\Sigma}(S^\Sigma[g_\sharp \rho], \widehat{\mu}_n) = 0$$

with $W_2^2 L_2 \lesssim n$ by Lemma 4.5. Finally, $\mathbb{E}[d_{\Gamma_\Sigma}(\widehat{\mu}_n, \mu)] \lesssim \frac{diam(\mathcal{X}_0)}{\sqrt{n}}$ and $\mathbb{E}[d_{\mathcal{D}_{NN}^\Sigma \circ \mathcal{G}_{NN}}(\rho, \widehat{\rho}_m)] \lesssim n^{-1}$ by Lemmas 4.7 and 4.8, respectively. ∎

9. Proof of Theorem 5.2 when $d = 1$.

Theorem 9.1 (1-dimensional submanifolds). *Let $\mathcal{X} = \Sigma \times \mathcal{X}_0 \subset \mathbb{R}^d$ and $\Gamma = Lip_H(\mathcal{X})$. Suppose $\mathcal{X}_0$ is diffeomorphic to some interval of finite length, the target distribution μ is Σ-invariant on $\mathcal{X}$, and the noise source distribution ρ is absolutely continuous on $\mathbb{R}$. Then there exists Σ-invariant discriminator architecture $\mathcal{D}_{NN}^\Sigma = S_\Sigma[\mathcal{D}_{NN}]$, where $\mathcal{D}_{NN} = \mathcal{NN}(W_1, L_1, N_1)$ as defined in (3.14) with $N_1 \lesssim n \log n$ and $L_1 \lesssim \log n$, and Σ-invariant generator architecture $\mathcal{G}_{NN}^\Sigma$, where $\mathcal{G}_{NN} = \mathcal{NN}(W_2, L_2)$, with $W_2^2 L_2 \lesssim n$, such that if $m \gtrsim n^4 \log^3 n$, we have*

$$\mathbb{E}\left[d_\Gamma(S^\Sigma[(g_{n,m}^*)_\sharp \rho], \mu) \right] \lesssim \frac{peri(\mathcal{X}_0)}{\sqrt{n}},$$

where $peri(\mathcal{X}_0)$ denotes the perimeter of $\mathcal{X}_0$ in $\mathbb{R}^d$.

Proof. It suffices to show that $\Delta_3 \lesssim \frac{peri(\mathcal{X}_0)}{\sqrt{n}}$, which directly follows the proof of Theorem XIV in [45]. ∎

REFERENCES

[1] M. ARJOVSKY, S. CHINTALA, AND L. BOTTOU, *Wasserstein generative adversarial networks*, in Proceedings of the International Conference on Machine Learning, PMLR, 2017, pp. 214–223.

[2] S. ARORA, R. GE, Y. LIANG, T. MA, AND Y. ZHANG, *Generalization and equilibrium in generative adversarial nets (GANs)*, in Proceedings of the International Conference on Machine Learning, PMLR, 2017, pp. 224–232.

[3] P. L. BARTLETT, D. J. FOSTER, AND M. J. TELGARSKY, *Spectrally-normalized margin bounds for neural networks*, Advances in Neural Information Processing Systems, 30 (2017).

[4] P. L. BARTLETT, N. HARVEY, C. LIAW, AND A. MEHRABIAN, *Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks*, J. Mach. Learn. Res., 20, 2019, pp. 1–17.

[5] M. BILOŠ AND S. GÜNNEMANN, *Scalable normalizing flows for permutation invariant densities*, in Proceedings of the International Conference on Machine Learning, PMLR, 2021, pp. 957–967.

[6] J. BIRRELL, P. DUPUIS, M. A. KATSOULAKIS, Y. PANTAZIS, AND L. REY-BELLET, (f,Γ)-*Divergences: Interpolating between f-divergences and integral probability metrics*, J. Mach. Learn. Res., 23(39), 2022, pp. 1–70.

