Statistical Guarantees for Unpaired Image-to-Image Cross-Domain Analysis using GANs

Saptarshi Chakraborty
UC Berkeley

Peter L. Bartlett
Google DeepMind and UC Berkeley

Abstract

The field of unpaired image-to-image translation has undergone a significant transformation with the introduction of Generative Adversarial Networks (GANs), with CycleGAN and DiscoGAN as prominent variants. While these models show impressive empirical performance, their statistical properties are under-studied. In this paper, we propose a framework for analyzing the generalization error in cross-domain deep generative models. Our findings reveal that when provided with independent and identically distributed (i.i.d.) samples from two domains, the translation error, measured under the Wasserstein-1 loss, scales as $\tilde{\mathcal{O}}\left(\min\{n,m\}^{-1/\max\{d,\tilde{d}\}}\right)$, provided that the true model possesses sufficient smoothness and the network sizes are chosen appropriately. Here, n and m represent the sizes of the sample sets, while d and $\tilde{d}$ denote the dimensions of the respective data domains. Furthermore, we highlight the importance of a cycle loss term for ensuring distributional cycle consistency. Additionally, we provide insights into the relationship between the network size and the number of data points. Notably, as the true model exhibits greater smoothness, it suffices to work with smaller networks.

1 Introduction

Generative Adversarial Networks (GANs) have emerged as a focal point within the machine learning community, owing to their remarkable capacity for generating realistic and high-fidelity images, alongside their inherent simplicity and computational efficiency (Goodfellow et al., 2014). One particularly significant application of GANs is in unsupervised and unpaired image-to-image (I2I) translation, facilitating the transformation between two distinct yet conceivably related sets of images. The domain of I2I style transfer has witnessed substantial advances, offering a plethora of transformative possibilities while maintaining desired attributes or styles. Notably, artistic style transfer leverages this technology to imbue regular photographs with the stylistic essence of renowned artworks, thereby engendering visually captivating compositions (Gatys et al., 2016). Moreover, the utility of this technique extends to the realms of photo enhancement and restoration, where it manifests its ability to enhance visual quality and reinstate missing details within images (Chen and Koltun, 2017). Beyond the realm of visual aesthetics, I2I style transfer has been instrumental in diverse applications such as virtual try-on systems and fashion design, empowering users to envisage themselves in various attire or experiment with diverse fashion styles (Han et al., 2018). Additionally, it has been harnessed for domain adaptation, cross-modal translation, medical image analysis, and augmented reality/virtual reality (AR/VR) applications, underscoring its versatility and utility across multiple domains (Isola et al., 2017; Zhu et al., 2017; Lv et al., 2018; Hou et al., 2017).

CycleGAN, proposed by Zhu et al. (2017), stands out as one of the most influential approaches in this category. It offers an unsupervised I2I translation framework for unpaired observations hailing from unrelated data spaces. In terms of architecture, both DualGAN (Yi et al., 2017) and DiscoGAN (Kim et al., 2017) are closely related to CycleGAN. The main idea for these approaches is to minimize the sum of a *translation loss* and a *cycle loss*. The translation loss determines the quality of the images after mapping through the corresponding generator, while the cycle loss ensures that the composition of the two generator mappings roughly results in the same image. GAN-based

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

cross-domain models, such as CycleGAN or DiscoGAN have achieved remarkable empirical success in various I2I translation tasks, showcasing their effectiveness in handling unpaired image datasets. The approach has demonstrated impressive results in tasks such as style transfer, where it can convert images from one artistic style to another (Zhu et al., 2017). CycleGAN has also shown promise in domain adaptation tasks, enabling the translation of images from one domain to another without the need for paired training data (Hoffman et al., 2018). Additionally, it has been utilized in applications like object transfiguration, where it can transform images of one object into the appearance of another while preserving the overall structure (Choi et al., 2018). Some other notable extensions of these models include Domain Transfer Network (DTN) (Taigman et al., 2016) and Unsupervised Image-to-Image Translation (UNIT) (Liu et al., 2017), which restructure the model assuming a shared latent space between the two domains, the SCAN model (Li et al., 2018), which utilizes a stacked architecture with multiple translator networks to achieve significant performance improvements, especially for high-resolution images, and U-GAT-IT (Kim et al., 2019) and other related approaches (Moriakov et al., 2019), which address the issue of tilt-shift in generated images by incorporating an additional identity loss.

Recent advancements in cross-domain image-to-image translation have emphasized realism, disentanglement, and task-specific flexibility. Torbunov et al. (2023) introduced UvcGAN, integrating U-Net and Vision Transformers for unpaired translations, while Mahpod et al. (2023) developed CtrGAN, leveraging cycle-transformers for gait transfer. Wu et al. (2024) proposed StegoGAN, incorporating steganography for non-bijective transformations. Palette, by Saharia et al. (2022), showcased diffusion models as a powerful alternative for diverse conditional image translation tasks. Ge et al. (2021) utilized Disentangled Cycle Consistency for realistic virtual try-ons by separating appearance and geometric features. Expanding to videos and specialized applications, Wu et al. (2023) tackled semi-supervised video inpainting, and Jang et al. (2023) applied unsupervised cycle consistency for live-cell contour tracking.

However, despite the remarkable empirical success of cycle-consistent GANs, very little attention has been given to studying their statistical properties. The work by Moriakov et al. (2019) has shown that the CycleGAN problem can have multiple solutions due to the presence of nontrivial automorphisms in the data space. Tiao et al. (2018) explored the connection between a cycle-consistency loss and the expected posterior log-likelihood within a Bayesian framework. However, these studies do not explore the statistical translation and cycle consistency properties of such models. A recent contribution by Chakrabarty and Das (2022) attempted to explore these properties under certain restrictive assumptions. Their work assumes that the densities of the two image distributions are smooth, and they further require that the push-forward of these densities through the networks also exhibit smoothness, which cannot be ensured in practice. Notably, they introduce the cycle-consistency error using a Total Variation distance, which greatly simplifies their analysis. However, it is important to underscore that this specific choice is not implemented in practical settings. Moreover, their theoretical framework only tackles the generalization gap, ignoring the misspecification error for models that are not realizable through neural networks. Furthermore, their analysis does not provide any recipe to determine the network sizes that achieve an optimal rate. Additionally, it is unclear as to what role the cycle loss plays in making these models more efficient. In order to address these questions, we make the following contributions:

- This work addresses fundamental challenges in cross-domain image-to-image translation by providing a rigorous theoretical framework that bridges empirical observations and formal guarantees. Unlike prior works, which often focus on empirical novelty or heuristic approaches, this paper establishes precise error decay patterns and demonstrates the necessity of cycle consistency loss through a non-parametric statistical analysis. By deriving sharp generalization bounds that depend on the data dimensions and problem-smoothness, we aim to provide insights into the generalization performance of cross-domain generative models

- We present a framework for analyzing the error in cycle-consistent GANs. Our results show that when given independent and identically distributed (i.i.d.) samples from two domains, the reconstruction error under the Wasserstein-1 losses scales as $\tilde{\mathcal{O}}\left(\min\{n,m\}^{-1/\max\{d,\tilde{d}\}}\right)$, where n and m represent the sizes of the sample sets, and d and $\tilde{d}$ denote the dimensions of the respective data domains.

- We provide a recipe for selecting the appropriate sizes of the corresponding networks with Rectified Linear Unit (ReLU) activations that achieve this optimal rate. These network sizes are expressed in terms of the number of samples and the smoothness of the underlying distributions. In particular, the optimal network sizes decrease as the true model becomes smoother.

- We delve into the intrinsic significance of the cycle loss, demonstrating theoretically that even in the scenario of a smooth model, the assurance of cycle consistency across distributions may not hold in the absence of the cycle loss term within the objective function. However, through the incorporation of the cycle loss, we demonstrate that cycle consistency holds.

2 Background

2.1 Related Works on GANs

Apart from the works discussed in the introduction of the paper (Section 1), there has been a significant amount of research dedicated to understanding deep generative models and their properties. Biau et al. (2020) analyzed the asymptotic properties of vanilla GANs along with parametric rates. Biau et al. (2021) also analyzed the asymptotic properties of WGANs. Liang (2021) explored the min-max rates for WGANs for different non-parametric density classes and under a sampling scheme from a kernel density estimate of the data distribution. Schreuder et al. (2021) studied the finite-sample rates under adversarial noise. Uppal et al. (2019) derived the convergence rates for Besov discriminator classes for WGANs. Luise et al. (2020) conducted a theoretical analysis of WGANs under an optimal transport-based paradigm. Asatryan et al. (2020) and Belomestny et al. (2021) improved upon the works of Biau et al. (2020) to understand the behavior of GANs for a Hölder class of density functions. Arora et al. (2017) showed that generalization might not hold in standard metrics, but that under a restricted "neural-net distance", the GAN is indeed guaranteed to generalize well. Arora et al. (2018) showed that GANs and their variants might not be well-equipped against mode collapse. Huang et al. (2022) expressed the generalization rates for GANs when the latent space is one-dimensional, while Dahal et al. (2022) derived convergence rates under the Wasserstein-1 distance in terms of the manifold dimension. Recently, Chakraborty and Bartlett (2024) showed that the convergence rate for WGANs only depends on the Wasserstein (Weed and Bach, 2019) dimension of the data, further sharpening the rates of Huang et al. (2022).

2.2 Notations

Before we go into the details of the theoretical results, we introduce some notation and recall some preliminary concepts. We use the notation $x \vee y := \max\{x, y\}$ and $x \wedge y := \min\{x, y\}$. $T_\sharp\mu$ denotes the push-forward of the measure μ by the map T, i.e. $T_\sharp\mu(B) = \mu(T^{-1}(B))$, for any Borel set B. For any function $f : \mathcal{S} \to \mathbb{R}$, and any measure γ on $\mathcal{S}$, let $\|f\|_{\mathbb{L}_p(\gamma)} := \left(\int_{\mathcal{S}} |f(x)|^p d\gamma(x)\right)^{1/p}$, if $0 < p < \infty$. Also let, $\|f\|_{\mathbb{L}_\infty(\gamma)} := \operatorname{ess\,sup}_{x \in \text{supp}(\gamma)} |f(x)|$ For any function class $\mathcal{F}$, and distributions P and Q, the $\mathcal{F}$-Integral Probability metric (IPM) is defined as, $\|P - Q\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left|\int f dP - \int f dQ\right|$. For function classes $\mathcal{F}_1$ and $\mathcal{F}_2$, $\mathcal{F}_1 \circ \mathcal{F}_2 = \{f_1 \circ f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$. We say $A_n \lesssim B_n$ (also written as $A_n = \mathcal{O}(B_n)$) if there exists $C > 0$, independent of n, such that $A_n \le C B_n$. Similarly, the notation, "$\precsim$" (also written as $A_n = \tilde{\mathcal{O}}(B_n)$) if $A_n \le C \log(en)^k B_n$, for some $C, k > 0$, essentially ignoring factors that do not depend on poly-log terms in n. We say $A_n \asymp B_n$, if $A_n \lesssim B_n$ and $B_n \lesssim A_n$. For any $k \in \mathbb{N}$, we let $[k] = \{1, \dots, k\}$. For two random variables X and Y, we say that $X \overset{d}{=} Y$, if the random variables have the same distribution. $\mathcal{W}_p(\cdot, \cdot)$ denotes the Wasserstein p-distance between probability distributions. δ_x denotes the probability measure which assigns probability 1 at x. We use notations, $X_n \xrightarrow{a.s.} X$ and $X_n \xrightarrow{d} X$ to denote almost sure and in distribution convergence of random variables (Karr, 1993). Additional notations used in this paper appear in the supplement.

Definition 1 (Neural networks)**.** Let $L \in \mathbb{N}$ and $\{N_i\}_{i \in [L]} \subset \mathbb{N}$. Then a L-layer neural network $f : \mathbb{R}^d \to \mathbb{R}^{N_L}$ is defined as,

$$f(x) = A_L \circ \sigma_{L-1} \circ A_{L-1} \circ \cdots \circ \sigma_1 \circ A_1(x) \tag{1}$$

Here, $A_i(y) = W_i y + b_i$, with $W_i \in \mathbb{R}^{N_i \times N_{i-1}}$ and $b_i \in \mathbb{R}^{N_{i-1}}$, with $N_0 = d$. Note that σ_j is applied component-wise. Here, $\{W_i\}_{1 \le i \le L}$ are known as weights, and $\{b_i\}_{1 \le i \le L}$ are known as biases. $\{\sigma_i\}_{1 \le i \le L-1}$ are known as the activation functions. Without loss of generality, one can take $\sigma_\ell(0) = 0, \forall \, \ell \in [L-1]$. We define the following quantities: (Depth) $\mathcal{L}(f) := L$ is known as the depth of the network; (Number of weights) The number of weights of the network f is denoted as $\mathcal{W}(f)$. The set of neural networks with depth bounded by L, width bounded by B and output bounded by R is denoted as:

$$\mathcal{NN}_{\{\sigma_i\}_{1 \le i \le L-1}}(L, W, R) = \Big\{f \text{ of the form (1)} : \mathcal{L}(f) \le L, \\ \mathcal{W}(f) \le W \text{ and } \sup_{x \in [0,1]^d} \|f(x)\|_\infty \le R\Big\}.$$

If $\sigma_j(x) = x \vee 0$, for all $j = 1, \dots, L-1$, we denote $\mathcal{NN}_{\{\sigma_i\}_{1 \le i \le L-1}}(L, W, R)$ as $\mathcal{RN}(L, W, R)$. We often drop R from the notation, when it is clear that R is finite.

Definition 2 (Hölder functions)**.** Let $f : \mathcal{S} \to \mathbb{R}$ be a function, where $\mathcal{S} \subseteq \mathbb{R}^d$. For a multi-index $\boldsymbol{s} = (s_1, \dots, s_d)$, let, $\partial^{\boldsymbol{s}} f = \frac{\partial^{|\boldsymbol{s}|} f}{\partial x_1^{s_1} \dots \partial x_d^{s_d}}$, denote the weak partial derivative of f, where, $|\boldsymbol{s}| = \sum_{\ell=1}^d s_\ell$.

We say that a function $f : \mathcal{S} \to \mathbb{R}$ is β-Hölder (for $\beta > 0$) if $\|f\|_{\mathcal{H}^\beta} := \sum_{\boldsymbol{s}:0\le|\boldsymbol{s}|\le\lfloor\beta\rfloor} \|\partial^{\boldsymbol{s}} f\|_\infty + \sum_{\boldsymbol{s}:|\boldsymbol{s}|=\lfloor\beta\rfloor} \sup_{x\neq y} \frac{\|\partial^{\boldsymbol{s}} f(x)-\partial^{\boldsymbol{s}} f(y)\|}{\|x-y\|^{\beta-\lfloor\beta\rfloor}} < \infty$. If $f : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$, then we define $\|f\|_{\mathcal{H}^\beta} = \sum_{j=1}^{\tilde{d}} \|f_j\|_{\mathcal{H}^\beta}$. We let, $\mathcal{H}^\beta(\mathcal{S}_1, \mathcal{S}_2, C) = \{f : \mathcal{S}_1 \to \mathcal{S}_2 : \|f\|_{\mathcal{H}^\beta} \le C\}$, where $\mathcal{S}_1$ and $\mathcal{S}_2$ are subsets of real vector spaces.

