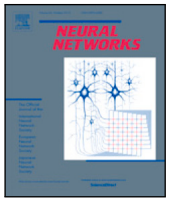
Neural Networks 181 (2025) 106770

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Full Length Article

A Wasserstein perspective of Vanilla GANs

Lea Kunkel *, Mathias Trabs

Karlsruhe Institute of Technology, Department of Mathematics, Germany

ARTICLE INFO

Keywords:
Generative adversarial networks
Rate of convergence
Oracle inequality
Wasserstein distance
Distribution estimation

ABSTRACT

The empirical success of Generative Adversarial Networks (GANs) caused an increasing interest in theoretical research. The statistical literature is mainly focused on Wasserstein GANs and generalizations thereof, which especially allow for good dimension reduction properties. Statistical results for Vanilla GANs, the original optimization problem, are still rather limited and require assumptions such as smooth activation functions and equal dimensions of the latent space and the ambient space. To bridge this gap, we draw a connection from Vanilla GANs to the Wasserstein distance. By doing so, existing results for Wasserstein GANs can be extended to Vanilla GANs. In particular, we obtain an oracle inequality for Vanilla GANs in Wasserstein distance. The assumptions of this oracle inequality are designed to be satisfied by network architectures commonly used in practice, such as feedforward ReLU networks. By providing a quantitative result for the approximation of a Lipschitz function by a feedforward ReLU network with bounded Hölder norm, we conclude a rate of convergence for Vanilla GANs as well as Wasserstein GANs as estimators of the unknown probability distribution.

1. Introduction

Generative Adversarial Networks (GANs) have attracted much attention since their introduction by Goodfellow et al. (2014), initially due to impressive results in the creation of photorealistic images. Meanwhile, the areas of application have expanded far beyond this, and GANs serve as a prototypical example of the rapidly evolving research area of generative models.

The *Vanilla GAN* as constructed by Goodfellow et al. (2014) relies on the minimax game

$$\inf_{G\in\mathcal{G}} \sup_{D\in D} \mathbb{E}\big[\log D(X) + \log\big(1 - D(G(Z))\big)\big], \tag{1}$$

to learn an unknown distribution $\mathbb{P}^*$ of the random variable X. The generator G chosen from a set $\mathcal{G}$, applied to the latent random variable Z aims to mimic the distribution of X as closely as possible. The discriminator D, chosen from a set D, has to distinguish between real and fake samples.

The optimization problem in (1) is motivated by the Jensen–Shannon divergence, see Goodfellow et al. (2014, Theorem 1). Generalizations of the underlying distance have led to various extensions of the original GAN, such as f-GANs (Nowozin, Cseke, & Tomioka, 2016). More famously, *Wasserstein GANs* (Arjovsky, Chintala, & Bottou, 2017), characterized by

$$\inf_{G\in\mathcal{G}} \sup_{W\in\mathrm{Lip}(1)} \mathbb{E}\big[W(X) - W(G(Z))\big], \tag{2}$$

are obtained by replacing the Jensen–Shannon divergence by the Kantorovich dual of the Wasserstein distance (see Section 2). Here, Lip(1) denotes the set of all Lipschitz continuous functions with a Lipschitz constant bounded by one. This approach can be generalized using *Integral Probability Metrics* (IPMs, Mueller, 1997). For the application to GANs, see for example Liang (2021).

The analysis of Wasserstein GANs can exploit the existing theory on the Wasserstein distance. The latter has a long record of research, particularly in the context of optimal transport (Villani, 2008) but also in machine learning, see Torres, Pereira, and Amini (2021) for an overview. In contrast, Vanilla GANs and the Jensen–Shannon divergence have been studied less extensively, and fundamental questions have not been settled. In particular, all statistical results for Vanilla GANs require the same dimension of the latent space and the target space which is in stark contrast to common practice. The Jensen–Shannon divergence between singular measures is by definition maximal. Therefore, we cannot expect proofs of convergence in a dimension reduction setting. In practice, however, Vanilla GANs have worked in a wide range settings. Another algorithmic drawback of Vanilla GANs highlighted by Arjovsky and Bottou (2017) is that an arbitrarily large discriminator class prevents the generator from learning. Thus, using neural networks as a discriminator class must be advantageous compared to the set of all measurable functions. This empirical fact is supported by the numerical results by Farnia and Tse (2018) who

* Corresponding author.
E-mail addresses: lea.kunkel@kit.edu (L. Kunkel), trabs@kit.edu (M. Trabs).

https://doi.org/10.1016/j.neunet.2024.106770
Received 11 April 2024; Received in revised form 27 September 2024; Accepted 28 September 2024
Available online 6 October 2024

impose a Lipschitz constraint on the discriminator class. In this work, we broaden the theoretical boundaries of Vanilla GANs to cope with the empirical evidence. To this end, we replace the Jensen–Shannon framework with a Wasserstein perspective.

A wide range of methods exists for measuring distances between probability measures with varying properties. In view of the manifold hypothesis for high-dimensional data, it is crucial that the selected metric can discriminate between different singular measures. This is not the case for the Total Variation distance, the Jensen–Shannon distance, or even stronger metrics where singular distributions always attain the maximal value of the distance. Conversely, it is advantageous to select a strong metric, as this yields immediate bounds in weaker metrics. In view of this tension we will analyze Vanilla GANs with respect to the Wasserstein-1 distance which is often used in the statistical as well as the machine learning literature. A comprehensive overview of the advantages of the Wasserstein-1 distance over other measures that metrize weak convergence, we direct the reader to Villani (2008, p. 98 f.). For an overview of distances between probability measures, we refer the reader to Gibbs and Su (2002, Figure 1).

Our contribution. Our work aims to bridge the gap in theoretical analysis between Vanilla GANs and Wasserstein GANs while addressing the theoretical limitations of the former ones. By imposing a Lipschitz condition on the discriminator class in (1), we recover Wasserstein GAN-like behavior. As a main result, we can derive an oracle inequality for the Wasserstein distance between the true data generating distribution and its Vanilla GAN estimate. In particular, this allows us to transfer key features, such as dimension reduction, known from the statistical analysis of Wasserstein GANs. We show that the statistical error of the modified Vanilla GAN depends only on the dimension of the latent space, independent of the potentially much larger dimension of the feature space $\mathcal{X}$. Thus Vanilla GANs can avoid the curse of dimensionality. Such properties are well known from practice, but cannot be verified by the classical Jensen–Shannon analysis. On the other hand the derived rate of convergence for the Vanilla GAN is slower than for Wasserstein GANs which is in line with the empirical advantage of Wasserstein GANs.

We then consider the most relevant case where the classes $\mathcal{G}$ and $\mathcal{D}$ are parameterized by neural networks. Using our previous results, we derive an oracle inequality that depends on the network approximation errors for the best possible generator and the optimal Lipschitz discriminator. To bound the approximation error of the discriminator, we replace the Lipschitz constraint on the networks with a less restrictive Hölder constraint. Building on Gühring, Kutyniok, and Petersen (2020), we prove a novel quantitative approximation theorem for Lipschitz functions using ReLU neural networks with bounded Hölder norm. As a result we obtain the rate of convergence $n^{-\alpha/2d^*}, \alpha \in (0,1)$, with latent space dimension $d^* \geq 2$ for sufficiently large classes of networks. Additionally, our approximation theorem allows for an explicit bound on the discriminator approximation error for Wasserstein-type GANs, which achieve the rate $n^{-\alpha/d^*}, \alpha \in (0,1)$.

We use a simple illustrative example to assess the practical implications of our theoretical results. This example allows us to quantify the rate depending on the number of observations, the dimension reduction property, and the stabilizing effect of a Lipschitz-constrained discriminator class.

Related work. The existence and uniqueness of the optimal generator for Vanilla GANs is shown by Biau, Cadre, Sangnier, and Tanielian (2020) under the condition that the class $\mathcal{G}$ is convex and compact. They also study the asymptotic properties of Vanilla GANs. Puchkin, Samsonov, Belomestny, Moulines, and Naumov (2024) have shown a non-asymptotic rate of convergence in the Jensen–Shannon divergence for Vanilla GANs with neural networks under the assumption that the density of $\mathbb{P}^*$ exists and that the generator functions are continuously differentiable.

In practice, however, the *Rectifier Linear Unit* activation function (ReLU activation function) is commonly used (Aggarwal, 2018, p.13). The resulting neural network generates continuous piecewise linear functions. Therefore, the convergence rate of Puchkin et al. (2024) combined with Belomestny, Naumov, Puchkin, and Samsonov (2023) is not applicable to this class of functions.

The statistical analysis of Wasserstein GANs is much better understood. Biau, Sangnier, and Tanielian (2021) have studied optimization and asymptotic properties. Liang (2021) has shown error decompositions with respect to the Kullback–Leibler divergence, the Hellinger distance and the Wasserstein distance. The case where the unknown distribution lies on a low-dimensional manifold is considered in Schreuder, Brunel, and Dalalyan (2021) as well as Tang and Yang (2023). The latter also derive minmax rates in a more general setting using the Hölder metric. Assuming that the density function of $\mathbb{P}^*$ exists, Liang (2017) has shown a rate of convergence in Wasserstein distance with ReLU activation function with a factor growing exponentially in the depth of the network. Theoretical results including sampling the latent distribution in addition to dimension reduction have been derived by Huang et al. (2022), who have also shown a rate of convergence in a slightly more general setting (using the Hölder metric) using ReLU networks whose Lipschitz constant grows exponentially in the depth. A rate of convergence using the total variation metric and leaky ReLU networks has been shown in Liang (2021).

Convergence rates with respect to the Wasserstein distance have been studied by Chen, Liao, Zha, and Zhao (2020) and Chae (2022). Up to a logarithmic factor, optimal rates in the Hölder metric were obtained by Stéphanovitch, Aamari, and Levrard (2023) using smooth networks. In a similar setting, Chakraborty and Bartlett (2024) discussed several methods for dimension reduction. Recently, Suh and Cheng (2024) have reviewed the theoretical advances in Wasserstein GANs.

Ensuring Lipschitz continuity of the discriminator class is the essential property of Wasserstein GANs. Lipschitz-constrained neural networks and their empirical success are subject of ongoing research ((Khromov & Singh, 2024), in context of GANs see Than and Vu (2021)). Implementations of the Lipschitz constrained discriminator have
evolved from weight clipping (Arjovsky et al., 2017) to penalizing the objective function (Asokan & Seelamantula, 2023; Gulrajani, Ahmed, Arjovsky, Dumoulin, & Courville, 2017; Miyato, Kataoka, Koyama, & Yoshida, 2018; Petzka, Fischer, & Lukovnikov, 2018; Wei, Liu, Wang, & Gong, 2018; Zhou et al., 2019), which heuristically leads to networks with bounded Lipschitz constants. Farnia and Tse (2018) use an objective function that combines Wasserstein and Vanilla GANs.

Outline. In Section 2 we introduce the Vanilla GAN distance, which characterizes the optimization problem (1). In Section 3, we investigate the relation between the Vanilla GAN distance and the Wasserstein distance. We show that the distances are compatible to each other while not being equivalent. Using this relation, we derive an oracle inequality for the Vanilla GAN in Section 4, where $\mathcal{G}$ is a nonempty compact set and $\mathcal{D}$ is a set of Lipschitz functions. We show that Vanilla GANs can avoid the curse of dimensionality. In Section 5 we consider the situation where $\mathcal{G}$ and $\mathcal{D}$ consist of neural networks. Here we relax the Lipschitz condition to a α-Hölder condition and prove a quantitative result for the approximation of a Lipschitz function by a feedforward ReLU network with bounded Hölder norm. We then prove a convergence rate for the Vanilla GAN with network generator and discriminator. In Section 6 we obtain a convergence rate for Wasserstein-type GANs with network generator and discriminator using our approximation result. This enables us to compare Vanilla GANs directly to Wasserstein GANs. In Section 7 we illustrate our theoretical results in a numerical example based on synthetic data. All proofs are deferred to the Appendix.

2. The Vanilla GAN distance

Let us first fix some notation. We equip $\mathbb{R}^d$ with the ℓ_p-norm $|x|_p$, $1 \le p \le \infty$, denote the number of nonzero entries of a $k \times l$ matrix A, where $k, l \in \mathbb{N}$, by $|A|_{\ell^0} := |\{(i,j) : A_{ij} \neq 0\}|$, and define the ceiling of $x \in \mathbb{R}$ as $\lceil x \rceil := \min\{k \in \mathbb{Z} \mid k \ge x\}$. For ease of notation we abbreviate for $x \in (0, \infty)$

$$[x]^{1;1/2} := \max\{x, \sqrt{x}\}. \tag{3}$$

For $\Omega \subset \mathbb{R}^d$ and $f : \Omega \to \mathbb{R}$ we define $\|f\|_{\infty,\Omega} := \operatorname{ess\,sup}\{|f(x)| : x \in \Omega\}$. We denote the set of bounded Lipschitz functions by

$$\mathrm{Lip}(L, B, \Omega) := \left\{ f : \Omega \to \mathbb{R} \,\middle|\, \|f\|_{\infty,\Omega} \le B, \ \frac{|f(x) - f(y)|}{|x - y|_p} \le L, x, y \in \Omega \right\}.$$

The set of unbounded Lipschitz functions is abbreviated by $\mathrm{Lip}(L, \Omega) := \mathrm{Lip}(L, \infty, \Omega)$. By Rademacher's theorem, a Lipschitz function is differentiable almost everywhere. For $\alpha \in (0, 1]$ we define the α-Hölder norm by

$$\|f\|_{\mathcal{H}^\alpha(\Omega)} := \max\left\{ \|f\|_\infty, \operatorname*{ess\,sup}_{x,y \in \Omega} \frac{|f(x) - f(y)|}{|x - y|_p^\alpha} \right\} \tag{4}$$

and the α-Hölder ball of functions with Hölder constant less or equal than $\Gamma > 0$ as

$$\mathcal{H}^\alpha(\Gamma, \Omega) := \left\{ f : \Omega \to \mathbb{R} \,\middle|\, \|f\|_{\mathcal{H}^\alpha(\Omega)} < \Gamma \right\}. \tag{5}$$

In particular, $\mathrm{Lip}(L, B, \Omega) \subseteq \mathcal{H}^\alpha(\max(L, 2B), \Omega)$ for any $\alpha \in (0, 1)$. We omit the domain Ω in our notation if $\Omega = (0,1)^d$.

We observe i.i.d. samples $X_1, \dots, X_n \sim \mathbb{P}^*$ with values in $\mathcal{X} := (0,1)^d$. On another space $\mathcal{Z} := (0,1)^{d^*}$, called the *latent* space, we choose a latent distribution $\mathbb{U}$. Unless otherwise specified, $X \sim \mathbb{P}^*$ and $Z \sim \mathbb{U}$. We further assume that $\mathbb{P}^*$ and $\mathbb{U}$ have finite first moments. Throughout, the generator class $\mathcal{G}$ is a nonempty set of measurable functions from $\mathcal{Z}$ to $\mathcal{X}$. For $G \in \mathcal{G}$ the distribution of the random variable $G(Z)$ is denoted by $\mathbb{P}^{G(Z)}$.

Typically the discriminator class consists of functions mapping to $\mathbb{R}$ concatenated to a sigmoid function that maps into $(0,1)$ to account for the classification task. This is especially the case for standard classification networks. The most common sigmoid function used for this purpose is the logistic function $x \mapsto (1 + e^{-x})^{-1}$, which we fix throughout. Together with a shift by $\log 4$, we can rewrite the Vanilla GAN optimization problem (1) as

$$\inf_{G \in \mathcal{G}} \mathsf{V}_{\mathcal{W}}(\mathbb{P}^*, \mathbb{P}^{G(Z)}) \tag{6}$$

in terms of the *Vanilla GAN distance* between probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$

$$\mathsf{V}_{\mathcal{W}}(\mathbb{P}, \mathbb{Q}) := \sup_{W \in \mathcal{W}} \mathbb{E}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}} \left[-\log\left(\frac{1 + e^{-W(X)}}{2} \right) - \log\left(\frac{1 + e^{W(Y)}}{2} \right) \right], \tag{7}$$

where $\mathcal{W}$ is a set of measurable functions $W : \mathcal{X} \to \mathbb{R}$. As long as $0 \in \mathcal{W}$, we have that $\mathsf{V}_{\mathcal{W}} \ge 0$.

