A Statistical Analysis for Supervised Deep Learning with Exponential Families for Intrinsically Low-dimensional Data

Saptarshi Chakraborty*[1] and Peter L. Bartlett†[1,2,3]

[1]Department of Statistics, UC Berkeley

[2]Department of Electrical Engineering and Computer Sciences, UC Berkeley

[3]Google DeepMind

Abstract

Recent advancements in the field of deep learning theory have revealed that the rate of convergence of the expected test error, in terms of the number of samples, decays as a function of the intrinsic dimension and *not* the nominal dimension of the full space. However, existing literature often defines this intrinsic dimension in terms of the Minkowski dimension or the manifold dimension of the support of the underlying probability measures, which often results in sub-optimal rates and unrealistic assumptions. Furthermore, the current theoretical framework predominantly focuses on regression models with an additive normal noise assumption. In this paper, we attempt to understand supervised deep learning when the responses, given the explanatory variable, are distributed according to an exponential family with a β-Hölder smooth mean function. By characterizing the intrinsic data-dimension through the entropic dimension, we demonstrate that when provided with n independent and identically distributed samples, the test error scales as $\tilde{\mathcal{O}}\left(n^{-\frac{2\beta}{2\beta+\bar{d}_{2\beta}(\lambda)}}\right)$, where $\bar{d}_{2\beta}(\lambda)$ is the 2β-entropic dimension of λ, the distribution of the explanatory variables. This surpasses the current best-known rates in the literature. Furthermore, under the assumption of an upper-bounded density of the explanatory variables, we explicitly delineate the rate of convergence as $\tilde{\mathcal{O}}\left(d^{\frac{2\lfloor\beta\rfloor(\beta+d)}{2\beta+d}}n^{-\frac{2\beta}{2\beta+d}}\right)$, establishing that the dependence on d is not exponential but at most polynomial. We also demonstrate that when the explanatory variable has a lower bounded density, this rate in terms of the number of data samples, is nearly optimal for learning the dependence structure for exponential families.

1 Introduction

The advent of deep learning has completely revolutionized both supervised and unsupervised machine learning learning to obtain super-human performance across all fields of modern science (LeCun et al., 2015). For the past decade, researchers have tried to come up with explanations for this super-human performance

*email: saptarshic@berkeley.edu

†email: peter@berkeley.edu

through developing statistical and optimization models and subjecting them to rigorous theoretical analyses. From an optimization viewpoint, the theory of implicit bias (Gunasekar et al., 2018; Vardi, 2023) attempts to characterize the directions in which different gradient-based methods, that are employed to train these networks converge. In contrast, the statistical theory of benign over-fitting (Bartlett et al., 2019; Tsigler and Bartlett, 2023) attempts to explain such over-parameterised models especially when the data lies in a high-dimensional space.

Another explanation that researchers have come up with is that most natural data, especially images are postulated to have an intrinsically low-dimensional structure despite the high-dimensional feature representation (Pope et al., 2020). Under this so-called *"manifold hypothesis"*, the recent theoretical developments in the generalization aspects of deep learning theory literature have revealed that the excess risk for different deep learning models, especially regression (Schmidt-Hieber, 2020; Suzuki, 2018) and generative models (Huang et al., 2022; Chakraborty and Bartlett, 2024a,b) exhibit a decay pattern that depends only on the intrinsic dimension of the data. Notably, Nakada and Imaizumi (2020), Huang et al. (2022) and Chakraborty and Bartlett (2024a) showed that the excess risk decays as $\mathcal{O}(n^{-1/\mathcal{O}(d_\mu)})$, where d_μ denotes the Minkowski dimension of the underlying distribution. For a supervised learning setting, this phenomenon has been proved for various deep regression models with additive Gaussian noise (Schmidt-Hieber, 2020; Nakada and Imaizumi, 2020; Suzuki, 2018; Suzuki and Nitanda, 2021).

Most of the aforementioned studies aim to describe this inherent dimensionality by utilizing the concept of the (upper) Minkowski dimension of the data's underlying support. However, it is worth noting that the Minkowski dimension primarily focuses on quantifying the rate of growth in the covering number of the support while neglecting to account for situations where the distribution may exhibit a higher concentration of mass within specific sub-regions of this support. Thus, the Minkowski dimension often overestimates the intrinsic dimension of the data distribution, resulting in slower rates of statistical convergence. On the other hand, some works (Chen et al., 2022, 2019; Jiao et al., 2021) try to impose a smooth Riemannian manifold structure for this support and characterize the rate through the dimension of this manifold. However, this assumption is not only very strong and unverifiable for all practical purposes but also ignores the fact that the data can be concentrated only on some sub-regions and can be thinly spread over the remainder, again resulting in an over-estimate. In contrast, recent insights from the optimal transport literature have introduced the concept of the Wasserstein dimension (Weed and Bach, 2019), which overcomes these limitations and offers a more accurate characterization of convergence rates when estimating a distribution through the empirical measure. Furthermore, recent advancements in this field have led to the introduction of the entropic dimension (Chakraborty and Bartlett, 2024b), which builds upon seminal work by Dudley (1969) and can be employed to describe the convergence rates for Bidirectional Generative Adversarial Networks (BiGANs) (Donahue et al., 2017). Remarkably, the entropic dimension is no larger than the Wasserstein

and Minkowski dimensions, resulting in faster convergence rates for the sample estimator. However, the application of this faster rate of convergence is limited to Generative Adversarial Networks (GANs) and their variants and it is not known as to whether such rates hold for supervised learning problems.

To overcome the aforementioned drawbacks in the current literature, we provide a statistical framework to understand deep supervised learning. Our approach involves modeling the conditional distribution of the response variable given the explanatory variable as a member of an exponential family with a smooth mean function. This framework accommodates a wide spectrum of scenarios, including standard regression and classification tasks, while also providing a statistical foundation for handling complex dependencies in real data settings. In this framework, the maximum likelihood estimates can be viewed as minimizing the canonical Bregman divergence loss between the predicted values and the actual responses. When the explanatory variables have a bounded density with respect to the d-dimensional Lebesgue measure, our analysis reveals that deep networks employing ReLU activation functions can achieve a test error on the order of $\tilde{\mathcal{O}}(n^{-2\beta/(2\beta+d)})$ provided that appropriately sized networks are selected. Here β denotes the Hölder smoothness of the true mean response function. This generalizes the known results in the literature for additive regression with Gaussian noise.

Another aspect overlooked by the current literature is that even when the explanatory variable is absolutely continuous, the rate of convergence of the sample estimator often exponentially increases with the ambient feature dimension, making the upper bound on the estimation error vacuous for high-dimensional data. In this paper, we prove that if the explanatory variable has a bounded density, the dependence, in terms of the ambient feature dimension, is not exponential but at most polynomial. Furthermore, we show that the derived rates for the test error are roughly minimax optimal, meaning that one cannot achieve a better rate of convergence through any estimator except for only potentially improving a logarithmic dependence on n. Lastly, when the data has an intrinsically low dimensional structure, we show that the test error can be improved to achieve a rate of roughly $\tilde{\mathcal{O}}(n^{-2\beta/(2\beta+\bar{d}_{2\beta}(\lambda))})$, where $\bar{d}_{2\beta}(\lambda)$ denotes the 2β-entropic dimension (see Definition 11) of λ, the distribution of the explanatory variables, thus, improving upon the rates in the current literature. This result not only extends the framework beyond additive Gaussian noise regression models but also improves upon the existing rates available in the literature (Nakada and Imaizumi, 2020; Schmidt-Hieber, 2020; Chen et al., 2022). The main results of this paper are summarized as follows:

- In Theorem 8, we demonstrate that when the explanatory variable has a bounded density, the test error for the learning problem scales as $\tilde{\mathcal{O}}\left(d^{\frac{2\lfloor\beta\rfloor(\beta+d)}{2\beta+d}} n^{-\frac{2\beta}{2\beta+d}}\right)$, showing explicit dependence on the problem dimension (d) and the number of samples (n)
- Theorem 10 establishes that the minimax rates scale as $\tilde{\mathcal{O}}\left(n^{-\frac{2\beta}{2\beta+d}}\right)$, ensuring that deep learners can attain the minimax optimal rate when network sizes are appropriately chosen. Notably, this theorem

recovers the seminal results of Yang and Barron (1999) as a special case.

- When the explanatory variable has an intrinsically low dimensional structure, we show that deep supervised learners can effectively achieve an error rate of $\tilde{\mathcal{O}}\left(n^{-\frac{2\beta}{2\beta+d_{2\beta}(\lambda)}}\right)$ in Theorem 12. This result provides the fastest known rates for deep supervised learners and encompasses many recent findings as special cases (Nakada and Imaizumi, 2020; Chen et al., 2022) for additive regression models.

- In the process, we are able to improve upon the recent $\mathbb{L}_p$-approximation results on ReLU networks, in Lemma 21, which might be of independent interest.

The remainder of the paper is organized as follows. After discussing the necessary preliminary background in Section 2, we discuss the exponential family learning framework in Section 3. Under this framework, we derive the learning rates (Theorem 8) when the explanatory variable is absolutely continuous in Section 4 and show that it is minimax optimal (Theorem 10). Next, we analyze the error rate (Theorem 12) when the explanatory variable has an intrinsically low dimensional structure in Section 5. The proofs of the main results are sketched in Section 6, followed by concluding remarks in Section 7.

2 Background

This section recalls some of the notations and background necessary for our theoretical analyses. We say $A \precsim B$ (for A, $B \geq 0$) if there exists an absolute constant $C > 0$ (independent of n and d), such that $A \leq CB$. Similarly, for non-negative functions f and g, we say $f(x) \precsim_x g(x)$ if there exists a constant C, which is independent of x such that $f(x) \leq Cg(x)$, for all x. $\text{sigmoid}(t) = 1/(1+e^{-t})$ denotes the sigmoid activation function. For any function $f : \mathcal{S} \to \mathbb{R}$, and any measure γ on $\mathcal{S}$, let $\|f\|_{\mathbb{L}_p(\gamma)} := \left(\int_{\mathcal{S}} |f(x)|^p d\gamma(x)\right)^{1/p}$, if $0 < p < \infty$. Also let, $\|f\|_{\mathbb{L}_\infty(\gamma)} := \text{ess}\sup_{x \in \text{supp}(\gamma)} |f(x)|$. We say $A_n = \tilde{\mathcal{O}}(B_n)$ if $A_n \leq B_n \times \log^C(en)$, for some factor constant $C > 0$. Moreover, $x \vee y = \max\{x, y\}$ and $x \wedge y = \min\{x, y\}$.

Definition 1 (Covering and packing numbers)**.** For a metric space (S, ϱ), the ϵ-covering number with respect to (w.r.t.) ϱ is defined as: $\mathcal{N}(\epsilon; S, \varrho) = \inf\{n \in \mathbb{N} : \exists\, x_1, \ldots x_n \text{ such that } \cup_{i=1}^n B_\varrho(x_i, \epsilon) \supseteq S\}$. An ϵ-cover of S is denoted as $\mathcal{C}(\epsilon; S, \varrho)$. Similarly, the ϵ-packing number is defined as: $\mathcal{M}(\epsilon; S, \varrho) = \sup\{m \in \mathbb{N} : \exists\, x_1, \ldots x_m \in S \text{ such that } \varrho(x_i, x_j) \geq \epsilon, \text{ for all } i \neq j\}$.

Definition 2 (Neural networks)**.** Let $L \in \mathbb{N}$ and $\{N_i\}_{i \in [L]} \subset \mathbb{N}$. Then a L-layer neural network $f : \mathbb{R}^d \to \mathbb{R}^{N_L}$ is defined as,

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

Here, $A_i(y) = W_i y + b_i$, with $W_i \in \mathbb{R}^{N_i \times N_{i-1}}$ and $b_i \in \mathbb{R}^{N_{i-1}}$, with $N_0 = d$. Note that σ_j is applied component-wise. Here, $\{W_i\}_{1 \leq i \leq L}$ are known as weights, and $\{b_i\}_{1 \leq i \leq L}$ are known as biases. $\{\sigma_i\}_{1 \leq i \leq L-1}$

are known as the activation functions. Without loss of generality, one can take $\sigma_\ell(0) = 0$, $\forall\, \ell \in [L-1]$. We define the following quantities: (Depth) $\mathcal{L}(f) := L$ is known as the depth of the network; (Number of weights) The number of weights of the network f is denoted as $\mathcal{W}(f)$; (maximum weight) $\mathcal{B}(f) = \max_{1\le j\le L}(\|b_j\|_\infty) \vee \|W_j\|_\infty$ to denote the maximum absolute value of the weights and biases.

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

If $\sigma_j(x) = x \vee 0$, for all $j = 1, \ldots, L-1$, we denote $\mathcal{NN}_{\{\sigma_i\}_{1\le i\le L-1}}(L, W, R)$ as $\mathcal{RN}(L, W, R)$.

Definition 3 (Hölder functions)**.** Let $f : \mathcal{S} \to \mathbb{R}$ be a function, where $\mathcal{S} \subseteq \mathbb{R}^d$. For a multi-index $\boldsymbol{s} = (s_1, \ldots, s_d)$, let, $\partial^{\boldsymbol{s}} f = \frac{\partial^{|\boldsymbol{s}|} f}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}$, where, $|\boldsymbol{s}| = \sum_{\ell=1}^d s_\ell$. We say that a function $f : \mathcal{S} \to \mathbb{R}$ is β-Hölder (for $\beta > 0$) if

$$\|f\|_{\mathcal{H}^\beta} := \sum_{\boldsymbol{s}:0\le|\boldsymbol{s}|\le\lfloor\beta\rfloor} \|\partial^{\boldsymbol{s}} f\|_\infty + \sum_{\boldsymbol{s}:|\boldsymbol{s}|=\lfloor\beta\rfloor} \sup_{x\ne y} \frac{\|\partial^{\boldsymbol{s}} f(x) - \partial^{\boldsymbol{s}} f(y)\|}{\|x-y\|^{\beta-\lfloor\beta\rfloor}} < \infty.$$

If $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, then we define $\|f\|_{\mathcal{H}^\beta} = \sum_{j=1}^{d_2} \|f_j\|_{\mathcal{H}^\beta}$. For notational simplicity, let, $\mathcal{H}^\beta(\mathcal{S}_1, \mathcal{S}_2, C) = \{f : \mathcal{S}_1 \to \mathcal{S}_2 : \|f\|_{\mathcal{H}^\beta} \le C\}$. Here, both $\mathcal{S}_1$ and $\mathcal{S}_2$ are both subsets of real vector spaces.