2.3 An Overview of CycleGANs

Suppose that the two domains of interest are $\mathcal{X}$ and $\tilde{\mathcal{X}}$. Since vectorized image data typically originates from bounded domains, for simplicity of exposition, we consider $\mathcal{X} = [0,1]^d$ and $\tilde{\mathcal{X}} = [0,1]^{\tilde{d}}$. Let μ be a distribution on $\mathcal{X}$ and ν be another distribution on $\tilde{\mathcal{X}}$. The goal is to learn functions (also known as generators), $G : \tilde{\mathcal{X}} \to \mathcal{X}$ and $F : \mathcal{X} \to \tilde{\mathcal{X}}$ such that $G_\sharp \nu \approx \mu$ and $F_\sharp \mu \approx \nu$. Learning such maps corresponds to style/domain transfer between the distributions μ and ν. The objective for the CycleGAN problem is given by,

$$V(\mu, \nu, G, F) = \underbrace{\|\mu - G_\sharp \nu\|_{\mathcal{D}} + \|\nu - F_\sharp \mu\|_{\tilde{\mathcal{D}}}}_{\text{Translation loss}} \tag{2}$$
$$+ \underbrace{\lambda \left(\int c(x, G \circ F(x)) d\mu(x) + \int \tilde{c}(\tilde{x}, F \circ G(\tilde{x})) d\nu(\tilde{x}) \right)}_{\text{Cycle loss}}.$$

Here, $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\ge 0}$ and $\tilde{c} : \tilde{\mathcal{X}} \times \tilde{\mathcal{X}} \to \mathbb{R}_{\ge 0}$ denote the loss functions in the respective domains, $\mathcal{X}$ and $\tilde{\mathcal{X}}$. In objective (2), the first term determines the Integral Probability Metric (IPM) w.r.t. the discriminator class $\mathcal{D}$ for measuring the difference between μ and the push-forward of ν by G. Similarly, the second term determines the error of estimating ν with $F_\sharp \mu$ w.r.t. the $\tilde{\mathcal{D}}$-IPM. Here $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are function classes on $\mathcal{X}$ and $\tilde{\mathcal{X}}$, respectively. Throughout this analysis, we take $\mathcal{D} = \mathcal{H}^\beta(\mathcal{X}, \mathbb{R}, 1)$ and $\tilde{\mathcal{D}} = \mathcal{H}^{\tilde{\beta}}\left(\tilde{\mathcal{X}}, \mathbb{R}, 1\right)$. The sum of the first two terms in (2) is known as the *translation loss*, where as the third term in (2) is known as the *reconstruction* or the *cycle* loss. This loss tries to make the compositions of the two generators, i.e. $G \circ F$ and $F \circ G$ close to the identity map, i.e. $id(x) = x$ w.r.t. the two loss functions. The introduction of this loss tries to ensure that the translation of the same point through the cycle of the two generators is mapped close to the initial value. Zhu et al. (2017) took both c and $\tilde{c}$ to be the ℓ_1-norm, whereas, Kim et al. (2017) considered the squared ℓ_2-norm. A pictorial representation of objective (2) is shown in Fig. 1.

In practice, one only has access to i.i.d. samples from μ and ν. We assume the following.

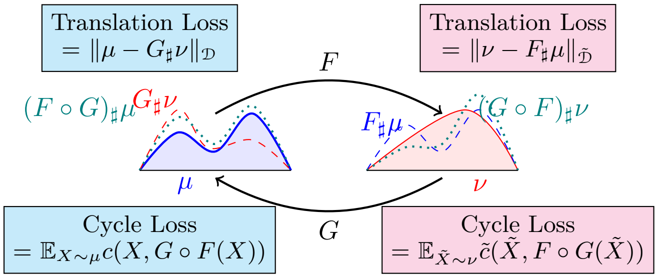

Translation Loss
$= \|\mu - G_\sharp \nu\|_{\mathcal{D}}$
Translation Loss
$= \|\nu - F_\sharp \mu\|_{\tilde{\mathcal{D}}}$
F
$(F \circ G)_\sharp \mu$
$G_\sharp \nu$
$F_\sharp \mu$
$(G \circ F)_\sharp \nu$
μ
ν
G
Cycle Loss
$= \mathbb{E}_{X\sim\mu} c(X, G \circ F(X))$
Cycle Loss
$= \mathbb{E}_{\tilde{X}\sim\nu} \tilde{c}(\tilde{X}, F \circ G(\tilde{X}))$

Figure 1: A pictorial representation of image-to-image translation using CycleGAN. The two image distributions are denoted as μ and ν. The transformation is achieved by employing the cycle-consistent mapping between the domains, which allows for the conversion of images from one domain to another while preserving their essential characteristics.

A 1. *$\{X\}_{i\in[n]}$ are independent and identically distributed according to the distribution μ. Furthermore, $\{\tilde{X}\}_{j\in[m]}$ are also i.i.d. and follow the distribution ν, with $\{X\}_{i\in[n]}$ and $\{\tilde{X}\}_{j\in[m]}$ being independent.*

The empirical counterpart of (2) is given by,

$$\begin{aligned} &V(\hat{\mu}_n, \hat{\nu}_m, G, F) \\ =&\|\hat{\mu}_n - G_\sharp \hat{\nu}_m\|_{\mathcal{D}} + \|\hat{\nu}_m - F_\sharp \hat{\mu}_n\|_{\tilde{\mathcal{D}}} \qquad (3) \\ &+ \frac{\lambda}{n} \sum_{i=1}^n c(X_i, G \circ F(X_j)) + \frac{\lambda}{m} \sum_{j=1}^m \tilde{c}(\tilde{X}_j, F \circ G(\tilde{X}_j)). \end{aligned}$$

The above objective is minimized w.r.t. the generators $G \in \mathcal{G} \equiv \mathcal{RN}(L_g, W_g)$ and $F \in \mathcal{F} \equiv \mathcal{RN}(L_f, W_f)$, where, $\mathcal{G}$ and $\mathcal{F}$ are taken to be a class of neural networks with ReLU activation. The empirical estimates of G and F are defined as:

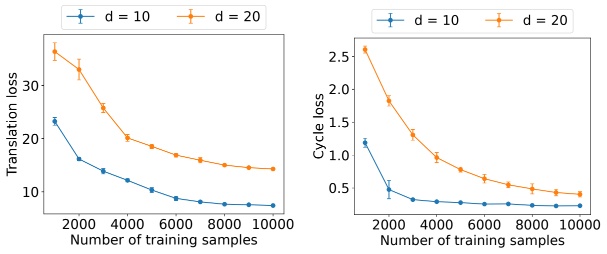
$$(\hat{G}, \hat{F}) \in \underset{G\in\mathcal{G},\, F\in\mathcal{F}}{\operatorname{argmin}}\ V(\hat{\mu}_n, \hat{\nu}_m, G, F). \tag{4}$$

In practice, one can only estimate $(\hat{G}, \hat{F})$ up to an optimization error. We measure the excess risk of the estimates as: $\mathfrak{R}(\hat{G}, \hat{F}) = V(\mu, \nu, \hat{G}, \hat{F})$. In Section 4, we derive bounds on this excess risk of the estimates under certain regularity assumptions.

Remark 3 (Identity Loss)**.** Moreover, to preserve the color composition and spatial structures in images, models like the extended CycleGAN (Zhu et al., 2017) and U-GAT-IT (Kim et al., 2017) introduce a constraint that discourages the translators from deviating significantly from the identity mapping:

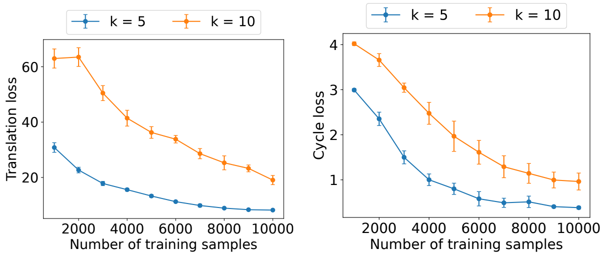
$$L_{\text{identity}} = \mathbb{E}_{X\sim\mu}\|X - F(X)\|_1 + \mathbb{E}_{\tilde{X}\sim\nu}\|\tilde{X} - G(\tilde{X})\|_1.$$

It is worth noting that such regularization is only applicable when the two data distributions have the same dimensionality. Additionally, while the primary objective of constructing F is to map μ to ν, as argued by Chakrabarty and Das (2022), the identity



(a) Translation Loss (b) Cycle Loss

Figure 2: (Experiment 1) Translation loss and cycle loss for CycleGAN for Gaussian to $\chi^2_{(4)}$ problem on $\mathbb{R}^d$, for different training sample sizes. The error bars denote the standard deviation out of 10 replications.



(a) Translation Loss (b) Cycle Loss

Figure 3: (Experiment 2) Translation loss and cycle loss for CycleGAN for Gaussian to $\chi^2_{(k)}$ problem on $\mathbb{R}^{10}$, for different training sample sizes. The error bars denote the standard deviation out of 10 replications.

loss, L_{identity} cannot be minimized beyond the inherent differences between μ and ν. Therefore, we opt not to include the identity loss in our objective (2), as it primarily functions as a regularizer to preserve structural consistency in images.

3 A Proof of Concept

Before we theoretically explore the problem, we discuss two experiments to demonstrate that the error rates for CycleGANs depend primarily only on the dimension of the data and the smoothness of the problem. The codes pertaining to the experiments are available at https://github.com/saptarshic27/CG.

Experiment 1 We take μ to be the standard Gaussian distribution on $\mathbb{R}^d$ and ν to be the distribution of the random vector in $\mathbb{R}^d$, whose coordinates are independent $\chi^2_{(4)}$ random variables. We take the discriminator and generators to be feed-forward neural networks to have three hidden layers with 128 nodes each. We use ReLU as the activation and the final output of the discriminator to have a sigmoid activation. The GAN-loss for the cycle GAN is taken to be the binary cross-entropy loss and the cycle loss is taken to have the ℓ_1-norm. We use the Adam optimizer with a learning rate of 0.0002 and $(\beta_1, \beta_2) = (0.5, 0.999)$. The penalty on the cycle loss is taken to be $\lambda = 10$. We run the experiments for $d \in \{10, 20\}$ for 30 epochs with a batch size of 100 with a total of 10 repetitions. The sample size is varied in $\{1000, 2000, \cdots, 10000\}$. The translation loss computed as $\mathcal{W}_2(\mu, \hat{G}_\sharp \nu) + \mathcal{W}_2(\nu, \hat{F}_\sharp \mu)$, using 2000 test samples as well as the test cycle loss is shown in Fig. 2. It is clear from Fig. 2 that the error rates for $d = 10$ are lower than for the case $d = 20$, showing that the hardness of the problem increases as the dimensions grow as predicted by the main result (Theorem 5). The codes of the experiment are provided in the supplement.

Experiemnt 2 Additionally, the test error, as a function of the training sample size, depends heavily on the smoothness of the problem as predicted by our theoretical analyses. We take a similar setting as described in Experiment 1. We fix $d = 10$ and take the two domains as $\mathbb{R}^d$. We take μ to be the standard Gaussian distribution on $\mathbb{R}^d$ and ν to be the distribution on $\mathbb{R}^d$, whose individual coordinates are independently distributed as $\chi^2_{(k)}$ random variables. We use the same settings as described in Experiment 1 by increasing the number of samples and varying $k \in \{5, 10\}$. It can be expected that the smoothness of the problem deteriorates when $k = 10$ compared to when $k = 5$. The experiments show that the decay patterns of the translation and cycle losses, as n is increased, are influenced by the value of k and the test losses decrease faster for $k = 5$ compared to the case $k = 10$.

4 Theoretical Analyses

4.1 Assumptions and Main Result

We begin by stating the necessary assumptions. We assume that there exists a set of true generators, i.e. $G^\star$ and $F^\star$, that are smooth in the sense that they are Hölder continuous. Formally,

A 2. *There exists $G^\star \in \mathcal{H}^\alpha\left(\tilde{\mathcal{X}}, \mathcal{X}, C\right)$ and $F^\star \in \mathcal{H}^{\tilde{\alpha}}\left(\mathcal{X}, \tilde{\mathcal{X}}, \tilde{C}\right)$, such that, $(X, F^\star(X)) \overset{d}{=} (G^\star(\tilde{X}), \tilde{X})$. Here, $X \sim \mu$ and $\tilde{X} \sim \nu$.*

One can think of α and $\tilde{\alpha}$ as a measure of smoothness of the underlying measures μ and ν. An easy consequence of Assumption A2 is that the maps $G^\star$ and $F^\star$ are inverses of each other, almost surely.

Lemma 4. *Under assumption A2, $F^\star \circ G^\star(\cdot) = id(\cdot)$, a.e. $[\mu]$ and $G^\star \circ F^\star(\cdot) = id(\cdot)$, a.e. $[\nu]$.*

It is important to note that neither the encoder nor the decoder maps are expected to be unique as there can be many realizations of the same random variable via different maps. For example, if X and $\tilde{X}$ are distributed uniformly on the unit hypercube $[0,1]^d$, both the pair of maps, $(G_1^\star(x) = x, F_1^\star(x) = x)$ and $(G_2^\star(x) = 1-x, F_2^\star(x) = 1-x)$ minimize (2). Thus, it does not make sense to study the behavior of $\|G^\star - \hat{G}\|$ or $\|F^\star - \hat{F}\|$ in some function norm but how well these maps can estimate the distributions, i.e. $\|\mu - \hat{G}_\sharp \nu\|_{\mathcal{D}}$ and $\|\nu - \hat{F}_\sharp \nu\|_{\tilde{\mathcal{D}}}$, which is captured by $\mathfrak{R}(\cdot,\cdot)$.

We also assume that the functions $c(\cdot,\cdot)$ and $\tilde{c}(\cdot,\cdot)$ are Lipschitz. This is true for the commonly used functions in practice, e.g. the ℓ_1 (Zhu et al., 2017) or the squared ℓ_2-loss (Kim et al., 2017) on compact domains. Clearly, due to the boundedness of the domains $\mathcal{X}$ and $\tilde{\mathcal{X}}$, the cost functions c and $\tilde{c}$ are also bounded due to their Lipschitzness.