[7] J. BIRRELL, M. KATSOULAKIS, L. REY-BELLET, AND W. ZHU, *Structure-preserving GANs*, in Proceedings of the International Conference on Machine Learning, PMLR, 2022, pp. 1982–2020.

[8] D. BOYDA, G. KANWAR, S. RACANIÈRE, D. J. REZENDE, M. S. ALBERGO, K. CRANMER, D. C. HACKETT, AND P. E. SHANAHAN, *Sampling using SU(N) gauge equivariant flows*, Phys. Rev. D, 103 (2021), 074504, https://doi.org/10.1103/PhysRevD.103.074504.

[9] M. CHEN, W. LIAO, H. ZHA, AND T. ZHAO, *Distribution Approximation and Statistical Estimation Guarantees of Generative Adversarial Networks*, preprint, arXiv:2002.03938, 2020.

[10] Z. CHEN, M. KATSOULAKIS, L. REY-BELLET, AND W. ZHU, *Sample complexity of probability divergences under group symmetry*, in Proceedings of the International Conference on Machine Learning, PMLR, 2023, pp. 4713–4734.

[11] Z. CHEN, M. A. KATSOULAKIS, AND B. J. ZHANG, *Equivariant Score-Based Generative Models Provably Learn Distributions with Symmetries Efficiently*, preprint, arXiv:2410.01244, 2024.

[12] T. COHEN AND M. WELLING, *Group equivariant convolutional networks*, in Proceedings of the International Conference on Machine Learning, PMLR, 2016, pp. 2990–2999.

[13] T. S. COHEN, M. GEIGER, AND M. WEILER, *A general theory of equivariant CNNs on homogeneous spaces*, Advances in Neural Information Processing Systems, 32 (2019).

[14] I. DAUBECHIES, R. DEVORE, S. FOUCART, B. HANIN, AND G. PETROVA, *Nonlinear approximation and (deep) ReLU networks*, Constr. Approx., 55 (2022), pp. 127–172, https://doi.org/10.1007/s00365-021-09548-z.

[15] N. DEY, A. CHEN, AND S. GHAFURIAN, *Group equivariant generative adversarial networks*, in Proceedings of the International Conference on Learning Representations, 2021.

[16] R. FLAMARY, N. COURTY, A. GRAMFORT, M. Z. ALAYA, A. BOISBUNON, S. CHAMBON, L. CHAPEL, A. CORENFLOS, K. FATRAS, N. FOURNIER, L. GAUTHERON, N. T. GAYRAUD, H. JANATI, A. RAKOTOMAMONJY, I. REDKO, A. ROLET, A. SCHUTZ, V. SEGUY, D. J. SUTHERLAND, R. TAVENARD, A. TONG, AND T. VAYER, *POT: Python Optimal Transport*, J. Mach. Learn. Res., 22 (2021), pp. 1–8, https://jmlr.org/papers/v22/20-451.html.

[17] R. FLAMARY, C. VINCENT-CUAZ, N. COURTY, A. GRAMFORT, O. KACHAIEV, H. QUANG TRAN, L. DAVID, C. BONET, N. CASSEREAU, T. GNASSOUNOU, E. TANGUY, J. DELON, A. COLLAS, S. MAZELET, L. CHAPEL, T. KERDONCUFF, X. YU, M. FEICKERT, P. KRZAKALA, T. LIU, AND E. FERNANDES MONTESUMA, *Pot Python Optimal Transport (Version 0.9.5)*, https://github.com/PythonOT/POT, 2024.

[18] G. B. FOLLAND, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed., Pure Appl. Math. (N. Y.), Wiley-Intersci. Publ., John Wiley & Sons, 1999.

[19] V. GARCIA SATORRAS, E. HOOGEBOOM, F. FUCHS, I. POSNER, AND M. WELLING, *E(n) equivariant normalizing flows*, Advances in Neural Information Processing Systems, 34 (2021), pp. 4181–4192.

[20] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE, AND Y. BENGIO, *Generative adversarial nets*, Advances in Neural Information Processing Systems, 27 (2014).

