A Statistical Analysis of Deep Federated Learning for Intrinsically Low-dimensional Data

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

[1]Department of Statistics, University of California, Berkeley

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

[3]Google DeepMind

Abstract

Federated Learning (FL) has become a revolutionary paradigm in collaborative machine learning, placing a strong emphasis on decentralized model training to effectively tackle concerns related to data privacy. Despite significant research on the optimization aspects of federated learning, the exploration of generalization error, especially in the realm of heterogeneous federated learning, remains an area that has been insufficiently investigated, primarily limited to developments in the parametric regime. This paper delves into the generalization properties of deep federated regression within a two-stage sampling model. Our findings reveal that the intrinsic dimension, characterized by the entropic dimension, plays a pivotal role in determining the convergence rates for deep learners when appropriately chosen network sizes are employed. Specifically, when the true relationship between the response and explanatory variables is described by a β-Hölder function and one has access to n independent and identically distributed (i.i.d.) samples from m participating clients, for participating clients, the error rate scales at most as $\tilde{\mathcal{O}}\left((mn)^{-2\beta/(2\beta+\bar{d}_{2\beta}(\lambda))}\right)$, whereas for non-participating clients, it scales as $\tilde{\mathcal{O}}\left(\Delta \cdot m^{-2\beta/(2\beta+\bar{d}_{2\beta}(\lambda))} + (mn)^{-2\beta/(2\beta+\bar{d}_{2\beta}(\lambda))}\right)$. Here $\bar{d}_{2\beta}(\lambda)$ denotes the corresponding 2β-entropic dimension of λ, the marginal distribution of the explanatory variables. The dependence between the two stages of the sampling scheme is characterized by Δ. Consequently, our findings not only explicitly incorporate the "closeness" of the clients, but also highlight that the convergence rates of errors of deep federated learners are not contingent on the nominal high dimensionality of the data but rather on its intrinsic dimension.

1 Introduction

Federated Learning (FL) stands at the forefront of collaborative machine learning techniques, revolutionizing data privacy and model decentralization in the digital landscape. This innovative approach, first introduced by Google in 2016 through its application in Gboard, has garnered substantial attention and research interest due to its potential to train machine learning models across distributed devices while preserving data privacy and security (McMahan et al., 2017). By enabling training on decentralized data sources such as mobile devices, FL addresses privacy concerns inherent in centralized model training paradigms (Zhang et al., 2021). This transformative framework allows devices to collaboratively learn a shared model while keeping sensitive information local, presenting a promising path forward for advancing technologies in a privacy-preserving manner.

From a theoretical standpoint, researchers have delved into comprehending the fundamental properties of Federated Learning (FL), mainly focusing on its optimization characteristics. While a substantial body of existing experimental and theoretical work centers on the convergence of optimization across training datasets (Li et al., 2020; Karimireddy et al., 2020; Mitra et al., 2021; Mishchenko et al., 2022; Yun et al., 2022), the exploration of generalization error, a crucial aspect in machine learning, appears to have received less meticulous scrutiny within the domain of heterogeneous federated learning. The existing research on the generalization error of FL primarily focuses on actively participating clients (Mohri et al., 2019; Qu et al., 2022), neglecting the disparities between these observed data distributions from actively participating clients and the unobserved distributions inherent in passively nonparticipating clients. In practical federated settings, a multitude of factors, such as network reliability or client availability, influence the likelihood of a client's participation in the training process. Consequently, the actual participation rate may be small, leading to a scenario where numerous clients never partake in the training phase (Kairouz et al., 2021; Yuan et al., 2021).

From a generalization viewpoint, when one has access to m participating clients and each client generates n many i.i.d. observations, Mohri et al. (2019) showed a rate of $\mathcal{O}(1/\sqrt{mn})$ holds for the excess risk of the participating clients. Chen et al. (2020) improved upon this risk bound by deriving fast rates of $\mathcal{O}(1/mn)$ for the expected excess risk for bounded losses. Recently, Hu et al. (2023) derived the generalization bounds for participating and nonparticipating clients under a generic unbounded loss under different model assumptions such as small-ball property or sub-Weibullness of the underlying distributions.

Despite the growing interest on the problem, the current literature suffers in many aspects. Firstly, the generalization bounds derived in the literature are parametric in nature and overlook the misspecification error inherent in the model assumptions. Secondly, the derived bounds do not take into account the size of the networks used and its impact on the generalization performance. Furthermore, one key aspect ignored by the current framework is that the data on which these models are trained are typically intrinsically low-dimensional in nature (Pope et al., 2020). However, the current bounds do not explore the dependence the intrinsic dimension of the data, questioning the efficacy of the current theoretical understanding concerning the real-world complexities of the problem.