A 3. *We assume that there exists a positive constant τ, such that $\|c(x,y) - c(x',y')\| \leq \tau\left(\|x-x'\|_2 + \|y-y'\|_2\right)$ and $\|\tilde{c}(\tilde{x},\tilde{y}) - \tilde{c}(\tilde{x}',\tilde{y}')\| \leq \tau\left(\|\tilde{x}-\tilde{x}'\|_2 + \|\tilde{y}-\tilde{y}'\|_2\right)$. Furthermore, $c(x,x) = \tilde{c}(\tilde{x},\tilde{x}) = 0$.*

Under Assumptions A1–3, one can bound the excess risk for the CycleGAN problem in terms of the sample-sizes n, m, the data dimensions d, $\tilde{d}$ and the smoothness parameters of the corresponding generator and discriminator networks. The main result of this paper is summarized as follows, with a proof sketch provided in the subsequent section.

Theorem 5 (Main Theorem)**.** *Under assumptions A1–3, we can find an $n_0 \in \mathbb{N}$ (that might depend on $d, \tilde{d}, \alpha, \tilde{\alpha}, \beta$ and $\tilde{\beta}$), such that if $n, m \geq n_0$, then,*

$$\mathbb{E}\,\mathfrak{R}(\hat{G},\hat{F}) \lesssim (n \wedge m)^{-\left(\max\left\{2+\frac{\tilde{d}}{\alpha(\tilde{\alpha}\wedge\beta\wedge 1)}, 2+\frac{d}{\tilde{\alpha}(\alpha\wedge\tilde{\beta}\wedge 1)}, \frac{d}{\beta}, \frac{\tilde{d}}{\tilde{\beta}}\right\}\right)^{-1}},$$

if $\mathcal{G} = \mathcal{RN}(L_g, W_g, 2C)$ and $\mathcal{F} = \mathcal{RN}(L_f, W_f, 2\tilde{C})$, where one can choose the network size as $L_g, L_f \lesssim \log(n \wedge m)$, $W_g \lesssim (n \wedge m)^{\tilde{d}/(\tilde{d}+2(\tilde{\alpha}\wedge\beta\wedge 1)\alpha)}$ and $W_f \lesssim (n \wedge m)^{d/(d+2(\alpha\wedge\tilde{\beta}\wedge 1)\tilde{\alpha})}$.

4.2 Inference from the Main Result

Smooth Models and Network Sizes Let us consider the specific case where we focus solely on the Wasserstein metric, which is equivalent to the 1-Hölder IPM on bounded domains. In addition, it is reasonable to assume that the dimensions d and $\tilde{d}$ of the respective domains are both greater than or equal to 3. We note that when the true model for the CycleGAN problem is assumed to be sufficiently smooth, i.e. if $\alpha \geq \frac{\tilde{d}}{\tilde{d}-2}$ and $\tilde{\alpha} \geq \frac{d}{d-2}$, then the excess risk for the CycleGAN problem roughly scales as $\tilde{\mathcal{O}}\left((n \wedge m)^{-1/(d \vee \tilde{d})}\right)$, barring the log-factors. This indicates that as the number of samples increases, the excess risk diminishes at a rate determined by the larger of the two dimensions d and $\tilde{d}$.

It is worth noting that according to Theorem 5, the depths of the generators in the CycleGAN framework can be selected on the order of $\log(n \wedge m)$. Additionally, the number of weights can be chosen as some exponent of $n \wedge m$. Importantly, the values of these exponents depend solely on the dimensions of the data and the smoothness properties of the true models. Moreover, the factors of these exponents are bounded by 1. Moreover, the bounds provided in Theorem 5 for the number of weights indicate that a smaller network is required when the smoothness parameters α and $\tilde{\alpha}$ increase. This implies that if the true generators exhibit high levels of smoothness and are well-behaved, the optimal choices for the number of weights decrease as α and $\tilde{\alpha}$ increase. This finding aligns with our expectations, as it suggests that less complex architectures are sufficient when dealing with target densities that are already sufficiently smooth.

We note that compared to the recently derived finite-sample error bounds for GANs, the error bounds for CycleGANs depend on the full data dimensions as opposed to the intrinsic dimension as shown in the GAN literature (Huang et al., 2022; Chakraborty and Bartlett, 2024). This is because, in GANs, one has the luxury to generate as many test samples as one wishes. This enables one to get better behaviors even when the range of the sample generator does not lie on an intrinsically low-dimensional structure. This enables one to increase the number of fake samples to match the non-parametric convergence rates for $\|\hat{\mu}_n - \mu\|_{\mathcal{H}^\beta}$, which appears in the oracle inequality to bound the error term (see Theorem 11). However, CycleGANs have no such luxury and the error rates depreciate because the sample generators, $\hat{G}$ and $\hat{F}$ cannot be restricted to a low-dimensional structure.

Model misspecification Although the analysis is conducted for a realizable setting, it can be extended to a non-realizable case with an additional misspecification error term on the right of Theorem 5. For simplicity, let us assume that A2 holds. However, one does not observe samples from μ and ν but samples from distributions μ' and ν' such that $\mathrm{TV}(\mu,\mu') + \mathrm{TV}(\nu,\nu') \leq \Delta$. Here $\mathrm{TV}(P,Q) := \sup_B |P(B) - Q(B)|$ denotes the total variation distance. Then, one only has to control the first term in Lemma 11 as the other terms can be bounded in the same way. Now, repeating the analysis, one will obtain an additional on the RHS of Theorem 5. We state the corresponding result as follows:

Corollary 6. *Suppose that* $X_1, \dots X_n \overset{i.i.d.}{\sim} \mu'$ *and* $\tilde{X}_1, \dots, \tilde{X}_m \overset{i.i.d.}{\sim} \nu'$, *such that,* $\mathrm{TV}(\mu, \mu') + \mathrm{TV}(\nu, \nu') \leq \Delta$. *Under the assumptions and choices of Theorem 5, if* $n, m \geq n_0$, *then,*

$$\mathbb{E}\,\mathfrak{R}(\hat{G}, \hat{F})$$
$$\precsim \Delta + (n \wedge m)^{-\left(\max\left\{2+\frac{\tilde{d}}{\alpha(\tilde{\alpha} \wedge \beta \wedge 1)}, 2+\frac{d}{\tilde{\alpha}(\alpha \wedge \tilde{\beta} \wedge 1)}, \frac{d}{\beta}, \frac{\tilde{d}}{\tilde{\beta}}\right\}\right)^{-1}}$$

4.3 Translation Consistency

One important consequence of Theorem 5 is that it guarantees that the model is translation consistent, i.e. $\hat{G}_\sharp \nu \xrightarrow{d} \mu$ and $\hat{F}_\sharp \mu \xrightarrow{d} \nu$, almost surely. For simplicity, we assume that $\beta = \tilde{\beta} = 1$, i.e., the discriminators are bounded and 1-Lispchitz. To show translation consistency, we first note that $V(\mu, \nu, \hat{G}, \hat{F})$ converges to 0, almost surely as $n \wedge m \to 0$. Formally,

Corollary 7. *Under Assumptions A1–3,* $\mathfrak{R}(\hat{G}, \hat{F}) \xrightarrow{a.s.} 0$, *as* $n \wedge m \to \infty$.

Corollary 7 directly leads to translation consistency of the estimates. The next proposition states and proves this result.

Proposition 8. *Suppose that* $\beta = \tilde{\beta} = 1$. *Then, under Assumptions A1–3,* $\hat{G}_\sharp \nu \xrightarrow{d} \mu$ *and* $\hat{F}_\sharp \mu \xrightarrow{d} \nu$, *almost surely, as* $n \wedge m \to \infty$.

Proof. We show that $\hat{G}_\sharp \nu \xrightarrow{d} \mu$, almost surely. To see this, we observe that $0 \leq \|\mu - \hat{G}_\sharp \nu\|_{\mathscr{D}} \leq V(\mu, \nu, \hat{G}, \hat{F}) \xrightarrow{a.s.} 0$. It is easy to note that $\{f : [0,1]^d \to \mathbb{R} : \|f\|_{\mathrm{Lip}} \leq d^{-1/2}\} \subset \mathscr{D}$. Thus, $\sup_{\|f\|_{\mathrm{Lip}} \leq d^{-1/2}} \left(\int f d\mu - \int f d(\hat{G}_\sharp \nu)\right) \xrightarrow{a.s.} 0$, which further implies that $\mathscr{W}_1(\mu, \hat{G}_\sharp \nu) \xrightarrow{a.s.} 0$. Since the Wasserstein-1 metric determines weak-convergence on bounded Polish spaces (Villani et al., 2009, Theorem 6.9), it implies that $\hat{G}_\sharp \nu \xrightarrow{d} \mu$, almost surely, as $n \wedge m \to \infty$. The other implication can be shown similarly. □

4.4 Cycle Consistency

Let us now explore the importance of having a cycle loss in objectives (2) and (3). We begin by defining the notion of cycle consistency of the distributions.

Definition 9 (Cycle Consistency)**.** *For the CycleGAN problem, we say that estimated functions* $\hat{G}$ *and* $\hat{F}$ *are cycle consistent if* $\|\mu - (\hat{G} \circ \hat{F})_\sharp \mu\|_{\mathscr{D}} \xrightarrow{a.s.} 0$ *and* $\|\nu - (\hat{F} \circ \hat{G})_\sharp \nu\|_{\tilde{\mathscr{D}}} \xrightarrow{a.s.} 0$.

Intuitively, the requirement states that the cyclic push-forwards, $(\hat{G} \circ \hat{F})_\sharp \mu$ and $(\hat{F} \circ \hat{G})_\sharp \nu$, should be close to the target distributions μ and ν, respectively, in the IPM defined by the corresponding discriminators on that space. In simpler terms, this means that the transformed distributions resulting from applying the mappings $\hat{G}$ and $\hat{F}$ should approximate the original distributions in a manner consistent with the evaluation of their proximity using the discriminators. Without the cycle loss in objectives (2) and (3), one cannot ensure the cycle consistency of the estimated generators, even under Assumption A2. To see this, we consider a simple counterexample with $\mathscr{X} = \tilde{\mathscr{X}} = \mathbb{R}$. Consider the distributions $\mu = \nu \equiv \mathrm{Unif}([0,1])$ be the uniform distribution on $[0,1]$. We take $\mathscr{D} = \tilde{\mathscr{D}} = \mathscr{H}^1(\mathbb{R}, \mathbb{R}, 1)$, i.e. the set of all 1-bounded 1-Lipschitz functions on $\mathbb{R}$. It is well known that $\mathscr{D}$ (or equivalently $\tilde{\mathscr{D}}$) metrizes in distribution convergence of real random variables. Clearly there exist maps $G^\star = F^\star = id(\cdot)$, i.e the identity map, such that $G^\star_\sharp \nu = \mu$ and $F^\star_\sharp \mu = \nu$. Furthermore both $G^\star$ and $F^\star$ are 1-Lipschitz. Thus, Assumption A2 is satisfied with $\beta = \tilde{\beta} = 1$ and $C = \tilde{C} = 1$. $F^n(x) = \frac{1}{n}\sum_{k=0}^{n-1} k \;\left(\left(x \in \left(\frac{k}{n}, \frac{k+1}{n}\right)\right)\right)$ and $G^n(x) = x - x\sum_{k=0}^{n-1} \;(x = k/n)$. Clearly, $F^n_\sharp \mu$ is uniformly distributed on $A_n = \{k/n : k = 0, \dots, n-1\}$ and $G^n_\sharp \nu$ is uniformly distributed on $[0,1]$ (since altering the outcomes on a zero measure set i.e. A_n does not change the density). It is well known that the discrete uniform distribution converges to the continuous one in law (for example see exercise 5.8 of Karr (1993)). Hence, $F^n_\sharp \mu \xrightarrow{d} \nu$ and $G^n_\sharp \nu \xrightarrow{d} \mu$. However, $G^n \circ F^n(x) = 0$, $\forall x \in [0,1]$, making $(G^n \circ F^n)_\sharp \mu \overset{d}{\not\to} \mu$. Thus, for this example, $(G^n \circ F^n)_\sharp \mu \overset{a.s.}{\not\to} \mu$. Hence the choices of G^n and F^n are not cycle-consistent.

However, under the use of cycle loss, i.e. for the case $\lambda > 0$, one can ensure the cycle consistency of the estimates $\hat{G}$ and $\hat{F}$. Using Corollary 7, we can state that the CycleGAN estimates $\hat{G}$ and $\hat{F}$ that are cycle-consistent. Since we are only interested in the CycleGAN estimates, we take c and $\tilde{c}$ to be the ℓ_1 distances on $\mathscr{X}$ and $\tilde{\mathscr{X}}$, respectively.

Proposition 10. *The CycleGAN estimates* $\hat{G}$ *and* $\hat{F}$ *defined in (4) are cycle-consistent.*

5 Proof Overview of the Main Result

As a preliminary step towards deriving bounds on the expected risks for the CycleGAN problem, we proceed by deriving the following oracle inequality. This inequality serves to bound the excess risk in terms of the approximation error and a generalization gap. The proof is provided in the supplement.

Lemma 11 (Oracle Inequality)**.** *For the estimates* $\hat{G}$

and $\hat{F}$,

$$
\begin{aligned}
& V(\mu, \nu, \hat{G}, \hat{F}) \\
\leq & \inf_{G \in \mathcal{G}, F \in \mathcal{F}} V(\mu, \nu, G, F) + 2\|\hat{\mu}_n - \mu\|_{\mathcal{D}} + 2\|\hat{\nu}_m - \nu\|_{\mathcal{D}\circ\mathcal{G}} \\
& + 2\|\hat{\nu}_m - \nu\|_{\tilde{\mathcal{D}}} + 2\|\hat{\mu}_n - \mu\|_{\tilde{\mathcal{D}}\circ\mathcal{F}} + 2\lambda\|\hat{\mu}_n - \mu\|_{\Phi_1} \\
& + 2\lambda\|\hat{\nu}_m - \nu\|_{\Phi_2}, \qquad (5)
\end{aligned}
$$

where, $\Phi_1 = \{c(\cdot, G \circ F(\cdot)) : G \in \mathcal{G},\, F \in \mathcal{F}\}$ and $\Phi_2 = \{\tilde{c}(\cdot, F \circ G(\cdot)) : G \in \mathcal{G},\, F \in \mathcal{F}\}$.

Building upon Lemma 11, we aim to bound each of the terms separately in the following sections. To control the misspecification errors, we develop an approximation result that deals with ReLU networks. We address the generalization gap by employing elements from the empirical process theory literature.