To choose the generator $\hat{G}_n$ as the empirical risk minimizer, the unknown distribution $\mathbb{P}^*$ in (6) must be replaced by the empirical distribution $\mathbb{P}_n$ based on the observations $X_1, \dots, X_n$. In practice, the expectation with respect to $Z \sim \mathbb{U}$ is replaced by an empirical mean too, which we omit for the sake of simplicity. Along Huang et al. (2022), the next and all subsequent results easily extend to the corresponding setting.

The following error bound in terms of the Vanilla GAN distance provides an error decomposition for the empirical risk minimizer of the Vanilla GAN.

Lemma 2.1. *Assume that $\mathcal{G}$ is chosen such that a minimum exists. Let $\mathcal{W}$ be symmetric, that is, $W \in \mathcal{W}$ implies $-W \in \mathcal{W}$. For*

$$\hat{G}_n \in \operatorname*{arg\,min}_{G \in \mathcal{G}} \mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)}) \tag{8}$$

we have that

$$\mathsf{V}_{\mathcal{W}}(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)}) \le \min_{G \in \mathcal{G}} \mathsf{V}_{\mathcal{W}}(\mathbb{P}^*, \mathbb{P}^{G(Z)}) + 2 \sup_{W \in \mathrm{Lip}(1) \circ \mathcal{W}} \frac{1}{n} \sum_{i=1}^n \big(W(X_i) - \mathbb{E}[W(X)] \big). \tag{9}$$

The first term in (9) is the error due to the approximation capabilities of the class $\mathcal{G}$. The second term refers to the stochastic error due to the amount of training data. As $\mathcal{W}$ is symmetric, the stochastic error is non-negative. Both error terms depend on the discriminator class $\mathcal{W}$. Large discriminator classes lead to finer discrimination between different probability distributions and thus to a larger approximation error term. Similarly, the stochastic error term will increase with the size of $\mathcal{W}$. The cost of small classes $\mathcal{W}$ is a less informative loss function on the left side of (9).

If $\mathcal{W}$ is the set of all measurable functions, the analysis by Goodfellow et al. (2014, Theorem 1) shows that the Vanilla GAN distance is equivalent to the Jensen–Shannon distance. Arjovsky and Bottou (2017) have elaborated on the theoretical and practical disadvantages of this case. Similar to the Total Variation distance or the Hellinger distance, the Jensen–Shannon divergence is not compatible with high-dimensional settings because it cannot distinguish between different singular measures. Therefore, we need a weaker distance and thus restrict $\mathcal{W}$.

The key insight of Wasserstein GANs (2) is that this particular drawback of the Jensen–Shannon distance can be solved by the Wasserstein distance. The latter is a metric on the space of probability distributions with finite first moment and meterizes weak convergence in this space (Villani, 2008, Theorem 6.9). Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions on the same measurable space $(\Omega, \mathcal{A})$, the *Wasserstein*-1 distance is defined as

$$\mathsf{W}_1(\mathbb{P}, \mathbb{Q}) := \sup_{\substack{W \in \mathrm{Lip}(1) \\ W(0)=0}} \mathbb{E}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}} [W(X) - W(Y)] = \sup_{W \in \mathrm{Lip}(1)} \mathbb{E}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}} [W(X) - W(Y)]. \tag{10}$$

Bounds for weaker metrics, such as the Kolmogorov or Levy metric, can be easily derived from the bounds in the Wasserstein metric under weak conditions, see e.g. Gibbs and Su (2002).

Therefore, we choose $\mathcal{W} = \mathrm{Lip}(L)$ for some $L \ge 1$ in Lemma 2.1. In this case the following result shows that the existence of an empirical risk minimizer is guaranteed as soon as $\mathcal{G}$ is compact.

Lemma 2.2. *Assume $\mathcal{G}$ is compact with respect to the supremum norm. The map $T : \mathcal{G} \to \mathbb{R}_{\ge 0}$, $T(G) := V_{\mathrm{Lip}(L)}(\mathbb{P}_n, \mathbb{P}^{G(z)})$ is continuous and $\operatorname{arg\,min}_{\mathcal{G}} V_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G(z)})$ is nonempty.*

Hence, we throughout assume the following:

Assumption 1. $\mathcal{G}$ is compact with respect to the supremum norm.

In the context of neural networks the compactness assumption is satisfied for all practically relevant implementations. Furthermore, it should be noted that the aforementioned assumption is only required for the use of the minimizing argument.

3. From Vanilla to Wasserstein and back

Our subsequent analysis builds on the following equivalence result between the Vanilla GAN distance and the Wasserstein distance with an additional L^2-penalty term on the discriminator.

Theorem 3.1. *For $L > 2$ and $B > 0$ we have for probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$*

$$\sup_{\substack{W \in \mathrm{Lip}(1, B') \\ W(\cdot) > -\log(2-2/L)}} \left\{ \mathbb{E}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}} [W(X) - W(Y)] - \frac{L(L-1)}{2} \mathbb{E}_{X \sim \mathbb{Q}} [W(X)^2] \right\}$$
$$\le \mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) \le \sup_{\substack{W \in \mathrm{Lip}(L,B) \\ W(\cdot) > -\log(2)}} \left\{ \mathbb{E}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}} [W(X) - W(Y)] \right.$$

$$- \frac{e^B}{(2e^B-1)^2}\mathbb{E}_{X\sim\mathbb{Q}}[W(X)^2] \Big\},$$

where $B' = \log((1+e^B)/2)$.

Theorem 3.1 reveals that the Vanilla GAN distance is indeed compatible with the Wasserstein distance and will allow us to prove rates of convergence of the Vanilla GAN with respect to the Wasserstein distance. In doing so, we need to investigate the consequences of the penalty term. An upper bound without the penalty term and independent of B can be shown as in the proof of Theorem 3.2. For the lower bound, a similar improvement cannot be expected in general as indicated in Example 3.3. However, the restriction to $\mathrm{Lip}(1,B')$ has far less severe consequences than the corresponding restriction in the upper bound.

We can deduce from Theorem 3.1 that the Vanilla GAN distance is bounded from above and below by the Wasserstein distance or the squared Wasserstein distance, respectively.

Theorem 3.2. *Let $L>2$ and $B\in[1,\infty]$. For probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$ we have*

$$\min\Big(c_1\mathsf{W}_1(\mathbb{P},\mathbb{Q}), c_2\mathsf{W}_1(\mathbb{P},\mathbb{Q})^2\Big) \le \mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P},\mathbb{Q}) \le L\mathsf{W}_1(\mathbb{P},\mathbb{Q}),$$

where $c_1 = \frac{1}{2}\frac{\log(2-2/L)}{d^{1/p}}$ *and* $c_2 = \frac{1}{2d^{2/p}}\frac{1}{L(L-1)}$, *setting* $1/p=0$ *if* $p=\infty$.

The assumption $L>2$ is not very restrictive. In practically relevant cases, such as neural network discriminators, the Lipschitz constant is typically quite large. A higher Lipschitz constraint on the discriminator will subsequently result in a less stringent constraint on the neural network. However, an arbitrarily large Lipschitz constant is also undesirable, as the upper bound grows linearly in L.

More importantly, we observe a gap between $\mathsf{W}_1(\mathbb{P},\mathbb{Q})^2$ in the lower bound and $\mathsf{W}_1(\mathbb{P},\mathbb{Q})$ in upper bound when $\mathsf{W}_1(\mathbb{P},\mathbb{Q})<1$ which is a consequence of the penalty term in Theorem 3.1. The following example indicates that this loss is unavoidable, by restricting the discriminator class to a subset of $\mathrm{Lip}(L)$.

Example 3.3. For $\varepsilon,\gamma>0, \gamma+\varepsilon<1$ let $\mathbb{P}=\frac{1}{2}(\delta_\gamma+\delta_{\gamma+\varepsilon})$ and $\mathbb{Q}=\frac{1}{2}(\delta_0+\delta_\varepsilon)$. The Wasserstein distance is then given by

$$\mathsf{W}_1(\mathbb{P},\mathbb{Q}) = \gamma.$$

We consider the Vanilla GAN distance using L-Lipschitz affine linear functions as discriminator, $\mathsf{V}_{a\cdot+b}(\mathbb{P},\mathbb{Q})$, with $a,b\in\mathbb{R}$ and $|a|\le L$. Note that the class of affine linear functions can be represented by one layer neural networks (for a definition see Section 5). The optimal b can be calculated explicitly, the optimal a can be determined numerically. Using the optimal slope a and b we obtain for $\gamma<\varepsilon$, $\varepsilon=\frac{1}{4}$ and $a>16$

$$\frac{\mathsf{W}_1(\mathbb{P},\mathbb{Q})^2}{2} \le \mathsf{V}_{a\cdot+b}(\mathbb{P},\mathbb{Q}) \le a\cdot\mathsf{W}_1(\mathbb{P},\mathbb{Q})^2.$$

If $\gamma\ge\varepsilon$, then the optimal a is $a=L$ and

$$\log(2)\cdot\mathsf{W}_1(\mathbb{P},\mathbb{Q}) \le \mathsf{V}_{a\cdot+b}(\mathbb{P},\mathbb{Q}) \le a\cdot\mathsf{W}_1(\mathbb{P},\mathbb{Q}).$$

See Appendix A.7 for more details on these calculations.

Wasserstein GANs, where the generator is chosen as the empirical risk minimizer of the Wasserstein distance (10), achieve optimal convergence rates (up to logarithmic factors) with respect to the Wasserstein distance as proved by Stéphanovitch et al. (2023), see also Section 6. In view of Theorem 3.2 we cannot hope that Vanilla GANs achieve the same rate even if we use a Lipschitz discriminator class. This is in line with the better performance of Wasserstein GANs in practice. However, Theorem 3.2 allows us to study the behavior of Vanilla GANs in settings where the dimension of the latent space is smaller than the dimension of the sample space, a setting that is excluded in all previous works on convergence rates for Vanilla GANs.

4. Oracle inequality for Vanilla GANs in Wasserstein distance

Our aim is to bound the Wasserstein distance between the unknown distribution $\mathbb{P}^*$ and the generated distribution $\mathbb{P}^{\hat{G}_n(Z)}$ using the empirical risk minimizer $\hat{G}_n$ of the Vanilla GAN. The following oracle inequality shows that imposing a Lipschitz constraint on the discriminator class does circumvent the theoretical limitations of Vanilla GANs which is caused by the Jensen–Shannon distance. Recall notation (3).

Theorem 4.1. *Let $L>2$ and $B\in[1,\infty]$. For the empirical risk minimizer $\hat{G}_n$ from (8) with $\mathcal{W}=\mathrm{Lip}(L,B)$ we have*

$$\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{\hat{G}_n(Z)}) \le c\Big[\inf_{G\in\mathcal{G}}\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G(Z)})\Big]^{1;1/2} + (1+c)[\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*)]^{1;1/2}, \quad (11)$$

for some constant $c>0$ depending on d,p and L.

Note that the discriminator class $\mathrm{Lip}(L,B)$ admits no finite dimensional parameterization and is therefore not feasible in practice. We will return to this issue in Section 5. The terms in (11) can be interpreted analogously to the interpretation of the bound in Lemma 2.1, but here we have an oracle inequality with respect to the Wasserstein distance. The first term is the approximation error. It is large when $\mathcal{G}$ is not flexible enough to provide a good approximation of $\mathbb{P}^*$ by $\mathbb{P}^{G(Z)}$ for some $G\in\mathcal{G}$. The second term refers to the stochastic error. With a growing number of observations the empirical measure $\mathbb{P}_n$ converges to $\mathbb{P}^*$ in Wasserstein distance, see Dudley (1969), and thus the stochastic error converges to zero. Together with the bounds on $\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*)$ by Schreuder (2020) we conclude the following:

Corollary 4.2. *Let $L>2, B\in[1,\infty]$. The empirical risk minimizer $\hat{G}_n$ from (8) with $\mathcal{W}=\mathrm{Lip}(L,B)$ satisfies for some constant $c>0$ depending on d,p and L that*

$$\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{\hat{G}_n(Z)})] \le \inf_{G^*:\mathcal{Z}\to\mathcal{X}}\Big\{c[\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G^*(Z)})]^{1;1/2} + c[\inf_{G\in\mathcal{G}}\|G-G^*\|_\infty]^{1;1/2}\Big\} + c\begin{cases} n^{-1/2d}, & d>2,\\ n^{-1/4}(\log n)^{1/2}, & d=2,\\ n^{-1/4}, & d=1,\end{cases}$$

where the infimum is taken over all Borel measurable functions $G^*:\mathcal{Z}\to\mathcal{X}$.

If there is some G^* such that $\mathbb{P}^*=\mathbb{P}^{G^*(Z)}$, which is commonly assumed in the GAN literature, see e.g. Stéphanovitch et al. (2023), the first term vanishes and the approximation error is bounded by $\inf_{G\in\mathcal{G}}[\|G-G^*\|_\infty^{1;1/2}]$. In the bound of the stochastic error we observe the curse of dimensionality: For large dimensions d the rate of convergence $n^{-1/2d}$ deteriorates.

To allow for a dimension reduction setting, we adopt the misspecified setting from Vardanyan, Minasyan, Hunanyan, Galstyan, and Dalalyan (2023, Theorem 1). In this scenario we can conclude statistical guarantees for Vanilla GANs that are comparable to the results obtained for Wasserstein GANs by Schreuder et al. (2021, Theorem 2). In view of Theorem 3.1 we expect a slower rate of convergence compared to Wasserstein GANs.

Theorem 4.3. *Let $L>2, B\in[1,\infty]$ and $M>0$. The empirical risk minimizer $\hat{G}_n$ from (8) with $\mathcal{W}=\mathrm{Lip}(L,B)$ satisfies*

$$\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{\hat{G}_n(Z)})] \le \inf_{G^*\in\mathrm{Lip}(M,\mathcal{Z})}\Big\{c[\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G^*(Z)})]^{1;1/2} + c[\inf_{G\in\mathcal{G}}\|G-G^*\|_\infty]^{1;1/2}\Big\} + c\begin{cases} n^{-1/2d^*}, & d^*>2,\\ n^{-1/4}(\log n)^{1/2}, & d^*=2,\\ n^{-1/4}, & d^*=1,\end{cases}$$

for some constant c depending d^*,d,p,L and M.

The Wasserstein distance $\mathsf{W}_1(\mathbb{P}^{G^*(Z)},\mathbb{P}^*)$ now includes an error due to the dimension reduction while the stochastic error is determined by the potentially much smaller dimension $d^*<d$ of the latent space.

Compared to Corollary 4.2, the only price for this improvement is the additional Lipschitz restriction on G^*. We observe a trade-off in the choice of d^*, since large latent dimensions reduce the approximation error for $\mathbb{P}^*$, but increase the stochastic error term. Additionally, there is a trade-off in M. A larger constant M results in a smaller value of $W_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})$, but increases the constant c. If the unknown distribution $\mathbb{P}^*$ is supported on a lower dimensional subspace and there exists a $G^* \in \mathrm{Lip}(M, \mathcal{Z})$ such that $\mathbb{P}^{G^*(Z)} = \mathbb{P}^*$, then the rate of convergence is solely determined by the dimension d^* of $\mathcal{Z}$. This is true for the smallest possible d^* for which a perfect G^* exists, as well as any d^{**} larger than d^*.