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, 2019) and Generative models (Huang et al., 2022) exhibit a decay pattern that depend only on the intrinsic dimension of the data. Notably, Nakada and Imaizumi (2020) and Huang et al. (2022) showed that the excess risk decays as $\mathcal{O}(n^{-1/\mathcal{O}(\overline{\dim}_M(\mu))})$, where $\overline{\dim}_M(\mu)$ denotes the Minkowski dimension of the underlying distribution (see Section 2.2 for a detailed overview). Nevertheless, it is important to highlight that the Minkowski dimension primarily focuses on measuring the growth rate in the covering number of the *entire* support, overlooking scenarios where the distribution may have higher concentrations of mass within a specific sub-regions. As a result, the Minkowski dimension often overestimates the intrinsic dimension of the data distribution, leading to slower rates of statistical convergence. In contrast, some studies (Chen et al., 2022, 2019; Jiao et al., 2021; Dahal et al., 2022) attempt to impose a smooth Riemannian manifold structure on this support and characterize the rate through the dimension of this manifold. However, this assumption is not only very strong and unverifiable in practical terms, but also ignores the possibility that the data may be concentrated only in certain sub-regions and thinly spread over the rest, again resulting in an overestimate.

Recent insights from the optimal transport literature introduce the Wasserstein dimension (Weed and Bach, 2019), overcoming these limitations and providing a more accurate characterization of convergence rates when estimating a distribution through the empirical measure. Furthermore, advancements in this field introduce the entropic dimension (Chakraborty and Bartlett, 2024), building upon Dudley's seminal work (Dudley, 1969), and can be applied to describe the convergence rates for Bidirectional Generative Adversarial Networks (BiGANs) (Donahue et al., 2017). Remarkably, this entropic dimension is no larger than the Wasserstein and Minkowski dimensions, resulting in faster rates of convergence for the sample estimator. However, there has been no developments in incorporating these fast rates available in the current deep learning theory literature for federated learning possibly due to the complicated heterogeneity between the clients and inter-dependencies among the data points generated by individual clients.

To address the gap between the theory and practice of federated learning as highlighted above, our work presents a comprehensive examination of the generalization error in a regression context, employing a two-level framework that effectively addresses the overlooked gaps within the current literature. This framework uniquely encapsulates both the diversity and interrelationships present among clients' distributions as well as addresses the misspecification error absent in the present literature. Furthermore, we characterize the low-dimensional nature of the data distribution through the entropic dimension, which is more efficient compared to the Minkowski dimension, which is the benchmark in the deep learning theory literature (Nakada and Imaizumi, 2020; Huang et al., 2022). Our utilization of the entropic dimension results in superior bounds, surpassing those derived from other dimensions like the Minkowski and Wasserstein dimensions.

Contributions: The main contributions of this paper can be summarized as follows:

- We study the generalization properties of deep federated learning in a two-stage Bayesian sampling setting when the relation between the response and explanatory variables can be expressed through a β-Hölder function and an additive sub-Gaussian noise.

- We show that when one has access to n i.i.d. samples from each of the m participating clients, the excess risk for the participating clients scales as $\tilde{\mathcal{O}}\left((mn)^{-2\beta/(2\beta+\bar{d}_{2\beta}(\lambda))}\right)$, where $\bar{d}_{2\beta}(\lambda)$ denotes the 2β-entropic dimension of λ, the marginal distribution of the explanatory variables.

- Furthermore, for nonparticipating clients, the error rate scales as $\tilde{\mathcal{O}}\left(\Delta \cdot m^{-(2\beta/\bar{d}_{2\beta}(\lambda)+2\beta)} + (mn)^{-(2\beta/\bar{d}_{2\beta}(\lambda)+2\beta)}\right)$, primarily depending on the number of participating clients, m, when a large amount of data is available for each participating client. Here $\Delta = \min\left\{\|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1}, 1\right\}$ characterizes the closeness of the client's distribution in terms of the Orlicz-1 norm of the discrepancy in terms of the χ^2-divergence.