5.1 Approximation Error

The approximation capabilities of neural networks have garnered significant attention over the past decade. Pioneering works by Cybenko (1989) and Hornik (1991) explored the universal approximation of networks with sigmoid-like activations, demonstrating that wide one-hidden-layer neural networks can approximate any continuous function on a compact set. More recently, researchers have investigated the approximation capabilities of deep neural networks (Lu et al., 2021; Petersen and Voigtlaender, 2018; Nakada and Imaizumi, 2020; Shen et al., 2022). Notably, Yarotsky (2017) showed that Sobolev functions can be approximated by ReLU networks. One can ensure that ReLU functions can ϵ-approximate β-Hölder functions when they have a depth of at most $\mathcal{O}(\log(1/\epsilon))$ and number of weight at most $\mathcal{O}(\epsilon^{-d/\beta}\log(1/\epsilon))$ as,

Lemma 12. *Suppose that $f \in \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$, for some $C > 0$. Then, we can find a constant α, that might depend on β, d and C, such that, for any $\epsilon \in (0,1)$, there exists a ReLU network, $\hat{f}$ with $\mathcal{L}(\hat{f}) \leq \alpha \log(1/\epsilon)$, $\mathcal{W}(\hat{f}) \leq \alpha \log(1/\epsilon)\epsilon^{-d/\beta}$, $\mathcal{B}(\hat{f}) \leq \alpha\epsilon^{-1/\beta}$ and $\mathcal{R}(\hat{f}) \leq 2C$, that satisfies, $\|f - \hat{f}\|_{\ell_\infty([0,1]^d)} \leq \epsilon$.*

Using Lemma 12, we can control the misspecification error for the CycleGAN problem as in the following lemma.

Lemma 13. *Suppose assumption A2 holds. Then, for any $\epsilon, \epsilon_2 \in (0,1)$, we can find networks with ReLU activation, $\mathcal{G} = \mathcal{RN}\,(L_g, W_g)$ and $\mathcal{F} = \mathcal{RN}\,(L_f, W_f)$, with $L_g \asymp \log(1/\epsilon_1)$, $W_g \asymp \epsilon_1^{-\tilde{d}/\alpha}\log(1/\epsilon_1)$, $L_f \asymp \log(1/\epsilon_2)$ and $W_f \asymp \epsilon_2^{-d/\tilde{\alpha}}\log(1/\epsilon_2)$, such that $V(\mu, \nu, G, F) \lesssim \epsilon_1^{\tilde{\alpha}\wedge\beta\wedge 1} + \epsilon_2^{\alpha\wedge\tilde{\beta}\wedge 1}$.*

5.2 Generalization Bound

In order to establish our generalization bounds, the next step is to effectively control the uniform concentration with respect to the function classes $\mathcal{D}$, $\tilde{\mathcal{D}}$, $\mathcal{D} \circ \mathcal{G}$, $\tilde{\mathcal{D}} \circ \mathcal{F}$, Φ_1, and Φ_2. Since the function classes $\mathcal{D}$ and $\tilde{\mathcal{D}}$ exhibit Hölder continuity, their corresponding generalization bounds can be addressed by drawing upon the seminal results of Kolmogorov and Tikhomirov (1961). Before we proceed, we recall the following definition:

Definition 14 (Covering Number)**.** For a metric space $(\mathcal{S}, \varrho)$, the ϵ-covering number w.r.t. ϱ is defined as: $\mathcal{N}(\epsilon; \mathcal{S}, \varrho) = \inf\{n \in \mathbb{N} : \exists x_1, \dots x_n$ such that $\cup_{i=1}^n B_\varrho(x_i, \epsilon) \supseteq \mathcal{S}\}$.

We also use the notation, $\mathcal{F}_{|_{X_{1:m}}} = \{[f(X_1) : \cdots : f(X_m)] : f \in \mathcal{F}\} \subseteq \mathbb{R}^{d' \times m}$, where $f : \mathbb{R}^d \to \mathbb{R}^{d'}$. One can we state the generalization bounds for $\|\hat{\mu}_n - \mu\|_{\mathcal{D}}$ and $\|\hat{\nu}_m - \nu\|_{\tilde{\mathcal{D}}}$ as follows.

Lemma 15. *Under Assumption A1, the followings hold:*

$$
\begin{aligned}
\|\hat{\mu}_n - \mu\|_{\mathcal{D}} &\lesssim n^{-\beta/d} \vee n^{-1/2}(\log n)^{\{d=2\beta\}}, \\
\|\hat{\nu}_m - \nu\|_{\tilde{\mathcal{D}}} &\lesssim m^{-\tilde{\beta}/\tilde{d}} \vee m^{-1/2}(\log m)^{\{\tilde{d}=2\tilde{\beta}\}}.
\end{aligned}
$$

The next step is to prove a generalization bound for the terms $\|\hat{\mu}_n - \mu\|_{\tilde{\mathcal{D}}\circ\mathcal{F}}$ and $\|\hat{\nu}_m - \nu\|_{\mathcal{D}\circ\mathcal{G}}$ by controlling the metric entropies of the corresponding function classes, $\tilde{\mathcal{D}} \circ \mathcal{F}$ and $\mathcal{D} \circ \mathcal{G}$, and then applying Dudley's chaining to bound these generalization terms. Lemmas 16 and 17 state these two results, respectively, with proofs in the supplement.

Lemma 16. *Suppose that $\mathcal{G} = \mathcal{RN}(L_g, W_g, 2C_g)$ and $\mathcal{F} = \mathcal{RN}(L_f, W_f, 2C_f)$, then, there exists a constant c, such that, if $m \geq cW_gL_g(\log W_g + L_g)$ and $n \geq cW_fL_f(\log W_f + L_f)$,*

$$
\log N_1 \lesssim \epsilon^{-d/\beta} + W_gL_g(\log W_g + L_g)\log(md/\epsilon), \qquad (6)
$$

$$
\log N_2 \lesssim \epsilon^{-\tilde{d}/\tilde{\beta}} + W_fL_f(\log W_f + L_f)\log(n\tilde{d}/\epsilon), \qquad (7)
$$

where $N_1 = \mathcal{N}\left(\epsilon; (\mathcal{D} \circ \mathcal{G})_{|_{\tilde{X}_{1:m}}}, \|\cdot\|_\infty\right)$ and $N_2 = \mathcal{N}\left(\epsilon; (\tilde{\mathcal{D}} \circ \mathcal{F})_{|_{X_{1:n}}}, \|\cdot\|_\infty\right)$.

Lemma 17. *Suppose that $\mathcal{G} = \mathcal{RN}(L_g, W_g, 2C_g)$ and $\mathcal{F} = \mathcal{RN}(L_f, W_f, 2C_f)$ with $L_d, L_g \geq 3$, $W_g \geq 6d + 2dL_g$ and $W_f \geq 6\tilde{d} + 2dL_f$. Then there is a constant c, such that, if $m \geq cW_gL_g(\log W_g + L_g)$ and $n \geq cW_fL_f(\log W_f + L_f)$, such that,*

$$
\mathbb{E}\|\hat{\nu}_m - \nu\|_{\mathcal{D}\circ\mathcal{G}} \lesssim m^{-\beta/d} \vee m^{-1/2}(\log m)^{\{d=2\beta\}}
$$

$$+ \sqrt{\frac{1}{m} W_g L_g (\log W_g + L_g) \log(md)},$$

$$\mathbb{E}\|\hat{\mu}_n - \mu\|_{\tilde{\mathfrak{D}} \circ \mathfrak{F}} \lesssim n^{-\tilde{\beta}/\tilde{d}} \vee n^{-1/2} (\log n)^{\{\tilde{d} = 2\tilde{\beta}\}}$$

$$+ \sqrt{\frac{1}{n} W_f L_f (\log W_f + L_f) \log(n\tilde{d})}.$$

We bound the uniform deviations w.r.t. the function classes Φ_1 and Φ_2, i.e., $\|\hat{\mu}_n - \mu\|_{\Phi_1}$ and $\|\hat{\nu}_m - \nu\|_{\Phi_2}$ in Lemma 18. Since the function classes Φ_1 and Φ_2 are ReLU networks, we can control their metric entropies, which in turn helps us control the uniform concentration.

Lemma 18. *Suppose that $W_g + W_f \geq 2(d \vee \tilde{d})(3 + L_g + L_f)$, and $m, n \geq \theta_2 (W_g + W_f)(L_g + L_f)(\log(W_g + W_f) + L_g + L_f)$, for a constant θ_2 (that might depend on d and $\tilde{d}$). Then,*

$$\mathbb{E}\|\hat{\mu}_n - \mu\|_{\Phi_1} \lesssim \sqrt{\frac{WL(\log W + L)\log(nd)}{n}} \quad (8)$$

$$\mathbb{E}\|\hat{\nu}_m - \nu\|_{\Phi_2} \lesssim \sqrt{\frac{WL(\log W + L)\log(m\tilde{d})}{m}}, \quad (9)$$

where $W = W_f + W_g$ and $L = W_f + W_g$.

5.3 Proof Sketch of Theorem 5

In light of the results presented in Lemmata 11, 13, 17 and 18, the key insight lies in selecting the sizes of the networks in a manner that minimizes the right-hand side of Equation (5). By strategically determining the network architectures, we aim to strike a balance between model misspecification and stochastic errors, thus achieving an optimal trade-off that minimizes the expected excess risk. The idea is to exploit Lemma 11 to obtain a bound on the expected excess risk as:

$$\begin{aligned}
&\mathbb{E}V(\mu, \nu, \hat{G}, \hat{F}) \\
\lesssim& \epsilon_1^{\tilde{\alpha} \wedge \beta \wedge 1} + \epsilon_2^{\alpha \wedge \tilde{\beta} \wedge 1} + n^{-\frac{\beta}{d}} \vee n^{-1/2} (\log n)^{\tau_1} \\
&+ m^{-\frac{\tilde{\beta}}{\tilde{d}}} \vee m^{-1/2} (\log m)^{\tau_2} + m^{-\frac{\beta}{d}} \vee m^{-1/2} (\log m)^{\tau_1} \\
&+ n^{-\frac{\tilde{\beta}}{\tilde{d}}} \vee n^{-1/2} (\log n)^{\tau_2} \\
&+ \sqrt{\frac{\log n}{n}} \left(\sqrt{WL(\log W + L)} + \sqrt{W_f L_f (\log W_f + L_f)} \right) \\
&+ \sqrt{\frac{\log m}{m}} \left(\sqrt{WL(\log W + L)} + \sqrt{W_g L_g (\log W_f + L_f)} \right)
\end{aligned}$$

Here, $\tau_i = \ \{d_i = 2\beta_i\}$. Since the network depths and widths are expressible in terms of ϵ_1 and ϵ_2, the proof follows by setting these to terms that minimize the right-hand side of the above inequality. We refer the reader to supplement for a detailed proof.

6 Conclusions

In this paper, we have introduced a framework for analyzing the error rates associated with learning cross-domain generative models, specifically focusing on models such as CycleGANs and DiscoGANs. Our framework allows us to analyze the accuracy of translation and reconstruction guarantees, i.e., how well the push-forward distributions approximate the target distributions, and how well the cyclic map of generators maps back the data points close enough to the initial values. To achieve this, we have developed an oracle inequality that characterizes the excess risk by quantifying the misspecification and generalization errors associated with the problem. This allows us to establish a balance between model-misspecification and stochastic errors, enabling us to determine appropriate network architectures that optimize this tradeoff based on the number of available samples. In the process, we are able to determine the appropriate sizes of the generator networks in terms of the number of data points. Additionally, we show that the incorporation of the cycle loss is crucial in order to obtain cycle consistency for the model as without such a loss, the estimates might fail to be cycle-consistent even under regularity assumptions.

While our findings provide insights into the theoretical characteristics of CycleGANs, it is important to acknowledge that accurately estimating the complete error of these generative models in practical applications requires accounting for an additional optimization error term. However, accurately quantifying this term poses a formidable challenge due to the non-convex and intricate nature of the optimization process. Nevertheless, it is worth noting that our error analysis is independent of this optimization error and can be seamlessly integrated with analyses focusing on the optimization aspects. Another interesting direction would be to explore whether the rates can be sharpened under stronger assumptions of low intrinsic dimensionality of both the data distributions. The current form of the oracle inequality does not allow for this consideration since one does not have direct control over the intrinsic dimensionality of the generators. As opposed to vanilla GANs, the problem cannot be mitigated just by increasing the number of fake samples (Chakraborty and Bartlett, 2024). Further study in this direction would be an interesting direction for future research.

Acknowledgements

We gratefully acknowledge the support of the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS2031883 and #814639, the NSF's support of FODSI through grant DMS-2023505, and the support of the ONR through MURI award N000142112431.

References

Anthony, M. and Bartlett, P. (2009). *Neural Network Learning: Theoretical Foundations*. cambridge university press.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. (2017). Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pages 224–232. PMLR.

Arora, S., Risteski, A., and Zhang, Y. (2018). Do GANs learn the distribution? some theory and empirics. In *International Conference on Learning Representations*.

Asatryan, H., Gottschalk, H., Lippert, M., and Rottmann, M. (2020). A convenient infinite dimensional framework for generative adversarial learning. *arXiv preprint arXiv:2011.12087*.

Athreya, K. B. and Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. Springer Texts in Statistics. Springer New York, NY, 1 edition.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight VC-dimension and Pseudodimension Nounds for Piecewise Linear Neural Networks. *The Journal of Machine Learning Research*, 20(1):2285–2301.

Belomestny, D., Moulines, E., Naumov, A., Puchkin, N., and Samsonov, S. (2021). Rates of convergence for density estimation with gans. *arXiv preprint arXiv:2102.00199*.

Biau, G., Cadre, B., Sangnier, M., and Tanielian, U. (2020). Some theoretical properties of gans. *The Annals of Statistics*, 48(3):1539–1566.

Biau, G., Sangnier, M., and Tanielian, U. (2021). Some theoretical insights into wasserstein gans. *The Journal of Machine Learning Research*, 22(1):5287–5331.

Chakrabarty, A. and Das, S. (2022). On translation and reconstruction guarantees of the cycle-consistent generative adversarial networks. *Advances in Neural Information Processing Systems*, 35:23607–23620.

Chakraborty, S. and Bartlett, P. L. (2024). On the statistical properties of generative adversarial models for low intrinsic data dimension. *arXiv preprint arXiv:2401.15801*.

Chen, Q. and Koltun, V. (2017). Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Dahal, B., Havrilla, A., Chen, M., Zhao, T., and Liao, W. (2022). On deep generative models for approximation and estimation of distributions on manifolds. *Advances in Neural Information Processing Systems*, 35:10615–10628.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423.

Ge, C., Song, Y., Ge, Y., Yang, H., Liu, W., and Luo, P. (2021). Disentangled cycle consistency for highly-realistic virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16928–16937.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Han, X., Wu, Z., Wu, Z., Yu, R., and Davis, L. S. (2018). Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

Hou, X., Shen, L., Sun, K., and Qiu, G. (2017). Deep feature consistent variational autoencoder. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 1133–1141. IEEE.