In many applications, the smallest possible d^* is unknown. Theorem 4.3 covers both over- and underestimation of the true dimension of the lower dimensional subspace. If the choice of d^* is too small, then $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})$ might not converge to 0, but the stochastic error still converges with the smaller rate d^*. If d^* is selected to be larger than the dimension of the lower dimensional subspace, then there could be a $G^* \in \mathrm{Lip}(M, \mathcal{Z})$ such that $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)}) = 0$, but the stochastic rates converges only with rate d^*. In the special case that a function $G^* \in \mathrm{Lip}(M, \mathcal{Z})$ exists such that $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)}) = 0$ the rate $n^{-1/2d^*}$ is slower to the rate n^{-1/d^*} obtained for the Wasserstein GAN by Schreuder et al. (2021). This is in line with Theorem 3.2 and Example 3.3.

However, Theorem 4.3 reveals why Vanilla GANs do perform well in high dimensions in the setting of an unknown distribution on a lower dimensional manifold. This phenomenon could not be explained in previous work on Vanilla GANs. Puchkin et al. (2024) and Biau et al. (2020) both obtain rates in the Jensen–Shannon distance.

5. Vanilla GANs with network discriminator

In practice, both $\mathcal{G}$ and $\mathcal{D}$ are sets of neural networks. Our conditions on the generator class $\mathcal{G}$ are compactness and good approximation properties of some G^* which is chosen such that $\mathbb{P}^{G^*(Z)}$ mimics $\mathbb{P}^*$. Since neural networks have a finite number of weights, and the absolute value of those weights is typically bounded, the compactness assumption is usually satisfied and neural networks enjoy excellent approximation properties, cf. DeVore, Hanin, and Petrova (2021).

The situation is more challenging for the discriminator class. So far, $\mathcal{D}$ was chosen as the set of Lipschitz functions concatenated to the logistic function. The Lipschitz property is crucial for proof of Theorem 4.1 and thus for all subsequent results.

Controlling the Lipschitz constant while preserving the approximation properties is an area of ongoing research and is far from trivial. Without further restrictions on the class of feedforward networks, the Lipschitz constant would be a term that depends exponentially on the size of the network, see Liang (2017, Theorem 3.2). Bounding the Lipschitz constant of a neural network is a problem that arises naturally in the implementation of Wasserstein GANs. Arjovsky et al. (2017) use weight clipping to ensure Lipschitz continuity. Later, other approaches such as gradient penalization (see Gulrajani et al. (2017), which was further developed by Wei et al. (2018), Zhou et al. (2019)), Lipschitz penalization (Petzka et al., 2018), or spectral penalization (Miyato et al., 2018) were introduced and have achieved improved performance in practice.

To extend the theory from the previous section to neural network discriminator classes, we first generalize Theorem 4.1 from $\mathcal{W} = \mathrm{Lip}(L, B)$ to subsets $\mathcal{W} \subseteq \mathrm{Lip}(L, B)$. As a result there is an additional approximation error term that accounts for the smaller discriminator class.

Theorem 5.1. *Let $L > 2, B \in [1, \infty]$. The empirical risk minimizer $\hat{G}_n$ from (8) with $\mathcal{W} \subseteq \mathrm{Lip}(L, B)$ satisfies*

$$\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)}) \le c\Big[\inf_{G\in\mathcal{G}} \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G(Z)})\Big]^{1;1/2} + c\Big[\inf_{W'\in\mathcal{W}} \sup_{W\in\mathrm{Lip}(L,B)} \|W - W'\|_\infty\Big]^{1;1/2} + c\big[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)\big]^{1;1/2},$$

for some constant $c > 0$ depending on d, p and L.

The approximation error of the discriminator depends on the supremum norm bound B of the functions in $\mathcal{W}$. While the statement remains true for $B = \infty$, when approximating the set $\mathrm{Lip}(L, B)$, this bound will be essential. To apply this result, we must ensure that the Lipschitz constant of a set of neural networks $\mathcal{W}$ is uniformly bounded by some constant L. Adding penalties to the objective function of the optimization problem does not guarantee a fixed bound on the Lipschitz constant. Approaches such as bounds on the spectral or row-sum norm of matrices in feedforward neural networks ensure a bound on the Lipschitz constant, but lead to a loss of expressiveness when considering ReLU networks, even in very simple cases such as the absolute value, see Huster, Chiang, and Chadha (2019) and Anil, Lucas, and Grosse (2019). On the other hand, Eckstein (2020) has shown that one-layer L Lipschitz networks are dense (with respect to the uniform norm) in the set of all L Lipschitz functions on bounded domains. While this implies that the discriminant approximation error converges to zero for growing network architectures, the density statement does not lead to a rate of convergence that depends on the size of the network.

Anil et al. (2019), motivated by Chernodub and Nowicki (2016), have introduced an adapted activation function, Group Sort, which leads to significantly improved approximation properties of the resulting networks. They show that networks using the Group Sort activation function are dense in the set of Lipschitz functions, but there is no quantitative approximation result. A discussion of the use of Group Sort in the context of Wasserstein GANs can be found in Biau et al. (2021).

To overcome this problem, we would like to approximate not only the optimal discriminating Lipschitz function from the Wasserstein optimization problem in the uniform norm, but also its (weak) derivative. This would allow us to keep the Lipschitz norm of the approximating neural network bounded. For networks with regular activation functions (Belomestny et al., 2023) have studied the simultaneous approximation of smooth functions and their derivatives. Gühring et al. (2020) have focused on ReLU networks and have derived quantitative approximation bounds in higher order Hölder and Sobolev spaces. As an intrinsic insight from approximation theory, the regularity of the function being approximated must exceed the regularity order of the norm used to derive approximation bounds. Therefore, we cannot expect to obtain quantitative approximation results for ReLU networks in Lipschitz norm without assuming the continuous differentiability of the approximated function.

Unfortunately, the maximizing function of the Wasserstein optimization problem is in general just Lipschitz continuous. Since we cannot increase the regularity of the target function, we instead relax the Lipschitz assumption of the discriminator in Theorem 5.2 to α-Hölder continuity for $\alpha \in (0, 1)$. This generalization in the context of Wasserstein GANs has recently been discussed by Stéphanovitch et al. (2023). Recall the definition of the Hölder ball from (5).

Theorem 5.2. *Let $L > 2, B \in [1, \infty)$, $\Gamma > \max(L, 2B)$ and $M > 0$. The empirical risk minimizer $\hat{G}_n$ from (8) with $\mathcal{W} \subseteq \mathcal{H}^\alpha(\Gamma)$ satisfies*

$$\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)})] \le \inf_{G^*\in\mathrm{Lip}(M,\mathcal{Z})} \Big\{ c\Big[\inf_{G\in\mathcal{G}} \|G^* - G\|_\infty^\alpha\Big]^{1;1/2} + c\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)^\alpha}\big]^{1;1/2} \Big\} + c\Big[\inf_{W\in\mathcal{W}} \sup_{W^*\in\mathrm{Lip}(L,B)} \|W - W^*\|_\infty\Big]^{1;1/2} + c \begin{cases} n^{-\alpha/2d^*}, & 2\alpha < d^*, \\ n^{-1/4}(\log n)^{1/2}, & 2\alpha = d^*, \\ n^{-1/4}, & 2\alpha > d^*, \end{cases}$$

for some constant c depending on d^*, d, p, L, M and Γ.

The lower bound on the Hölder constant of the discriminator class $\mathcal{W}$ is not overly restrictive when employing neural networks for this function class. Since c is increasing in Γ, it is advantageous to control the value of Γ.

It remains to show that there are ReLU networks that satisfy the assumptions of Theorem 5.2. To this end, we build on and extend the approximation results by Gühring et al. (2020).

To fix the notation we give a general definition of feedforward neural networks. Let $d, K, N_1, \dots, N_K \in \mathbb{N}$. A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is a neural network with K layers and $N_1 + \cdots + N_K$ neurons if it results for an argument $x \in \mathbb{R}^d$ from the following scheme:

$$\begin{aligned} x_0 &:= x, \\ x_k &:= \sigma\left(A_k x_{k-1} + b_k\right), \quad \text{for } k = 1, \dots, K-1, \\ \Phi(x) = x_K &:= A_K x_{K-1} + b_K, \end{aligned} \tag{12}$$

where for $k \in \{1, \dots, K\}$, $A_k \in \mathbb{R}^{N_k \times N_{k-1}}$ and $b_k \in \mathbb{R}^{N_k}$. $\sigma : \mathbb{R} \to \mathbb{R}$ is an element-wise applied arbitrary activation function. The number of nonzero weights of all A_k, b_k is given by $\sum_{j=1}^{K} \left(|A_j|_{\ell^0} + |b_j|_{\ell^0} \right)$. We focus on the ReLU activation function $\sigma(x) = \max(0, x)$.

Theorem 5.3. *Let $L, B > 0$, and $0 < \alpha < 1$. Then there are constants $C', C'', C''' > 0$ depending on d, L, α and B with the following properties: For any $\varepsilon \in (0, 1/2)$ and any $f \in \mathrm{Lip}(L, B)$, there is a ReLU neural network Φ_ε with no more than $\lceil C' \log_2(\varepsilon^{-\frac{1}{1-\alpha}}) \rceil$ layers, $\lceil C'' \varepsilon^{-\frac{d}{1-\alpha}} \log_2^2(\varepsilon^{-\frac{1}{1-\alpha}}) \rceil$ nonzero weights and $\lceil C''' \varepsilon^{-\frac{d}{1-\alpha}} (\log_2^2(\varepsilon^{-\frac{1}{1-\alpha}}) \vee \log_2(\varepsilon^{-\frac{1}{1-\alpha}})) \rceil$ neurons such that*

$$\left\| \Phi_\varepsilon - f \right\|_\infty \le \varepsilon \qquad \text{and} \qquad \Phi_\varepsilon \in \mathcal{H}^\alpha\left(\max(L, 2B) + \varepsilon\right).$$

Since there are many different reasons why controlling the Hölder constant of neural networks is interesting (with stability probably being the most prominent one), Theorem 5.3 is of interest on its own. Combining Theorems 5.2 and 5.3 with a standard approximation result for the generator approximation error, such as Yarotsky (2017, Theorem 1), leads to a rate of convergence. The networks in $\mathcal{G}$ approximating the function $G^* \in \mathrm{Lip}(M, \mathcal{Z})$ are only required to be measurable without any additional smoothness assumption.

Corollary 5.4. *For $0 < \alpha < 1$, $\Gamma > 5$, $M > 0$, $d^* > 2\alpha$ and $n > 2^{\frac{2d^*}{\alpha}}$ choose $\mathcal{G}$ as the set of ReLU networks with at most $\lceil c \cdot \log(n) \rceil$ layers, $\lceil c \cdot n \log(n) \rceil$ nonzero weights and $\lceil c \cdot n \log(n) \rceil$ neurons and $\mathcal{W}'$ as the set of ReLU networks with at most $\lceil c \cdot \log(n) \rceil$ layers, $\lceil c \cdot n^{\frac{\alpha}{2(1-\alpha)}} \log^2(n) \rceil$ nonzero weights and $\lceil c \cdot n^{\frac{\alpha}{2(1-\alpha)}} \log^2(n) \rceil$ neurons, where c is a constant depending on d, d^*, Γ, M and α. Then the empirical risk minimizer $\hat{G}_n$ from (8) with $\mathcal{W} = \mathcal{W}' \cap \mathcal{H}^\alpha(\Gamma)$ satisfies*

$$\mathbb{E}\left[\mathsf{W}_1\left(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)} \right) \right] \le c \cdot n^{-\alpha/2d^*} + c \left[\inf_{G^* \in \mathrm{Lip}(M, \mathcal{Z})} \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})^\alpha \right]^{1;1/2}.$$

From Theorem 5.3 we know that the set $\mathcal{W}' \cap \mathcal{H}^\alpha(\Gamma)$ of ReLU networks of finite width and depth is nonempty. In practice, this corresponds to a discriminator network with a controlled Hölder constant. On a bounded domain, any Lipschitz function is a Hölder function. Corollary 5.4 shows that Vanilla GANs with a Hölder regular discriminator class are theoretically advantageous. The Hölder parameter α can be chosen arbitrarily close to one. On the one hand this reveals why a Lipschitz regularization as implemented for Wasserstein GANs also improves the Vanilla GAN. An empirical confirmation can be found in Zhou et al. (2019) and Section 7. On the other hand the corollary then requires more neurons in the discriminator than in the generator class which coincides with common practice.

The width of the generator networks in Corollary 5.4 can be improved by replacing $G^* \in \mathrm{Lip}(M, \mathcal{Z})$ by with $G^* \in \mathrm{C}^{n-1}(\mathcal{Z}), n \in \mathbb{N}$, whose $(n-1)$-th derivative is Lipschitz continuous with Lipschitz constant M. Once more, this results in a trade-off, as $\left[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})^\alpha \right]^{1;1/2}$ increases when G^* is selected from a smaller set of functions.

6. Wasserstein GAN

The same analysis can be applied to Wasserstein-type GANs. The constrained on the Hölder constant can be weakened, as we do not need Theorem 3.2. Note that this does not impact the rate, but the constant. Define the Wasserstein-type distance with discriminator class $\mathcal{W}$ as

$$\mathsf{W}_{\mathcal{W}}(\mathbb{P}, \mathbb{Q}) = \sup_{W \in \mathcal{W}} \mathbb{E}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}} [W(X) - W(Y)].$$

The following theorem shows that by using Hölder continuous ReLU networks as the discriminator class, Wasserstein-type GANs can avoid the curse of dimensionality. Furthermore, this avoids the difficulties arising from the Lipschitz assumption of the neural network, as pointed out by Huang et al. (2022).

Theorem 6.1. *For $0 < \alpha < 1$, $\Gamma > 1, M > 0$ and $d > 2\alpha$ and $n > 2^{\frac{d}{\alpha}}$ choose $\mathcal{G}$ as the set of ReLU networks with at most $\lceil c \cdot \log(n) \rceil$ layers, $\lceil c \cdot n \log(n) \rceil$ nonzero weights and $\lceil c \cdot n \log(n) \rceil$ neurons and $\mathcal{W}'$ as the set of ReLU networks with at most $\lceil c \cdot \log(n) \rceil$ layers, $\lceil c \cdot n^{\frac{\alpha}{(1-\alpha)}} \log^2(n) \rceil$ nonzero weights and $\lceil c \cdot n^{\frac{\alpha}{(1-\alpha)}} \log^2(n) \rceil$ neurons, where c is a constant depending on d, d^*, Γ, M and α. The empirical risk minimizer with $\mathcal{W} = \mathcal{W}' \cap \mathcal{H}^\alpha(\Gamma)$*

$$\hat{G}_n \in \operatorname*{argmin}_{G \in \mathcal{G}} \mathsf{W}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)})$$

satisfies

$$\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)})] \le c \cdot n^{-\frac{\alpha}{d^*}} + \inf_{G^* \in \mathrm{Lip}(M, \mathcal{Z})} \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)}).$$

Compared to Corollary 5.4 the rate improves to $n^{-\alpha/d^*}$ for any $\alpha < 1$. The number of observations necessary for the theorem to hold, the size of the discriminator network and the lower bound for Γ decrease. Note that Γ does not effect the rate, but the constants. In case there exists a G^* such that $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)}) = 0$, this upper bound coincides with the lower bound in Tang and Yang (2023, Theorem 1) up to an arbitrary small polynomial factor.

Our rate does not depend exponentially on the number of layers like the results of Liang (2017), Huang et al. (2022) and we use non-smooth simple ReLU networks compared to smooth ReQU networks in Stéphanovitch et al. (2023) or group sort networks in Biau et al. (2021).