Definition 4 (Smoothness and strong convexity)**.** We say a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is H-smooth if $\|\nabla f(x) - \nabla f(y)\|_2 \le H\|x-y\|_2$. Similarly, we say that f is α-strongly convex if $f(tx + (1-t)y) \le tf(x) + (1-t)f(y) - \frac{\alpha t(1-t)}{2}\|x-y\|_2^2$.

Definition 5 (Bregman divergences)**.** A differentiable, convex function $\phi : \mathbb{R}^p \to \mathbb{R}$ generates the Bregman divergence $d_\phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_{\ge 0}$ defined by $d_\phi(x\|y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x-y\rangle$.

It is evident that $d_\phi(x, y) \ge 0$ holds for all $x, y \in \mathbb{R}^p$ due to the fact that $\phi(x) \ge \phi(y) + \langle \nabla\phi(y), x-y\rangle$ is equivalent to the convex nature of the function ϕ. From a geometric standpoint, one can conceptualize $d_\phi(x\|y)$ as the separation between $\phi(x)$ and the linear approximation of $\phi(x)$ centered around $\phi(y)$. Put simply, this can be described as the distance between $\phi(x)$ and the value obtained by evaluating the tangent line to $\phi(y)$ at the point x. Some prominent examples of Bregman divergences include the squared Euclidean distance, Kullback-Leibler (KL) divergence, Mahalanobis distance, etc. We refer the reader to Banerjee et al. (2005, Table 1) for more examples of the Bregman family. Bregman divergences have a direct association with standard exponential families, as elaborated in the upcoming section, rendering them particularly suitable for modeling and learning from various common data types that originate from exponential family distributions.

3 Learning Framework

To discuss our framework, let us first recall the definition of exponential families (Lehmann and Casella, 2006, Chapter 1, Section 5). We suppose that $\boldsymbol{\theta} \in \Theta$ is the corresponding natural parameter. We say that

$\boldsymbol{X}$ is distributed according to an exponential family, $\mathcal{F}_\Psi$ if the density of $\boldsymbol{X}$ w.r.t. some dominating measure ν, is given by,

$$p_{\Psi,\theta}(d\boldsymbol{x}) = h(\boldsymbol{x}) \exp\left(\langle \boldsymbol{\theta}, T(\boldsymbol{x})\rangle - \Psi(\boldsymbol{\theta})\right) \nu(d\boldsymbol{x}).$$

Here, $T(\cdot)$ is called the natural statistic. Often, it is assumed that the exponential family is expressed in its regular form, i.e. the components of $T(\cdot)$ are affinely independent, i.e. there exists no $\boldsymbol{\upsilon}$, such that $\langle \boldsymbol{\upsilon}, T(\boldsymbol{x})\rangle = c$ (a constant), for all $\boldsymbol{x}$. Popular examples of exponential families include Gaussian, binomial, Poisson, exponential, and many other distributions commonly used in the Statistics literature. Given an exponential family, one can express it in its natural form. Formally,

Definition 6 (Natural form of Exponential families)**.** A multivariate parametric family $\mathcal{F}_\Psi$ of distributions $\{p_{\Psi,\boldsymbol{\theta}} | \boldsymbol{\theta} \in \Theta = \text{int}(\Theta) = \text{dom}(\Psi) \subseteq \mathbb{R}^{d_\theta}\}$ is called a regular exponential family provided that each probability density, w.r.t. some dominating measure ν, is of the form,

$$p_{\Psi,\boldsymbol{\theta}}(d\boldsymbol{x}) = \exp\left(\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle - \Psi(\boldsymbol{\theta})\right) h(\boldsymbol{x}) \nu(d\boldsymbol{x})$$

for all $\boldsymbol{x} \in \mathbb{R}^d$, where $\boldsymbol{x}$ represents a minimal sufficient statistic for the family.

It is well known that $\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{X} \sim p_{\Psi,\boldsymbol{\theta}}} \boldsymbol{X} = \nabla \Psi(\boldsymbol{\theta})$. For simplicity, we assume that Ψ is proper, closed, convex, and differentiable. The conjugate of Ψ, denoted as ϕ is defined as, $\phi(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta} \in \Theta} \{\langle \boldsymbol{\theta}, \boldsymbol{\mu}\rangle - \Psi(\boldsymbol{\theta})\}$. It is well known (Banerjee et al., 2005, Theorem 4) that $p_{\Psi,\boldsymbol{\theta}}$ can be expressed as,

$$p_{\Psi,\boldsymbol{\theta}}(d\boldsymbol{x}) = \exp(-d_\phi(\boldsymbol{x} \| \boldsymbol{\mu}(\boldsymbol{\theta})) b_\phi(\boldsymbol{x}) \nu(d\boldsymbol{x}), \tag{2}$$

where $d_\phi(\cdot\|\cdot)$ denotes the Bergman divergence corresponding to ϕ. In this paper, we are interested in the supervised learning problem when the response, given the explanatory variable, is distributed according to an exponential family. For simplicity, we assume that the responses are real-valued. We assume that there exists a "true" predictor function, $f_0 : \mathbb{R}^d \to \mathbb{R}$, such that

$$y | \boldsymbol{x} \sim p_{\Psi, f_0(\boldsymbol{x})} \quad \text{and} \quad \boldsymbol{x} \sim \lambda.$$

Thus, the joint distribution of $(\boldsymbol{X}, Y)$ is given by,

$$\mathcal{P}(d\boldsymbol{x}, dy) \propto \exp(-d_\phi(y \| \mu(f_0(\boldsymbol{x}))) b_\phi(\boldsymbol{x}) \lambda(d\boldsymbol{x}) \nu(dy) \tag{3}$$

By definition, we observe that $\mathbb{E}(y|\boldsymbol{x}) = \mu(f_0(\boldsymbol{x}))$. We will assume that the data is independent and identically distributed according to the distribution $\mathcal{P}$. We also assume that the distribution of $\boldsymbol{x}$ is bounded in the compact set, $[0,1]^d$. Formally,

A 1. *We assume that the data* $\{(\boldsymbol{x}_i, y_i)\}_{i \in [n]}$ *are independent and identically distributed according to the distribution* $\mathcal{P}$*, defined in* (3)*. Furthermore,* $\lambda\left([0,1]^d\right) = 1$.

In the classical Statistics, literature, one estimates f_0 by finding its maximum likelihood estimates (m.l.e.) as,

$$\operatorname*{argmax}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \log \mathcal{P}(\boldsymbol{x}_i, y_i).$$

Plugging in the form of $\mathcal{P}$ as in (3), it is easy to see that the above optimization problem is equivalent to

$$\operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} d_\phi\left(y_i \| \mu(f(\boldsymbol{x}_i))\right) \tag{4}$$

Here, $\mu : \mathbb{R} \to \mathbb{R}$ is known as the link function. In practice, we take $\mathcal{F}$ to be some sort of neural network class, with the final output passing through the activation function μ. The empirical minimizer of (4) is denoted as $\hat{f}$. To show that this framework covers a wide range of supervised learning problems, we consider the classical example for the case when, $y|\boldsymbol{x} \sim \text{Normal}(f_0(\boldsymbol{x}), \sigma^2)$, for some unknown σ^2. In this case, it is well known that $\mu(\cdot)$ is the identity map and $d_\phi(\cdot\|\cdot)$ becomes the classical squared Euclidean distance. Thus, the m.l.e. problem (4) becomes the classical regression problem, i.e. $\operatorname{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2$.

Another example is the case of logistic regression. We assume that $y|\boldsymbol{x}$ is a Bernoulli random variable. This makes, $\mu(z) = \frac{1}{1+e^{-z}}$, i.e. the sigmoid activation function. Furthermore, an easy calculation (Banerjee et al., 2005, Table 2) shows that $d_\phi(x\|y) = x\log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right)$. Plugging in the values of μ and d_ϕ into (4), we note that the estimation problem becomes,

$$\operatorname*{argmin}_{f\in\mathcal{F}} -\frac{1}{n}\sum_{i=1}^{n} \left(y_i \log\circ\,\text{sigmoid}(f(\boldsymbol{x}_i)) + (1-y_i)\log\circ(1-\text{sigmoid}(f(\boldsymbol{x}_i)))\right).$$

Thus, the problem reduces to the classical two-class learning problem with the binary cross-entropy loss and with sigmoid activation at the final layer. For simplicity, we assume that all activations, excluding that of the final layer of f, are realized by the ReLU activation function. The choice of ReLU activation is a natural choice for practitioners and enables us to harness the massive approximation theory of ReLU networks developed throughout the recent literature (Yarotsky, 2017; Uppal et al., 2019). However, using a leaky ReLU network will also result in a similar analysis, changing only the constants in the main theorems.

To facilitate the theoretical analysis, we will assume that the problem is smooth in terms of the learning function f_0. As a notion of smoothness, we will use Hölder smoothness. This has been a popular choice in the recent literature (Nakada and Imaizumi, 2020; Schmidt-Hieber, 2020; Chen et al., 2022) and covers a vast array of functions commonly modeled in the literature.

A2. *f_0 is β-Hölder continuous, i.e. $f_0 \in \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$.*

We make the additional assumption that the function Ψ is well-behaved. In particular, we assume that Ψ possesses both smoothness and strong convexity properties. It is important to note that these assumptions are widely employed in the existing literature (Telgarsky and Dasgupta, 2013; Paul et al., 2021). Though A3 is not applicable for the classification problem, as in that case, $\Psi(x) = x\ln x + (1-x)\ln(1-x)$, which

does not satisfy A3. However for all practical purposes, one clips the output network (which is often done in practice to ensure smooth training), i.e. ensures that $\epsilon \leq f, f_0 \leq 1-\epsilon$, for some positive ϵ, A3 is satisfied. The assumption is formally stated as follows.

A3. *We assume that Ψ is σ_1-smooth and σ_2-strongly convex.*

A direct implication of A3 is that by Kakade et al. (2009, Theorem 6), ϕ is τ_2-smooth and τ_1- strongly convex. Here $\tau_i = 1/\sigma_i$. Also, since Ψ is σ_1-smooth, $\mu(\cdot) = \nabla\Psi(\cdot)$ is σ_1-Lipschitz. This fact will be useful for the proofs of the main results. In the subsequent sections, under the above assumptions, we derive probabilistic error bounds for the excess risk of $\hat{f}$.

4 Optimal Rates for Distributions with Bounded Densities

We begin the analysis of the test error for the problem (4) when λ, the distribution of the explanatory variable has a bounded density on $[0,1]^d$. First note that the excess risk is upper bounded by the estimation error for f_0 in the $\mathbb{L}_2(\lambda)$-norm. The excess risk for the problem is given by

$$\mathfrak{R}(\hat{f}) = \mathbb{E}_{(y,\boldsymbol{x})\sim\mathcal{P}} d_\phi(y\|\mu(\hat{f}(\boldsymbol{x}))) - \mathbb{E}_{(y,\boldsymbol{x})\sim\mathcal{P}} d_\phi(y\|\mu(f_0(\boldsymbol{x}))).$$

The following lemma ensures that $\mathfrak{R}(\hat{f}) \asymp \|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)}$ and hence, it is enough to prove bounds on $\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)}$ to derive upper and lower bounds on the excess risk.

Lemma 7. *For any $\hat{f} \in \mathcal{F}$, $\frac{\sigma_2}{\sigma_1}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \leq \mathfrak{R}(\hat{f}) \leq \frac{\sigma_1}{\sigma_2}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)}$.*

As already mentioned, we assume that λ admits a density w.r.t. the Lebesgue measure, and this density is upper bounded. Formally,

A4. *Suppose that λ admits an upper-bounded density p_λ w.r.t. the Lebesgue measure on $[0,1]^d$, i.e. $\|p_\lambda\|_\infty \leq \bar{b}_\lambda$, almost surely (under the Lebesgue measure).*

Under these assumptions, i.e. A1–4, we observe that with high probability, $\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)}$ and there by, $\mathfrak{R}(\hat{f})$ scales at most as $\tilde{\mathcal{O}}\left(d^{2\lfloor\beta\rfloor} n^{-\frac{2\beta}{2\beta+d}}\right)$, ignoring log-terms in n, for large n, if the network sizes are chosen properly. This rate of decrease aligns consistently with prior findings reported by Nakada and Imaizumi (2020) for additive Gaussian-noise regression. It is important to underscore that the existing literature predominantly investigates the rate of decrease in $\mathfrak{R}(\hat{f})$ solely with regard to the sample size n, overlooking terms dependent upon the data dimensionality. These dimension-dependent terms harbor the potential for exponential growth with respect to the dimensionality of the explanatory variables, and may therefore attain substantial magnitudes, making such bounds inefficient in high-dimensional statistical learning contexts. This analysis shows that the dependence in d is not exponential and can be made to increase at a polynomial

rate only under the assumption of the existence of a bounded density of the explanatory variable. The main upper bound for this case is stated in Theorem 8, with a proof outline appearing in Section 6.1.

Theorem 8. *Suppose assumptions A1–4 holds. Then we can choose* $\mathcal{F} = \mathcal{RN}(L, W, 2C)$ *with* $L \precsim \log n$ *and* $W \precsim n^{\frac{d}{2\beta+d}} \log n$*, such that, with probability at least* $1 - 3\exp\left(-n^{\frac{d}{2\beta+d}}\right)$*,*

$$\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \precsim d^{\frac{2\lfloor\beta\rfloor(\beta+d)}{2\beta+d}} n^{-\frac{2\beta}{2\beta+d}} (\log n)^5,$$

for $n \geq n_0$*, where,* n_0 *might depend on* d*.*

From the bound on the network size, i.e. W in Theorem 8, it is clear that when f_0 is smooth i.e. for large β, one requires a network of smaller size compared to when f_0 is less smooth i.e. when β is small. Similarly, in cases where the dimensionality of the explanatory variables, represented by d is substantial, a larger network is required as compared to situations where d is relatively small. This observation aligns with the intuitive expectation that solving more intricate and complex problems in higher dimensions demands the utilization of larger networks.