[21] L.-A. GOTTLIEB, A. KONTOROVICH, AND R. KRAUTHGAMER, *Efficient regression in metric spaces via approximate Lipschitz extension*, IEEE Trans. Inform. Theory, 63 (2017), pp. 4838–4849, https://doi.org/10.1109/TIT.2017.2713820.

[22] I. GULRAJANI, F. AHMED, M. ARJOVSKY, V. DUMOULIN, AND A. C. COURVILLE, *Improved training of Wasserstein GANs*, Advances in Neural Information Processing Systems, 30 (2017).

[23] E. HOOGEBOOM, V. G. SATORRAS, C. VIGNAC, AND M. WELLING, *Equivariant diffusion for molecule generation in 3D*, in Proceedings of the International Conference on Machine Learning, PMLR, 2022, pp. 8867–8887.

[24] J. HUANG, Y. JIAO, Z. LI, S. LIU, Y. WANG, AND Y. YANG, *An Error analysis of generative adversarial networks for learning distributions*, J. Mach. Learn. Res., 23(116), 2022, pp. 1–43.

[25] C. JACOBS, *Symmetries as a Guide to the Structure of Physical Quantities*, Ph.D. thesis, University of Oxford, 2021.

[26] M. KANG, J.-Y. ZHU, R. ZHANG, J. PARK, E. SHECHTMAN, S. PARIS, AND T. PARK, *Scaling up GANs for text-to-image synthesis*, in Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2023, pp. 10124–10134.

[27] M. KIRSZBRAUN, *Über die zusammenziehende und Lipschitzsche Transformationen*, Fund. Math., 22 (1934), pp. 77–108, https://doi.org/10.4064/fm-22-1-77-108.

[28] J. KÖHLER, L. KLEIN, AND F. NOÉ, *Equivariant flows: Exact likelihood generative learning for symmetric densities*, in Proceedings of the International Conference on Machine Learning, PMLR, 2020, pp. 5361–5370.

[29] A. N. KOLMOGOROV AND V. M. TIKHOMIROV, *ϵ-entropy and ϵ-capacity of sets in function spaces*, Uspekhi Matematicheskikh Nauk, 14 (1959), pp. 3–86.

[30] S.-H. LI AND L. WANG, *Neural network renormalization group*, Phys. Rev. Lett., 121 (2018), 260601, https://doi.org/10.1103/PhysRevLett.121.260601.

[31] T. LIANG, *How well generative adversarial networks learn distributions*, J. Mach. Learn. Res., 22 (2021), pp. 1–41.

[32] H. LU, S. SZABADOS, AND Y. YU, *Structure Preserving Diffusion Models*, preprint, arXiv:2402.19369, 2024.

[33] E. MARTINEZ, R. JACOME, A. HERNANDEZ-ROJAS, AND H. ARGUELLO, *LD-GAN: Low-dimensional generative adversarial network for spectral image generation with variance regularization*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2023, pp. 265–275.

[34] A. MÜLLER, *Integral probability metrics and their generating classes of functions*, Adv. Appl. Probab., 29 (1997), pp. 429–443, https://doi.org/10.2307/1428011.

[35] K. A. NICOLI, S. NAKAJIMA, N. STRODTHOFF, W. SAMEK, K.-R. MÜLLER, AND P. KESSEL, *Asymptotically unbiased estimation of physical observables with neural samplers*, Phys. Rev. E, 101 (2020), 023304, https://doi.org/10.1103/PhysRevE.101.023304.

[36] W. G. NOID, *Perspective: Coarse-grained models for biomolecular systems*, J. Chem. Phys., 139 (2013), 090901, https://doi.org/10.1063/1.4818908.

[37] F. OHLSSON, J. BORGQVIST, AND M. CVIJOVIC, *Symmetry structures in dynamic models of biochemical systems*, J. R. Soc. Interface, 17 (2020), 20200204, https://doi.org/10.1098/rsif.2020.0204.