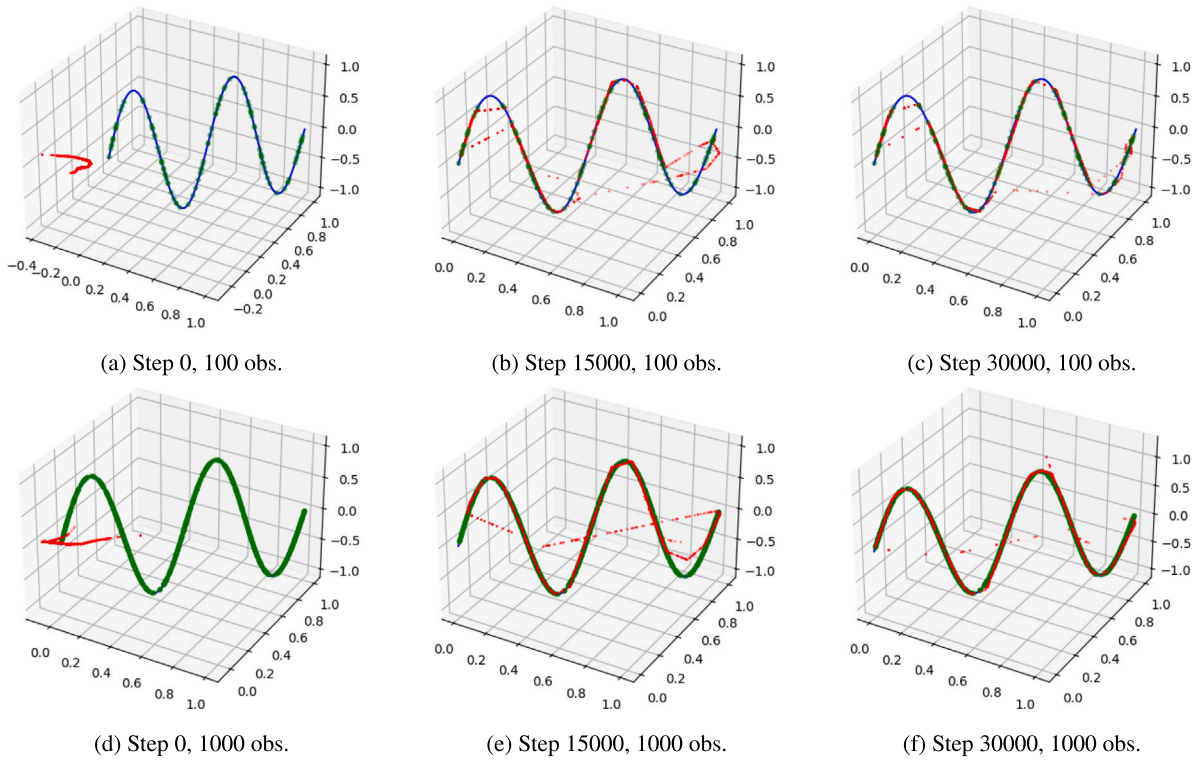
7. Numerical illustration

The results in Sections 5 and 6 were obtained under the assumption that the discriminator class consists of Lipschitz networks. In the context of image generation, these findings align with the results of Zhou et al. (2019), Miyato et al. (2018), Kodali, Abernethy, Hays, and Kira (2017), and Fedus et al. (2017). Furthermore, Fedus et al. (2017) demonstrated in a two-dimensional experiment that a Vanilla GAN with a gradient penalty (and, consequently, a lower Lipschitz constant) can be effective in scenarios where the measures $\mathbb{P}^*$ and $\mathbb{P}^{\hat{G}(Z)}$ are singular.

This section presents a transparent and accessible example that confirms our theoretical findings and especially demonstrates how imposing a Lipschitz constant on the discriminator stabilizes the Vanilla GAN. Additionally, it demonstrates the capacity of the Vanilla GAN to detect a lower dimensional manifold. In order to monitor rates of convergence, it is necessary to at least approximately evaluate $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)})$. Therefore, we study the numerical performance of the Vanilla GAN in a simulation setting, where the true data distribution is known by construction.

In this work, the Wasserstein distance is employed as the metric for measuring the rate of convergence. In practice, the Wasserstein distance is only computable in the one-dimensional case. To investigate multivariate distributions, we approximated the Wasserstein distance by averaging the Wasserstein distance on the marginals.

In order to model the distribution $\mathbb{P}^*$ of a lower dimensional manifold, we employed a one-dimensional uniform distribution on the graph of the function $x \mapsto \sin(4\pi x)$ on the diagonal of the two-dimensional unit cube, resulting in a three-dimensional distribution. For the latent



(a) Step 0, 100 obs. (b) Step 15000, 100 obs. (c) Step 30000, 100 obs.

(d) Step 0, 1000 obs. (e) Step 15000, 1000 obs. (f) Step 30000, 1000 obs.

Fig. 1. Training of Vanilla GAN with weight clip using 100 observations (first row) or 1000 observations (second row). Red dots show 1000 generated samples, green dots show the observations used for the training. The blue line is the one-dimensional manifold.

distribution, we used the one-dimensional normal distribution. Consequently, the dimensions of the lower dimensional manifold and the latent space are identical.

For the discriminator, we used a neural network with four layers of width 128 concatenated to a sigmoid function. For the generator, we used a neural network with three layers of width 64. In order to preserve as much alignment as possible with the theoretical result, we used plain ReLU activations. Each training consisted of 30000 training iterations. We used the Adam optimizer (Kingma & Ba, 2014) with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a learning rate of $\gamma = 0.0005$. When the number of observations exceeded 512, we used minibatches of that size in each iteration. We updated generator and discriminator alternating.

Three snapshots of the training of the Vanilla GAN for samples sizes $n = 100$ and $n = 1000$ are given in Fig. 1, respectively. The difference between Figs. 1(c) and 1(f) is solely due to the number of observations. The observations in Figs. 1(a) to 1(c) cover the manifold to a lesser extent than the observations in Figs. 1(d) to 1(f). This corresponds to a larger stochastic error.

To maintain the Lipschitz constant within a controllable range, we implemented the simple weight clipping mechanism of Arjovsky et al. (2017), limiting each weight to a value of 0.5. It is important to note that the network used in the unclipped case is also Lipschitz continuous, however, we do not have control over this Lipschitz constant. Given the width and depth parameters used in this study, it is evident that the Lipschitz constant of the clipped network remains relatively high and is considerably distinct from the theoretical value typically employed in Wasserstein GANs. However, a smaller Lipschitz constant requires an adjustment to the learning rate. Otherwise the weights are likely to remain at their maximum absolute value. This affects the experiment in several other ways. To ensure a fair and accurate comparison between the clipped and unclipped scenarios, we kept the learning rate consistent.

The results are summarized in Fig. 2. As predicted by our theory, the averaged marginal Wasserstein distance between the generated distribution and the true data distribution decays approximately as $n^{-1/2}$ for $n \in \{10, 100, 1000\}$. While we see a clear improvement with 10,000 observations, the additional gain is limited by the optimization error, since the manifold is already densely covered for 1000 observations.

It is apparent that controlling the Lipschitz constant overall stabilizes the training process, resulting in less variability in the results. In certain cases, the GAN without weight clipping can achieve the same level of effectiveness. This does not negate the outcome. Since the discriminator without clipped weights is still Lipschitz continuous (with a large Lipschitz constant), the theoretical limitations of Vanilla GANs without restricted discriminator classes do not directly translate to practice. This, combined with the finite nature of the implementations, ultimately resulted in the empirical success of these models. The variability between different simulation runs is described by the first to the third quartile in Fig. 2 which again confirms a more stable behavior of the clipped algorithm.

Fig. 3 demonstrates the high degree of precision with which the generated samples concentrate on the low-dimensional support of the true data distribution. Our experiments show that this concentration holds true across all sample sizes and can be observed in both the clipped and unclipped case. However, a high concentration does not necessarily indicate that the generated distribution is an accurate imitation of the unknown distribution with respect to the Wasserstein distance. Consequently, Fig. 3 is only informative in conjunction with Fig. 2.

Additionally, we investigated the use of a space $\mathbb{U}$ of the same dimension as the ambient space. Our observations indicated that the Vanilla GAN is still capable of identifying the lower dimensional subspace with reasonable efficacy.

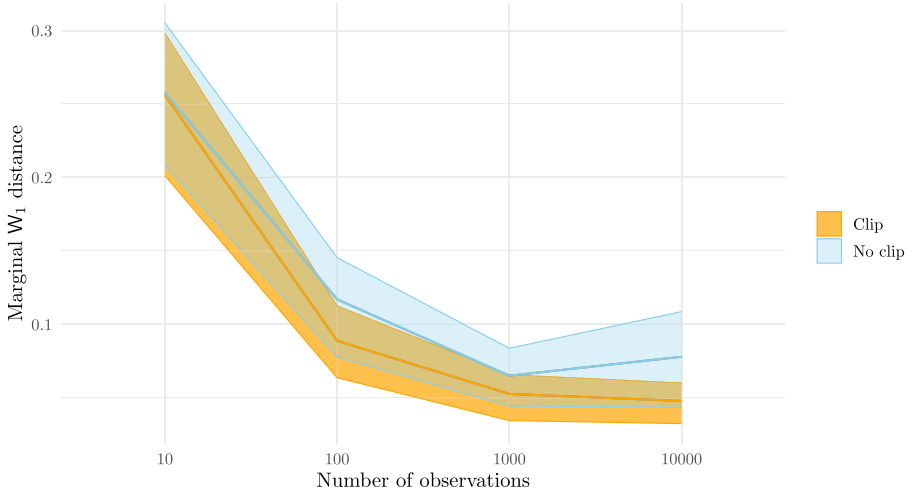


Fig. 2. Marginal W_1 distance depending on number of observations. Thick line shows the average over 50 independent runs, ribbons show the first to third quartile.

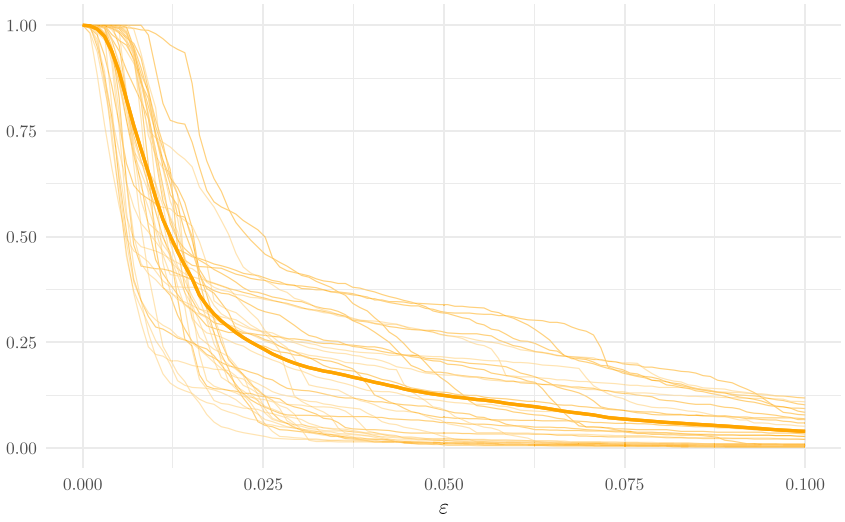


Fig. 3. Percentage of generated samples with euclidean distance to manifold greater than ε using 1000 observations and a discriminator with 0.5 clip. Transparent lines show the individual runs, thick line shows the average over 50 runs.

8. Discussion and limitations

Our analysis demonstrates that GANs originally built on too sensitive distribution distances, such as the Jensen–Shannon distance, can be improved by a Lipschitz constraint in the discriminator class. This insight might also be applicable to other GANs, e.g. f-GANs of Nowozin et al. (2016), which rely on a divergence that cannot discriminate between different singular distributions and thus is not suitable for a dimension reduction setting. Overall, we conclude that the choice of the discriminator class is much more important for the data generation capabilities than the choice of the loss function, which is typically dictated by some distance. Moreover, our analysis of the discriminator approximation error is not limited to Vanilla GANs, but is also applicable to optimal transport based GANs as demonstrated for the Wasserstein GAN.

There are several potential avenues for further development of the results presented in this paper. In particular, these include the above-mentioned potential implications for other types of GANs. While our analysis was limited to feedforward ReLU networks, one advancement in neural network research is the use of more sophisticated network architectures whose statistical analysis is not yet settled. In the context of Wasserstein GANs, see for example Radford, Metz, and Chintala (2015).

Furthermore, the inclusion of a bound on the Lipschitz constant (and not only the Hölder constant) would enable a direct application of Theorem 5.1, thereby eliminating the need to include the parameter α and thus improving the rates. Additionally, it would be interesting, whether there are conditions that allow for a faster rate of convergence for the Vanilla GAN in some cases (excluding scenarios as in Example 3.3).

The experiments also demonstrated that the GAN is capable of detecting data from a lower dimensional manifold if the latent space is of the same dimension as the ambient space. The proof of Theorem 4.3 is contingent upon the dimension of the latent space. If the dimension of the latent space is chosen to be too small, then $\inf_{G^* \in \mathrm{Lip}(M,\mathcal{Z})} \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})$ will be large. If the dimension of the latent space is chosen too large, $\inf_{G^* \in \mathrm{Lip}(M,\mathcal{Z})} \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})$ does not deteriorate, but the corresponding rate depends on the higher latent dimension. Therefore rates that are adaptive to the unknown intrinsic dimension, potentially benefiting from results like Berenfeld and Hoffmann (2021), would be interesting.

Bounds in other distances suitable for dimension reduction are also of high interest. For example, the Wasserstein-2 metric is slightly

stronger than the Wasserstein-1 metric (in the sense that $\mathsf{W}_1 \le \mathsf{W}_2$, see Villani (2008, Remark 6.6)). Our proofs rely on the duality of the Wasserstein-1 distance, hence they cannot be translated directly to the Wasserstein-2 distance.

Finally, the objective of this study was to examine statistical perspectives, and thus, the optimization problem was not addressed. In the proofs, we employ the global minimizer and maximizer. Since we face a non-convex optimization problem, gradient based methods may suffer from a considerable optimization error, especially for high-dimensional parameter spaces. Incorporating this optimization error would be more consistent with real-world scenarios.