The high probability bound in Theorem 8 ensures that the expected test error also scales with the same rate of convergence. This result is shown in Corollary 9

Corollary 9. *Under the assumptions and choices of Theorem 8,* $\mathbb{E}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \precsim d^{\frac{2\lfloor\beta\rfloor(\beta+d)}{2\beta+d}} n^{-\frac{2\beta}{2\beta+d}} (\log n)^5$.

To understand whether deep supervised learning can achieve the optimal rates for the learning problem, we derive the minimax rates for estimating f_0. The minimax expected risk for this problem is given by,

$$\mathfrak{M}_n = \inf_{\hat{f}} \sup_{f \in \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)} \mathbb{E}_f \|\hat{f} - f\|^2_{\mathbb{L}_2(\lambda)},$$

With the notation $\mathbb{E}_f(\cdot)$ we denote the expectation w.r.t. the measure (3), with f_0, replaced with f. Here the infimum is taken over all measurable estimates of $\hat{f}$, based on the data. Minimax lower bounds are used to understand the theoretical limits of any statistical estimation problem. The aim of this analysis is to show that deep learning with ReLU networks for the exponential family dependence is (almost) minimax optimal. To facilitate the theoretical analysis, we assume that the density of λ is lower bounded by a positive constant. Formally,

A5. λ *admits a lower-bounded density* p_λ *w.r.t. the Lebesgue measure on* $[0,1]^d$*, i.e.* $\|p_\lambda\|_\infty \geq \underline{b}_\lambda$*, almost surely (under the Lebesgue measure).*

Theorem 10 provides a characterization of this minimax lower bound for estimating f. It is important to note that the seminal works of Yang and Barron (1999) for the normal-noise regression problem is a special case of Theorem 10.

Theorem 10 (Minimax lower bound). *Suppose that assumptions A1–3 and A5 hold. Then, we can find an* $n_0 \in \mathbb{N}$*, such that, if* $n \geq n_0$*,*

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)} \mathbb{E}_f \|\hat{f} - f\|^2_{\mathbb{L}_2(\lambda)} \gtrsim_n n^{-\frac{2\beta}{2\beta+d}}$$

Thus, from Theorems 8 and 10 it is clear that deep supervised estimators for the exponential family dependence can achieve this minimax optimal rate with high probability, barring an excess poly-log factor of n.

5 Rates for Low Intrinsic Dimension

Frequently, it is posited that real-world data, especially vision data, resides within a lower-dimensional structure embedded in a high-dimensional feature space (Pope et al., 2020). To quantify this intrinsic dimensionality of the data, researchers have introduced various measures of the effective dimension of the underlying probability distribution assumed to generate the data. Among these approaches, the most commonly used ones involve assessing the rate of growth of the covering number, in a logarithmic scale, for most of the support of this data distribution. Let us consider a compact Polish space denoted as $(\mathcal{S}, \varrho)$, with γ representing a probability measure defined on it. For the remainder of this paper, we will assume that ϱ corresponds to the ℓ_∞-norm. The simplest measure of the dimension of a probability distribution is the upper Minkowski dimension of its support, defined as follows:

$$\overline{\dim}_M(\gamma) = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}(\epsilon; \mathrm{supp}(\gamma), \ell_\infty)}{\log(1/\epsilon)}.$$

This dimensionality concept relies solely on the covering number of the support and does not assume the existence of a smooth mapping to a lower-dimensional Euclidean space. Consequently, it encompasses not only smooth Riemannian manifolds but also encompasses highly non-smooth sets like fractals. The statistical convergence properties of various estimators concerning the upper Minkowski dimension have been extensively explored in the literature. Kolmogorov and Tikhomirov (1961) conducted a comprehensive study on how the covering number of different function classes depends on the upper Minkowski dimension of the support. Recently, Nakada and Imaizumi (2020) demonstrated how deep learning models can leverage this inherent low-dimensionality in data, which is also reflected in their convergence rates. Nevertheless, a notable limitation associated with utilizing the upper Minkowski dimension is that when a probability measure covers the entire sample space but is concentrated predominantly in specific regions, it may yield a high dimensionality estimate, which might not accurately reflect the underlying dimension.

To overcome the aforementioned difficulty, as a notion of the intrinsic dimension of a measure γ, Chakraborty and Bartlett (2024b) introduced the notion of α-entropic dimension of a measure. Before we pro-

ceed, we recall the (ϵ, τ)-cover of a measure (Posner et al., 1967) as: $\mathcal{N}_\epsilon(\gamma, \tau) = \inf\{\mathcal{N}(\epsilon; S, \varrho) : \gamma(S) \geq 1-\tau\}$, i.e. $\mathcal{N}_\epsilon(\gamma, \tau)$ counts the minimum number of ϵ-balls required to cover a set S of probability at least $1-\tau$.

Definition 11 (Entropic dimension, Chakraborty and Bartlett, 2024b)**.** For any $\alpha > 0$, we define the α-entropic dimension of γ as:

$$\bar{d}_\alpha(\gamma) = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\gamma, \epsilon^\alpha)}{\log(1/\epsilon)}.$$

This notion extends Dudley's notion of entropic dimension (Dudley, 1969) to characterize the convergence rate for the BiGAN problem (Donahue et al., 2017). Chakraborty and Bartlett (2024b) showed that the entropic dimension is not larger than the upper Minkowski dimension and the upper Wasserstein dimension (Weed and Bach, 2019). Furthermore, strict inequality holds even for simplistic examples for measures on the unit hypercube. We refer the reader to Section 3 of Chakraborty and Bartlett (2024b). Chakraborty and Bartlett (2024b) showed that the entropic dimension is a more efficient way of characterizing the intrinsic dimension of the data distributions compared to the popular measures such as the upper Minkowski dimension or the Wasserstein dimension (Weed and Bach, 2019) as the entropic dimension enables us to derive a faster rate of convergence of the estimates.

As an intrinsically low-dimensional probability measure is not guaranteed to be dominated by the Lebesgure measure, we remove assumption A4. Under only assumptions A1–3, if the network sizes are properly chosen, the rate of convergence of $\hat{f}$ to f_0 under the $\mathbb{L}_2(\lambda)$-norm decays at a rough rate of $\tilde{\mathcal{O}}\left(n^{-2\beta/(2\beta+\bar{d}_{2\beta}(\lambda))}\right)$, as shown by Theorem 12.

Theorem 12. *Suppose assumptions A1–3 holds and let $d^\star > \bar{d}_{2\beta}(\lambda)$. Then we can choose $\mathcal{F} = \mathcal{RN}(L, W, 2C)$ with $L \precsim \log n$ and $W \precsim n^{\frac{d^\star}{2\beta+d^\star}} \log n$, such that, with probability at least $1 - 3\exp\left(-n^{\frac{d^\star}{2\beta+d^\star}}\right)$,*

$$\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \precsim_n n^{-\frac{2\beta}{2\beta+d^\star}} (\log n)^5,$$

for $n \geq n_0$, where, n_0 depends on d and $\mathcal{P}$.

Since the normal-noise regression model with β-Hölder f_0 is a special case of our model in (3), Theorem 12 derives a faster rate compared to Nakada and Imaizumi (2020, Theorem 7), who show a rate of $\tilde{\mathcal{O}}\left(n^{-\frac{2\beta}{2\beta+\dim_M(\lambda)}}\right)$. This is because the upper Minkowski dimension is at least the 2β-entropic dimension by Chakraborty and Bartlett (2024b, Proposition 8(c)), i.e. $\bar{d}_{2\beta}(\lambda) \leq \overline{\dim}_M(\lambda)$.

An immediate corollary of Theorem 12 is that the expected test-error rate follows the same rate of decay. The proof of this result can be done following the proof of Corollary 9.

Corollary 13. *Under the assumptions and choices of Theorem 12, $\mathbb{E}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \precsim n^{-\frac{2\beta}{2\beta+d^\star}} (\log n)^5$.*

One can state that a rate similar to that observed in Theorem 12 holds when the support of λ is regular. We recall the definition (Weed and Bach, 2019, Definition 6) of regular sets in $[0,1]^d$ as follows.

Definition 14 (Regular sets). *We call a set $\mathbb{M}$ is $\tilde{d}$-regular w.r.t. the $\tilde{d}$-dimensional Hausdorff measure $\mathbb{H}^{\tilde{d}}$, if $\mathbb{H}^{\tilde{d}}(B_\varrho(x,r)) \asymp r^{\tilde{d}}$, for all $x \in \mathbb{M}$. Recall that the d-Hausdroff measure of a set S is defined as, $\mathbb{H}^d(S) := \liminf_{\epsilon \downarrow 0} \left\{ \sum_{k=1}^{\infty} r_k^d : S \subseteq \sum_{k=1}^{\infty} B_\varrho(x_k, r_k), r_k \leq \epsilon, \forall k \right\}$.*

Examples of regular sets include compact $\tilde{d}$-dimensional differentiable manifolds; nonempty, compact convex set spanned by an affine space of dimension $\tilde{d}$; the relative boundary of a nonempty, compact convex set of dimension $\tilde{d}+1$; or a self-similar set with similarity dimension $\tilde{d}$. When the support of λ is $\tilde{d}$-regular, it can be shown that $\bar{d}_\alpha(\lambda) \leq \tilde{d}$. Formally,

Lemma 15. *Suppose that the support of γ is $\tilde{d}$-regular. Then, $\bar{d}_\alpha(\gamma) \leq \tilde{d}$, for any $\alpha > 0$. Furthermore, if $\gamma \ll \mathbb{H}^{\tilde{d}}$, $\bar{d}_\alpha(\gamma) = \tilde{d}$.*

Thus, applying Theorem 12 and Lemma 15, we note that $\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)}$ decays at most at a rate of $\tilde{\mathcal{O}}\left(n^{-2\beta/(2\beta+\tilde{d})}\right)$, resulting in the follwoing corollary.

Corollary 16. *Suppose assumptions A1–3 and the support of λ is $\tilde{d}$-regular. Let $d^\star > \tilde{d}$, then we can choose $\mathcal{F} = \mathcal{RN}(L, W, 2C)$ with $L \precsim \log n$ and $W \precsim n^{\frac{d^\star}{2\beta+d^\star}} \log n$, such that, with probability at least $1 - 3\exp\left(-n^{\frac{d^\star}{2\beta+d^\star}}\right)$, $\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \precsim_n n^{-\frac{2\beta}{2\beta+d^\star}} (\log n)^5$, for $n \geq n_0$, where, n_0 depends on d and $\mathcal{P}$.*

Since compact $\tilde{d}$-dimensional differentiable manifolds are a special case of $\tilde{d}$-regular sets, Corollary 16 recovers the results by Chen et al. (2022) as a special case i.e., an additive Gaussian-noise regression model. Importantly, this recovery is achieved without imposing assumptions about uniform sharpness on the manifold, as done by Chen et al. (2022, Assumption 2).

6 Proof of the Main Results

This section discusses the proof of the main results of this paper, i.e, Theorems 8, 10 and 12, with proofs of auxiliary supporting lemmata appearing in the appendix. The proof of the main upper bounds (Theorems 8 and 12) are presented in Section 6.1, while the minimax lower bound is proved in Section 6.2.

6.1 Proof of the Upper Bounds

In order to prove Theorem 8, we first decompose the error through an oracle inequality through the following lemma. For any vector $v \in \mathbb{R}^q$, we denote $|||v|||_{p,q} = \left(\frac{1}{q}\sum_{i=1}^q |v_i|^p\right)^{1/p}$.

Lemma 17 (Oracle inequality). *Let $f^* \in \mathcal{F}$. Suppose that $\xi_i = y_i - \mu(f_0(\boldsymbol{x}_i))$, $\hat{\Delta}_i = \mu(\hat{f}(\boldsymbol{x}_i)) - \mu(f_0(\boldsymbol{x}_i))$ and $\tilde{\Delta}_i = \nabla\phi(\mu(f^*(\boldsymbol{x}_i))) - \nabla\phi(\mu(f_0(\boldsymbol{x}_i)))$. Then,*

$$\tau_1 \left|\left|\left| \hat{\Delta} \right|\right|\right|^2_{2,n} \leq \tau_2 \|\mu(f^*) - \mu(f_0)\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{1}{n}\sum_{i=1}^n \xi_i \tilde{\Delta}_i. \tag{5}$$

The first term in the right-hand side (RHS) of (5) is analogous to an approximation error while the second term is akin to a generalization gap. It is worth noting that while taking a large network reduces the approximation error, can potentially give rise to a large generalization gap and vice versa. The key idea is to select a network of appropriate size that ensures that both these errors are small enough. In the following two sections, we control these terms individually.

6.1.1 Generalization Error

To effectively control the generalization error, we employ localization techniques as described by Wainwright (2019, Chapter 14). These techniques are instrumental in achieving rapid convergence of the sample estimator to the population estimator within the framework of the $\mathbb{L}_2(\lambda)$ norm. It is important to note that in some cases, the true function, denoted as f_0, may not be precisely representable by a ReLU network. Our next best approach is to establish a high-probability bound for the squared $\mathbb{L}_2(\lambda)$ norm difference between our estimated function $\hat{f}$ and f^*, where we will take f^* to belong in the neural network function class, close enough to f_0. Our strategy revolves around a two-step process: firstly, we derive a local complexity bound, as outlined in Lemma 18 and subsequently, we leverage this local complexity bound to derive an estimate for $\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)}$, as elucidated in Lemma 19. Here λ_n denotes the empirical distribution of the explanatory variables. We then use this result to control $\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)}$ in Lemma 22 for large n. We state these results subsequently with proof appearing in Appendix A.