Huang, J., Jiao, Y., Li, Z., Liu, S., Wang, Y., and Yang, Y. (2022). An error analysis of generative adversarial networks for learning distributions. *Journal of Machine Learning Research*, 23(116):1–43.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE con-*

ference on computer vision and pattern recognition, pages 1125–1134.

Jang, J., Lee, K., and Kim, T.-K. (2023). Unsupervised contour tracking of live cells by mechanical and cycle consistency losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 227–236.

Karr, A. F. (1993). *Probability*. Springer Texts in Statistics. Springer New York, NY, 1 edition.

Kim, J., Kim, M., Kang, H., and Lee, K. (2019). U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*.

Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR.

Kolmogorov, A. N. and Tikhomirov, V. M. (1961). ϵ-entropy and ϵ-capacity of sets in function spaces. *Translations of the American Mathematical Society*, 17:277–364.

Li, M., Huang, H., Ma, L., Liu, W., Zhang, T., and Jiang, Y. (2018). Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 184–199.

Liang, T. (2021). How Well Generative Adversarial Networks Learn Distributions. *The Journal of Machine Learning Research*, 22(1):10366–10406.

Liu, M.-Y., Breuel, T., and Kautz, J. (2017). Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30.

Lu, J., Shen, Z., Yang, H., and Zhang, S. (2021). Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506.

Luise, G., Pontil, M., and Ciliberto, C. (2020). Generalization properties of optimal transport gans with latent distribution learning. *arXiv preprint arXiv:2007.14641*.

Lv, J., Chen, W., Li, Q., and Yang, C. (2018). Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7948–7956.

Mahpod, S., Gaash, N., Hoffman, H., and Ben-Artzi, G. (2023). Ctrgan: Cycle transformers gan for gait transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 371–381.

Moriakov, N., Adler, J., and Teuwen, J. (2019). Kernel of cyclegan as a principal homogeneous space. In *International Conference on Learning Representations*.

Nakada, R. and Imaizumi, M. (2020). Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38.

Petersen, P. and Voigtlaender, F. (2018). Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. (2022). Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10.

Schreuder, N., Brunel, V.-E., and Dalalyan, A. (2021). Statistical guarantees for generative models without domination. In *Algorithmic Learning Theory*, pages 1051–1071. PMLR.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press.

Shen, Z., Yang, H., and Zhang, S. (2022). Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135.

Taigman, Y., Polyak, A., and Wolf, L. (2016). Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*.

Tiao, L. C., Bonilla, E. V., and Ramos, F. (2018). Cycle-consistent adversarial learning as approximate bayesian inference. *arXiv preprint arXiv:1806.01771*.

Torbunov, D., Huang, Y., Yu, H., Huang, J., Yoo, S., Lin, M., Viren, B., and Ren, Y. (2023). Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 702–712.

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

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

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620 – 2648.

Wu, S., Chen, Y., Mermet, S., Hurni, L., Schindler, K., Gonthier, N., and Landrieu, L. (2024). Stegogan: Leveraging steganography for non-bijective image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7922–7931.

Wu, Z., Xuan, H., Sun, C., Guan, W., Zhang, K., and Yan, Y. (2023). Semi-supervised video inpainting with cycle consistency constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22586–22595.

Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114.

Yi, Z., Zhang, H., Tan, P., and Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Contents

1 Introduction 1

2 Background 3
- 2.1 Related Works on GANs . . . 3
- 2.2 Notations . . . 3
- 2.3 An Overview of CycleGANs . . . 4

3 A Proof of Concept 5

4 Theoretical Analyses 5
- 4.1 Assumptions and Main Result . . . 5
- 4.2 Inference from the Main Result . . . 6
- 4.3 Translation Consistency . . . 7
- 4.4 Cycle Consistency . . . 7

5 Proof Overview of the Main Result 7
- 5.1 Approximation Error . . . 8
- 5.2 Generalization Bound . . . 8
- 5.3 Proof Sketch of Theorem 5 . . . 9

6 Conclusions 9

A Additional Notations 14

B Proofs from Section 4 14
- B.1 Proofs from Section 4.1 . . . 14
  - B.1.1 Proof of Lemma 4 . . . 14
  - B.1.2 Proof of Theorem 5 . . . 14
- B.2 Proofs from Section 4.2 . . . 15
  - B.2.1 Proof of Corollary 6 . . . 15
- B.3 Proofs from Section 4.3 . . . 15
  - B.3.1 Proof of Corollary 7 . . . 15
- B.4 Proofs from Section 4.4 . . . 16
  - B.4.1 Proof of Proposition 10 . . . 16

C Proofs from Section 5 16
- C.1 Proof of Lemma 11 . . . 16
- C.2 Proofs from Section 5.1 . . . 17
  - C.2.1 Proof of Lemma 12 . . . 17

C.2.2 Proof of Lemma 13 18
C.3 Proofs from Section 5.2 19
C.3.1 Proof of Lemma 15 19
C.3.2 Proof of Lemma 16 20
C.3.3 Proof of Lemma 17 20
C.3.4 Proof of Lemma 18 21
D Supporting Results from the Literature **22**

A Additional Notations

The major notations used in the paper are summarized in Table 1.

Notation	Meaning
$\mathcal{N}(\epsilon; A, \varrho)$	ϵ-covering number of the set A w.r.t. the metric ϱ
$\Vert P - Q \Vert_{\mathcal{F}}$	$\sup_{f \in \mathcal{F}} \lvert \int f dP - \int f dQ \rvert$
$\mathcal{L}(f)$	Depth of the network f
$\mathcal{W}(f)$	Number of weights of the network f
$\mathcal{B}(f)$	$\max_{1 \le j \le L}(\Vert b_j \Vert_\infty) \vee \Vert W_j \Vert_\infty$, the maximum absolute value of the weights and biases
$\mathcal{R}(f)$	$\sup_{x \in [0,1]^d} \Vert f(x) \Vert_\infty$, the maximum value of the output of the network
$X_{1:n}$	$\{X_1, \dots, X_n\}$, the dataset
$\mathcal{F}_{\lvert X_{1:m}}$	$\{[f(X_1) : \dots : f(X_m)] : f \in \mathcal{F}\} \subseteq \mathbb{R}^{d' \times m}$, where $f : \mathbb{R}^d \to \mathbb{R}^{d'}$
$\mathfrak{R}(X_{1:n}, \mathcal{F})$	$\frac{1}{n}\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i f(X_i)$ denotes the empirical Rademacher complexity.
$\text{Pdim}(\mathcal{F})$	Pseudo dimension of the function class $\mathcal{F}$. See Definition

Table 1: Table of Notations

B Proofs from Section 4

B.1 Proofs from Section 4.1

B.1.1 Proof of Lemma 4

Lemma 4. *Under assumption A2, $F^\star \circ G^\star(\cdot) = id(\cdot)$, a.e. $[\mu]$ and $G^\star \circ F^\star(\cdot) = id(\cdot)$, a.e. $[\nu]$.*

Proof. To show the first implication, it is enough to show that $\int \|x - F^\star \circ G^\star(x)\|_2^2 d\mu(x) = 0$. To observe this, we note that,

$$\int \|x - F^\star \circ G^\star(x)\|_2^2 d\mu(x) = \int \|F^\star(z) - F^\star(z)\|_2^2 d\nu(z) = 0.$$

Here the first equality follows from the assumption $(X, G^\star(X)) \overset{d}{=} (F^\star(Z), Z)$. The other implication follows similarly. □

B.1.2 Proof of Theorem 5

Theorem 5 (Main Theorem)**.** *Under assumptions A1–3, we can find an $n_0 \in \mathbb{N}$ (that might depend on d, $\tilde{d}, \alpha$, $\tilde{\alpha}$, β and $\tilde{\beta}$), such that if $n, m \ge n_0$, then,*

$$\mathbb{E}\,\mathfrak{R}(\hat{G}, \hat{F}) \precsim (n \wedge m)^{-\left(\max\left\{2 + \frac{\tilde{d}}{\alpha(\tilde{\alpha} \wedge \beta \wedge 1)}, 2 + \frac{d}{\tilde{\alpha}(\alpha \wedge \tilde{\beta} \wedge 1)}, \frac{d}{\beta}, \frac{\tilde{d}}{\tilde{\beta}}\right\}\right)^{-1}},$$

if $\mathcal{G} = \mathcal{RN}(L_g, W_g, 2C)$ and $\mathcal{F} = \mathcal{RN}(L_f, W_f, 2\tilde{C})$, where one can choose the network size as $L_g, L_f \lesssim \log(n \wedge m)$, $W_g \precsim (n \wedge m)^{\tilde{d}/(\tilde{d} + 2(\tilde{\alpha} \wedge \beta \wedge 1)\alpha)}$ and $W_f \precsim (n \wedge m)^{d/(d + 2(\alpha \wedge \tilde{\beta} \wedge 1)\tilde{\alpha})}$.

Proof. For notational simplicity, let $\tau_i = \quad \{d_i = 2\beta_i\}$. Form Lemma 11, we observe that,

$$\begin{aligned}
&\mathbb{E}V(\mu,\nu,\hat{G},\hat{F})\\
\lesssim &\inf_{G\in\mathcal{G},F\in\mathcal{F}} V(\mu,\nu,G,F) + \mathbb{E}\|\hat{\mu}-\mu\|_{\mathcal{D}} + \mathbb{E}\|\hat{\nu}-\nu\|_{\mathcal{D}\circ\mathcal{G}}\\
&+\mathbb{E}\|\hat{\nu}-\nu\|_{\tilde{\mathcal{D}}} + \mathbb{E}\|\hat{\mu}-\mu\|_{\tilde{\mathcal{D}}\circ\mathcal{F}} + \mathbb{E}\|\hat{\mu}-\mu\|_{\Phi_1} + \mathbb{E}\|\hat{\nu}-\nu\|_{\Phi_2}\\
\lesssim &\epsilon_1^{\tilde{\alpha}\wedge\beta\wedge 1} + \epsilon_2^{\alpha\wedge\tilde{\beta}\wedge 1} + n^{-\frac{\beta}{d}} \vee n^{-1/2}(\log n)^{\tau_1} + m^{-\frac{\tilde{\beta}}{\tilde{d}}} \vee m^{-1/2}(\log m)^{\tau_2} + m^{-\frac{\beta}{d}} \vee m^{-1/2}(\log m)^{\tau_1}\\
&+ n^{-\frac{\tilde{\beta}}{\tilde{d}}} \vee n^{-1/2}(\log n)^{\tau_2}\\
&+\sqrt{\frac{W_g L_g(\log W_g + L_g)\log m}{m}} + \sqrt{\frac{WL\,(\log W + L)\log n}{n}} + \sqrt{\frac{W_f L_f(\log W_f + L_f)\log n}{n}}\\
&+\sqrt{\frac{WL\,(\log W + L)\log m}{m}} \qquad (10)\\
\lesssim &\epsilon_1^{\tilde{\alpha}\wedge\beta\wedge 1} + \epsilon_2^{\alpha\wedge\tilde{\beta}\wedge 1} + (m\wedge n)^{-\frac{\beta}{d}} \vee (m\wedge n)^{-1/2}\left(\log(m\wedge n)\right)^{\tau_1} + (m\wedge n)^{-\frac{\tilde{\beta}}{\tilde{d}}} \vee (n\wedge m)^{-1/2}\left(\log(n\wedge m)\right)^{\tau_2}\\
&+\log^{3/2}\left(\frac{1}{\epsilon_1}\right)\frac{\epsilon_1^{-\frac{\tilde{d}}{2\alpha}}\log^{1/2} m}{m^{1/2}} + \log^{3/2}\left(\frac{1}{\epsilon_2}\right)\frac{\epsilon_2^{-\frac{d}{2\tilde{\alpha}}}\log^{1/2} n}{n^{1/2}} + \log^{3/2}\left(\frac{1}{\epsilon_1\wedge\epsilon_2}\right)\sqrt{\epsilon_1^{-\frac{\tilde{d}}{\alpha}} + \epsilon_2^{-\frac{d}{\tilde{\alpha}}}}\left(\frac{\log(n\wedge m)}{n\wedge m}\right)^{1/2}\\
\precsim &(n\wedge m)^{-\frac{\beta}{d}\vee\frac{\tilde{\beta}}{\tilde{d}}} + (n\wedge m)^{-\frac{1}{2+\max\left\{\frac{\tilde{d}}{\alpha(\tilde{\alpha}\wedge\beta\wedge 1)},\frac{d}{\tilde{\alpha}(\alpha\wedge\tilde{\beta}\wedge 1)}\right\}}} \qquad (11)\\
\precsim &(n\wedge m)^{-\frac{1}{\max\left\{2+\frac{\tilde{d}}{\alpha(\tilde{\alpha}\wedge\beta\wedge 1)},2+\frac{d}{\tilde{\alpha}(\alpha\wedge\tilde{\beta}\wedge 1)},\frac{d}{\beta},\frac{\tilde{d}}{\tilde{\beta}}\right\}}}.
\end{aligned}$$

Here, (10) follows from applying Lemmata 13, 17 and 18. Inequality (11) follows from taking $\epsilon_1 \asymp (n\wedge m)^{-\frac{1}{2(\beta\wedge 1)+\frac{\tilde{d}}{\alpha}}}$ and $\epsilon_2 \asymp (n\wedge m)^{-\frac{1}{2(\tilde{\beta}\wedge 1)+\frac{d}{\tilde{\alpha}}}}$. □

B.2 Proofs from Section 4.2

B.2.1 Proof of Corollary 6

Corollary 6. *Suppose that* $X_1,\dots X_n \overset{i.i.d.}{\sim} \mu'$ *and* $\tilde{X}_1,\dots,\tilde{X}_m \overset{i.i.d.}{\sim} \nu'$, *such that,* $\mathrm{TV}(\mu,\mu') + \mathrm{TV}(\nu,\nu') \le \Delta$. *Under the assumptions and choices of Theorem 5, if* $n, m \ge n_0$, *then,*

$$\begin{aligned}
&\mathbb{E}\,\mathfrak{R}(\hat{G},\hat{F})\\
&\precsim \Delta + (n\wedge m)^{-\left(\max\left\{2+\frac{\tilde{d}}{\alpha(\tilde{\alpha}\wedge\beta\wedge 1)},2+\frac{d}{\tilde{\alpha}(\alpha\wedge\tilde{\beta}\wedge 1)},\frac{d}{\beta},\frac{\tilde{d}}{\tilde{\beta}}\right\}\right)^{-1}}
\end{aligned}$$

Proof. The result follows by noting that, $|V(\mu',\nu',\hat{G},\hat{F}) - V(\mu,\nu,\hat{G},\hat{F})| \lesssim \Delta$ and repeating the proof of Theorem 5. □

B.3 Proofs from Section 4.3

B.3.1 Proof of Corollary 7

Corollary 7. *Under Assumptions A1–3,* $\mathfrak{R}(\hat{G},\hat{F}) \xrightarrow{a.s.} 0$, *as* $n\wedge m\to\infty$.