CRediT authorship contribution statement

Lea Kunkel: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Mathias Trabs:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix

A.1. Proof for Section 2

Proof of Lemma 2.1. Let $X \sim \mathbb{P}^*$, $\hat{X} \sim \mathbb{P}_n$ and $Z \sim \mathbb{U}$. The symmetry of $\mathcal{W}$ and the Lipschitz continuity of $x \mapsto \log(1+e^{-x})$ yields for any $G \in \mathcal{G}$

$$
\begin{aligned}
&\mathsf{V}_{\mathcal{W}}(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n}(Z)) \\
&= \sup_{W\in\mathcal{W}} \mathbb{E}\Big[-\log\Big(\frac{1+e^{-W(X)}}{2}\Big) + \log\Big(\frac{1+e^{-W(\hat{X})}}{2}\Big) - \log\Big(\frac{1+e^{-W(\hat{X})}}{2}\Big) \\
&\quad - \log\Big(\frac{1+e^{W(\hat{G}_n(Z))}}{2}\Big)\Big] \\
&\le \sup_{W\in\mathcal{W}} \mathbb{E}\big[-\log(1+e^{-W(X)}) + \log(1+e^{-W(\hat{X})})\big] + \mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) \\
&\le \sup_{W\in\mathcal{W}} \mathbb{E}\big[-\log(1+e^{-W(X)}) + \log(1+e^{-W(\hat{X})})\big] + \mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)}) \\
&= \sup_{W\in\mathcal{W}} \mathbb{E}[-\log(1+e^{-W(X)}) + \log(1+e^{-W(\hat{X})})] \\
&\quad + \sup_{W\in\mathcal{W}} \mathbb{E}[-\log(1+e^{-W(\hat{X})}) + \log(1+e^{-W(X)})] + \mathsf{V}_{\mathcal{W}}(\mathbb{P}^*, \mathbb{P}^{G(Z)}) \\
&\le 2 \sup_{W\in\mathrm{Lip}(1)\circ\mathcal{W}} \mathbb{E}[W(X) - W(\hat{X})] + \mathsf{V}_{\mathcal{W}}(\mathbb{P}^*, \mathbb{P}^{G(Z)}).
\end{aligned}
$$

The bound for $\hat{G}_n$ from (8) follows since $G \in \mathcal{G}$ was arbitrary. □

Proof of Lemma 2.2. Let $(G_n)_{n\in\mathbb{N}} \in \mathcal{G}$ be a sequence that converges to $G \in \mathcal{G}$. If $\mathsf{V}_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G(Z)}) \ge \mathsf{V}_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G_n(Z)})$, then

$$
\begin{aligned}
&\mathsf{V}_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G(Z)}) - \mathsf{V}_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G_n(Z)}) \\
&\le \sup_{W\in\mathrm{Lip}(L)} \mathbb{E}\Big[\log\Big(\frac{1+e^{-W(G_n(Z))}}{2}\Big) - \log\Big(\frac{1+e^{-W(G(Z))}}{2}\Big)\Big] \\
&\le \sup_{W\in\mathrm{Lip}(L)} \mathbb{E}\Big[W(G_n(Z)) - W(G(Z))\Big] \\
&\le L\|G_n - G\|_\infty.
\end{aligned}
$$

The case $\mathsf{V}_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G(Z)}) < \mathsf{V}_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G_n(Z)})$ can be bounded analogously. Therefore, T is continuous and there is at least one minimizer if $\mathcal{G}$ is compact. □

A.2. Proofs for Section 3

Before we prove the main results from Section 3 we require an auxiliary lemma:

Lemma A.1. *For $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$ and an arbitrary set of measurable functions $\mathcal{W}$ we have that*

$$
\mathsf{V}_{\mathcal{W}}(\mathbb{P}, \mathbb{Q}) \le \sup_{W\in\mathcal{W}} \mathbb{E}[-\log\big(1+e^{-W(X)}\big) + \log\big(1+e^{-W(Y)}\big)].
$$

Proof. Since

$$
\log(1+e^{x}) + \log(1+e^{-x}) \ge \log(4) \quad \text{for all } x \in \mathbb{R},
$$

we can bound

$$
\begin{aligned}
&\sup_{W\in\mathcal{W}} \mathbb{E}\Big[-\log\big(1+e^{-W(X)}\big) - \log\big(1+e^{W(Y)}\big)\Big] \\
&\quad = \sup_{W\in\mathcal{W}} \mathbb{E}[-\log\big(1+e^{-W(X)}\big) + \log\big(1+e^{-W(Y)}\big) \\
&\qquad - \log\big(1+e^{-W(Y)}\big) - \log\big(1+e^{W(Y)}\big)] \\
&\quad \le \sup_{W\in\mathcal{W}} \mathbb{E}[-\log\big(1+e^{-W(X)}\big) + \log\big(1+e^{-W(Y)}\big)] \\
&\qquad - \inf_{W\in\mathcal{W}} \mathbb{E}[\log\big(1+e^{-W(Y)}\big) + \log\big(1+e^{W(Y)}\big)] \\
&\quad \le \sup_{W\in\mathcal{W}} \mathbb{E}[-\log\big(1+e^{-W(X)}\big) + \log\big(1+e^{-W(Y)}\big)] - \log(4). \quad \square
\end{aligned}
$$

Proof of Theorem 3.1. Defining

$$
\psi : \mathbb{R} \to \mathbb{R},\ \psi(x) := -\log\Big(\frac{1+e^{-x}}{2}\Big),
$$

we can rewrite

$$
\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) = \sup_{W\in\mathrm{Lip}(L,B)} \mathbb{E}[\psi(W(X)) + \psi(-W(Y))].
$$

The function $f : [-\log(2-2/L), \infty) \to \mathbb{R},\ f(x) = \log(2e^x - 1)$ is bijective and Lipschitz continuous with Lipschitz constant L and satisfies $\psi(-f(x)) = x$ for all $x \ge -\log(2-2/L)$. Therefore, we obtain a lower bound

$$
\begin{aligned}
\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) &\ge \sup_{\substack{W\in\mathrm{Lip}(1,\log((1+e^B)/2)) \\ W(\cdot)\ge -\log(2-2/L)}} \mathbb{E}[\psi(f(W(X))) + \psi(-f(W(Y)))] \\
&= \sup_{\substack{W\in\mathrm{Lip}(1,B') \\ W(\cdot)\ge -\log(2-2/L)}} \mathbb{E}[\psi(f(W(X))) - W(Y)].
\end{aligned}
$$

Since $f^{-1} \in \mathrm{Lip}(1, \mathbb{R})$, we can estimate $\mathsf{V}_{\mathrm{Lip}(L,B)}$ from above by

$$
\begin{aligned}
\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) &= \sup_{W\in\mathrm{Lip}(L,B)} \mathbb{E}[\psi(f(f^{-1}(W(X)))) + \psi(-f(f^{-1}(W(Y))))] \\
&\le \sup_{\substack{W\in\mathrm{Lip}(L,B) \\ W(\cdot)>-\log(2)}} \mathbb{E}[\psi(f(W(X))) + \psi(-f(W(Y)))] \\
&= \sup_{\substack{W\in\mathrm{Lip}(L,B) \\ W(\cdot)>-\log(2)}} \mathbb{E}[\psi(f(W(X))) - W(Y)].
\end{aligned}
$$

A Taylor approximation at zero of the function $\psi\circ f(x) = \log(2-e^{-x})$ yields that for every $x \in (-\log(2), \infty)$ there exists a ξ between x and 0 such that

$$
\psi\circ f(x) = x - \frac{e^\xi}{(2e^\xi - 1)^2} x^2.
$$

For the lower bound, we thus conclude

$$
\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) \ge \sup_{\substack{W\in\mathrm{Lip}(1,B') \\ W(\cdot)\ge -\log(2-2/L)}} \mathbb{E}[W(X) - W(Y)] - \frac{L(L-1)}{2}\mathbb{E}[W(X)^2].
$$

For the upper bound, we get

$$
\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) \le \sup_{\substack{W\in\mathrm{Lip}(L,B) \\ W(\cdot)>-\log(2)}} \mathbb{E}[W(X) - W(Y)] - \frac{e^B}{(2e^B - 1)^2}\mathbb{E}[W(X)^2]. \quad \square
$$

Note that, using the function $g : (-\infty, \log(2-2/L)) \to \mathbb{R},\ g(x) = -\log(2e^{-x} - 1)$, we obtain lower and upper bounds with a penalty term depending on $\mathbb{E}[W(Y)^2]$ instead of $\mathbb{E}[W(X)^2]$.

Proof of Theorem 3.2. We prove the lower bound first. Theorem 3.1 yields

$$\begin{aligned}\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P},\mathbb{Q}) &\geq \sup_{\substack{W\in\mathrm{Lip}(1,B')\\ W(\cdot)>-\log(2-2/L)}} \mathbb{E}[W(X)-W(Y)] - \frac{L(L-1)}{2}\mathbb{E}[W(X)^2]\\ &\geq \sup_{W\in\mathrm{Lip}(1,\log(2-2/L))} \mathbb{E}[W(X)-W(Y)] - \frac{L(L-1)}{2}\mathbb{E}[W(X)^2].\end{aligned}$$

Let $W^* \in \operatorname*{arg\,max}_{W\in\mathrm{Lip}(1,\log(2-2/L))} \mathbb{E}[W(X)-W(Y)]$. This element exists by Villani (2008, Theorem 5.10 (iii)). Then $\delta W^* \in \mathrm{Lip}(1,\log(2-2/L))$ for all $\delta\in(0,1]$ and we can conclude

$$\begin{aligned}&\sup_{W\in\mathrm{Lip}(1,\log(2-2/L))} \mathbb{E}[W(X)-W(Y)] - \frac{L(L-1)}{2}\mathbb{E}[W(X)^2]\\ &\quad\geq \sup_{\delta\in(0,1]}\Big\{\mathbb{E}[\delta W^*(X)-\delta W^*(Y)] - \frac{L(L-1)}{2}\mathbb{E}[(\delta W^*(X))^2]\Big\}\\ &\quad= \sup_{\delta\in(0,1]}\Big\{\delta\mathbb{E}[W^*(X)-W^*(Y)] - \delta^2\frac{L(L-1)}{2}\mathbb{E}[(W^*(X))^2]\Big\},\end{aligned}$$

which is independent from B. In case $\Delta := \mathbb{E}[W^*(X)-W^*(Y)] < L(L-1)\mathbb{E}[W^*(X)^2]$ we have for $\delta = \frac{\Delta}{\mathbb{E}[W^*(X)^2]L(L-1)} \in (0,1)$

$$\begin{aligned}&\sup_{W\in\mathrm{Lip}(1,\log(2-2/L))} \mathbb{E}[W(X)-W(Y)] - \frac{L(L-1)}{2}\mathbb{E}[W(X)^2]\\ &\quad\geq \frac{\Delta^2}{\mathbb{E}[W^*(X)^2]L(L-1)} - \frac{\Delta^2}{2\mathbb{E}[W^*(X)^2]L(L-1)}\\ &\quad= \frac{\Delta^2}{2\mathbb{E}[W^*(X)^2]L(L-1)}\\ &\quad\geq \frac{\Delta^2}{2\log(2-2/L)^2L(L-1)},\end{aligned}$$

where we used $|W^*(x)| \leq \log(2-2/L)$ in the last step. In case $\Delta \geq L(L-1)\mathbb{E}[W^*(X)^2]$ we obtain

$$\mathbb{E}[W^*(X)-W^*(Y)] - \frac{L(L-1)}{2}\mathbb{E}[W^*(X)^2] \geq \frac{1}{2}\mathbb{E}[W^*(X)-W^*(Y)].$$

Using the boundedness of $[0,1]^d$, we get

$$\begin{aligned}\Delta &= \sup_{W\in\mathrm{Lip}(1,\log(2-2/L))} \mathbb{E}[W(X)-W(Y)]\\ &\geq \sup_{W\in\mathrm{Lip}(\log(2-2/L)d^{-1/p},\infty)} \mathbb{E}[W(X)-W(Y)]\\ &= \frac{\log(2-2/L)}{d^{1/p}}\mathsf{W}_1(\mathbb{P},\mathbb{Q}).\end{aligned}$$

Hence we can conclude the claimed lower bound for

$$c_1 = \frac{1}{2}\frac{\log(2-2/L)}{d^{1/p}}, \quad c_2 = \frac{1}{2d^{2/p}L(L-1)}.$$

For the upper bound we use Lemma A.1 with $\mathcal{W} = \mathrm{Lip}(L)$. Since for $W\in\mathrm{Lip}(L)$ the function $-\log\big(1+e^{-W(\cdot)}\big) \in \mathrm{Lip}(L)$ we conclude

$$\begin{aligned}\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P},\mathbb{Q}) &\leq \sup_{W\in\mathrm{Lip}(L)} \mathbb{E}[\psi(W(X))+\psi(W(Y))]\\ &\leq \sup_{W\in\mathrm{Lip}(L)} \mathbb{E}[-\log\big(1+e^{-W(X)}\big)+\log\big(1+e^{-W(Y)}\big)]\\ &\leq \sup_{W\in\mathrm{Lip}(L)} \mathbb{E}[W(X)-W(Y)]\\ &= L\sup_{W\in\mathrm{Lip}(1)} \mathbb{E}[W(X)-W(Y)]. \quad \square\end{aligned}$$

A.3. Proofs for Section 4

Proof of Theorem 4.1. Using Theorem 3.2 and the triangle inequality for the Wasserstein distance, we deduce for every $G\in\mathcal{G}$ and $c = \max(c_1^{-1}, c_2^{-1/2})$ that

$$\begin{aligned}\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{\hat{G}_n(Z)}) &\leq \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}_n) + \mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^{\hat{G}_n(Z)})\\ &\leq \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}_n) + c\big[\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}_n,\mathbb{P}^{\hat{G}_n(Z)})\big]^{1;1/2}\\ &\leq \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}_n) + c\big[\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}_n,\mathbb{P}^{G(Z)})\big]^{1;1/2}\\ &\leq \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}_n) + cL\big[\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^{G(Z)})\big]^{1;1/2}\\ &\leq (1+cL)\big[\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}_n)\big]^{1;1/2} + cL\big[\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G(Z)})\big]^{1;1/2}.\end{aligned}$$

As $G\in\mathcal{G}$ was arbitrary, we can choose the infimum over $\mathcal{G}$. $\square$

Proof of Corollary 4.2. For every measurable $G^*:\mathcal{Z}\to\mathcal{X}$ and any $G\in\mathcal{G}$ we have

$$\begin{aligned}\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{\hat{G}_n(Z)}) &\leq \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G^*(Z)}) + \mathsf{W}_1(\mathbb{P}^{G^*(Z)},\mathbb{P}^{G(Z)})\\ &= \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G^*(Z)}) + \sup_{W\in\mathrm{Lip}(1)} \mathbb{E}[W(G^*(Z))-W(G(Z))]\\ &\leq \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G^*(Z)}) + \mathbb{E}[|G^*(Z)-G(Z)|_p]\\ &\leq \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G^*(Z)}) + \|G^*-G\|_\infty.\end{aligned}$$

Since G^* was arbitrary Theorem 4.1 yields for some constant c

$$\begin{aligned}\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G(Z)})] \leq{}& c\cdot\mathbb{E}[\max(\sqrt{\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*)},\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*))]\\ &+ c\cdot\inf_{G^*:\mathcal{Z}\to\mathcal{X}}\Big\{[\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G^*(Z)})]^{1;1/2} + [\inf_{G\in\mathcal{G}}\|G-G^*\|_\infty^{1;1/2}]\Big\}\end{aligned}$$

Here the infimum can be used as we can increase the constant c multiplied to both terms by an arbitrary small $\varepsilon>0$ to account for the possibly infinitesimal smaller value. Using Jensen's inequality, we can bound the stochastic error term by

$$\begin{aligned}\mathbb{E}[\max(\sqrt{\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*)},\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*))] &\leq \mathbb{E}[\sqrt{\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*)}] + \mathbb{E}[\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*)]\\ &\leq \sqrt{\mathbb{E}[\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*)]} + \mathbb{E}[\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*)].\end{aligned}$$

From Schreuder (2020, Theorem 4) we know

$$\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}_n)] \leq c'\begin{cases}n^{-1/d}, & d>2\\ n^{-1/2}\log(n), & d=2\\ n^{-1/2}, & d=1.\end{cases}$$

where c depends only on d. Since $(\log n)/\sqrt{n}\leq 1$, we conclude

$$\sqrt{\mathbb{E}[\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*)]} + \mathbb{E}[\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*)] \leq 2c'\begin{cases}n^{-1/2d}, & d>2\\ n^{-1/4}(\log n)^{1/2}, & d=2\\ n^{-1/4}, & d=1.\end{cases}\quad\square$$

Proof of Theorem 4.3. With the same reasoning as in the proof of Corollary 4.2, there exists some c such that for any measurable $G^*:\mathcal{Z}\to\mathcal{X}$ and any $G\in\mathcal{G}$

$$\begin{aligned}\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G(Z)})] \leq c\Big(&\sqrt{\mathbb{E}[\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*)]} + \mathbb{E}[\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*)]\\ &+ [\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G^*(Z)})]^{1;1/2} + [\inf_{G\in\mathcal{G}}\|G^*-G\|_\infty]^{1,1/2}\Big)\end{aligned}$$

By the triangle inequality

$$\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^*) \leq \mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^{G^*(Z)}) + \mathsf{W}_1(\mathbb{P}^{G^*(Z)},\mathbb{P}^*).$$