Lemma 18. *Suppose that* $\mathcal{G}_\delta = \left\{\nabla\phi(\mu(f)) - \nabla\phi(\mu(f')) : \|f - f'\|_{\mathbb{L}_\infty(\lambda_n)} \leq \delta \text{ and } f, f' \in \mathcal{F}\right\}$, *with* $\delta \leq 1/e$. *Also let,* $n \geq \mathrm{Pdim}(\mathcal{F})$. *Then, for any* $t > 0$, *with probability (conditioned on* $x_{1:n}$*) at least* $1 - e^{-nt^2/\delta^2}$,

$$\sup_{g \in \mathcal{G}_\delta} \frac{1}{n} \sum_{i=1}^{n} \xi_i g(x_i) \precsim t + \delta \sqrt{\frac{\mathrm{Pdim}(\mathcal{F}) \log(n/\delta)}{n}}. \tag{6}$$

Here, $\mathrm{Pdim}(\mathcal{F})$ *denotes the pseudo-dimension of the function class* $\mathcal{F}$ *(Anthony and Bartlett, 2009).*

Lemma 19. *Suppose* $\alpha \in (0, 1/2)$ *and* $n \geq \max\left\{e^{1/\alpha}, \mathrm{Pdim}(\mathcal{F})\right\}$. *Then, for any* $f^* \in \mathcal{F}$, *with probability at least,* $1 - \exp\left(-n^{1-2\alpha}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)} \precsim n^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{1}{n} \mathrm{Pdim}(\mathcal{F}) \log n \tag{7}$$

Lemma 20. *For* $\alpha \in (0, 1/2)$, *if* $n \geq \max\left\{e^{1/\alpha}, \mathrm{Pdim}(\mathcal{F})\right\}$, *with probability at least* $1 - 3\exp\left(-n^{1-2\alpha}\right)$

$$\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \precsim n^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda)} + \frac{1}{n} \mathrm{Pdim}(\mathcal{F}) \log^2 n + \frac{1}{n} \log\log n, \tag{8}$$

for any $f^* \in \mathcal{F}$.

6.1.2 Approximation Error

To effectively bound the overall error in Lemma 17, one needs to control the approximation error, denoted by the first term of (5). Exploring the approximating potential of neural networks has witnessed substantial interest in the research community in the past decade or so. Pioneering studies such as those by Cybenko (1989) and Hornik (1991) have extensively examined the universal approximation properties of networks utilizing sigmoid-like activations. These foundational works demonstrated that wide, single-hidden-layer neural networks possess the capacity to approximate any continuous function within a bounded domain. In light of recent advancements in deep learning, there has been a notable surge in research dedicated to exploring the approximation capabilities of deep neural networks. Some important results in this direction include those by Yarotsky (2017); Lu et al. (2021); Petersen and Voigtlaender (2018); Shen et al. (2019); Schmidt-Hieber (2020) among many others. All of the aforementioned results indicate that when ϵ-approximating a β-Hölder function in the ℓ_∞-norm, one requires a network of depth $\mathcal{O}(\log(1/\epsilon))$ with at most $\mathcal{O}(\epsilon^{-d/\beta}\log(1/\epsilon))$-many weights for the approximating network. However, the constants in the expressions of the upper bound of the number of weights and depth of the network can potentially increase exponentially with d. Shen et al. (2022) showed that if one approximates in the $\mathbb{L}_2(\text{Leb})$-norm, this exponential dependence can be mitigated for the case, $\beta \leq 1$. Here $\text{Leb}(\cdot)$ denotes the Lebesgue measure on $[0,1]^d$. Lemma 21 generalizes this result to include all $\beta > 0$ to achieve a precise dependence on d. The proof is provided in Appendix B.

Lemma 21. *Suppose that $f \in \mathcal{H}^\beta(\mathbb{R}, \mathbb{R}, C)$. Then, we can find a ReLU network, $\hat{f}$, with $\mathcal{L}(\hat{f}) \leq \vartheta\lceil\log_2(8/\eta)\rceil + 4$ and $\mathcal{W}(\hat{f}) \leq \left\lceil \frac{1}{2(\eta/20)^{1/\beta}} \right\rceil^d \left(\frac{3}{\beta}\right)^\beta (d+\lfloor\beta\rfloor)^{\lfloor\beta\rfloor} \left(\vartheta\left\lceil \log_2\left(\frac{8}{\eta d^{\lfloor\beta\rfloor}}\right)\right\rceil + 8d + 4\lfloor\beta\rfloor\right)$, and a constant $\eta_0 \in (0,1)$ (that might depend on β and d) such that $\|f - \hat{f}\|_{\mathbb{L}_p(Leb)} \leq C d^{\lfloor\beta\rfloor}\eta$, for all $\eta \in (0, \eta_0]$. Here, ϑ is an absolute constant.*

We now provide formal proofs of Theorems 8 and 12 by combining the results in Sections 6.1.1 and 6.1.2.

6.1.3 Proof of Theorem 8

Proof. We take $\mathcal{F} = \mathcal{RN}(L_\epsilon, W_\epsilon, 2C)$, with $L_\epsilon \precsim \log(1/\epsilon)$ and $W_\epsilon \precsim d^{\lfloor\beta\rfloor}\epsilon^{-d/\beta}\log(1/\epsilon)$. Then, by Lemma 21, we can find $f^* \in \mathcal{F}$, such that, $\|f^* - f_0\|_{\mathbb{L}_2(\lambda)} \precsim d^{\lfloor\beta\rfloor}\epsilon$. Furthermore, by Lemma 20, we observe that with probability at least $1 - 3\exp\left(-n^{1-2\alpha}\right)$,

$$\begin{aligned}
\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \precsim & n^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda)} + \frac{1}{n}\,\text{Pdim}(\mathcal{F})\log^2 n + \frac{\log\log n}{n} \\
\leq & n^{-2\alpha} + d^{2\lfloor\beta\rfloor}\epsilon^2 + \frac{1}{n}\,\text{Pdim}(\mathcal{F})\log^2 n + \frac{\log\log n}{n} \\
\precsim & n^{-2\alpha} + d^{2\lfloor\beta\rfloor}\epsilon^2 + \frac{\log^2 n}{n} W_\epsilon L_\epsilon \log(W_\epsilon) + \frac{\log\log n}{n} \\
\precsim & n^{-2\alpha} + d^{2\lfloor\beta\rfloor}\epsilon^2 + \frac{d^{\lfloor\beta\rfloor}\log^2 n}{n}\epsilon^{-d/\beta}\log^3(1/\epsilon) + \frac{\log\log n}{n}. \qquad (9)
\end{aligned}$$

Here (9) follows from the following calculations. Suppose α_2 be the constant that honours $W_\epsilon \precsim d^{\lfloor\beta\rfloor}\epsilon^{-d/\beta}\log(1/\epsilon)$, i.e. $W_\epsilon \leq \alpha_2 d^{\lfloor\beta\rfloor}\epsilon^{-d/\beta}\log(1/\epsilon)$. Then,

$$\log W_\epsilon \leq \log\alpha_2 + \lfloor\beta\rfloor\log d + \frac{d}{\beta}\log(1/\epsilon) + \log\log(1/\epsilon) \leq \frac{3d}{\beta}\log(1/\epsilon),$$

when ϵ is small enough. Taking $\epsilon \asymp (nd^{\lfloor\beta\rfloor})^{-\frac{\beta}{2\beta+d}}$ and $\alpha = \frac{\beta}{2\beta+d}$, we note that with probability at least $1 - 3\exp\left(-n^{\frac{d}{2\beta+d}}\right)$,

$$\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \precsim d^{\frac{2\lfloor\beta\rfloor(\beta+d)}{2\beta+d}} n^{-\frac{2\beta}{2\beta+d}} (\log n)^5.$$

Note that for the above bounds to hold, one requires $n \geq \mathrm{Pdim}(\mathcal{F})$ and $\epsilon \leq \epsilon_0$, which holds when n is large enough. □

6.1.4 Proof of Theorem 12

Proof. We take $\mathcal{F} = \mathcal{RN}(L_\epsilon, W_\epsilon, 2C)$, with $L_\epsilon \precsim_\epsilon \log(1/\epsilon)$ and $W_\epsilon \precsim_\epsilon \epsilon^{-d^\star/\beta}\log(1/\epsilon)$. Then, by Chakraborty and Bartlett (2024b, Theorem 18), we can find $f^* \in \mathcal{F}$, such that, $\|f^* - f_0\|_{\mathbb{L}_2(\lambda)} \leq \epsilon$. Furthermore, by Lemma 20, we observe that with probability at least $1 - 3\exp\left(-n^{1-2\alpha}\right)$,

$$\begin{aligned}
\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \precsim& n^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda)} + \frac{1}{n}\mathrm{Pdim}(\mathcal{F})\log^2 n + \frac{\log\log n}{n} \\
\leq& n^{-2\alpha} + \epsilon^2 + \frac{1}{n}\mathrm{Pdim}(\mathcal{F})\log^2 n + \frac{\log\log n}{n} \\
\precsim& n^{-2\alpha} + \epsilon^2 + \frac{1}{n}W_\epsilon L_\epsilon \log(W_\epsilon)\log^2 n + \frac{\log\log n}{n} \qquad (10) \\
\precsim_\epsilon& n^{-2\alpha} + \epsilon^2 + \frac{\log^2 n}{n}\epsilon^{-d^\star/\beta}\log^3(1/\epsilon) + \frac{\log\log n}{n}.
\end{aligned}$$

Here, (10) follows from (Bartlett et al., 2019, Theorem 6). Taking $\epsilon \asymp n^{-\frac{\beta}{2\beta+d^\star}}$ and $\alpha = \frac{\beta}{2\beta+d^\star}$, we note that, with probability at least $1 - 3\exp\left(-n^{\frac{d^\star}{2\beta+d^\star}}\right)$, $\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \precsim_n n^{-\frac{2\beta}{2\beta+d^\star}}(\log n)^5$. Note that for the above bounds to hold, one requires $n \geq \mathrm{Pdim}(\mathcal{F})$ and $\epsilon \leq \epsilon_0$, which holds when n is large enough. □

6.2 Proof of the Minimax Rates

In this section, we give a formal proof of Theorem 10. We use the standard technique of Fano's method (Wainwright, 2019, Chapter 15) to construct hypotheses that are well separated in $\mathbb{L}_2(\lambda)$ sense but difficult to distinguish in the KL-divergence.

Proof of Theorem 10: Let $b(x) = \exp\left(\frac{1}{x^2-1}\right)\mathbb{1}\{|x| \leq 1\}$ be the standard bump function on $\mathbb{R}$. For any $x \in \mathbb{R}^d$ and $\delta \in (0, 1]$, we let, $h_\delta(x) = a\delta^\beta \prod_{j=1}^d b(x_j/\delta)$. Here a is such that $ab(x) \in \mathcal{H}^\beta(\mathbb{R}, \mathbb{R}, C)$. It is easy to observe that $h_\delta \in \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$. In what follows, we take, $\delta = 1/m$. Let,

$$\mathcal{F}_\delta = \left\{ f_{\boldsymbol{\omega}}(x) = \sum_{\boldsymbol{\xi}\in[m]^d} \omega_{\boldsymbol{\xi}} h_\delta\left(x - \frac{1}{m}(\xi_i - 1/2)\right) : \boldsymbol{\omega} \in \{0,1\}^{m^d} \right\}.$$

Since each element of $\mathcal{F}_\delta$ is a sum of members in $\mathcal{H}^\beta(\mathbb{R},\mathbb{R},C)$ with disjoint support, $\mathcal{F}_\delta \subseteq \mathcal{H}^\beta(\mathbb{R},\mathbb{R},C)$. By the Varshamov-Gilbert bound (Tsybakov, 2009, Lemma 2.9), we can construct a subset of $\Omega = \{\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_M\}$ of $\{0,1\}^{m^d}$ with $\|\boldsymbol{\omega}_i - \boldsymbol{\omega}_j\|_1 \geq \frac{m^d}{8}$, for all $i \neq j$ and $M \geq 2^{m^d/8}$. We note that for any $\boldsymbol{\omega}$, $\boldsymbol{\omega}' \in \Omega$,

$$\begin{aligned}\|f_{\boldsymbol{\omega}} - f_{\boldsymbol{\omega}'}\|^2_{\mathbb{L}_2(\lambda)} \geq \underline{b}_\lambda \|f_{\boldsymbol{\omega}} - f_{\boldsymbol{\omega}'}\|^2_{\mathbb{L}_2(\mathrm{Leb})} = \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1 \int h^2_\delta(x)dx &= \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1 \times a^2\delta^{2\beta+d}\int b^2(x)dx \\ &\gtrsim m^d\delta^{2\beta+d} \\ &= \delta^{2\beta}.\end{aligned}$$

Let $P_{\boldsymbol{\omega}}$ denote the the distribution of the form (3) with f_0 replaced with $f_{\boldsymbol{\omega}}$. Thus,

$$\begin{aligned}\mathrm{KL}(P^{\otimes n}_{\boldsymbol{\omega}} \| P^{\otimes n}_{\boldsymbol{\omega}'}) = n\,\mathrm{KL}(P_{\boldsymbol{\omega}} \| P_{\boldsymbol{\omega}'}) &= n\mathbb{E}_{\boldsymbol{x}} d_\phi\left(f_{\boldsymbol{\omega}}(\boldsymbol{x}) \| f_{\boldsymbol{\omega}'}(\boldsymbol{x})\right) \\ &\leq n\tau_2\bar{b}_\lambda \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1 \times a^2\delta^{2\beta+d}\int b^2(x)dx \\ &\lesssim nm^d\delta^{2\beta+d}\end{aligned} \tag{11}$$

Here, (11) follows from Lemma 24. Choosing $m \asymp n^{1/(2\beta+d)}$, we can make, $\mathrm{KL}(P^{\otimes n}_{\boldsymbol{\omega}} \| P^{\otimes n}_{\boldsymbol{\omega}'}) \leq \frac{m^d}{1000}$. Thus, from Wainwright (2019, equation 15.34), $I(Z;J) \leq \frac{1}{M^2}\sum_{\boldsymbol{\omega},\boldsymbol{\omega}'\in\Omega} \mathrm{KL}(P^{\otimes n}_{\boldsymbol{\omega}} \| P^{\otimes n}_{\boldsymbol{\omega}'}) \leq \frac{m^d}{1000}$. Here $I(Z_1;Z_2)$ denotes the mutual information between the random variables Z_1 and Z_2 (Cover and Thomas, 2005, Section 2.3). Hence, $\frac{I(Z;J)+\log 2}{\log M} \leq 8\frac{m^d/1000+\log 2}{m^d \log 2} \leq 1/2$, if n is large enough. Thus, applying Proposition 15.2 of Wainwright (2019), we note that, $\inf_{\hat{f}}\sup_{f\in\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},C)} \mathbb{E}_f\|\hat{f} - f\|^2_{\mathbb{L}_2(\lambda)} \gtrsim \delta^{2\beta} \asymp n^{-\frac{2\beta}{2\beta+d}}$.