[38] A. J. PAK AND G. A. VOTH, *Advances in coarse-grained modeling of macromolecular complexes*, Curr. Opin. Struc. Biol., 52 (2018), pp. 119–126, https://doi.org/10.1016/j.sbi.2018.11.005.

[39] E. PRETTI, V. K. SHEN, J. MITTAL, AND N. A. MAHYNSKI, *Symmetry-based crystal structure Enumeration in two dimensions*, J. Phys. Chem. A, 124 (2020), pp. 3276–3285, https://doi.org/10.1021/acs.jpca.0c00846.

[40] S. REED, Z. AKATA, X. YAN, L. LOGESWARAN, B. SCHIELE, AND H. LEE, *Generative adversarial text to image synthesis*, in Proceedings of the International Conference on Machine Learning, PMLR, 2016, pp. 1060–1069.

[41] A. SAUER, T. KARRAS, S. LAINE, A. GEIGER, AND T. AILA, *StyleGAN-t: Unlocking the power of GANs for fast large-scale text-to-image synthesis*, in Proceedings of the International Conference on Machine Learning, PMLR, 2023, pp. 30105–30118.

[42] J. SOKOLIC, R. GIRYES, G. SAPIRO, AND M. RODRIGUES, *Generalization error of invariant classifiers*, in Proceedings of Artificial Intelligence and Statistics, PMLR, 2017, pp. 1094–1103.

[43] B. K. SRIPERUMBUDUR, K. FUKUMIZU, A. GRETTON, B. SCHÖLKOPF, AND G. R. LANCKRIET, *On the empirical estimation of integral probability metrics*, Electron. J. Statist., 6 (2012), pp. 1550–1599, https://doi.org/10.1214/12-EJS722.

[44] B. TAHMASEBI AND S. JEGELKA, *Sample complexity bounds for estimating probability divergences under invariances*, in Proceedings of the Forty-First International Conference on Machine Learning, 2024, pp. 47396–47417, https://dl.acm.org/doi/10.5555/3692070.3694001.

[45] V. M. TIKHOMIROV, *ε-entropy and ε-capacity of sets in functional spaces*, in Selected Works of A.N. Kolmogorov: Volume III: Information Theory and the Theory of Algorithms, Math. Appl. (Soviet Ser.), 27, Kluwer Academic Publishers Group, 1993, pp. 86–170.

[46] R. VAN HANDEL, *Probability in High Dimension*, Technical report, Princeton University, Princeton, NJ, 2014.

[47] R. VERSHYNIN, *High-dimensional Probability: An Introduction with Applications in Data Science*, Cambridge Ser. Statist. Probab. Math. 47, Cambridge University Press, 2018.

[48] M. WEILER AND G. CESA, *General E(2)-equivariant steerable CNNs*, Advances in Neural Information Processing Systems, 32 (2019).

[49] Y. YANG, Z. LI, AND Y. WANG, *On the capacity of deep generative networks for approximating distributions*, Neural Networks, 145 (2022), pp. 144–154, https://doi.org/10.1016/j.neunet.2021.10.012.

[50] D. YAROTSKY, *Error bounds for approximations with deep ReLU networks*, Neural Networks, 94 (2017), pp. 103–114, https://doi.org/10.1016/j.neunet.2017.07.002.

[51] J. YU, Z. LIN, J. YANG, X. SHEN, X. LU, AND T. S. HUANG, *Generative image inpainting with contextual attention*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 5505–5514.

[52] D.-X. ZHOU, *Universality of deep convolutional neural networks*, Appl. Comput. Harmon. Anal., 48 (2020), pp. 787–794, https://doi.org/10.1016/j.acha.2019.06.004.

[53] J.-Y. ZHU, T. PARK, P. ISOLA, AND A. A. EFROS, *Unpaired image-to-image translation using cycle-consistent adversarial networks*, in Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2017, pp. 2223–2232.

[54] Z. ZHU, T. XU, L. LI, AND Y. WANG, *Noise Dimension of GAN: An Image Compression Perspective*, preprint, arXiv:2403.09196, 2024.