Let $Z_i\sim\mathbb{U}$ be i.i.d. random variables and denote the corresponding empirical measure by $\mathbb{U}_n$. For $G^*\in\mathrm{Lip}(M,\mathcal{Z})$ we can then bound the first term by

$$\begin{aligned}\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^{G^*(Z)}) &= \sup_{W\in\mathrm{Lip}(1)}\frac{1}{n}\sum_{i=1}^n W(X_i) - \mathbb{E}[W\circ G^*(Z)]\\ &\leq \sup_{W\in\mathrm{Lip}(1)}\frac{1}{n}\sum_{i=1}^n |W(X_i) - W\circ G^*(Z_i)|\\ &\quad+ \sup_{W\in\mathrm{Lip}(1)}\frac{1}{n}\sum_{i=1}^n W\circ G^*(Z_i) - \mathbb{E}[W\circ G^*(Z)]\\ &\leq \frac{1}{n}\sum_{i=1}^n |X_i - G^*(Z_i)|_p + \sup_{f\in\mathrm{Lip}(M)}\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z)]\\ &= \frac{1}{n}\sum_{i=1}^n |X_i - G^*(Z_i)|_p + M\cdot\mathsf{W}_1(\mathbb{U}_n,\mathbb{U})\end{aligned} \tag{13}$$

Hence,

$$\mathbb{E}\big[\mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^{G^*(Z)})\big] \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[|X_i - G^*(Z_i)|_p] + M\cdot\mathbb{E}[W_1(\mathbb{U}_n,\mathbb{U})].$$

Note that $\mathbb{E}[|X_i - G^*(Z_i)|_p] = \mathsf{W}_1(\mathbb{P}^{G^*(Z)}, \mathbb{P}^*)$ by the duality formula of W_1 used in this work, see Villani (2008, Definition 6.2 and Remark 6.5). For $\mathbb{E}[W_1(\mathbb{U}_n, \mathbb{U})]$, we can exploit the convergence rate for the empirical distribution as in Corollary 4.2, but now in the d^*-dimensional latent space $\mathcal{Z}$. Therefore, there exists a c' such that

$$\sqrt{\mathbb{E}[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)]} + \mathbb{E}[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)] \le c'[\mathsf{W}_1(\mathbb{P}^{G^*(Z)}, \mathbb{P}^*)]^{1;1/2} + c' \begin{cases} n^{-1/2d^*}, & d^* > 2 \\ n^{-1/4}(\log n)^{1/2}, & d^* = 2 \\ n^{-1/4}, & d^* = 1. \end{cases} \quad \square$$

A.4. Proofs of Theorems 5.1 and 5.2

Proof of Theorem 5.1. First, we verify that for any two nonempty sets $\mathcal{W}_1$ and $\mathcal{W}_2$ we have

$$\mathsf{V}_{\mathcal{W}_1}(\mathbb{P}, \mathbb{Q}) \le \mathsf{V}_{\mathcal{W}_2}(\mathbb{P}, \mathbb{Q}) + 2 \inf_{W \in \mathcal{W}_2} \sup_{W^* \in \mathcal{W}_1} \|W - W^*\|_\infty. \tag{14}$$

Indeed, the difference $\mathsf{V}_{\mathcal{W}_1}(\mathbb{P}, \mathbb{Q}) - \mathsf{V}_{\mathcal{W}_2}(\mathbb{P}, \mathbb{Q})$ is bounded by

$$\begin{aligned}
&\inf_{W \in \mathcal{W}_2} \sup_{W^* \in \mathcal{W}_1} \Big\{ \mathbb{E}\Big[-\log\Big(\frac{1+e^{-W^*(X)}}{2}\Big)\Big] - \log\Big(\frac{1+e^{W^*(Y)}}{2}\Big) \\
&\qquad - \mathbb{E}\Big[-\log\Big(\frac{1+e^{-W(X)}}{2}\Big)\Big] - \log\Big(\frac{1+e^{W(Y)}}{2}\Big)\Big] \Big\} \\
&\le \inf_{W \in \mathcal{W}_2} \sup_{W^* \in \mathcal{W}_1} \Big\{ \mathbb{E}\Big[\Big|-\log\Big(\frac{1+e^{-W^*(X)}}{2}\Big) + \log\Big(\frac{1+e^{-W(X)}}{2}\Big)\Big|\Big] \\
&\qquad + \mathbb{E}\Big[\Big|-\log\Big(\frac{1+e^{W^*(Y)}}{2}\Big) + \log\Big(\frac{1+e^{W(Y)}}{2}\Big)\Big|\Big] \Big\} \\
&\le \inf_{W \in \mathcal{W}_2} \sup_{W^* \in \mathcal{W}_1} \big\{ \mathbb{E}[|W^*(X) - W(X)|] + \mathbb{E}[|W^*(Y) - W(Y)|] \big\} \\
&\le 2 \inf_{W \in \mathcal{W}_2} \sup_{W^* \in \mathcal{W}_1} \|W^* - W\|_\infty,
\end{aligned}$$

due to Lipschitz continuity of $x \mapsto -\log((1+e^x)/2)$.

From (14) we deduce for $\mathcal{W} \subset \mathrm{Lip}(L, B)$

$$\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) \le \mathsf{V}_{\mathcal{W}}(\mathbb{P}, \mathbb{Q}) + 2 \inf_{W' \in \mathcal{W}} \sup_{W \in \mathrm{Lip}(L,B)} \|W - W'\|_\infty.$$

We abbreviate $\Delta_{\mathcal{W}} := \inf_{W' \in \mathcal{W}} \sup_{W \in \mathrm{Lip}(L,B)} \|W - W'\|_\infty$. Now we can proceed as in Theorem 4.1. In particular, it is sufficient to bound $\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})$. Due to Theorem 3.2 there is some constant $c > 0$ such that for every $G \in \mathcal{G}$

$$\begin{aligned}
\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) &\le c[\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})]^{1;1/2} \\
&\le c[\mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) + 2\Delta_{\mathcal{W}}]^{1;1/2} \\
&\le c[\mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})]^{1;1/2} + 2c[\Delta_{\mathcal{W}}]^{1;1/2} \\
&\le c[\mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)})]^{1;1/2} + 2c[\Delta_{\mathcal{W}}]^{1;1/2}
\end{aligned}$$

Because $\mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)}) \le \mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}_n, \mathbb{P}^{G(Z)})$ due to $\mathcal{W} \subset \mathrm{Lip}(L, B)$, the rest of the proof is identical to the proof of Theorem 4.1. $\square$

Proof of Theorem 5.2. Since $\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)$ can be estimated as in Theorem 4.3, we only need to bound $\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})$. For $\Gamma > \max(L, 2B)$, we have $\mathrm{Lip}(L, B) \subset \mathcal{H}^\alpha(\Gamma)$, $\alpha \in (0, 1)$, and the assumptions of Theorem 3.2 are satisfied. Therefore for every $\alpha \in (0, 1)$

$$\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) \le c\big[\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})\big]^{1;1/2} \le c\big[\mathsf{V}_{\mathcal{H}^\alpha(\Gamma)}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})\big]^{1;1/2}.$$

Now, (14) yields

$$\mathsf{V}_{\mathcal{H}^\alpha(\Gamma)}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) \le \mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) + 2\Delta_{\mathcal{W}} \quad \text{for} \quad \Delta_{\mathcal{W}} := \inf_{W \in \mathcal{W}} \sup_{W^* \in \mathcal{H}^\alpha(\Gamma)} \|W^* - W\|_\infty.$$

Using that $\hat{G}_n$ is the empirical risk minimizer and $\mathcal{W} \subseteq \mathcal{H}^\alpha(\Gamma)$, we thus have

$$\begin{aligned}
\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) &\le c\big[\mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})\big]^{1;1/2} + c[\Delta_{\mathcal{W}}]^{1;1/2} \\
&\le c\big[\mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)})\big]^{1;1/2} + c[\Delta_{\mathcal{W}}]^{1;1/2} \\
&\le c\big[\mathsf{V}_{\mathcal{H}^\alpha(\Gamma)}(\mathbb{P}_n, \mathbb{P}^{G(Z)})\big]^{1;1/2} + c[\Delta_{\mathcal{W}}]^{1;1/2}
\end{aligned}$$

To bound the first term, we apply Lemma A.1 and $\{-\log(1 + e^{-W(\cdot)}) \mid W \in \mathcal{H}^\alpha(\Gamma)\} \subset \mathcal{H}^\alpha(\Gamma)$ to obtain

$$\begin{aligned}
\mathsf{V}_{\mathcal{H}^\alpha(\Gamma)}(\mathbb{P}_n, \mathbb{P}^{G(Z)}) &\le \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{\hat{X} \sim \mathbb{P}_n}\Big[-\log\Big(1 + e^{-W(\hat{X})}\Big) + \log\Big(1 + e^{-W(G(Z))}\Big)\Big] \\
&\le \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{\hat{X} \sim \mathbb{P}_n}[W(\hat{X}) - W(G(Z))] \qquad (15) \\
&\le \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{\hat{X} \sim \mathbb{P}_n}[W(\hat{X}) - W(X)] + \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}[W(X) - W(G(Z))].
\end{aligned}$$

For the second term we have by Hölder continuity, Jensens inequality and the duality formula of W_1 as used in the proof of Theorem 4.3 that

$$\begin{aligned}
\sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}[W(X) - W(G(Z))] &\le \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}[|W(X) - W(G^*(Z))|] \\
&\quad + \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}[|W(G^*(Z)) - W(G(Z))|] \\
&\le \Gamma \mathbb{E}[|X - G^*(Z)|_p^\alpha] + \Gamma \|G^* - G\|_\infty^\alpha \\
&\le \Gamma \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})^\alpha + \Gamma \|G^* - G\|_\infty^\alpha.
\end{aligned}$$

Hence, we have for any $G \in \mathcal{G}$ and any measurable $G^* : \mathcal{Z} \to \mathcal{X}$ for some constant $c > 0$

$$\begin{aligned}
\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) &\le c\Big[\sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{\hat{X} \sim \mathbb{P}_n}[W(\hat{X}) - W(X)]\Big]^{1;1/2} \\
&\quad + c\Big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})^\alpha + \|G^* - G\|_\infty^\alpha\Big]^{1;1/2} + c[\Delta_{\mathcal{W}}]^{1;1/2}.
\end{aligned}$$

For the remaining stochastic error term, we first note that

$$\begin{aligned}
\sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{\hat{X} \sim \mathbb{P}_n}[W(\hat{X}) - W(X)] &\le \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{X_n \sim \mathbb{P}_n}[W(X_n) - W(G^*(Z))] \\
&\quad + \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}[W(G^*(Z)) - W(X)] \\
&\le \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{X_n \sim \mathbb{P}_n}[W(X_n) - W(G^*(Z))] \\
&\quad + \Gamma \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})^\alpha
\end{aligned}$$

and as in (13) together with Schreuder (2020, Theorem 4) we obtain

$$\begin{aligned}
\mathbb{E}\Big[\sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{X_n \sim \mathbb{P}_n}[W(X_n) - W(G^*(Z))]\Big] &\le \mathbb{E}\Big[\sup_{W \in \mathcal{H}^\alpha(\Gamma)} |X - G^*(Z)|_p^\alpha\Big] \\
&\quad + \mathbb{E}\Big[\sup_{f \in \mathcal{H}^\alpha(M \cdot \Gamma)} \frac{1}{n} \sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z)]\Big] \\
&\le c\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})^\alpha + c \begin{cases} n^{-\alpha/d^*}, & 2\alpha < d^*, \\ n^{-1/2}\ln(n), & 2\alpha = d^*, \\ n^{-1/2}, & 2\alpha > d^*. \end{cases}
\end{aligned}$$

For the expectation of the first term we use Jensen's inequality

$$\mathbb{E}[|X_i - G^*(Z_i)|_p^\alpha] \le \mathbb{E}[|X_i - G^*(Z_i)|_p]^\alpha = \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})^\alpha. \quad \square$$

A.5. Proof of Theorem 5.3

To prove Theorem 5.3 some additional notation is required. The set of locally integrable functions is given by

$$L^1_{\mathrm{loc}}(\Omega) := \{f : \Omega \to \mathbb{R} \Big| \int_K |f(x)|\, dx < \infty, \text{ for all compact } K \subset \Omega^\circ\}.$$

A function $f \in L^1_{\mathrm{loc}}(\Omega)$ has a weak derivative, $D^\alpha_w f$, provided there exists a function $g \in L^1_{\mathrm{loc}}(\Omega)$ such that

$$\int_\Omega g(x)\phi(x)dx = (-1)^{|\alpha|} \int_\Omega f(x)\phi^{(\alpha)}(x)dx \quad \text{for all } \phi \in C^\infty(\Omega) \text{ with compact support.}$$

If such a g exists, we define $D^\alpha_w f := g$. For $f \in L^1_{\mathrm{loc}}(\Omega)$ and $k \in \mathbb{N}_0$ the Sobolev norm is

$$\|f\|_{W^{k,\infty}(\Omega)} := \max_{|\alpha| \le k} \big\|D^\alpha_w f\big\|_{\infty,\Omega}.$$

The Sobolev space $W^{k,\infty}(\Omega) := \{f \in L^1_{\mathrm{loc}}(\Omega) : \|f\|_{W^k_p(\Omega)} < \infty\}$ is a Banach space (Brenner & Scott, 2008, Theorem 1.3.2). For $f \in W^{k,\infty}(\Omega)$, define the Sobolev semi norm by

$$|f|_{W^{k,\infty}(\Omega)} := \max_{|\alpha| = k} \big\|D^\alpha_w f\big\|_{\infty,\Omega}.$$

Note that $\mathrm{Lip}(L,B,\Omega)\subset W^{1,\infty}(\Omega)$, since $\|f\|_{W^{1,\infty}}\le\max(L,B)$ for any $f\in\mathrm{Lip}(L,B,\Omega)$. For two normed spaces $(A,\|\cdot\|_A),(B,\|\cdot\|_B)$ we denote the operator norm of a linear operator $T:A\to B$ by

$$\|T\| := \sup\big\{\|Tx\|_B \mid x\in A,\|x\|_A\le 1\big\}.$$

Theorem 5.3 is very close to by Gühring et al. (2020, Theorem 4.1), which however applies only to functions f which are at least twice (weakly) differentiable. Our proof can thus build on numerous auxiliary results and arguments from Gühring et al. (2020). We basically keep the proof structure of Gühring et al. (2020) which in turn relies on Yarotsky (2017).

Let $d,N\in\mathbb{N}$. For $m\in\{0,\dots,N\}^d$, define the functions $\phi_m:\mathbb{R}^d\to\mathbb{R}$,

$$\phi_m(x)=\prod_{\ell=1}^{d}\psi\Big(3N\Big(x_\ell-\frac{m_\ell}{N}\Big)\Big),\quad\text{where}\quad \psi(x)=\begin{cases}1, & |x|<1,\\ 0, & |x|>2,\\ 2-|x|, & 1\le|x|\le 2\end{cases}$$

By definition, we have $\|\phi_m\|_\infty=1$ for all m and

$$\operatorname{supp}\phi_m\subset\Big\{x:\Big|x_k-\frac{m_k}{N}\Big|<\frac{1}{N}\forall k\Big\}=:B_{\frac1N,|\cdot|_\infty}(\tfrac{m}{N}). \tag{16}$$

Gühring et al. (2020, Lemma C.3 (iv)) have verified that $\|\phi_m\|_{W^{1,\infty}(\mathbb{R}^d)}\le cN$ for some constant $c>0$.