7 Conclusion

In this paper, we discussed a statistical framework to understand the finite sample properties for supervised deep learning that encompasses both standard deep regression and classification settings. In particular, we modeled the dependence of the response given the explanatory variable through a exponential families and showed that the maximum likelihood estimates can be achieved by minimizing the corresponding Bregman loss and incorporating the mean function as the activation for the final layer. Under the assumption of the existence of a bounded density for the explanatory variable, we show that deep ReLU networks can achieve the minimax optimal rate when the network size is chosen properly. Furthermore, when the explanatory variable has an intrinsically low dimensional structure, the convergence rate of the sample estimator, in terms of the sample size, only depends on the entropic dimension of the underlying distribution of the explanatory variable, resulting in better convergence rates compared to the existing literature for both classification and regression problems.

While our findings offer insights into the theoretical aspects of deep supervised learning, it is crucial to recognize that assessing the complete error of models in practical applications necessitates the consideration

of an optimization error component. Regrettably, the accurate estimation of this component remains a formidable challenge in the non-overparametrized regime due to the non-convex and intricate nature of the optimization problem. Nevertheless, it is worth emphasizing that our error analyses operate independently of the optimization process and can be readily integrated with optimization analyses.

Appendices

Contents

A Proofs of Main Lemmata

A.1 Proof of Lemma 7

Lemma 7. *For any* $\hat{f} \in \mathcal{F}$, $\frac{\sigma_2}{\sigma_1}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \leq \mathfrak{R}(\hat{f}) \leq \frac{\sigma_1}{\sigma_2}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)}$.

Proof. To prove Lemma 7, we first make the following observation.

$$
\begin{aligned}
&\mathbb{E}d_\phi(y\|\mu(\hat{f}(\boldsymbol{x}))) - \mathbb{E}d_\phi(y\|\mu(f_0(\boldsymbol{x})))\\
=&\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{y|\boldsymbol{x}}\left(\phi(\mu(f_0(\boldsymbol{x}))) - \phi(\mu(\hat{f}(\boldsymbol{x}))) - \left\langle \nabla\phi(\mu(\hat{f}(\boldsymbol{x}))), y - \mu(\hat{f}(\boldsymbol{x}))\right\rangle + \langle \nabla\phi(\mu(f_0(\boldsymbol{x}))), y - \mu(f_0(\boldsymbol{x}))\rangle\right)\\
=&\mathbb{E}_{\boldsymbol{x}} d_\phi\left(\mu(f_0(\boldsymbol{x}))\|\mu(\hat{f}(\boldsymbol{x}))\right)\\
\leq&\tau_2\mathbb{E}_{\boldsymbol{x}}\|\mu(f_0(\boldsymbol{x})) - \mu(\hat{f}(\boldsymbol{x}))\|_2^2 \qquad (12)\\
\leq&\tau_2\sigma_1\mathbb{E}_{\boldsymbol{x}}\|f_0(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\|_2^2 \qquad (13)\\
=&\frac{\sigma_1}{\sigma_2}\|f_0 - \hat{f}\|^2_{\mathbb{L}_2(\lambda)}.
\end{aligned}
$$

Here (12) follows from Lemma S.4 of the supplement. Inequality (13) follows from the fact that $\mu(\cdot)$ is σ_1-Lipschitz. We also note that,

$$
\begin{aligned}
\mathbb{E}d_\phi(y\|\hat{f}(\boldsymbol{x})) - \mathbb{E}d_\phi(y\|\mu(f_0(\boldsymbol{x}))) =&\mathbb{E}_{\boldsymbol{x}} d_\phi\left(\mu(f_0(\boldsymbol{x}))\|\mu(\hat{f}(\boldsymbol{x}))\right)\\
\geq&\tau_1\mathbb{E}_{\boldsymbol{x}}\|\mu(f_0(\boldsymbol{x})) - \mu(\hat{f}(\boldsymbol{x}))\|_2^2 \qquad (14)\\
\geq&\tau_1\sigma_2\mathbb{E}_{\boldsymbol{x}}\|f_0(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\|_2^2 \qquad (15)\\
=&\frac{\sigma_2}{\sigma_1}\|f_0 - \hat{f}\|^2_{\mathbb{L}_2(\lambda)}.
\end{aligned}
$$

As before, (14) follows from the fact that ϕ is τ_1-strongly convex and applying Lemma S.4 of the supplement. Inequality (15) follows from the fact that $\mu'(\cdot) = \Psi''(\cdot) \geq \sigma_2$, due to the strong convexity of Ψ and a simple application of the mean value theorem. □

A.2 Proof of Lemma 15

Lemma 15. *Suppose that the support of* γ *is* $\tilde{d}$*-regular. Then,* $\bar{d}_\alpha(\gamma) \leq \tilde{d}$*, for any* $\alpha > 0$*. Furthermore, if* $\gamma \ll \mathbb{H}^{\tilde{d}}$, $\bar{d}_\alpha(\gamma) = \tilde{d}$.

Proof. We note that $\bar{d}_\alpha(\gamma) \leq \overline{\dim}_M(\gamma) = \tilde{d}$, by Weed and Bach (2019, Proposition 7). When $\mu \ll \mathbb{H}^{\tilde{d}}$, again by Weed and Bach (2019, Proposition 8), it is known that $d_*(\gamma) = \tilde{d}$, where $d_*(\gamma)$ denotes the lower Wasserstein dimension of γ (Weed and Bach, 2019, Definition 4). The result now follows from Proposition 8 of Chakraborty and Bartlett (2024b). □

A.3 Proof of Lemma 17

Lemma 17 (Oracle inequality)**.** *Let $f^* \in \mathcal{F}$. Suppose that $\xi_i = y_i - \mu(f_0(\boldsymbol{x}_i))$, $\hat{\Delta}_i = \mu(\hat{f}(\boldsymbol{x}_i)) - \mu(f_0(\boldsymbol{x}_i))$ and $\tilde{\Delta}_i = \nabla\phi(\mu(f^*(\boldsymbol{x}_i))) - \nabla\phi(\mu(f_0(\boldsymbol{x}_i)))$. Then,*

$$\tau_1 \left|\!\left|\!\left| \hat{\Delta} \right|\!\right|\!\right|_{2,n}^2 \leq \tau_2 \|\mu(f^*) - \mu(f_0)\|_{\mathbb{L}_2(\lambda_n)}^2 + \frac{1}{n}\sum_{i=1}^n \xi_i \tilde{\Delta}_i. \tag{5}$$

Proof. Since $\hat{f}$ is the global minimizer of $\sum_{i=1}^n d_\phi(y_i \| \mu(f(\boldsymbol{x}_i)))$, we note that,

$$\sum_{i=1}^n d_\phi(y_i \| \mu(\hat{f}(\boldsymbol{x}_i))) \leq \sum_{i=1}^n d_\phi(y_i \| \mu(f(\boldsymbol{x}_i))), \tag{16}$$

for any $f \in \mathcal{F}$. A little algebra shows that (16) is equivalent in saying that,

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n d_\phi(\mu(f_0(\boldsymbol{x}_i)) \| \mu(\hat{f}(\boldsymbol{x}_i))) \\ \leq &\frac{1}{n}\sum_{i=1}^n d_\phi(\mu(f_0(\boldsymbol{x}_i)) \| \mu(f(\boldsymbol{x}_i))) + \frac{1}{n}\sum_{i=1}^n \left\langle \nabla\phi(\mu(\hat{f}(\boldsymbol{x}_i))) - \nabla\phi(\mu(f(\boldsymbol{x}_i))), \xi_i \right\rangle \end{aligned} \tag{17}$$

From (17), applying Lemma S.4. we observe that,

$$\begin{aligned} &\frac{\tau_1}{n}\sum_{i=1}^n (\mu(f_0(\boldsymbol{x}_i)) - \mu(\hat{f}(\boldsymbol{x}_i)))^2 \\ \leq &\frac{\tau_2}{n}\sum_{i=1}^n (\mu(f_0(\boldsymbol{x}_i)) - \mu(f(\boldsymbol{x}_i)))^2 + \frac{1}{n}\sum_{i=1}^n \left\langle \nabla\phi(\mu(\hat{f}(\boldsymbol{x}_i))) - \nabla\phi(\mu(f(\boldsymbol{x}_i))), \xi_i \right\rangle \end{aligned} \tag{18}$$

Plugging in $f \leftarrow f^*$, we get the desired result. □

A.4 Proof of Lemma 18

Lemma 18. *Suppose that $\mathcal{G}_\delta = \left\{\nabla\phi(\mu(f)) - \nabla\phi(\mu(f')) : \|f - f'\|_{\mathbb{L}_\infty(\lambda_n)} \leq \delta \text{ and } f, f' \in \mathcal{F}\right\}$, with $\delta \leq 1/e$. Also let, $n \geq \mathrm{Pdim}(\mathcal{F})$. Then, for any $t > 0$, with probability (conditioned on $x_{1:n}$) at least $1 - e^{-nt^2/\delta^2}$,*

$$\sup_{g \in \mathcal{G}_\delta} \frac{1}{n}\sum_{i=1}^n \xi_i g(x_i) \precsim t + \delta\sqrt{\frac{\mathrm{Pdim}(\mathcal{F})\log(n/\delta)}{n}}. \tag{6}$$

Here, $\mathrm{Pdim}(\mathcal{F})$ denotes the pseudo-dimension of the function class $\mathcal{F}$ (Anthony and Bartlett, 2009).

Proof. From the definition of $\mathcal{G}_\delta$, it is clear that $\log \mathcal{N}(\epsilon; \mathcal{G}_\delta, \|\cdot\|_{\infty,n}) \leq 2\log\mathcal{N}(\epsilon/2; \mathcal{F}, \|\cdot\|_{\infty,n})$. Let $Z_f = \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(x_i)$. Clearly, $\mathbb{E}_\xi Z_f = 0$. Furthermore, applying Lemma S.2 of the supplement, we observe that,

$$\mathbb{E}_\xi \exp(\lambda(Z_f - Z_g)) = \mathbb{E}_\xi \exp\left(\frac{\lambda}{\sqrt{n}}\sum_{i=1}^n \xi_i(f(x_i) - g(x_i))\right) = \exp\left(\frac{\lambda^2}{2}\|f - g\|_{\mathbb{L}_2(\lambda_n)}^2 \sigma_1\right).$$

Thus, $(Z_f - Z_g)$ is $\|f-g\|^2_{\mathbb{L}_2(\lambda_n)}\sigma_1$-subGaussian. Furthermore,

$$\begin{aligned}
\sup_{f,g\in\mathcal{G}_\delta}\|f-g\|_{\mathbb{L}_2(\lambda_n)} =& \sup_{f,f'\in\mathcal{F}:\|f-f'\|_{\mathbb{L}_\infty(\lambda_n)}\le\delta}\|\nabla\phi(\mu(f))-\nabla\phi(\mu(f'))\|_{\mathbb{L}_2(\lambda_n)}\\
\le&\tau_1\sigma_1\sup_{f,f'\in\mathcal{F}:\|f-f'\|_{\mathbb{L}_\infty(\lambda_n)}\le\delta}\|f-f'\|_{\mathbb{L}_2(\lambda_n)}\\
\le&\sup_{f,f'\in\mathcal{F}:\|f-f'\|_{\mathbb{L}_\infty(\lambda_n)}\le\delta}\|f-f'\|_{\mathbb{L}_\infty(\lambda_n)}\\
\le&\delta.
\end{aligned}$$

From Wainwright (2019, Proposition 5.22),

$$\begin{aligned}
\mathbb{E}_\xi\sup_{g\in\mathcal{G}_\delta}\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i g(x_i) = \mathbb{E}_\xi\sup_{g\in\mathcal{G}_\delta}Z_g =&\mathbb{E}_\xi\sup_{g\in\mathcal{G}_\delta}(Z_g - Z_{g'})\\
\le&\mathbb{E}_\xi\sup_{g,g'\in\mathcal{G}_\delta}(Z_g-Z_{g'})\\
\le&32\int_0^\delta\sqrt{\log\mathcal{N}(\epsilon;\mathcal{G}_\delta,\mathbb{L}_2(\lambda_n))}d\epsilon\\
\precsim&\int_0^\delta\sqrt{\log\mathcal{N}(\epsilon/(2\sigma_1);\mathcal{F},\mathbb{L}_\infty(\lambda_n))}d\epsilon\\
\precsim&\int_0^\delta\sqrt{\mathrm{Pdim}(\mathcal{F})\log(n/\epsilon)}d\epsilon\\
\le&\delta\sqrt{\mathrm{Pdim}(\mathcal{F})\log n}+\sqrt{\mathrm{Pdim}(\mathcal{F})}\int_0^\delta\sqrt{\log(1/\epsilon)}d\epsilon\\
\le&\delta\sqrt{\mathrm{Pdim}(\mathcal{F})\log n}+2\sqrt{\mathrm{Pdim}(\mathcal{F})}\delta\sqrt{\log(1/\delta)} \qquad (19)\\
\precsim&\delta\sqrt{\mathrm{Pdim}(\mathcal{F})\log(n/\delta)} \qquad (20)
\end{aligned}$$

(19) follows from Lemma S.1 of the supplement. Thus, $\mathbb{E}\sup_{g\in\mathcal{G}_\delta}\frac{1}{n}\sum_{i=1}^n\xi_i g(x_i)\precsim\delta\sqrt{\frac{\mathrm{Pdim}(\mathcal{F})\log(n/\delta)}{n}}$ Applying Lemma S.3 of the supplement, we note that for $t>0$, with probability at least $1-e^{-nt^2/\delta^2}$,

$$\sup_{g\in\mathcal{G}_\delta}\frac{1}{n}\sum_{i=1}^n\xi_i g(x_i)\precsim t+\delta\sqrt{\frac{\mathrm{Pdim}(\mathcal{F})\log(n/\delta)}{n}}. \qquad (21)$$

□

A.5 Proof of Lemma 19

Lemma 19. *Suppose $\alpha\in(0,1/2)$ and $n\ge\max\{e^{1/\alpha},\mathrm{Pdim}(\mathcal{F})\}$. Then, for any $f^*\in\mathcal{F}$, with probability at least, $1-\exp(-n^{1-2\alpha})$,*

$$\|\hat{f}-f^*\|^2_{\mathbb{L}_2(\lambda_n)}\precsim n^{-2\alpha}+\|f^*-f_0\|^2_{\mathbb{L}_2(\lambda_n)}+\frac{1}{n}\mathrm{Pdim}(\mathcal{F})\log n \qquad (7)$$

Proof. We take $\delta=\max\left\{n^{-\alpha},2\|\hat{f}-f_0\|_{\mathbb{L}_2(\lambda_n)}\right\}$ and let $t=n^{-2\alpha}$. We consider two cases as follows.