Proof. We note that $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are bounded function classes. Thus, a simple application of the bounded difference inequality yields that with probability at least $1-\delta/2$,

$$\begin{aligned}
\|\hat{\mu}_n-\mu\|_{\mathcal{D}} + \|\hat{\mu}_n-\mu\|_{\mathcal{D}\circ\mathcal{G}} + \|\hat{\mu}_n-\mu\|_{\tilde{\mathcal{D}}\circ\mathcal{F}} \le &\mathbb{E}\left(\|\hat{\mu}_n-\mu\|_{\mathcal{D}} + \|\hat{\mu}_n-\mu\|_{\mathcal{D}\circ\mathcal{G}} + \|\hat{\mu}_n-\mu\|_{\tilde{\mathcal{D}}\circ\mathcal{F}}\right)\\
&+\theta_1\sqrt{\frac{\log(1/\delta)}{n}}\\
\|\hat{\nu}_m-\nu\|_{\mathcal{D}} + \|\hat{\nu}_m-\nu\|_{\mathcal{D}\circ\mathcal{G}} + \|\hat{\nu}_m-\nu\|_{\tilde{\mathcal{D}}\circ\mathcal{F}} \le &\mathbb{E}\left(\|\hat{\nu}_m-\nu\|_{\mathcal{D}} + \|\hat{\nu}_m-\nu\|_{\mathcal{D}\circ\mathcal{G}} + \|\hat{\nu}_m-\nu\|_{\tilde{\mathcal{D}}\circ\mathcal{F}}\right)\\
&+\theta_1\sqrt{\frac{\log(1/\delta)}{m}},
\end{aligned}$$

for some positive constant θ_1, by a simple application of the bounded difference inequality. Thus, with probability at least $1-\delta$,

$$V(\mu,\nu,\hat{G},\hat{F}) \lesssim (n \wedge m)^{-\frac{1}{\max\left\{2+\frac{\tilde{d}}{\alpha(\tilde{\alpha}\wedge\beta\wedge 1)},2+\frac{d}{\tilde{\alpha}(\alpha\wedge\tilde{\beta}\wedge 1)},\frac{d}{\beta},\frac{\tilde{d}}{\tilde{\beta}}\right\}}} + \theta_2\sqrt{\frac{\log(1/\delta)}{n \wedge m}},$$

for some positive constant θ_2. From the above, $\mathbb{P}(V(\mu,\nu,\hat{G},\hat{F}) > \epsilon) \le e^{-\frac{n\wedge m\epsilon^2}{\theta_3}}$. This implies that $\sum_{(n\wedge m)\ge 1}\mathbb{P}(V(\mu,\nu,\hat{G},\hat{F}) > \epsilon) < \infty$. A simple application of the first Borel-Cantelli Lemma yields (see Proposition 5.7 of Karr (1993)) that this implies that $V(\mu,\nu,\hat{G},\hat{E}) \to 0$, almost surely. □

B.4 Proofs from Section 4.4

B.4.1 Proof of Proposition 10

Proposition 10. *The CycleGAN estimates $\hat{G}$ and $\hat{F}$ defined in (4) are cycle-consistent.*

Proof. Since,

$$0 \le \int \|x - \hat{G}\circ\hat{F}(x)\|_1 d\mu(x),\ \int \|\tilde{x} - \hat{F}\circ\hat{G}(\tilde{x})\|_1 d\nu(\tilde{x}) \lesssim \mathfrak{R}(\hat{G},\hat{F}),$$

it is clear that

$$\int \|x - \hat{G}\circ\hat{F}(x)\|_1 d\mu(x),\ \int \|\tilde{x} - \hat{F}\circ\hat{G}(\tilde{x})\|_1 \xrightarrow{a.s.} 0.$$

Thus, for any such ω in the sample space, for which both the convergence holds,

$$\|X - \hat{G}_\omega\circ\hat{F}_\omega(X)\|_1 \xrightarrow{\ell_1(\mu)} 0 \text{ and } \|\tilde{X} - \hat{F}_\omega\circ\hat{G}_\omega(\tilde{X})\|_1 \xrightarrow{\ell_1(\nu)} 0,$$

where, $X \sim \mu$ and $\tilde{X} \sim \nu$. Here we use the subscript ω to denote that the estimated generators are dependent on the point of the sample space. Thus,

$$\begin{aligned}\|\mu - (\hat{G}_\omega\circ\hat{F}_\omega)_\sharp\mu\|_{\mathcal{D}} = \sup_{D\in\mathcal{D}}\int \Big(D(x) - (D\circ\hat{G}_\omega\circ\hat{F}_\omega)(x)\Big)d\mu(x) &\le \int \|x - (\hat{G}_\omega\circ\hat{F}_\omega)(x)\|_2^{\beta\wedge 1}d\mu(x)\\ &\le \left(\int \|x - (\hat{G}_\omega\circ\hat{F}_\omega)(x)\|_2 d\mu(x)\right)^{\beta\wedge 1} \qquad (12)\\ &\lesssim \left(\int \|x - (\hat{G}_\omega\circ\hat{F}_\omega)(x)\|_1 d\mu(x)\right)^{\beta\wedge 1} \to 0,\end{aligned}$$

as $m \wedge n \to 0$. In the above calculations, (12) follows from Jensen's inequality (Athreya and Lahiri, 2006, Proposition 6.2.6). Thus, $\|\mu - (\hat{G}\circ\hat{F})_\sharp\mu\|_{\mathcal{D}} \xrightarrow{a.s.} 0$. Similarly, $\|\nu - (\hat{F}\circ\hat{G})_\sharp\nu\|_{\tilde{\mathcal{D}}} \xrightarrow{a.s.} 0$. This proves that the generator estimates are cycle-consistent. □

C Proofs from Section 5

C.1 Proof of Lemma 11

Lemma 11 (Oracle Inequality)**.** *For the estimates $\hat{G}$ and $\hat{F}$,*

$$\begin{aligned}&V(\mu,\nu,\hat{G},\hat{F})\\ \le &\inf_{G\in\mathcal{G},F\in\mathcal{F}} V(\mu,\nu,G,F) + 2\|\hat{\mu}_n - \mu\|_{\mathcal{D}} + 2\|\hat{\nu}_m - \nu\|_{\mathcal{D}\circ\mathcal{G}}\\ &+ 2\|\hat{\nu}_m - \nu\|_{\tilde{\mathcal{D}}} + 2\|\hat{\mu}_n - \mu\|_{\tilde{\mathcal{D}}\circ\mathcal{F}} + 2\lambda\|\hat{\mu}_n - \mu\|_{\Phi_1}\\ &+ 2\lambda\|\hat{\nu}_m - \nu\|_{\Phi_2}, \qquad (5)\end{aligned}$$

where, $\Phi_1 = \{c(\cdot, G\circ F(\cdot)) : G\in\mathcal{G},\ F\in\mathcal{F}\}$ and $\Phi_2 = \{\tilde{c}(\cdot, F\circ G(\cdot)) : G\in\mathcal{G},\ F\in\mathcal{F}\}$.

Proof. With a simple application of triangle inequality, we note the following:

$$
\begin{aligned}
& V(\mu, \nu, \hat{G}, \hat{F}) \\
= & \|\mu - \hat{G}_\sharp \nu\|_{\mathscr{D}} + \|\nu - \hat{F}_\sharp \mu\|_{\tilde{\mathscr{D}}} + \lambda \int c(x, G \circ F(x)) d\mu(x) + \lambda \int \tilde{c}(\tilde{x}, F \circ G(\tilde{x})) d\nu(\tilde{x}) \\
\le & V(\hat{\mu}, \hat{\nu}, \hat{G}, \hat{F}) + \|\hat{\mu} - \mu\|_{\mathscr{D}} + \|\hat{\nu} - \nu\|_{\mathscr{D} \circ \mathscr{G}} + \|\hat{\nu} - \nu\|_{\tilde{\mathscr{D}}} + \|\hat{\mu} - \mu\|_{\tilde{\mathscr{D}} \circ \mathscr{F}} + \lambda \|\hat{\mu} - \mu\|_{\Phi_1} + \lambda \|\hat{\nu} - \nu\|_{\Phi_2} \\
\le & V(\hat{\mu}, \hat{\nu}, G, F) + \|\hat{\mu} - \mu\|_{\mathscr{D}} + \|\hat{\nu} - \nu\|_{\mathscr{D} \circ \mathscr{G}} + \|\hat{\nu} - \nu\|_{\tilde{\mathscr{D}}} + \|\hat{\mu} - \mu\|_{\tilde{\mathscr{D}} \circ \mathscr{F}} + \lambda \|\hat{\mu} - \mu\|_{\Phi_1} + \lambda \|\hat{\nu} - \nu\|_{\Phi_2} \\
\le & V(\mu, \nu, G, F) + 2\|\hat{\mu} - \mu\|_{\mathscr{D}} + 2\|\hat{\nu} - \nu\|_{\mathscr{D} \circ \mathscr{G}} + 2\|\hat{\nu} - \nu\|_{\tilde{\mathscr{D}}} + 2\|\hat{\mu} - \mu\|_{\tilde{\mathscr{D}} \circ \mathscr{F}} + 2\lambda \|\hat{\mu} - \mu\|_{\Phi_1} + 2\lambda \|\hat{\nu} - \nu\|_{\Phi_2}.
\end{aligned}
$$

Taking infimum on both sides, w.r.t. G and F gives us the desired result. □

C.2 Proofs from Section 5.1

C.2.1 Proof of Lemma 12

Lemma 12. *Suppose that $f \in \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$, for some $C > 0$. Then, we can find a constant α, that might depend on β, d and C, such that, for any $\epsilon \in (0, 1)$, there exists a ReLU network, $\hat{f}$ with $\mathcal{L}(\hat{f}) \le \alpha \log(1/\epsilon)$, $\mathcal{W}(\hat{f}) \le \alpha \log(1/\epsilon)\epsilon^{-d/\beta}$, $\mathcal{B}(\hat{f}) \le \alpha \epsilon^{-1/\beta}$ and $\mathcal{R}(\hat{f}) \le 2C$, that satisfies, $\|f - \hat{f}\|_{\ell_\infty([0,1]^d)} \le \epsilon$.*

Proof. The proof can be done by replicating the proof of Theorem 18 of Chakraborty and Bartlett (2024). The proof is provided here for completeness. Let $K = \lceil \frac{1}{\epsilon} \rceil$. For any $\boldsymbol{i} \in [K]^d$, let $\theta^{\boldsymbol{i}} = (i_1 \epsilon, \dots, i_d \epsilon)$. For $0 < b \le a$, let,

$$
\xi_{a,b}(x) = \mathrm{ReLU}\left(\frac{x+a}{a-b}\right) - \mathrm{ReLU}\left(\frac{x+b}{a-b}\right) - \mathrm{ReLU}\left(\frac{x-b}{a-b}\right) + \mathrm{ReLU}\left(\frac{x-a}{a-b}\right).
$$

Thus, $\mathcal{L}(\xi_{a,b}) = 2$ and $\mathcal{W}(\xi_{a,b}) = 12$. Suppose that $\delta = \epsilon/3$ and let, $\zeta(x) = \prod_{\ell=1}^d \xi_{\epsilon+\delta,\delta}(x_\ell)$. Clearly, $\mathcal{B}(\xi_{\epsilon+\delta,\delta}) \le \frac{1}{\delta}$. It is easy to observe that $\{\zeta(\cdot - \theta^{\boldsymbol{i}}) : \boldsymbol{i} \in [K]^d\}$ forms a partition of unity on $[0,1]^d$, i.e. $\sum_{\boldsymbol{i} \in [K]^d} \zeta(x - \theta^{\boldsymbol{i}}) = 1, \forall x \in [0,1]^d$.

We consider the Taylor approximation of f around θ as,

$$
P_\theta(x) = \sum_{|\boldsymbol{s}| \le \lfloor \beta \rfloor} \frac{\partial^{\boldsymbol{s}} f(\theta)}{\boldsymbol{s}!} (x - \theta)^{\boldsymbol{s}}.
$$

Note that for any $x \in [0,1]^d$, $f(x) - P_\theta(x) = \sum_{\boldsymbol{s}:|\boldsymbol{s}| = \lfloor \beta \rfloor} \frac{(x-\theta)^{\boldsymbol{s}}}{\boldsymbol{s}!} (\partial^{\boldsymbol{s}} f(y) - \partial^{\boldsymbol{s}} f(\theta))$, for some y, which is a convex combination of x and θ. Thus,

$$
\begin{aligned}
f(x) - P_\theta(x) = & \sum_{\boldsymbol{s}:|\boldsymbol{s}| = \lfloor \beta \rfloor} \frac{(x-\theta)^{\boldsymbol{s}}}{\boldsymbol{s}!} (\partial^{\boldsymbol{s}} f(y) - \partial^{\boldsymbol{s}} f(\theta)) \\
\le & \|x - \theta\|_\infty^{\lfloor \beta \rfloor} \sum_{\boldsymbol{s}:|\boldsymbol{s}| = \lfloor \beta \rfloor} \frac{1}{\boldsymbol{s}!} |\partial^{\boldsymbol{s}} f(y) - \partial^{\boldsymbol{s}} f(\theta)| \\
\le & 2C \|x - \theta\|_\infty^{\lfloor \beta \rfloor} \|y - \theta\|_\infty^{\beta - \lfloor \beta \rfloor} \\
\le & 2C \|x - \theta\|_\infty^\beta.
\end{aligned}
\tag{13}
$$

Next we define $\tilde{f}(x) = \sum_{\boldsymbol{i} \in [K]^d} \zeta(x - \theta^{\boldsymbol{i}}) P_{\theta^{\boldsymbol{i}}}(x)$. Thus, if $x \in [0,1]^d$,

$$
\begin{aligned}
|f(x) - \tilde{f}(x)| = \left| \sum_{\boldsymbol{i} \in [K]^d} \zeta(x - \theta^{\boldsymbol{i}})(f(x) - P_{\theta^{\boldsymbol{i}}}(x)) \right| \le & \sum_{\boldsymbol{i} \in [K]^d : \|x - \theta^{\boldsymbol{i}}\|_\infty \le 2\epsilon} |f(x) - P_{\theta^{\boldsymbol{i}}}(x)| \\
\le & C 2^{d+1} (2\epsilon)^\beta \\
= & C 2^{d+\beta+1} \epsilon^\beta.
\end{aligned}
\tag{14}
$$