A direct consequence of Lemma 2.11, Lemma C.3, Lemma C.5 and Lemma C.6 by Gühring et al. (2020) is the following approximation result for the localizing functions ϕ_m via ReLU networks:

Lemma A.2. *For any $\varepsilon\in(0,1/2)$ and any $m\in\{0,\dots,N\}^d$ there is a network Ψ_ε with ReLU activation function, not more than $C_1\log_2(\varepsilon^{-1})$ layers and no more than $C_2(N+1)^d\log_2^2(\varepsilon^{-1})$ nonzero weights and no more than neurons $C_3(N+1)^d(\log_2^2(\varepsilon^{-1})\vee\log_2(\varepsilon^{-1}))$ such that for $k\in\{0,1\}$*

$$\|\Psi_\varepsilon-\phi_m\|_{W^{k,\infty}}\le cN^k\varepsilon,$$

where C_1,C_2,C_3 and c are constants independent of m and ε. Additionally,

$$\phi_m(x)=0\Longrightarrow\Psi_\varepsilon(x)=0,$$

and therefore $\operatorname{supp}\Psi_\varepsilon\subset B_{\frac1N,|\cdot|_\infty}(\frac{m}{N})$.

Next we approximate a bounded Lipschitz function using linear combinations of the set $\{\phi_m:m\in\{1,\dots,N\}^d\}$. The approximation error will be measured in the Hölder norm from (4).

Lemma A.3. *Let $0<\alpha<1$. There exists a constant $C_1>0$ such that for any $f\in W^{1,\infty}((0,1)^d)$ there are constants $c_{f,m}$ for $m\in\{0,\dots,N\}^d$ such that*

$$\Big\|f-\sum_{m\in\{0,\dots,N\}^d}c_{f,m}\phi_m\Big\|_{\mathcal{H}^\alpha}\le C_1\Big(\frac1N\Big)^{1-\alpha}\|f\|_{W^{1,\infty}}.$$

The coefficients satisfy for a $C_2>0$

$$|c_{f,m}|\le C_2\|\tilde f\|_{W^{1,\infty}(\Omega_{m,N})},$$

where $\Omega_{m,N}:=B_{\frac1N,|\cdot|_\infty}(\frac{m}{N})$ and $\tilde f\in W^{1,\infty}(\mathbb{R})$ is an extension of f.

Proof. Let $E:W^{1,\infty}((0,1)^d)\to W^{1,\infty}(\mathbb{R})$ be the continuous linear extension operator from Stein (1970, Theorem 5) and set $\tilde f:=Ef$. As E is continuous there exists a $C_E>0$ such that

$$\|\tilde f\|_{W^{1,\infty}(\mathbb{R}^d)}\le C_E\|f\|_{W^{1,\infty}}.$$

Step 1 (Choice of $c_{f,m}$): For each $m\in\{0,\dots,N\}^d$ we define

$$c_{f,m}=\int_{B_{m,N}}\tilde f(y)\rho(y)\,\mathrm{d}y\ \text{for}\ B_{m,N}:=B_{\frac{3}{4N},|\cdot|}\Big(\frac{m}{N}\Big)$$

and an arbitrary cut-off function ρ supported in $B_{m,N}$, i.e.

$\rho\in C_c^\infty(\mathbb{R}^d)$ with $\rho(x)\ge0$ for all $x\in\mathbb{R}^d$, $\operatorname{supp}\rho=B_{m,N}$ and $\int_{\mathbb{R}^n}\rho(x)dx=1$.

Then

$$\begin{aligned}|c_{m,f}|&=\Big|\int_{B_{m,N}}\tilde f(y)\rho(y)\,\mathrm{d}y\Big|\le\|\tilde f\|_{\infty,\Omega_{m,N}}\int_{B_{m,N}}\rho(y)\,\mathrm{d}y\\&=\|\tilde f\|_{\infty,\Omega_{m,N}}\le C_E\|f\|_{W^{1,\infty}(\Omega_{m,N})}.\end{aligned}$$

Step 2 (Local estimates in $\|\cdot\|_{W^{k,p}}$): The coefficients $c_{m,f}$ are the averaged Taylor polynomials in the sense of Brenner and Scott (2008, Definition 4.1.3) of order 1 averaged over $B_{m,N}$. As Gühring et al. (2020, Proof of Lemma C.4, Step 2) showed, the conditions of the Bramble-Hilbert-Lemma (Brenner & Scott, 2008, Theorem 4.3.8) are satisfied. Hence for $k\in\{0,1\}$

$$|\tilde f-c_{m,f}|_{W^{k,\infty}(\Omega_{m,N})}\le C_1\Big(\frac{2\sqrt d}{N}\Big)^{1-k}|\tilde f|_{W^{1,\infty}(\Omega_{m,n})}\le C_2\Big(\frac1N\Big)^{1-k}\|\tilde f\|_{W^{1,\infty}(\Omega_{m,n})}.$$

Now using ϕ_m as defined above, we get

$$\Big\|\phi_m\Big(\tilde f-c_{f,m}\Big)\Big\|_{\infty,\Omega_{m,N}}\le\|\phi_m\|_{\infty,\Omega_{m,N}}\cdot\Big\|\tilde f-c_{f,m}\Big\|_{\infty,\Omega_{m,N}}\le C_2\frac1N\|\tilde f\|_{W^{1,\infty}(\Omega_{m,N})} \tag{17}$$

Due to the product inequality for weak derivatives (Gühring et al., 2020, Lemma B.6) there is a constant $C'>0$ such that the supremum norm of the weak derivative is bounded by

$$\begin{aligned}\Big|\phi_m\Big(\tilde f-c_{f,m}\Big)\Big|_{W^{1,\infty}(\Omega_{m,N})}\le&C'|\phi_m|_{W^{1,\infty}(\Omega_{m,N})}\cdot\Big\|\tilde f-c_{f,m}\Big\|_{\infty,\Omega_{m,N}}\\&+C'\|\phi_m\|_{\infty,\Omega_{m,N}}\cdot|\tilde f-c_{f,m}|_{W^{1,\infty}(\Omega_{m,N})}\\\le&C'\cdot cN\cdot C_2\frac1N\|\tilde f\|_{W^{1,\infty}(\Omega_{m,N})}+C'\cdot C_3\|\tilde f\|_{W^{1,\infty}(\Omega_{m,N})}\\=&C_4\|\tilde f\|_{W^{1,\infty}(\Omega_{m,N})}.\end{aligned} \tag{18}$$

Combining (17) and (18) we get

$$\Big\|\phi_m\Big(\tilde f-c_{f,m}\Big)\Big\|_{W^{1,\infty}(\Omega_{m,N})}\le C_5\|\tilde f\|_{W^{1,\infty}(\Omega_{m,N})}.$$

Step 3 (Global estimate in $\|\cdot\|_{W^{k,p}}$): As $\sum_{m\in\{0,\dots,N\}^d}\phi_m=1$, we have that

$$\tilde f(x)=\sum_{m\in\{0,\dots,N\}^d}\phi_m(x)\tilde f(x),\qquad\text{for a.e. } x\in(0,1)^d.$$

As $\tilde f\big|_{(0,1)^d}=f$ we have for $k\in\{0,1\}$

$$\begin{aligned}\Big\|f-\sum_{m\in\{0,\dots,N\}^d}\phi_mc_{f,m}\Big\|_{W^{k,\infty}((0,1)^d)}&=\Big\|\tilde f-\sum_{m\in\{0,\dots,N\}^d}\phi_mc_{f,m}\Big\|_{W^{k,\infty}((0,1)^d)}\\&=\Big\|\sum_{m\in\{0,\dots,N\}^d}\phi_m\big(\tilde f-c_{f,m}\big)\Big\|_{W^{k,\infty}((0,1)^d)}\\&\le\sup_{\widetilde m\in\{0,\dots,N\}^d}\Big\|\sum_{m\in\{0,\dots,N\}^d}\phi_m\big(\tilde f-c_{f,m}\big)\Big\|_{W^{k,\infty}(\Omega_{\widetilde m,N})}\end{aligned} \tag{19}$$

where the last step follows from $(0,1)^d\subset\bigcup_{\widetilde m\in\{0,\dots,N\}^d}\Omega_{\widetilde m,N}$. Now we obtain for each $\widetilde m\in\{0,\dots,N\}^d$ using (16), (17) and (18)

$$\begin{aligned}\Big\|\sum_{m\in\{0,\dots,N\}^d}\phi_m\big(\tilde f-c_{f,m}\big)\Big\|_{W^{k,\infty}(\Omega_{\widetilde m,N})}&\le\sup_{\substack{m\in\{0,\dots,N\}^d\\|m-\widetilde m|_\infty\le1}}\Big\|\phi_m\big(\tilde f-c_{f,m}\big)\Big\|_{W^{k,\infty}(\Omega_{m,N})}\\&\le\sup_{\substack{m\in\{0,\dots,N\}^d\\|m-\widetilde m|_\infty\le1}}\Big\|\phi_m\big(\tilde f-c_{f,m}\big)\Big\|_{W^{k,\infty}(\Omega_{m,N})}\\&\le C_6\Big(\frac1N\Big)^{1-k}\sup_{\substack{m\in\{0,\dots,N\}^d,\\|m-\widetilde m|_\infty\le1}}\|\tilde f\|_{W^{1,\infty}(\Omega_{m,N})}.\end{aligned}$$

Plugging this into (19), we obtain for $k\in\{0,1\}$

$$\begin{aligned}\Big\|f-\sum_{m\in\{0,\dots,N\}^d}\phi_mc_{f,m}\Big\|_{W^{k,\infty}((0,1)^d)}&\le C_6\Big(\frac1N\Big)^{(1-k)}\sup_{\widetilde m\in\{0,\dots,N\}^d}\Big(\sup_{\substack{m\in\{0,\dots,N\}^d,\\|m-\widetilde m|_\infty\le1}}\|\tilde f\|_{W^{1,\infty}(\Omega_{m,N})}\Big)\\&\le C_7\Big(\frac1N\Big)^{(1-k)}\sup_{\widetilde m\in\{0,\dots,N\}^d}\|\tilde f\|_{W^{1,\infty}(\Omega_{\widetilde m,N})}\\&\le C_8\Big(\frac1N\Big)^{(1-k)}\|\tilde f\|_{W^{1,\infty}(\mathbb{R}^d)}\\&\le C_9\Big(\frac1N\Big)^{(1-k)}\|f\|_{W^{1,\infty}((0,1)^d)}.\end{aligned} \tag{20}$$

Step 4 (Interpolation): Define the linear operators $T_0 : W^{1,\infty}((0,1)^d) \to L^\infty((0,1)^d), T_\alpha : W^{1,\infty}((0,1)^d) \to \mathcal{H}^\alpha((0,1)^d)$ and $T_1 : W^{1,\infty}((0,1)^d) \to W^{1,\infty}((0,1)^d)$ via

$$T_k(f) = f - \sum_{m\in\{0,\dots,N\}^d} \phi_m c_{f,m}, \quad k \in \{0,\alpha,1\}$$

Note that the linearity follows from the definition of the constants $c_{f,m}$. Using Lunardi (2018, Theorem 1.6), for the nontrivial interpolation couple see Lunardi (2018, p.11 f.), leads to

$$\|T_\alpha\| \le \|T_0\|^{1-\alpha}\|T_1\|^\alpha.$$

Note that $\|\cdot\|_{\mathcal{H}^\alpha}$ is equivalent to $\|\cdot\|_{W^{s,\infty}(\Omega)}$ in Gühring et al. (2020). Using (20) we conclude

$$\Big\|f - \sum_{m\in\{0,\dots,N\}^d} \phi_m c_{f,m}\Big\|_{\mathcal{H}^\alpha} \le C_{10}\Big(\frac{1}{N}\Big)^{1-\alpha}\|f\|_{W^{1,\infty}}. \quad \square$$

Now we want to approximate the function $\sum_{m\in\{0,\dots,N\}^d} c_{f,m}\phi_m$ in Hölder norm using a ReLU network.

Lemma A.4. *For any $\varepsilon \in (0,1/2)$ there is a neural network Φ_ε with ReLU activation function such that for $(c_{f,m})_m$ from Lemma A.3, there is a constant $C > 0$ such that*

$$\Big\|\sum_{m\in\{0,\dots,N\}^d} \phi_m c_{f,m} - \Phi_\varepsilon\Big\|_{\mathcal{H}^\alpha} \le C\|f\|_{W^{1,\infty}} N^\alpha \varepsilon,$$

the number of layers is at most $\lceil C_1 \log_2(\varepsilon^{-1})\rceil$, the number of nonzero weights is at most $\lceil C_2(N+1)^d \log_2^2(\varepsilon^{-1})\rceil$ and the number of neurons is at most $\lceil C_3(N+1)^d(\log_2^2(\varepsilon^{-1}) \vee \log_2(\varepsilon^{-1}))\rceil$, with C_1, C_2 and C_3 from Lemma A.2.

Proof. From Lemma A.2 we know that there are neural networks $\Psi_{\varepsilon,m}$ with at most $\lceil C_1 \log_2(\varepsilon^{-1})\rceil$ layers, $\lceil C_2(N+1)^d \log_2^2(\varepsilon^{-1})\rceil$ nonzero weights and $\lceil C_3(N+1)^d(\log_2^2(\varepsilon^{-1})\vee\log_2(\varepsilon^{-1}))\rceil$ neurons that approximate ϕ_m such that for $k \in \{0,1\}$

$$\|\phi_m - \Psi_{\varepsilon,m}\|_{W^{k,\infty}} \le c' N^k \varepsilon.$$

Now we parallelize these networks and multiply with the coefficients $c_{f,m}$ afterwards. Hereby, we construct a network Φ_ε with $1 + \lceil C_1 \log_2(\varepsilon^{-1})\rceil$ layers, $N^d + \lceil C_2(N+1)^d \log_2^2(\varepsilon^{-1})\rceil$ nonzero weights and $1 + \lceil C_3(N+1)^d(\log_2^2(\varepsilon^{-1}) \vee \log_2(\varepsilon^{-1}))\rceil$ neurons such that

$$\Phi_\varepsilon = \sum_{m\in\{0,\dots,N\}^d} c_{f,m}\Psi_{\varepsilon,m}. \tag{21}$$

For each $m \in \{0,\dots,N\}^d$ denote $\Omega_{m,N} = B_{\frac{1}{N},|\cdot|_\infty}(\frac{m}{N})$ as above. For $k \in \{0,1\}$ we get

$$\begin{aligned}
\Big\|\Phi_\varepsilon - \sum_{m\in\{0,\dots,N\}^d} c_{f,m}\phi_m\Big\|_{W^{k,\infty}((0,1)^d)} &= \Big\|\sum_{m\in\{0,\dots,N\}^d} c_{f,m}(\Psi_{\varepsilon,m} - \phi_m)\Big\|_{W^{k,\infty}((0,1)^d)} \\
&\le \sup_{\widetilde{m}\in\{0,\dots,N\}^d} \Big\|\sum_{m\in\{0,\dots,N\}^d} c_{f,m}(\Psi_{\varepsilon,m} - \phi_m)\Big\|_{W^{k,\infty}(\Omega_{\widetilde{m},N}\cap(0,1)^d)} \\
&\le 3^d \sup_{\widetilde{m}\in\{0,\dots,N\}^d}\sup_{m\in\{0,\dots,N\}^d} \Big\|c_{f,m}(\Psi_{\varepsilon,m} - \phi_m)\Big\|_{W^{k,\infty}(\Omega_{\widetilde{m},N}\cap(0,1)^d)} \\
&\le 3^d \sup_{\widetilde{m}\in\{0,\dots,N\}^d}\sup_{m\in\{0,\dots,N\}^d} |c_{f,m}|\Big\|(\Psi_{\varepsilon,m} - \phi_m)\Big\|_{W^{k,\infty}(\Omega_{\widetilde{m},N}\cap(0,1)^d)} \\
&\le 3^d \sup_{\widetilde{m}\in\{0,\dots,N\}^d}\sup_{m\in\{0,\dots,N\}^d} \|\tilde f\|_{W^{1,\infty}(\Omega_{m,N})}\|\Psi_{\varepsilon,m} - \phi_m\|_{W^{k,\infty}(\Omega_{\widetilde{m},N}\cap(0,1)^d)} \\
&\le CN^k\varepsilon\|f\|_{W^{1,\infty}}.
\end{aligned}$$

The second to last inequality follows from the fact that on $\Omega_{\widetilde{m},N}$ is within the support of ϕ_m only for $|m-\widetilde{m}|_\infty \le 1$. The last inequality follows from (21) and the continuity of the extension operator, see Stein (1970, Theorem 5). As in Step 4 of Lemma A.3, we conclude using Lunardi (2018, Theorem 1.6)

$$\Big\|\Phi_\varepsilon - \sum_{m\in\{0,\dots,N\}^d} \phi_m c_{f,m}\Big\|_{\mathcal{H}^\alpha} \le CN^\alpha\varepsilon\|f\|_{W^{1,\infty}}. \quad \square$$

Now we are ready to proof Theorem 5.3.