Case 1: $\|\hat{f}-f^*\|_{\mathbb{L}_2(\lambda_n)}\le\delta$

Then, by Lemma 18, with probability at least $1 - \exp\left(-n^{1-2\alpha}\right)$

$$\begin{aligned}
\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)} \leq & 2\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda_n)} + 2\|f_0 - f^*\|^2_{\mathbb{L}_2(\lambda_n)} \\
\precsim & \|f_0 - f^*\|^2_{\mathbb{L}_2(\lambda_n)} + \|\mu(\hat{f}) - \mu(f_0)\|^2_{\mathbb{L}_2(\lambda_n)} && (22) \\
\precsim & \|f_0 - f^*\|^2_{\mathbb{L}_2(\lambda_n)} + \sup_{g \in \mathcal{G}_\delta} \frac{1}{n} \sum_{i=1}^n \xi_i g(x_i) && (23) \\
\precsim & \|f_0 - f^*\|^2_{\mathbb{L}_2(\lambda_n)} + t + \delta \sqrt{\frac{\mathrm{Pdim}(\mathcal{F}) \log(n/\delta)}{n}} && (24)
\end{aligned}$$

In the above calculations, (22) follows from the fact that $\mu(\cdot)$ is strongly convex and (23) follows from Lemma 17. Inequality (24) follows from Lemma 18. Let $\alpha_1 \geq 1$ be the corresponding constant that honors the inequality in (24). Then using the upper bound on δ, we observe that,

$$\begin{aligned}
& \|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)} \\
\leq & \alpha_1 \|f_0 - f^*\|^2_{\mathbb{L}_2(\lambda_n)} + \alpha_1 \delta \sqrt{\frac{\mathrm{Pdim}(\mathcal{F}) \log(n/\delta)}{n}} + n^{-2\alpha} \\
\leq & \alpha_1 \|f_0 - f^*\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{\delta^2}{16} + \frac{4\alpha_1^2}{n} \mathrm{Pdim}(\mathcal{F}) \log(n/\delta) + n^{-2\alpha} && (25) \\
\leq & \alpha_1 \|f_0 - f^*\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{9n^{-2\alpha}}{8} + \frac{1}{4}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{4(1+\alpha)\alpha_1^2}{n} \mathrm{Pdim}(\mathcal{F}) \log(n) \\
\leq & \alpha_1 \|f_0 - f^*\|^2_{\mathbb{L}_2(\lambda_n)} + 2n^{-2\alpha} + \frac{1}{2}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{1}{2}\|f^* - f_0\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{4(1+\alpha)\alpha_1^2}{n} \mathrm{Pdim}(\mathcal{F}) \log(n)
\end{aligned}$$

Here, (25) follows from the fact that $\sqrt{xy} \leq \frac{x}{16\alpha_1} + 4\alpha_1 y$, from the AM-GM inequality and taking $x = \delta^2$ and $y = \frac{\mathrm{Pdim}(\mathcal{F}) \log(n/\delta)}{n}$. Thus,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)} \precsim n^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{1}{n} \mathrm{Pdim}(\mathcal{F}) \log(n). \tag{26}$$

Case 2: $\|\hat{f} - f^*\|_{\mathbb{L}_2(\lambda_n)} \geq \delta$

It this case, we note that $\|\hat{f} - f^*\|_{\mathbb{L}_2(\lambda_n)} \geq 2\|\hat{f} - f_0\|_{\mathbb{L}_2(\lambda_n)}$. Thus,

$$\begin{aligned}
\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)} \leq & 2\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda_n)} + 2\|f_0 - f^*\|^2_{\mathbb{L}_2(\lambda_n)} \leq \frac{1}{2}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)} + 2\|f_0 - f^*\|^2_{\mathbb{L}_2(\lambda_n)} \\
\implies & \|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)} \precsim \|f_0 - f^*\|^2_{\mathbb{L}_2(\lambda_n)}
\end{aligned}$$

Thus, from the above two cases, combining equations (24) and (26), with probability at least, $1 - \exp\left(-n^{1-2\alpha}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)} \precsim n^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{1}{n} \mathrm{Pdim}(\mathcal{F}) \log(n) \tag{27}$$

From equation (27), we note that, for some constant B_4,

$$\mathbb{P}\left(\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)} \leq B_4 \left(n^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{1}{n} \mathrm{Pdim}(\mathcal{F}) \log(n) \right) \middle| x_{1:n} \right) \geq 1 - \exp\left(-n^{1-2\alpha}\right)$$

Integrating both sides w.r.t. the measure $\mu^{\otimes_n}$, i.e. the joint distribution of $\boldsymbol{x}_{1:n}$, we observe that, unconditionally, with probability at least, $1 - \exp\left(-n^{1-2\alpha}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)} \precsim n^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{1}{n} \mathrm{Pdim}(\mathcal{F}) \log(n) \tag{28}$$

□

A.6 Proof of Lemma 20

To prove Lemma 20, we first state and prove the following result.

Lemma 22. *For $\alpha \in (0, 1/2)$, if $n \geq \max\left\{e^{1/\alpha}, \mathrm{Pdim}(\mathcal{F})\right\}$, with probability at least $1 - 2\exp\left(-n^{1-2\alpha}\right)$,*

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \precsim n^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{1}{n}\mathrm{Pdim}(\mathcal{F})\log^2 n + \frac{\log\log n}{n}$$

Proof. From Lemma 23, we note that if $n \geq \mathrm{Pdim}(\mathcal{F})$, then,

$$\mathbb{E}_\epsilon \sup_{h\in\mathcal{H}_r} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(\boldsymbol{x}_i) \precsim \sqrt{\frac{r\,\mathrm{Pdim}(\mathcal{F})\log n}{n}} \tag{29}$$

$$\leq \sqrt{\frac{(\mathrm{Pdim}(\mathcal{F}))^2\log n}{n^2} + r\frac{\mathrm{Pdim}(\mathcal{F})\log(n/e\,\mathrm{Pdim}(\mathcal{F}))\log n}{n}} \tag{30}$$

Here, (29) follows from Lemma 23 and (30) follows from the fact that for all $x, y > 0$, $x\log x \leq y + x\log(1/ye)$. It is easy to see that we note that, the RHS of (30) has a fixed point of r^* and $r^* \precsim \frac{\mathrm{Pdim}(\mathcal{F})\log^2 n}{n}$. Then, by Theorem 6.1 of Bousquet (2002), we note that with probability at least $1 - e^{-x}$,

$$\int h d\lambda \leq B_3\left(\int h d\lambda_n + \frac{\mathrm{Pdim}(\mathcal{F})\log^2 n}{n} + \frac{x}{n} + \frac{\log\log n}{n}\right), \forall h \in \mathcal{H}, \tag{31}$$

for some absolute constant B_3. Now, taking $x = n^{1-2\alpha}$ in (31), we note that, with probability at least $1 - \exp\left(-n^{1-2\alpha}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \precsim n^{-2\alpha} + \|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{1}{n}\mathrm{Pdim}(\mathcal{F})\log^2 n + \frac{\log\log n}{n} \tag{32}$$

Combining (32) with Lemma 19, we observe that with probability at least $1 - 2\exp\left(-n^{1-2\alpha}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \precsim n^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda_n)} + \frac{1}{n}\mathrm{Pdim}(\mathcal{F})\log^2 n + \frac{\log\log n}{n} \tag{33}$$

□

A.6.1 Proof of Lemma 20

Lemma 20. *For $\alpha \in (0, 1/2)$, if $n \geq \max\left\{e^{1/\alpha}, \mathrm{Pdim}(\mathcal{F})\right\}$, with probability at least $1 - 3\exp\left(-n^{1-2\alpha}\right)$*

$$\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \precsim n^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda)} + \frac{1}{n}\mathrm{Pdim}(\mathcal{F})\log^2 n + \frac{1}{n}\log\log n, \tag{8}$$

for any $f^* \in \mathcal{F}$.

Proof. Finally, let $Z_i = (f^*(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))^2 - \mathbb{E}(f^*(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))^2$. Since $\|f_0\|_\infty, \|f^*\|_\infty \leq B$, for some constant B, it is easy to see that,

$$\sigma^2 = \sum_{i=1}^n \mathbb{E}Z_i^2 = \sum_{i=1}^n \mathrm{Var}\left((f^*(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))^2\right) \leq \sum_{i=1}^n \mathbb{E}\left(f^*(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right)^4 \leq 4nB^2\|f^* - f_0\|^2_{\mathbb{L}_2(\lambda)}.$$

Taking $u = v \vee \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda)}$, we note that, $\sigma^2 \leq 4nB^2u$. Applying Bernstein's inequality (Vershynin, 2018, Theorem 2.8.4), with $t = nu$,

$$\mathbb{P}\left(\left|\|f^* - f_0\|^2_{\mathbb{L}(\lambda_n)} - \|f^* - f_0\|^2_{\mathbb{L}(\lambda)}\right| \geq u\right) \leq \exp\left(-\frac{n^2u^2/2}{4nB^2u + Bnu/3}\right) \leq \exp\left(-\frac{nu}{8B^2 + 2B/3}\right) = \exp\left(-\frac{nu}{B_5}\right) \leq \exp\left(-\frac{nv}{B_5}\right)$$

with $B_5 = 8B^2 + 2B/3$. Thus, with probability, at least $1 - \exp\left(-\frac{nv}{B_5}\right)$,

$$\|f^* - f_0\|^2_{\mathbb{L}(\lambda_n)} \leq \|f^* - f_0\|^2_{\mathbb{L}(\lambda)} + u \leq 2\|f^* - f_0\|^2_{\mathbb{L}(\lambda)} + v.$$

Taking $v = B_5 n^{-2\alpha}$, we observe that, with probability at least $1 - \exp\left(-n^{1-2\alpha}\right)$,

$$\|f^* - f_0\|^2_{\mathbb{L}(\lambda_n)} \precsim \|f^* - f_0\|^2_{\mathbb{L}(\lambda)} + n^{-2\alpha}. \tag{34}$$

Combining (34) with Lemma 22, we observe that, with probability at least $1 - 3\exp\left(-n^{1-2\alpha}\right)$

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \precsim n^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda)} + \frac{1}{n}\operatorname{Pdim}(\mathcal{F})\log^2 n + \frac{\log\log n}{n}.$$

The theorem now follows from observing that $\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \leq 2\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} + 2\|f_0 - f^*\|^2_{\mathbb{L}_2(\lambda)}$. □

B Proof of Approximation Results (Lemma 21)

Lemma 21. *Suppose that $f \in \mathcal{H}^\beta(\mathbb{R}, \mathbb{R}, C)$. Then, we can find a ReLU network, $\hat{f}$, with $\mathcal{L}(\hat{f}) \leq \vartheta\lceil\log_2(8/\eta)\rceil + 4$ and $\mathcal{W}(\hat{f}) \leq \left\lceil\frac{1}{2(\eta/20)^{1/\beta}}\right\rceil^d \left(\frac{3}{\beta}\right)^\beta (d + \lfloor\beta\rfloor)^{\lfloor\beta\rfloor}\left(\vartheta\left\lceil\log_2\left(\frac{8}{\eta d^{\lfloor\beta\rfloor}}\right)\right\rceil + 8d + 4\lfloor\beta\rfloor\right)$, and a constant $\eta_0 \in (0,1)$ (that might depend on β and d) such that $\|f - \hat{f}\|_{\mathbb{L}_p(Leb)} \leq Cd^{\lfloor\beta\rfloor}\eta$, for all $\eta \in (0, \eta_0]$. Here, ϑ is an absolute constant.*

Proof. We first fix any $\epsilon \in (0,1)$ and let, $K = \lceil\frac{1}{2\epsilon}\rceil$. For any $\boldsymbol{i} \in [K]^d$, let $\boldsymbol{\theta^i} = (\epsilon + 2(i_1 - 1)\epsilon, \ldots, \epsilon + 2(i_d - 1)\epsilon)$. Clearly, $\{\boldsymbol{\theta^i} : \boldsymbol{i} \in [K]^d\}$ constitutes an ϵ-net of $[0,1]^d$, w.r.t. the ℓ_∞-norm. We let,

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

for any $0 < b \leq a$. For $0 < \delta \leq \epsilon/3$, we define,

$$\zeta_{\epsilon,\delta}(\boldsymbol{x}) = \prod_{j=1}^d \xi_{\epsilon,\delta}(x_j)$$

We define the region $\mathcal{Q}_{\epsilon,\delta} = \cup_{\boldsymbol{i}\in[K]^d} B_{\ell_\infty}(\boldsymbol{\theta^i}, \delta)$. It is easy to observe that $\text{Leb}([0,1]^d \setminus \mathcal{Q}_{\epsilon,\delta}) \leq 2d\delta$. Here $\text{Leb}(\cdot)$ denotes the Lebesgue measure on $\mathbb{R}^d$. Clearly, $\zeta_{\epsilon,\delta}\left(\cdot - \boldsymbol{\theta^i}\right) = 1$ on $\mathcal{Q}_{\epsilon,\delta}$.