We note that,

$$
\tilde{f}(x) = \sum_{\boldsymbol{i} \in [K]^d} \zeta(x - \theta^{\boldsymbol{i}}) P_{\theta^{\boldsymbol{i}}}(x) = \sum_{\boldsymbol{i} \in [K]^d} \sum_{|\boldsymbol{s}| \le \lfloor \beta \rfloor} \frac{\partial^{\boldsymbol{s}} f(\theta^{\boldsymbol{i}})}{\boldsymbol{s}!} \zeta(x - \theta^{\boldsymbol{i}}) \left(x - \theta^{\boldsymbol{i}}\right)^{\boldsymbol{s}}.
$$

Let $a_{\boldsymbol{i},\boldsymbol{s}} = \frac{\partial^{\boldsymbol{s}} f(\theta^{\boldsymbol{i}})}{\boldsymbol{s}!}$ and

$$\begin{aligned}&\hat{f}_{\boldsymbol{i},\boldsymbol{s}}(x)\\ =&\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}(\xi_{\epsilon_1,\delta_1}(x_1-\theta^{\boldsymbol{i}}_1),\dots,\xi_{\epsilon_d,\delta_d}(x_d-\theta^{\boldsymbol{i}}_d),\underbrace{(x_1-\theta^{\boldsymbol{i}}_1),\dots,(x_1-\theta^{\boldsymbol{i}}_1)}_{s_1\text{ times}},\dots,\underbrace{(x_1-\theta^{\boldsymbol{i}}_d),\dots,(x_d-\theta^{\boldsymbol{i}}_d)}_{s_d\text{ times}}),\end{aligned}$$

where $\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}(\cdot)$ is defined in Lemma 27. We note that $\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}$ has at most $d+|\boldsymbol{s}| \le d+\lfloor\beta\rfloor$ many inputs. By Lemma 27, $\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}$ can be implemented by a ReLU network with $\mathcal{L}(\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}), \mathcal{W}(\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}) \le c_3 m$ and $\mathcal{B}(\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}) \le 4 \vee (d-1)^2$. Thus, $\mathcal{L}(\hat{f}_{\boldsymbol{i},\boldsymbol{s}}) \le c_3 m + 2$ and $\mathcal{W}(\hat{f}_{\boldsymbol{i},\boldsymbol{s}}) \le c_3 m + 8d + 4|s| \le c_3 m + 8d + 4\lfloor\beta\rfloor$. Furthermore, $\mathcal{B}(\hat{f}_{\boldsymbol{i},\boldsymbol{s}}) \le 4 \vee \frac{1}{\delta} \le 1/\delta$, when δ is small enough. With this $\hat{f}_{\boldsymbol{i},\boldsymbol{s}}$, we observe from Lemma 27 that,

$$\left|\hat{f}_{\boldsymbol{i},\boldsymbol{s}}(x) - \zeta(x-\theta^{\boldsymbol{i}})\left(x-\theta^{\boldsymbol{i}}\right)^{\boldsymbol{s}}\right| \le \frac{d+\lfloor\beta\rfloor}{2^{2m-1}}, \forall x \in S. \tag{15}$$

Finally, let, $\hat{f}(x) = \sum_{\boldsymbol{i}\in[K]^d}\sum_{|\boldsymbol{s}|\le\lfloor\beta\rfloor} a_{\boldsymbol{i},\boldsymbol{s}}\hat{f}_{\boldsymbol{i},\boldsymbol{s}}(x)$. Clearly, $\mathcal{L}(\hat{f}_{\boldsymbol{i},\boldsymbol{s}}) \le c_3 m + 3$ and $\mathcal{W}(\hat{f}_{\boldsymbol{i},\boldsymbol{s}}) \le \lfloor\beta\rfloor^d\left(c_3 m + 8d + 4\lfloor\beta\rfloor\right)$. This implies that,

$$\begin{aligned}|\hat{f}(x)-\tilde{f}(x)| \le & \sum_{\boldsymbol{i}\in[K]^d:\|x-\theta^{\boldsymbol{i}}\|_\infty\le 2\epsilon}\ \sum_{|\boldsymbol{s}|\le\lfloor\beta\rfloor}|a_{\boldsymbol{i},\boldsymbol{s}}|\zeta(x-\theta^{\boldsymbol{i}})|\hat{f}_{\boldsymbol{i}\boldsymbol{s}}(x)-\left(x-\theta^{\boldsymbol{i}}\right)^{\boldsymbol{s}}|\\ \le & 2^d\sum_{|\boldsymbol{s}|\le\lfloor\beta\lfloor\beta\rfloor}|a_{\theta,\boldsymbol{s}}|\left|\hat{f}_{\theta^{\boldsymbol{i}(x)},\boldsymbol{s}}(x)-\zeta_{\boldsymbol{\epsilon},\boldsymbol{\delta}}(x-\theta^{(\boldsymbol{i}(x))})\left(x-\theta^{\boldsymbol{i}(x)}\right)^{\boldsymbol{s}}\right|\\ \le & \frac{(d+\lfloor\beta\rfloor)C}{2^{2m-d-1}}.\end{aligned}$$

We thus get that if $x \in [0,1]^d$,

$$|f(x)-\hat{f}(x)| \le |f(x)-\tilde{f}(x)| + |\hat{f}(x)-\tilde{f}(x)| \le C2^{d+\beta+1}\epsilon^\beta + \frac{(d+\lfloor\beta\rfloor)C}{2^{2m-d-1}}. \tag{16}$$

Taking $\epsilon \asymp \eta^{1/\beta}$ and $m \asymp \log(1/\eta)$ ensures that $\|f-\hat{f}\|_{\ell_\infty([0,1]^d)} \le \eta$. We note that $\hat{f}$ has $\mathcal{N}(\epsilon; S, \ell_\infty) \le \epsilon^{-d}$ many networks with depth $c_3 m + 3$ and number of weights $\lfloor\beta\rfloor^d\left(c_3 m + 8d + 4\lfloor\beta\rfloor\right)$. Thus, $\mathcal{L}(\hat{f}) \le c_3 m + 4$ and $\mathcal{W}(\hat{f}) \le \epsilon^{-d}(6\lfloor\beta\rfloor)^d\left(c_3 m + 8d + 4\lfloor\beta\rfloor\right)$. we thus get,

$$\mathcal{L}(\hat{f}) \le c_3 m + 4 \le c_4 \log\left(\frac{1}{\eta}\right),$$

where c_4 is a function of δ, $\lfloor\beta\rfloor$ and d. Similarly,

$$\mathcal{W}(\hat{f}) \le \epsilon^{-d}(6\lfloor\beta\rfloor)^d\left(c_3 m + 8d + 4\lfloor\beta\rfloor\right) \le c_6\log(1/\eta)\eta^{-d/\beta}.$$

Taking $\alpha = c_4 \vee c_6$ gives the result. Furthermore, by construction, we note that,

$$\mathcal{B}(\hat{f}) \lesssim 1/\delta = 3/\epsilon \lesssim \eta^{-1/\beta}.$$

□

C.2.2 Proof of Lemma 13

Lemma 13. *Suppose assumption A2 holds. Then, for any $\epsilon, \epsilon_2 \in (0,1)$, we can find networks with ReLU activation, $\mathcal{G} = \mathcal{RN}(L_g, W_g)$ and $\mathcal{F} = \mathcal{RN}(L_f, W_f)$, with $L_g \asymp \log(1/\epsilon_1)$, $W_g \asymp \epsilon_1^{-\tilde{d}/\alpha}\log(1/\epsilon_1)$, $L_f \asymp \log(1/\epsilon_2)$ and $W_f \asymp \epsilon_2^{-d/\tilde{\alpha}}\log(1/\epsilon_2)$, such that $V(\mu,\nu,G,F) \lesssim \epsilon_1^{\tilde{\alpha}\wedge\beta\wedge 1} + \epsilon_2^{\alpha\wedge\tilde{\beta}\wedge 1}$.*

Proof.

$$\mathbb{E}c(X, G\circ F(X)) = \int c(G^\star\circ F^\star(x), G\circ F(x))d\mu(x)$$

$$
\begin{aligned}
&\lesssim \int \|G^\star \circ F^\star(x) - G \circ F(x)\|_2 d\mu(x) \\
&\leq \int \|G^\star \circ F^\star(x) - G^\star \circ F(x)\|_2 d\mu(x) + \int \|G^\star \circ F(x) - G \circ F(x)\|_2 d\mu(x) \\
&\leq \int \|F^\star(x) - F(x)\|_2^{\alpha \wedge 1} d\mu(x) + \|G^\star - G\|_{\ell_\infty(\tilde{\mathcal{X}})} \\
&\leq \|F^\star - F\|^{\alpha \wedge 1}_{\ell_\infty(\mathcal{X})} + \|G^\star - G\|_{\ell_\infty(\tilde{\mathcal{X}})} \qquad (17)
\end{aligned}
$$

Similarly,

$$\mathbb{E}\tilde{c}(\tilde{X}, F \circ G(\tilde{X})) \lesssim \|G^\star - G\|^{\tilde{\alpha} \wedge 1}_{\ell_\infty(\tilde{\mathcal{X}})} + \|F^\star - F\|_{\ell_\infty(\mathcal{X})}. \qquad (18)$$

We also note that,

$$\|\mu - G_\sharp \nu\|_{\mathscr{D}} = \|G^\star_\sharp \nu - G_\sharp \nu\|_{\mathscr{D}} \leq \|G^\star - G\|^{\beta \wedge 1}_{\ell_\infty(\tilde{\mathcal{X}})} \qquad (19)$$

and

$$\|\nu - F_\sharp \mu\|_{\mathscr{D}} = \|F^\star_\sharp \mu - F_\sharp \mu\|_{\tilde{\mathscr{D}}} \leq \|F^\star - F\|^{\tilde{\beta} \wedge 1}_{\ell_\infty(\mathcal{X})} \qquad (20)$$

From equations (17)-(20), we note that,

$$
\begin{aligned}
V(\mu, \nu, G, F) &\lesssim \|G^\star - G\|^{\tilde{\alpha} \wedge \beta \wedge 1}_{\ell_\infty(\tilde{\mathcal{X}})} + \|F^\star - F\|^{\alpha \wedge \tilde{\beta} \wedge 1}_{\ell_\infty(\mathcal{X})} + \|G^\star - G\|_{\ell_\infty(\tilde{\mathcal{X}})} + \|F^\star - F\|_{\ell_\infty(\mathcal{X})} \\
&\lesssim \epsilon^{\tilde{\alpha} \wedge \beta \wedge 1} + \epsilon_2^{\alpha \wedge \tilde{\beta} \wedge 1}
\end{aligned}
$$

□

C.3 Proofs from Section 5.2

The approach involves leveraging Lemma 19 from Kolmogorov and Tikhomirov (1961) to control the metric entropies of these function classes. Subsequently, Dudley's chaining is employed to effectively control the expected differences between the empirical and actual distributions with respect to the $\mathscr{D}$ and $\tilde{\mathscr{D}}$ IPMs.

Lemma 19 (Kolmogorov and Tikhomirov (1961))**.** *The ϵ-covering number of $\mathscr{H}^\beta([0,1]^d, \mathbb{R}, 1)$ can be bounded as,*

$$\log \mathcal{N}\left((\epsilon; \mathscr{H}^\beta([0,1]^d, \mathbb{R}, 1), \|\cdot\|_\infty\right) \lesssim \epsilon^{-d/\beta}.$$

C.3.1 Proof of Lemma 15

Lemma 15. *Under Assumption A1, the followings hold:*

$$
\begin{aligned}
\|\hat{\mu}_n - \mu\|_{\mathscr{D}} &\lesssim n^{-\beta/d} \vee n^{-1/2} (\log n)^{\ \{d = 2\beta\}}, \\
\|\hat{\nu}_m - \nu\|_{\tilde{\mathscr{D}}} &\lesssim m^{-\tilde{\beta}/\tilde{d}} \vee m^{-1/2} (\log m)^{\ \{\tilde{d} = 2\tilde{\beta}\}}.
\end{aligned}
$$

Proof. From Dudley's chaining, we recall that,

$$
\begin{aligned}
\|\hat{\mu}_n - \mu\|_{\mathscr{D}} &\lesssim \inf_{\delta \in (0,1)} \left(\delta + \int_\delta^1 \sqrt{\frac{1}{n} \log \mathcal{N}(\epsilon; \mathscr{H}^\beta([0,1]^d, \mathbb{R}, 1, \|\cdot\|_\infty)} d\epsilon\right) \\
&\lesssim n^{-\beta/d} \vee n^{-1/2} (\log n)^{\ \{d = 2\beta\}},
\end{aligned}
$$

where the last inequality follows from Lemma 21. This proves the first inequality. The latter result also follows from a similar calculation. □

C.3.2 Proof of Lemma 16

We state and prove the following result, which will help prove Lemma 16.

Lemma 20. *Suppose that $Z_1, \dots, Z_n \in \mathbb{R}^d$ are independent and identically distributed according to the law γ and suppose that $\hat{\gamma}_n$ be the empirical distribution of ν. Then,*

$$\mathbb{E}\|\hat{\gamma}_n - \gamma\|_{\mathcal{H}^\beta([0,1]^d)} \lesssim (n^{-\beta/d} \vee n^{-1/2})(\log n)^{\{d=2\beta\}}$$

Proof. By Dudley's chaining argument, it is easy to observe that

$$\begin{aligned}
\mathbb{E}\|\hat{\gamma}_n - \gamma\|_{\mathcal{H}^\beta([0,1]^d)} \lesssim & \inf_{0<\delta\leq B} \left(\delta + \frac{1}{\sqrt{n}} \int_\delta^{1/2} \sqrt{\log \mathcal{N}\left((\epsilon; \mathcal{H}^\beta([0,1]^d), \|\cdot\|_\infty\right) d\epsilon}\right) \\
\lesssim & \inf_{0<\delta\leq B} \left(\delta + \frac{1}{\sqrt{n}} \int_\delta^{1/2} \epsilon^{-\frac{d}{2\beta}} d\epsilon\right) && (21) \\
\lesssim & (n^{-\beta/d} \vee n^{-1/2})(\log n)^{\{d=2\beta\}}. && (22)
\end{aligned}$$

Here, (21) follows from Lemma 19 and (22) follows from Lemma 21. □

Lemma 16. *Suppose that $\mathcal{G} = \mathcal{RN}(L_g, W_g, 2C_g)$ and $\mathcal{F} = \mathcal{RN}(L_f, W_f, 2C_f)$, then, there exists a constant c, such that, if $m \geq cW_gL_g(\log W_g + L_g)$ and $n \geq cW_fL_f(\log W_f + L_f)$,*

$$\log N_1 \lesssim \epsilon^{-d/\beta} + W_gL_g(\log W_g + L_g)\log(md/\epsilon), \tag{6}$$

$$\log N_2 \lesssim \epsilon^{-\tilde{d}/\tilde{\beta}} + W_fL_f(\log W_f + L_f)\log(n\tilde{d}/\epsilon), \tag{7}$$

where $N_1 = \mathcal{N}\left(\epsilon; (\mathcal{D} \circ \mathcal{G})_{|\tilde{X}_{1:m}}, \|\cdot\|_\infty\right)$ and $N_2 = \mathcal{N}\left(\epsilon; (\tilde{\mathcal{D}} \circ \mathcal{F})_{|X_{1:n}}, \|\cdot\|_\infty\right)$.