Proof of Theorem 5.3. Combining Lemmas A.3 and A.4 with $\|f\|_{W^{1,\infty}} \le B$ yields for a constant $C > 0$ for any $\tilde\varepsilon \in (0,1/2)$ that

$$\begin{aligned}
\|f - \Phi_{\tilde\varepsilon}\|_{\mathcal{H}^\alpha} &\le \Big\|f - \sum_{m\in\{0,\dots,N\}^d} c_{f,m}\phi_m\Big\|_{\mathcal{H}^\alpha} + \Big\|\sum_{m\in\{0,\dots,N\}^d} c_{f,m}\phi_m - \Phi_{\tilde\varepsilon}\Big\|_{\mathcal{H}^\alpha} \\
&\le CB\Big(\Big(\frac{1}{N}\Big)^{1-\alpha} + N^\alpha\tilde\varepsilon\Big),
\end{aligned} \tag{22}$$

where $\tilde\varepsilon$ determines the approximation accuracy in Lemma A.4. For

$$N := \Big\lceil\Big(\frac{\varepsilon}{2CB}\Big)^{-1/(1-\alpha)}\Big\rceil,$$

we get for the first term in (22)

$$\Big(\frac{1}{N}\Big)^{1-\alpha} \le \frac{\varepsilon}{2CB}.$$

Choosing

$$\tilde\varepsilon = \frac{\varepsilon}{2CB}\Big(\Big(\frac{\varepsilon}{2CB}\Big)^{-\frac{1}{1-\alpha}} + 1\Big)^{-\alpha} \tag{23}$$

leads to

$$\|f - \Phi_{\tilde\varepsilon}\|_{\mathcal{H}^\alpha} \le \varepsilon.$$

From Lemma A.2 we know that there is a ReLU network with no more than $1 + \lceil C_1 \log_2(\tilde\varepsilon^{-1})\rceil$ layers, $N^d + \lceil C_2(N+1)^d \log_2^2(\tilde\varepsilon^{-1})\rceil$ nonzero weights and $1 + \lceil C_3(N+1)^d(\log_2^2(\tilde\varepsilon^{-1}) \vee \log_2(\tilde\varepsilon^{-1}))\rceil$ neurons with the required properties. Inserting (23) and assuming $CB > \frac{1}{2}$ yields

$$\log_2(\tilde\varepsilon^{-1}) \le \log_2\Big(\frac{2CB}{\varepsilon}2^\alpha\Big(\frac{\varepsilon}{2CB}\Big)^{-\frac{\alpha}{1-\alpha}}\Big) \le C'\log_2(\varepsilon^{-\frac{1}{1-\alpha}}).$$

Thus there are C', C'' and C''' such that the ReLU network has no more than $1 + \lceil C'\log_2(\varepsilon^{-\frac{1}{1-\alpha}})\rceil$ layers, $\lceil C''\varepsilon^{-\frac{d}{1-\alpha}}\log_2^2(\varepsilon^{-\frac{1}{1-\alpha}})\rceil$ nonzero weights and $1 + \lceil C'''\varepsilon^{-\frac{d}{1-\alpha}}(\log_2^2(\varepsilon^{-\frac{1}{1-\alpha}}) \vee \log_2(\varepsilon^{-\frac{1}{1-\alpha}}))\rceil$ neurons.

Since $f \in \mathrm{Lip}(L,B) \subseteq \mathcal{H}^\alpha(\Gamma)$ for $\Gamma = \max(L,2B)$, we conclude

$$\|\Phi_{\tilde\varepsilon}\|_{\mathcal{H}^\alpha} \le \|f\|_{\mathcal{H}^\alpha} + \|\Phi_{\tilde\varepsilon} - f\|_{\mathcal{H}^\alpha} \le \Gamma + \varepsilon.$$

Corollary 5.4 is a straightforward combination of Theorems 5.2 and 5.3. $\square$

A.6. Proof of Theorem 6.1

Proof of Theorem 6.1. First we note that for $\Gamma > 1$, there is an $L > 0$, such that there is a $B > 0$ with $2B < \Gamma - 1$ and with $\hat X \sim \mathbb{P}^*$

$$\sup_{W\in\mathrm{Lip}(L)} \mathbb{E}[W(\hat X) - W(\hat G_n(Z))] = \sup_{W\in\mathrm{Lip}(L,2B)} \mathbb{E}[W(\hat X) - W(\hat G_n(Z))].$$

This $L > 0$ exists as $[0,1]^d$ is bounded and adding a constant to any function $W \in \mathrm{Lip}(L)$ will not change the value of $\mathbb{E}[W(\hat X) - W(\hat G_n(Z))]$.

Then we get for every $G \in \mathcal{G}$ with the same reasoning as in the proof of Theorem 5.1

$$\begin{aligned}
\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat G_n(Z)}) &\le \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n) + \mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat G_n(Z)}) \\
&= \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n) + \frac{1}{L}\mathsf{W}_L(\mathbb{P}_n, \mathbb{P}^{\hat G_n(Z)}) \\
&\le \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n) + \frac{1}{L}\mathsf{W}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{\hat G_n(Z)}) + \frac{2}{L}\inf_{W\in\mathcal{W}}\sup_{W'\in\mathrm{Lip}(L,2B)}\|W - W'\|_\infty \\
&\le \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n) + \frac{1}{L}\mathsf{W}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)}) + \frac{2}{L}\inf_{W\in\mathcal{W}}\sup_{W'\in\mathrm{Lip}(L,2B)}\|W - W'\|_\infty \\
&\le \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n) + \frac{1}{L}\mathsf{W}_{\mathcal{H}^\alpha}(\mathbb{P}_n, \mathbb{P}^{G(Z)}) + \frac{2}{L}\inf_{W\in\mathcal{W}}\sup_{W'\in\mathrm{Lip}(L,2B)}\|W - W'\|_\infty
\end{aligned}$$

The bound on $\mathsf{W}_{\mathcal{H}^\alpha}(\mathbb{P}_n, \mathbb{P}^{G(Z)})$ depending on the intrinsic dimension d^* was already derived in Theorem 5.2 (starting with Eq. (15)). The bound on $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n)$ depending on the intrinsic dimension d^* was already derived in Corollary 4.2. $\square$

A.7. Calculations for Example 3.3

For the Wasserstein distance we get $\mathsf{W}_1(\mathbb{P}, \mathbb{Q}) = \gamma$. The Vanilla GAN distance using all Lipschitz L affine functions as discriminator yields in

this example $\mathsf{V}_{a\cdot+b}(\mathbb{P},\mathbb{Q}) = \max_{\substack{a,b\in\mathbb{R},\\ |a|\le L}} f(a,b)$ for

$$f(a,b) := \frac{1}{2}\big(-\log\big(1+e^{-a\gamma-b}\big)-\log\big(1+e^{-a(\gamma+\varepsilon)-b}\big)-\log\big(1+e^{b}\big)-\log\big(1+e^{a\varepsilon+b}\big)\big)+\log(4).$$

Standard calculus yields for fixed a the unique maximizer $b^* = -\frac{a(\varepsilon+\gamma)}{2}$ and

$$f(a,b^*) = -\log\Big(1+e^{-\frac{a(\varepsilon+\gamma)}{2}}\Big) - \log\Big(1+e^{\frac{a(\varepsilon-\gamma)}{2}}\Big) + \log(4).$$

Since

$$\frac{\partial}{\partial a} f(a,b^*) = \frac{\varepsilon+\gamma}{2(e^{\frac{a(\gamma+\varepsilon)}{2}}+1)} - \frac{\varepsilon-\gamma}{2(e^{-\frac{a(\varepsilon-\gamma)}{2}}+1)},$$

for $\varepsilon \le \gamma$, the maximizing a is maximal $a^* = L$. This coincides with the intuitive choice: as the support of $\mathbb{P}^X$ and the support of $\mathbb{P}^Y$ can be separated by a single point on $\mathbb{R}$, we expect the optimal discriminator to be affine linear. Standard calculus yields the linear upper and lower bound for $\varepsilon = \frac{1}{4}$.

For $\varepsilon > \gamma$, the unrestricted maximizing a^* solves the equation

$$(\varepsilon-\gamma)e^{\frac{a^*(\varepsilon+\gamma)}{2}} - (\varepsilon+\gamma)e^{-\frac{a^*(\varepsilon-\gamma)}{2}} = 2\gamma.$$

While there is no closed form solution, a numerical approximation (for $\varepsilon = \frac{1}{4}$) yields for $\gamma < \varepsilon$ and $L > 16$ such that a^* is feasible

$$\frac{\mathsf{W}_1(\mathbb{P}^X,\mathbb{P}^Y)^2}{2} \le \mathsf{V}_{a\cdot+b}(\mathbb{P}^X,\mathbb{P}^Y) \le a\cdot \mathsf{W}_1(\mathbb{P}^X,\mathbb{P}^Y)^2.$$

References

Aggarwal, C. (2018). *Neural networks and deep learning: A textbook*. Springer Cham.

Anil, C., Lucas, J., & Grosse, R. (2019). Sorting out Lipschitz function approximation. In *International conference on machine learning* (pp. 291–301). PMLR.

Arjovsky, M., & Bottou, L. (2017). Towards principled methods for training generative adversarial networks. In *International conference on learning representations*.

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

Asokan, S., & Seelamantula, C. S. (2023). Euler-Lagrange analysis of generative adversarial networks. *Journal of Machine Learning Research*, *24*(126), 1–100.

Belomestny, D., Naumov, A., Puchkin, N., & Samsonov, S. (2023). Simultaneous approximation of a smooth function and its derivatives by deep neural networks with piecewise-polynomial activations. *Neural Networks*, *161*, 242–253.

Berenfeld, C., & Hoffmann, M. (2021). Density estimation on an unknown submanifold. *Electronic Journal of Statistics*, *15*(1), 2179–2223.

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

Biau, G., Sangnier, M., & Tanielian, U. (2021). Some theoretical insights into Wasserstein GANs. *Journal of Machine Learning Research*, *22*(119), 1–45.

Brenner, S. C., & Scott, L. R. (2008). The mathematical theory of finite element methods. In *Texts in applied mathematics*: *vol. 15*, Springer.

Chae, M. (2022). Rates of convergence for nonparametric estimation of singular distributions using generative adversarial networks. arXiv preprint arXiv:2202.02890.

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

Chen, M., Liao, W., Zha, H., & Zhao, T. (2020). Distribution approximation and statistical estimation guarantees of generative adversarial networks. arXiv preprint arXiv:2002.03938.

Chernodub, A., & Nowicki, D. (2016). Norm-preserving orthogonal permutation linear unit activation functions (oplu). arXiv preprint arXiv:1604.02313.

DeVore, R. A., Hanin, B., & Petrova, G. (2021). Neural network approximation. *Acta Numerica*, *30*, 327–444.

Dudley, R. (1969). The speed of mean Glivenko–Cantelli convergence. *The Annals of Mathematical Statistics*, *40*(1), 40–50.

Eckstein, S. (2020). Lipschitz neural networks are dense in the set of all Lipschitz functions. arXiv preprint arXiv:2009.13881.

Farnia, F., & Tse, D. (2018). A convex duality framework for GANs. *Advances in Neural Information Processing Systems*, *31*.

Fedus, W., Rosca, M., Lakshminarayanan, B., Dai, A. M., Mohamed, S., & Goodfellow, I. (2017). Many paths to equilibrium: GANs do not need to decrease a divergence at every step. In *International conference on learning representations*.

Gibbs, A. L., & Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, *70*(3), 419–435.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

Gühring, I., Kutyniok, G., & Petersen, P. (2020). Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, *18*(05), 803–859.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, *30*.

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

Huster, T., Chiang, C. Y. J., & Chadha, R. (2019). Limitations of the Lipschitz constant as a defense against adversarial examples. In *ECML PKDD 2018 workshops: Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and green data mining 2018, Dublin, Ireland, September 10-014, 2018, proceedings 18* (pp. 16–29). Springer.

Khromov, G., & Singh, S. P. (2024). Some fundamental aspects about Lipschitz continuity of neural networks. In *International conference on learning representations*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In *International conference on learning representations*.

Kodali, N., Abernethy, J., Hays, J., & Kira, Z. (2017). On convergence and stability of GANs. arXiv preprint arXiv:1705.07215.

Liang, T. (2017). How well can generative adversarial networks learn densities: A nonparametric view. arXiv preprint arXiv:1712.08244.

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

Lunardi, A. (2018). Interpolation theory. *CRM series ; 16lecture notes (Scuola normale superiore) ; 16springer eBook CollectionSpringerLink*, Pisa: Edizioni della Normale.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In *International conference on learning representations*.

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

Nowozin, S., Cseke, B., & Tomioka, R. (2016). f-GAN: Training generative neural samplers using variational divergence minimization. *Advances in Neural Information Processing Systems*, *29*.

Petzka, H., Fischer, A., & Lukovnikov, D. (2018). On the regularization of Wasserstein GANs. In *International conference on learning representations*.

Puchkin, N., Samsonov, S., Belomestny, D., Moulines, E., & Naumov, A. (2024). Rates of convergence for density estimation with generative adversarial networks. *Journal of Machine Learning Research*, *25*(29), 1–47.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. In *International conference on learning representations*.

Schreuder, N. (2020). Bounding the expectation of the supremum of empirical processes indexed by Hölder classes. *Mathematical Methods of Statistics*, *29*(1), 76–86.

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

Stein, E. M. (1970). *Singular integrals and differentiability properties of functions (PMS-30)*. Princeton University Press.

Stéphanovitch, A., Aamari, E., & Levrard, C. (2023). Wasserstein GANs are minimax optimal distribution estimators. arXiv preprint arXiv:2311.18613.

Suh, N., & Cheng, G. (2024). A survey on statistical theory of deep learning: Approximation, training dynamics, and generative models. arXiv preprint arXiv:2401.07187.

Tang, R., & Yang, Y. (2023). Minimax rate of distribution estimation on unknown submanifolds under adversarial losses. *The Annals of Statistics*, *51*(3), 1282–1308.

Than, K., & Vu, N. (2021). Generalization of GANs and overparameterized models under Lipschitz continuity. arXiv preprint arXiv:2104.02388.

Torres, L. C., Pereira, L. M., & Amini, M. H. (2021). A survey on optimal transport for machine learning: Theory and applications. arXiv preprint arXiv:2106.01963.

Vardanyan, E., Minasyan, A., Hunanyan, S., Galstyan, T., & Dalalyan, A. (2023). Guaranteed optimal generative modeling with maximum deviation from the empirical distribution. arXiv preprint arXiv:2307.16422.

Villani, C. (2008). Optimal transport: Old and new. In *Grundlehren der mathematischen Wissenschaften*, Springer Berlin Heidelberg.

Wei, X., Liu, Z., Wang, L., & Gong, B. (2018). Improving the improved training of Wasserstein GANs. In *International conference on learning representations*.

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

Zhou, Z., Liang, J., Song, Y., Yu, L., Wang, H., Zhang, W., et al. (2019). Lipschitz generative adversarial nets. In *International conference on machine learning* (pp. 7584–7593). PMLR.