Consider the Taylor expansion of f around $\boldsymbol{\theta}$ as, $P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{|\boldsymbol{s}|\leq\lfloor\beta\rfloor} \frac{\partial^{\boldsymbol{s}} f(\boldsymbol{\theta})}{\boldsymbol{s}!}(\boldsymbol{x} - \boldsymbol{\theta})^{\boldsymbol{s}}$.

$$\text{Clearly, } f(\boldsymbol{x}) - P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{|\boldsymbol{s}|=\lfloor\beta\rfloor} \frac{(\boldsymbol{x} - \boldsymbol{\theta})^{\boldsymbol{s}}}{\boldsymbol{s}!}\left(\partial^{\boldsymbol{s}} f(\boldsymbol{y}) - \partial^{\boldsymbol{s}} f(\boldsymbol{\theta})\right) \leq \|\boldsymbol{x} - \boldsymbol{\theta}\|_\infty^{\lfloor\beta\rfloor} \sum_{|\boldsymbol{s}|=\lfloor\beta\rfloor} \frac{1}{\boldsymbol{s}!}\left|\partial^{\boldsymbol{s}} f(\boldsymbol{y}) - \partial^{\boldsymbol{s}} f(\boldsymbol{\theta})\right|$$

$$\leq \frac{Cd^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\|\boldsymbol{x} - \boldsymbol{\theta}\|_\infty^\beta. \tag{35}$$

In the above calculations, $\boldsymbol{y}$ lies on the line segment joining $\boldsymbol{x}$ and $\boldsymbol{\theta}$. Inequality (35) follows from the fact that $|\partial^{\boldsymbol{s}} f(\boldsymbol{y}) - \partial^{\boldsymbol{s}} f(\boldsymbol{\theta})| \leq C\|\boldsymbol{y}-\boldsymbol{\theta}\|_\infty^{\beta-\lfloor\beta\rfloor} \leq C\|\boldsymbol{x}-\boldsymbol{\theta}\|_\infty^{\beta-\lfloor\beta\rfloor}$ and the identity that $\frac{d^k}{k!} = \sum_{|\boldsymbol{s}|=k} \frac{1}{\boldsymbol{s}!}$. Next, we suppose that $\tilde{f}(\boldsymbol{x}) = \sum_{\boldsymbol{i}\in[K]^d} \zeta_{\epsilon,\delta}(\boldsymbol{x}-\boldsymbol{\theta}^{\boldsymbol{i}}) P_{\boldsymbol{\theta}^{\boldsymbol{i}}}(\boldsymbol{x})$. Thus, if $\boldsymbol{x} \in \mathcal{Q}_{\epsilon,\delta}$,

$$|f(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x})| \leq \max_{\boldsymbol{i}\in[K]^d} \sup_{\boldsymbol{x}\in B_{\ell_\infty}(\boldsymbol{\theta}^{\boldsymbol{i}},\delta)} |f(\boldsymbol{x}) - P_{\boldsymbol{\theta}^{\boldsymbol{i}}}(\boldsymbol{x})| \leq \frac{Cd^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\delta^\beta. \tag{36}$$

Here, (36) follows from (35). Thus, $\|f - \tilde{f}\|_{\mathbb{L}_\infty(\mathcal{Q}_{\epsilon,\delta})} \leq \frac{Cd^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\delta^\beta$. Furthermore, by definition, $\|\tilde{f}\|_\infty \leq C + \frac{Cd^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\epsilon^\beta$. Let $a_{\boldsymbol{i},\boldsymbol{s}} = \frac{\partial^{\boldsymbol{s}} f(\theta^{\boldsymbol{i}})}{\boldsymbol{s}!}$ and

$$\hat{f}_{\boldsymbol{i},\boldsymbol{s}}(\boldsymbol{x}) = \mathrm{prod}_m^{(d+|\boldsymbol{s}|)}(\xi_{\epsilon_1,\delta_1}(x_1-\theta_1^{\boldsymbol{i}}),\ldots,\xi_{\epsilon_d,\delta_d}(x_d-\theta_d^{\boldsymbol{i}}),\underbrace{(x_1-\theta_1^{\boldsymbol{i}}),\ldots,(x_1-\theta_1^{\boldsymbol{i}})}_{s_1 \text{ times}},\ldots,\underbrace{(x_1-\theta_d^{\boldsymbol{i}}),\ldots,(x_d-\theta_d^{\boldsymbol{i}})}_{s_d \text{ times}}).$$

Here, $\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}$ has at most $d+|\boldsymbol{s}| \leq d+\lfloor\beta\rfloor$ many inputs. By Chakraborty and Bartlett (2024b, Lemma 40) of the supplement, $\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}$ can be implemented by a ReLU network with $\mathcal{L}(\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}), \mathcal{W}(\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}) \leq c_3 m$, where c_3 is an absolute constant. Thus, $\mathcal{L}(\hat{f}_{\boldsymbol{i},\boldsymbol{s}}) \leq c_3 m + 2$ and $\mathcal{W}(\hat{f}_{\boldsymbol{i},\boldsymbol{s}}) \leq c_3 m + 8d + 4|s| \leq c_3 m + 8d + 4\lfloor\beta\rfloor$. From Chakraborty and Bartlett (2024b, Lemma 40), we observe that,

$$\left|\hat{f}_{\boldsymbol{i},\boldsymbol{s}}(\boldsymbol{x}) - \zeta(x-\theta^{\boldsymbol{i}})\left(x-\theta^{\boldsymbol{i}}\right)^{\boldsymbol{s}}\right| \leq \frac{1}{2^m}, \forall x \in S. \tag{37}$$

Here, $m \geq \max \frac{1}{2}(\log_2(4d) - 1)$. Finally, let $\hat{f}(\boldsymbol{x}) = \sum_{\boldsymbol{i}\in[K]^d} \sum_{|\boldsymbol{s}|\leq\lfloor\beta\rfloor} a_{\boldsymbol{i},\boldsymbol{s}} \hat{f}_{\boldsymbol{i},\boldsymbol{s}}(\boldsymbol{x})$. Clearly, $\mathcal{L}(\hat{f}) \leq c_3 m + 3$ and $\mathcal{W}(\hat{f}) \leq \binom{d+\lfloor\beta\rfloor}{\lfloor\beta\rfloor}(c_3 m + 8d + 4\lfloor\beta\rfloor)$. This implies that,

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathcal{Q}_{\epsilon,\delta}} |\hat{f}(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x})| &\leq \max_{\boldsymbol{i}\in[K]^d} \sup_{\boldsymbol{x}\in B_{\ell_\infty}(\boldsymbol{\theta}^{\boldsymbol{i}},\delta)} \sum_{|\boldsymbol{s}|\leq\lfloor\beta\rfloor} |a_{\boldsymbol{i},\boldsymbol{s}}| \zeta(x-\theta^{\boldsymbol{i}}) |\hat{f}_{\boldsymbol{i}\boldsymbol{s}}(x) - \left(x-\theta^{\boldsymbol{i}}\right)^{\boldsymbol{s}}| \\
&\leq \sum_{|\boldsymbol{s}|\leq k} |a_{\theta,\boldsymbol{s}}| \left|\hat{f}_{\theta^{\boldsymbol{i}(x)},\boldsymbol{s}}(x) - \zeta_{\boldsymbol{\epsilon},\boldsymbol{\delta}}(x-\theta^{\boldsymbol{i}(x)})\left(x-\theta^{\boldsymbol{i}(x)}\right)^{\boldsymbol{s}}\right| \\
&\leq \frac{C}{2^m}.
\end{aligned} \tag{38}$$

From (36) and (38), we thus get that if $\boldsymbol{x} \in \mathcal{Q}_{\epsilon,\delta}$,

$$|f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})| \leq |f(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x})| + |\hat{f}(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x})| \leq \frac{Cd^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\delta^\beta + \frac{C}{2^m}. \tag{39}$$

Furthermore, it is easy to observe that, $\|\hat{f}\|_{\mathbb{L}_\infty([0,1]^d)} \leq C + \frac{Cd^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\epsilon^\beta + \frac{C}{2^m}$. Hence,

$$\begin{aligned}
\|f-\hat{f}\|_{\mathbb{L}_p(\mathrm{Leb})}^p &= \int_{\mathcal{Q}_{\epsilon,\delta}} |f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})|^p d\mathrm{Leb}(\boldsymbol{x}) + \int_{\mathcal{Q}_{\epsilon,\delta}^{\complement}} |f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})|^p d\mathrm{Leb}(\boldsymbol{x}) \\
&\leq \left(\frac{Cd^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\delta^\beta + \frac{C}{2^m}\right)^p \mathrm{Leb}\left(\mathcal{Q}_{\epsilon,\delta}\right) + \left(2C + \frac{Cd^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\epsilon^\beta + \frac{C}{2^m}\right)^p \mathrm{Leb}(\mathcal{Q}_{\epsilon,\delta}^{\complement}) \\
&\leq \left(\frac{Cd^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\delta^\beta + \frac{C}{2^m}\right)^p + 2\left(2C + \frac{Cd^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\epsilon^\beta + \frac{C}{2^m}\right)^p d\delta \\
\implies \|f-\hat{f}\|_{\mathbb{L}_p(\mathrm{Leb})} &\leq \frac{2Cd^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\delta^\beta + \frac{2C}{2^m} + 4C(d\delta)^{1/p} \leq \frac{2Cd^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\epsilon^\beta + \frac{2C}{2^m} + 4C\epsilon^\beta \leq 10Cd^{\lfloor\beta\rfloor}\epsilon^\beta + \frac{2C}{2^m},
\end{aligned}$$

taking $\delta = \frac{1}{d}\epsilon^{p\beta} \wedge (\epsilon/3)$. We take $m = \left\lceil \log_2 \left(\frac{8}{\eta d^{\lfloor \beta \rfloor}} \right) \right\rceil$ and $\epsilon = (\eta/20)^{1/\beta}$. Thus, $\|f - \hat{f}\|_{\mathbb{L}_p(\mathrm{Leb})} \leq C d^{\lfloor \beta \rfloor} \eta$. We note that $\hat{f}$ has at most K^d-many networks of depth $c_3 m + 3$ and number of weights $\binom{d+\lfloor \beta \rfloor}{\lfloor \beta \rfloor} (c_3 m + 8d + 4\lfloor \beta \rfloor)$. Thus, $\mathcal{L}(\hat{f}) \leq c_3 m + 4$ and $\mathcal{W}(\hat{f}) \leq K^d \binom{d+\lfloor \beta \rfloor}{\lfloor \beta \rfloor} (c_3 m + 8d + 4\lfloor \beta \rfloor)$. We thus get,

$$\mathcal{L}(\hat{f}) \leq c_3 m + 4 \leq c_3 \left\lceil \log_2 \left(\frac{8}{\eta d^{\lfloor \beta \rfloor}} \right) \right\rceil + 4.$$

Similarly,

$$\begin{aligned}
\mathcal{W}(\hat{f}) \leq & K^d \binom{d + \lfloor \beta \rfloor}{\lfloor \beta \rfloor} (c_3 m + 8d + 4\lfloor \beta \rfloor) \\
\leq & \left\lceil \frac{1}{2(\eta/20)^{1/\beta}} \right\rceil^d \binom{d + \lfloor \beta \rfloor}{\lfloor \beta \rfloor} \left(c_3 \left\lceil \log_2 \left(\frac{8}{\eta d^{\lfloor \beta \rfloor}} \right) \right\rceil + 8d + 4\lfloor \beta \rfloor \right) \\
\leq & \left\lceil \frac{1}{2(\eta/20)^{1/\beta}} \right\rceil^d \left(\frac{3}{\beta} \right)^{\beta} (d + \lfloor \beta \rfloor)^{\lfloor \beta \rfloor} \left(c_3 \left\lceil \log_2 \left(\frac{8}{\eta d^{\lfloor \beta \rfloor}} \right) \right\rceil + 8d + 4\lfloor \beta \rfloor \right).
\end{aligned}$$

The proof is now complete by replacing c_3 with ϑ. □

C Proof of Corollary 9

Corollary 9. *Under the assumptions and choices of Theorem 8,* $\mathbb{E}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \precsim d^{\frac{2\lfloor \beta \rfloor (\beta + d)}{2\beta + d}} n^{-\frac{2\beta}{2\beta + d}} (\log n)^5$.

Proof. Suppose that κ be the constant that honors the inequality. Then

$$\begin{aligned}
\mathbb{E}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} = & \mathbb{E}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \mathbb{1}\left\{ \|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \leq \kappa d^{\frac{2\lfloor \beta \rfloor (\beta + d)}{2\beta + d}} n^{-\frac{2\beta}{2\beta + d^\star}} (\log n)^5 \right\} \\
& + \mathbb{E}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \mathbb{1}\left\{ \|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} > \kappa d^{\frac{2\lfloor \beta \rfloor (\beta + d)}{2\beta + d}} n^{-\frac{2\beta}{2\beta + d^\star}} (\log n)^5 \right\} \\
\leq & \kappa n^{-\frac{2\beta}{2\beta + d^\star}} (\log n)^5 + 9C^2 \mathbb{P}\left(\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} > n^{-\frac{2\beta}{2\beta + d^\star}} (\log n)^5 \right) \\
\leq & \kappa n^{-\frac{2\beta}{2\beta + d^\star}} (\log n)^5 + 27C^2 \exp\left(-n^{\frac{d^\star}{2\beta + d^\star}} \right) \\
\precsim & n^{-\frac{2\beta}{2\beta + d^\star}} (\log n)^5.
\end{aligned}$$

□

D Supporting Lemmata

Lemma 23. *Let* $\mathcal{H}_r = \{h = (f - f')^2 : f, f' \in \mathcal{F} \text{ and } \lambda_n h \leq r\}$ *with* $\sup_{f \in \mathcal{F}} \|f\|_{\mathbb{L}_\infty(\lambda_n)} < \infty$. *Then, we can find* $r_0 > 0$, *such that if* $0 < r \leq r_0$ *and* $n \geq \mathrm{Pdim}(\mathcal{F})$,

$$\mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_r} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(\boldsymbol{x}_i) \precsim \sqrt{\frac{r \log(1/r) \, \mathrm{Pdim}(\mathcal{F}) \log n}{n}}.$$