Proof. We only give a proof of inequality (6) of the main paper; the proof of inequality (7) can be derived similarly. Let $\mathcal{V} = \{v_1, \dots, v_r\}$ be an optimal ℓ_∞ ϵ-cover of $\mathcal{G}_{|\tilde{X}_{1:m}}$. Clearly, $\log r \lesssim W_gL_g(\log W_g + L_g)\log\left(\frac{md}{\epsilon}\right)$, by Lemma 24. Suppose that $\mathcal{D}_1^\delta = \{f_1, \dots, f_k\}$ be an optimal ℓ_∞ δ-cover of $\mathcal{D}$. By Lemma 19, we know that, $\log k \leq \delta^{-\frac{d}{\beta}}$. We note that, for any $f \in \mathcal{D}$, we can find $\tilde{f} \in \mathcal{D}_1^\delta$, such that, $\|f - \tilde{f}\|_\infty \leq \delta$. For any $\tilde{X} = \tilde{X}_{1:m}$, let $\|\tilde{X} - v^{\tilde{X}}\|_\infty \leq \epsilon$, with $v^{\tilde{X}} \in \mathcal{V}$.

$$|f(\tilde{X}_j) - \tilde{f}(v_j^{\tilde{X}})| \leq |f(\tilde{X}_j) - f(v_j^{\tilde{X}})| + |f(v_j^{\tilde{X}}) - \tilde{f}(v_j^{\tilde{X}})| \lesssim \|\tilde{X}_j - v_j^{\tilde{X}}\|_\infty^{\beta\wedge 1} + \|f - \tilde{f}\|_\infty \lesssim \epsilon^{\beta\wedge 1} + \delta.$$

Taking $\delta \asymp \eta/2$ and $\epsilon \asymp \left(\frac{\eta}{2}\right)^{1/(\beta\wedge 1)}$, we conclude that $\max_{1\leq j\leq m} |f(Z_j) - \tilde{f}(v_j^Z)| \leq \eta$. Thus, $\{(\tilde{f}(v_1), \dots, \tilde{f}(v_m)) : \tilde{f} \in \mathcal{D}_1^\delta,\ v \in \mathcal{V}\}$ constitutes an η-net of $(\mathcal{D} \circ \mathcal{G})_{|\tilde{X}_{1:m}}$. Hence,

$$\begin{aligned}
\log \mathcal{N}(\eta; (\mathcal{D} \circ \mathcal{G})_{|\tilde{X}_{1:m}}, \ell_\infty) \leq & \log\left(N(\delta; \mathcal{D}, \ell_\infty) \times \mathcal{N}(\epsilon; \mathcal{G}_{|Z_{1:m}}, \ell_\infty)\right) \\
\lesssim & \eta^{-\frac{d}{\beta}} + W_gL_g(\log W_g + L_g)\log\left(\frac{md}{\eta^{1/(\beta\wedge 1)}}\right) \\
\lesssim & \eta^{-\frac{d}{\beta}} + W_gL_g(\log W_g + L_g)\log\left(\frac{md}{\eta}\right).
\end{aligned}$$

Replacing η with ϵ gives us the desired (6). □

C.3.3 Proof of Lemma 17

Lemma 17 Suppose that $\mathcal{G} = \mathcal{RN}(L_g, W_g, 2C_g)$ and $\mathcal{F} = \mathcal{RN}(L_f, W_f, 2C_f)$ with $L_d, L_g \geq 3$, $W_g \geq 6d + 2dL_g$ and $W_f \geq 6\tilde{d} + 2dL_f$. Then there is a constant c, such that, if $m \geq cW_gL_g(\log W_g + L_g)$ and $n \geq cW_fL_f(\log W_f + L_f)$, such that,

$$\mathbb{E}\|\hat{\nu}_m - \nu\|_{\mathcal{D}\circ\mathcal{G}} \lesssim m^{-\beta/d} \vee m^{-1/2}(\log m)^{\{d=2\beta\}} + \sqrt{\frac{1}{m}W_gL_g(\log W_g + L_g)\log(md)},$$

$$\mathbb{E}\|\hat{\mu}_n - \mu\|_{\tilde{\mathcal{D}}\circ\mathcal{F}} \lesssim n^{-\tilde{\beta}/\tilde{d}} \vee n^{-1/2}(\log n)^{\{\tilde{d}=2\tilde{\beta}\}} + \sqrt{\frac{1}{n}W_fL_f(\log W_f + L_f)\log(n\tilde{d})}.$$

Proof. We define the $\|\cdot\|_n$-norm between two function f and g as,

$$\|f-g\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f(X_i)-g(X_i))^2.$$

Since, $\|f-g\|_n \le \max_{i\in[n]}|f(X_i)-g(X_i)|$, $\mathcal{N}(\epsilon;\mathcal{F},\|\cdot\|_n) \le \mathcal{N}(\epsilon;\mathcal{F}_{|X_{1:n}},\ell_\infty)$. From Dudley's chaining argument, we note that,

$$\begin{aligned}
\mathbb{E}\|\hat{\nu}_m-\nu\|_{\mathcal{D}\circ\mathcal{G}} \lesssim & \mathbb{E}\inf_{0<\delta\le 1/2}\left(\delta+\frac{1}{\sqrt{m}}\int_\delta^{1/2}\sqrt{\log\mathcal{N}(\epsilon;\mathcal{D}\circ\mathcal{G},\|\cdot\|_m)}d\epsilon\right)\\
\le & \inf_{0<\delta\le 1/2}\left(\delta+\frac{1}{\sqrt{m}}\int_\delta^{1/2}\sqrt{\log\mathcal{N}(\epsilon;(\mathcal{D}\circ\mathcal{G})_{|Z_{1:m}},\ell_\infty)}d\epsilon\right)\\
\lesssim & \inf_{0<\delta\le 1/2}\left(\delta+\frac{1}{\sqrt{m}}\int_\delta^{1/2}\sqrt{\epsilon^{-\frac{d}{\beta}}+W_gL_g(\log W_g+L_g)\log\left(\frac{md}{\epsilon}\right)}d\epsilon\right) && (23)\\
\le & \inf_{0<\delta\le 1/2}\left(\delta+\frac{1}{\sqrt{m}}\int_\delta^{1/2}\left(\sqrt{\epsilon^{-\frac{d}{\beta}}}+\sqrt{W_gL_g(\log W_g+L_g)\log\left(\frac{md}{\epsilon}\right)}\right)d\epsilon\right)\\
\le & \inf_{0<\delta\le 1/2}\left(\delta+\frac{1}{\sqrt{m}}\int_\delta^{1/2}\epsilon^{-\frac{d}{2\beta}}d\epsilon+\frac{1}{\sqrt{m}}\int_0^{1/2}\sqrt{W_gL_g(\log W_g+L_g)\log\left(\frac{md}{\epsilon}\right)}d\epsilon\right)\\
\lesssim & \inf_{0<\delta\le 1/2}\left(\delta+\frac{1}{\sqrt{m}}\int_\delta^{1/2}\epsilon^{-\frac{d}{2\beta}}d\epsilon+\sqrt{\frac{W_gL_g(\log W_g+L_g)\log(md)}{m}}\right)\\
\lesssim & m^{-\beta/d}\vee m^{-1/2}(\log m)^{\{d=2\beta\}}+\sqrt{\frac{W_gL_g(\log W_g+L_g)\log(md)}{m}}. && (24)
\end{aligned}$$

In the above calculations, (23) follows from Lemma 16 and (24) follows from Lemma 21. The latter part of the lemma follows from a similar calculation as above. □

C.3.4 Proof of Lemma 18

Lemma 18. *Suppose that $W_g+W_f\ge 2(d\vee\tilde{d})(3+L_g+L_f)$, and $m,n\ge\theta_2(W_g+W_f)(L_g+L_f)(\log(W_g+W_f)+L_g+L_f)$, for a constant θ_2 (that might depend on d and $\tilde{d}$). Then,*

$$\mathbb{E}\|\hat{\mu}_n-\mu\|_{\Phi_1}\lesssim\sqrt{\frac{WL\,(\log W+L)\log(nd)}{n}} \qquad (8)$$

$$\mathbb{E}\|\hat{\nu}_m-\nu\|_{\Phi_2}\lesssim\sqrt{\frac{WL\,(\log W+L)\log(m\tilde{d})}{m}}, \qquad (9)$$

where $W=W_f+W_g$ and $L=W_f+W_g$.

Proof. We only give a proof of the first. The proof of the latter can be done similarly. We begin by observing that

$$\begin{aligned}
\mathbb{E}\|\hat{\mu}_n-\mu\|_{\Phi_1}\le & 2\mathbb{E}\mathcal{R}(X_{1:n},\Phi) && (25)\\
\lesssim & \mathbb{E}\mathcal{R}(X_{1:n},\mathcal{G}\circ\mathcal{F}) && (26)\\
\lesssim & \mathbb{E}\inf_{0<\delta\le 1/2}\left(\delta+\frac{1}{\sqrt{m}}\int_\delta^{1/2}\sqrt{\log\mathcal{N}(\epsilon;\mathcal{G}\circ\mathcal{F},\|\cdot\|_n)}d\epsilon\right) && (27)\\
\lesssim & \sqrt{\frac{(W_g+W_f)(L_g+L_f)\left(\log(W_g+W_f)+(L_g+L_f)\right)\log(nd)}{n}}. && (28)
\end{aligned}$$

Equation (25) follows from symmetrization, whereas, (26) follows from Lemma 26.9 of Shalev-Shwartz and Ben-David (2014). Inequality (27) follows from Dudley's entropy integral and (28) follows by applying Lemma 24. □

D Supporting Results from the Literature

Lemma 21 (Lemma 41 of Chakraborty and Bartlett, 2024)**.**

$$\inf_{0<\delta\le B}\left(\delta + \frac{1}{\sqrt{n}}\int_{\delta}^{B}\epsilon^{-\tau}d\epsilon\right) \lesssim \begin{cases} n^{-1/2}, & \textit{if } \tau < 1 \\ n^{-1/2}\log n, & \textit{if } \tau = 1 \\ n^{-\frac{1}{2\tau}}, & \textit{if } \tau > 1. \end{cases}$$

Lemma 22 (Theorem 12.2 of Anthony and Bartlett, 2009)**.** *Assume for all $f \in \mathcal{F}$, $\|f\|_\infty \le M$. Denote the pseudo-dimension of $\mathcal{F}$ as $Pdim(\mathcal{F})$, then for $n \ge Pdim(\mathcal{F})$, we have for any ϵ and any $X_1, \dots, X_n$,*

$$\mathcal{N}(\epsilon; \mathscr{F}_{|X_{1:n}}, \ell_\infty) \le \left(\frac{2eMn}{\epsilon Pdim(\mathcal{F})}\right)^{Pdim(\mathcal{F})}.$$

Lemma 23 (Theorem 6 of Bartlett et al., 2019)**.** *Consider the function class computed by a feed-forward neural network architecture with W many weight parameters and U many computation units arranged in L layers. Suppose that all non-output units have piecewise-polynomial activation functions with $p+1$ pieces and degrees no more than d, and the output unit has the identity function as its activation function. Then the VC-dimension and pseudo-dimension are upper-bounded as*

$$VCdim(\mathcal{F}), Pdim(\mathcal{F}) \le C \cdot LW\log(pU) + L^2 W \log d.$$

Lemma 24 (Lemma 21 of Chakraborty and Bartlett, 2024)**.** *Suppose that $n \ge 6$ and $\mathcal{F}$ be a class neural network with depth at most L and number of weights at most W. Furthermore, the activation functions are piece-wise polynomial activation with a number of pieces and degree at most $k \in \mathbb{N}$. Then, there is a constant θ (that might depend on d and d'), such that, if $n \ge \theta(W + 6d' + 2d'L)(L+3)\left(\log(W + 6d' + 2d'L) + L + 3\right)$,*

$$\log \mathcal{N}(\epsilon; \mathcal{F}_{|X_{1:n}}, \ell_\infty) \lesssim (W + 6d' + 2d'L)(L+3)\left(\log(W + 6d' + 2d'L) + L + 3\right)\log\left(\frac{nd'}{\epsilon}\right),$$

where d' is the output dimension of the networks in $\mathcal{F}$.

Lemma 25 (Proposition 2 of Yarotsky, 2017)**.** *The function $f(x) = x^2$ on the segment $[0,1]$ can be approximated with any error by a ReLU network, $sq_m(\cdot)$, such that,*

1. *$\mathcal{L}(sq_m)$, $\mathcal{W}(sq_m) \le c_1 m$.*
2. *$sq_m\left(\frac{k}{2^m}\right) = \left(\frac{k}{2^m}\right)^2$, for all $k = 0, 1, \dots, 2^m$.*
3. *$\|sq_m - x^2\|_{\mathcal{L}_\infty([0,1])} \le \frac{1}{2^{2m+2}}$.*
4. *$\mathcal{B}(sq_m) \le 4$.*

Lemma 26 (Lemma 39 of Chakraborty and Bartlett, 2024)**.** *Let $M \ge 1$, then we can find a ReLU network $prod_m^{(2)}$, such that,*

1. *$\mathcal{L}(prod_m^{(2)}), \mathcal{W}(prod_m^{(2)}) \le c_2 m$, for some absolute constant c_2.*
2. *$\|prod_m^{(2)} - xy\|_{\mathcal{L}_\infty([-M,M]\times[-M,M])} \le \frac{M^2}{2^{2m+1}}$.*
3. *$\mathcal{B}(prod_m^{(2)}) \le 4 \vee M^2$.*

Lemma 27 (Lemma 40 of Chakraborty and Bartlett, 2024)**.** *For any $m \ge \frac{1}{2}(\log_2(4d) - 1)$, we can construct a ReLU network $prod_m^{(d)} : \mathbb{R}^d \to \mathbb{R}$, such that for any $x_1, \dots, x_d \in [-1,1]$, $\|prod_m^{(d)}(x_1, \dots, x_d) - x_1 \dots x_d\|_{\mathcal{L}_\infty([-1,1]^d)} \le \frac{1}{2^m}$. Furthermore,*

1. *$\mathcal{L}(prod_m^{(d)}) \le c_3 m$, $\mathcal{W}(prod_m^{(d)}) \le c_3 m$.*
2. *$\mathcal{B}(prod_m^{(d)}) \le 4$.*