Proof. Let $B = 4 \sup_{f \in \mathcal{F}} \|f\|^2_{\mathbb{L}_\infty(\lambda_n)}$. We first fix $\epsilon \leq \sqrt{2Br}$ and let $h = f - f'$ be a member of $\mathcal{H}_r$ with $f, f' \in \mathcal{F}$. We use the notation $\mathcal{F}_{|\boldsymbol{x}_{1:n}} = \{(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))^\top : f \in \mathcal{F}\}$. Suppose that $\boldsymbol{v}^f, \boldsymbol{v}^{f'} \in \mathcal{C}(\epsilon; \mathcal{F}_{|\boldsymbol{x}_{1:n}}, \|\cdot\|_\infty)$

be such that $|\boldsymbol{v}_i^f - f(\boldsymbol{x}_i)|, |\boldsymbol{v}_i^{f'} - f'(\boldsymbol{x}_i)| \leq \epsilon$, for all i. Here $\mathcal{C}(\epsilon; \mathcal{F}_{|_{\boldsymbol{x}_{1:n}}}, \|\cdot\|_\infty)$ denotes the ϵ cover of $\mathcal{F}_{|_{\boldsymbol{x}_{1:n}}}$ w.r.t. the ℓ_∞-norm. Let $\boldsymbol{v} = \boldsymbol{v}^f - \boldsymbol{v}^{f'}$ Then

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (h(\boldsymbol{x}_i) - \boldsymbol{v}_i^2)^2 =& \frac{1}{n}\sum_{i=1}^n ((f(\boldsymbol{x}_i) - f'(\boldsymbol{x}_i))^2 - (v_i^f - v_i^{f'})^2)^2 \\
\leq& \frac{2}{n}\sum_{i=1}^n ((f(\boldsymbol{x}_i) - f'(\boldsymbol{x}_i))^2 + (v_i^f - v_i^{f'})^2)((f(\boldsymbol{x}_i) - f'(\boldsymbol{x}_i)) - (v_i^f - v_i^{f'}))^2 \qquad (40)\\
\precsim& \epsilon^2. \qquad (41)
\end{aligned}$$

Here (40) follows from the fact that $(t^2 - r^2)^2 = (t+r)^2(t-r)^2 \leq 2(t^2+r^2)(t-r)^2$, for any $t, r \in \mathbb{R}$. Hence, from the above calculations, $\mathcal{N}(\epsilon; \mathcal{H}_r, \mathbb{L}_2(\lambda_n)) \leq \left(\mathcal{N}\left(a_1\epsilon; \mathcal{F}, \mathbb{L}_\infty(\lambda_n)\right)\right)^2$, for some absolute constant a_1.

$$\operatorname{diam}^2(\mathcal{H}_r, \mathbb{L}_2(\lambda_n)) = \sup_{h,h' \in \mathcal{H}_r} \frac{1}{n}\sum_{i=1}^n (h(\boldsymbol{x}_i) - h'(\boldsymbol{x}_i))^2 \leq 2 \sup_{h \in \mathcal{H}_r} \frac{1}{n}\sum_{i=1}^n h^2(\boldsymbol{x}_i) \leq 2B \sup_{h \in \mathcal{H}_r} \frac{1}{n}\sum_{i=1}^n h(\boldsymbol{x}_i) \leq 2Br.$$

Hence, $\operatorname{diam}(\mathcal{H}_r, \mathbb{L}_2(\lambda_n)) \leq \sqrt{2Br}$. Thus from Wainwright (2019, Theorem 5.22)

$$\begin{aligned}
\mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_r} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(\boldsymbol{x}_i) \precsim& \int_0^{\sqrt{2Br}} \sqrt{\frac{1}{n}\log \mathcal{N}(\epsilon; \mathcal{H}_r, \mathbb{L}_2(\lambda_n))}d\epsilon \\
\leq& \int_0^{\sqrt{2Br}} \sqrt{\frac{2\operatorname{Pdim}(\mathcal{F})}{n}\log\left(\frac{a_2 n}{\epsilon}\right)}d\epsilon \\
\precsim& \sqrt{2Br}\sqrt{\frac{\operatorname{Pdim}(\mathcal{F})\log n}{n}} + \int_0^{\sqrt{2Br}} \sqrt{\frac{\operatorname{Pdim}(\mathcal{F})}{n}\log(a_2/\epsilon)}d\epsilon \\
\precsim& \sqrt{\frac{r\log(1/r)\operatorname{Pdim}(\mathcal{F})\log n}{n}}. \qquad (42)
\end{aligned}$$

Here, (42) follows from Lemma S.1 of the supplement. □

Lemma 24. $\mathrm{KL}(p_{\Psi,\boldsymbol{\theta}} \| p_{\Psi,\boldsymbol{\theta}'}) = d_\phi\left(\boldsymbol{\mu}(\boldsymbol{\theta}) \| \boldsymbol{\mu}(\boldsymbol{\theta}')\right).$

Proof. To prove this result, we make the following observations.

$$\begin{aligned}
\mathrm{KL}(p_{\Psi,\boldsymbol{\theta}} \| p_{\Psi,\boldsymbol{\theta}'}) =& \mathbb{E}_{\boldsymbol{x} \sim p_{\Psi,\boldsymbol{\theta}}}\left(d_\phi(\boldsymbol{x} \| \boldsymbol{\mu}(\boldsymbol{\theta}') - d_\phi(\boldsymbol{x} \| \boldsymbol{\mu}(\boldsymbol{\theta})\right) \\
=& \mathbb{E}_{\boldsymbol{x} \sim p_{\Psi,\boldsymbol{\theta}}}\left(\phi(\boldsymbol{x}) - \phi(\boldsymbol{\mu}(\boldsymbol{\theta}')) - \langle \nabla\phi(\boldsymbol{\mu}(\boldsymbol{\theta}'), \boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}')\rangle - (\phi(\boldsymbol{x}) - \phi(\boldsymbol{\mu}(\boldsymbol{\theta})) - \langle \nabla\phi(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})\rangle)\right) \\
=& \mathbb{E}_{\boldsymbol{x} \sim p_{\Psi,\boldsymbol{\theta}}}\left(\phi(\boldsymbol{\mu}(\boldsymbol{\theta})) - \phi(\boldsymbol{\mu}(\boldsymbol{\theta}')) - \langle \nabla\phi(\boldsymbol{\mu}(\boldsymbol{\theta}'), \boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}')\rangle + \langle \nabla\phi(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})\rangle\right) \\
=& d_\phi\left(\boldsymbol{\mu}(\boldsymbol{\theta}) \| \boldsymbol{\mu}(\boldsymbol{\theta}')\right). \qquad (43)
\end{aligned}$$

Here, (43) follows from noting that $E_{\boldsymbol{x} \sim p_{\Psi,\boldsymbol{\theta}}} \boldsymbol{x} = \boldsymbol{\mu}(\boldsymbol{\theta})$. □

Lemma 25. *For any $\delta \leq 1/e$, $\int_0^\delta \sqrt{\log(1/\epsilon)}d\epsilon \leq 2\delta\sqrt{\log(1/\delta)}$.*

Proof. We start by making a transformation $x = \log(1/\epsilon)$ and observe that,

$$\int_0^\delta \sqrt{\log(1/\epsilon)}d\epsilon = \int_{\log(1/\delta)}^\infty \sqrt{x}e^{-x}dx = \int_{\log(1/\delta)}^\infty \sqrt{x}e^{-x/2}e^{-x/2}dx$$

$$
\begin{aligned}
&\leq \sqrt{\log(1/\delta)} e^{-\frac{1}{2}\log(1/\delta)} \int_{\log(1/\delta)}^{\infty} e^{-x/2} dx \qquad (44)\\
&= 2\sqrt{\log(1/\delta)} e^{-\frac{1}{2}\log(1/\delta)} e^{-x/2}\big|_{\log(1/\delta)}^{\infty}\\
&= 2\sqrt{\log(1/\delta)} e^{\log(1/\delta)}\\
&= 2\delta\sqrt{\log(1/\delta)}.
\end{aligned}
$$

In the above calculations, (44) follows from the fact that the function $\sqrt{x}e^{-x/2}$ is decreasing when $x \geq 1$. □

Lemma 26. *Let $Z \sim p_{\Psi,\theta}$, with $\theta \in \Theta = \mathbb{R}$. Then, $Z - \mathbb{E}Z$ is σ_1-SubGaussian.*

Proof. We observe the following,

$$
\begin{aligned}
\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} = e^{-\lambda\nabla\Psi(\theta)} \int e^{\lambda z} e^{\theta z - \Psi(\theta)} h(z) d\tau(z) &= e^{-\lambda\nabla\Psi(\theta) - \Psi(\theta)} \int e^{(\theta+\lambda)z} h(z) d\tau(z)\\
&= e^{\Psi(\theta+\lambda) - \Psi(\theta) - \lambda\nabla\Psi(\theta)} \leq e^{\frac{\sigma_1 \lambda^2}{2}}.
\end{aligned}
$$

□

Lemma 27. *Suppose that $Z_1, \ldots, Z_n$ are independent and identically distributed sub-Gaussian random variables with variance proxy σ^2 and suppose that $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$. Then with probability at least $1 - \delta$,*

$$
\frac{1}{n} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} Z_i f(\boldsymbol{x}_i) - \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} Z_i f(\boldsymbol{x}_i) \precsim b\sigma\sqrt{\frac{\log(1/\delta)}{n}}.
$$

Proof. Recall that for a random variable, Z, $\|Z\|_{\psi_2} = \sup_{p\geq 1} \frac{(\mathbb{E}|Z|^p)^{1/p}}{\sqrt{p}}$. Let $g(Z) = \frac{1}{n}\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} Z_i f(\boldsymbol{x}_i)$. Using the notations of Maurer and Pontil (2021), we note that

$$
\begin{aligned}
\|g_k(Z)\|_{\psi_2} &= \frac{1}{n} \left\| \sup_{f\in\mathcal{F}} \left(\sum_{i\neq k} z_i f(\boldsymbol{x}_i) + Z_k f(\boldsymbol{x}_k) \right) - \mathbb{E}_{Z_k'} \sup_{f\in\mathcal{F}} \left(\sum_{i\neq k} z_i f(\boldsymbol{x}_i) + Z_k' f(\boldsymbol{x}_k) \right) \right\|_{\psi_2}\\
&\leq \frac{1}{n} \left\| \mathbb{E}_{Z_k'} |Z_k - Z_k' f(\boldsymbol{x}_k)| \right\|_{\psi_2}\\
&\leq \frac{b}{n} \left\| \mathbb{E}_{Z_k'} |Z_k - Z_k'| \right\|_{\psi_2} \qquad (45)\\
&\leq \frac{b}{n} \|Z_k - Z_k'\|_{\psi_2}\\
&\leq \frac{2b}{n} \|Z_k\|_{\psi_2}\\
&\precsim \frac{b\sigma}{n}.
\end{aligned}
$$

Here, (45) follows from (Maurer and Pontil, 2021, Lemma 6). Thus, $\left\| \sum_{k=1}^{n} \|g_k(Z)\|_{\psi_2}^2 \right\|_\infty \precsim b^2\sigma^2/n$. Hence applying (Maurer and Pontil, 2021, Theorem 3), we note that with probability at least $1 - \delta$,

$$
\frac{1}{n} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} Z_i f(\boldsymbol{x}_i) - \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} Z_i f(\boldsymbol{x}_i) \precsim b\sigma\sqrt{\frac{\log(1/\delta)}{n}}.
$$

□

Lemma 28 (Lemma B.1 of Telgarsky and Dasgupta (2013)). *If differentiable f is r_1 strongly convex, then $B_f(x\|y) \geq r_1\|x-y\|_2^2$. Furthermore, if differentiable f has Lipschitz gradients with parameter r_2 with respect to $\|\cdot\|_2$, then $B_f(x\|y) \leq r_2\|x-y\|_2^2$.*

Acknowledgment

We gratefully acknowledge the support of the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and #814639 and the NSF's support of FODSI through grant DMS-2023505.

References

Anthony, M. and Bartlett, P. (2009). *Neural network learning: Theoretical foundations*. Cambridge University Press.

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, 6(10).

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301.

Bousquet, O. (2002). *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Biologische Kybernetik.

Chakraborty, S. and Bartlett, P. (2024a). A statistical analysis of wasserstein autoencoders for intrinsically low-dimensional data. In *The Twelfth International Conference on Learning Representations*.

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

Chen, M., Jiang, H., Liao, W., and Zhao, T. (2019). Efficient approximation of deep relu networks for functions on low dimensional manifolds. *Advances in neural information processing systems*, 32.

Chen, M., Jiang, H., Liao, W., and Zhao, T. (2022). Nonparametric regression on low-dimensional manifolds using deep relu networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253.

Cover, T. M. and Thomas, J. A. (2005). *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ.

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

Donahue, J., Krähenbühl, P., and Darrell, T. (2017). Adversarial feature learning. In *International Conference on Learning Representations*.

Dudley, R. M. (1969). The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. (2018). Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31.

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

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

Jiao, Y., Shen, G., Lin, Y., and Huang, J. (2021). Deep nonparametric regression on approximately low-dimensional manifolds. *arXiv preprint arXiv:2104.06708*.

Kakade, S., Shalev-Shwartz, S., Tewari, A., et al. (2009). On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09. pdf*, 2(1):35.

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

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lehmann, E. L. and Casella, G. (2006). *Theory of Point Estimation*. Springer Science & Business Media.

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

Maurer, A. and Pontil, M. (2021). Concentration inequalities under sub-gaussian and sub-exponential conditions. *Advances in Neural Information Processing Systems*, 34:7588–7597.

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

Paul, D., Chakraborty, S., Das, S., and Xu, J. (2021). Uniform concentration bounds toward a unified framework for robust clustering. *Advances in Neural Information Processing Systems*, 34:8307–8319.

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

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. (2020). The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*.

Posner, E. C., Rodemich, E. R., and Rumsey Jr, H. (1967). Epsilon entropy of stochastic processes. *The Annals of Mathematical Statistics*, pages 1000–1020.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897.

Shen, Z., Yang, H., and Zhang, S. (2019). Nonlinear approximation via compositions. *Neural Networks*, 119:74–84.

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

Suzuki, T. (2018). Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*.

Suzuki, T. and Nitanda, A. (2021). Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. *Advances in Neural Information Processing Systems*, 34:3609–3621.

Telgarsky, M. J. and Dasgupta, S. (2013). Moment-based uniform deviation bounds for k-means and friends. *Advances in Neural Information Processing Systems*, 26.

Tsigler, A. and Bartlett, P. L. (2023). Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24:123–1.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, Springer New York, NY, 1 edition. Published: 26 November 2008.

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

Vardi, G. (2023). On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6):86–93.

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

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

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

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599.

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