Organization: The remainder of the paper is organized as follows. In Section 2, we introduce the necessary notations and background. In Section 3 we introduce the problem at hand, followed by a simulation study in Section 4 to understand the effect of the intrinsic dimension on the error bounds. In Section 5, we discuss the main theoretical results of the paper along with the necessary assumptions. We then give a brief proof overview of the main results in Section 6, followed by concluding remarks and discussions in Section 7.

2 Background

2.1 Notations and Definitions

This section recalls some of the notations and background necessary for our theoretical analyses. We say $A \lesssim B$ (for $A, B \geqslant 0$) if there exists a constant $C > 0$, independent of n, such that $A \leqslant CB$. For a function $f : \mathcal{S} \to \mathbb{R}$ (with $\mathcal{S}$ being Polish) and a probability measure ν on $\mathcal{S}$, $\operatorname{ess\,sup}^{\nu}_{x \in \mathcal{S}} f(x) = \inf\left\{a : \nu\left(f^{-1}\left((a,\infty)\right)\right) = 0\right\}$. For any function $f : \mathcal{S} \to \mathbb{R}$, and any measure γ on $\mathcal{S}$, let $\|f\|_{\mathbb{L}_p(\gamma)} := \left(\int_{\mathcal{S}} |f(x)|^p d\gamma(x)\right)^{1/p}$, if $0 < p < \infty$. Also let, $\|f\|_{\mathbb{L}_\infty(\gamma)} := \operatorname{ess\,sup}^{\gamma}_{x \in \mathcal{S}} |f(x)|$. We say $A_n = \tilde{\mathcal{O}}(B_n)$ if $A_n \leqslant B_n \times \text{polylog}(n)$, for some factor polylog(n) that is a polynomial in $\log n$.

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

Definition 2 (Neural networks)**.** Let $L \in \mathbb{N}$ and $\{N_i\}_{i\in[L]} \in \mathbb{N}$. 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, $\{W_i\}_{1\leqslant i\leqslant L}$ are known as weights, and $\{b_i\}_{1\leqslant i\leqslant L}$ are known as biases. $\{\sigma_i\}_{1\leqslant i\leqslant L-1}$ 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) $\mathcal{B}(f) = \max_{1\leqslant j\leqslant L}(\|b_j\|_\infty) \vee \|W_j\|_\infty$ to denote the maximum absolute value of the weights and biases.

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

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

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 $\|f\|_{\mathcal{H}^\beta} := \sum_{\boldsymbol{s}:0\leqslant|\boldsymbol{s}|\leqslant\lfloor\beta\rfloor} \|\partial^{\boldsymbol{s}} f\|_\infty + \sum_{\boldsymbol{s}:|\boldsymbol{s}|=\lfloor\beta\rfloor} \sup_{x\neq y} \frac{\|\partial^{\boldsymbol{s}} f(x) - \partial^{\boldsymbol{s}} f(y)\|_\infty}{\|x-y\|_\infty^{\beta-\lfloor\beta\rfloor}} < \infty$. 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, $\mathcal{H}^\beta(\mathcal{S}_1, \mathcal{S}_2, C) = \{f : \mathcal{S}_1 \to \mathcal{S}_2 : \|f\|_{\mathcal{H}^\beta} \leqslant C\}$. Here, both $\mathcal{S}_1$ and $\mathcal{S}_2$ are both subsets of real vector spaces.

Definition 4 (χ^2-divergence)**.** Suppose that P and Q are distributions on $[0,1]^d$. Then,

$$\chi^2(P, Q) = \begin{cases} \int \left(\frac{dP}{dQ} - 1\right)^2 dQ & \text{if } P \ll Q \\ \infty & \text{Otherwise.} \end{cases}$$

Definition 5 (Orlicz norm, Vershynin, 2018)**.** For a random variable X, the ψ_p-Orlicz norm is defined as:

$$\|X\|_{\psi_p} = \inf\{t > 0 : \mathbb{E}\exp(|X^p|/t^p) \leqslant 2\}.$$

2.2 Intrinsic Dimension

It is hypothesized that real-world data, particularly vision data, is mostly constrained within a lower-dimensional structure embedded in a high-dimensional feature space (Pope et al., 2020). To quantify this reduced dimensionality, researchers have introduced various metrics to gauge the effective dimension of the underlying probability distribution that generates the data. Among these methods, the most commonly employed ones involve assessing the rate of growth of the covering number, on a logarithmic scale, for the majority of the support of this data distribution.

Let us consider a compact Polish space (Villani, 2021) denoted as $(\mathcal{S}, \varrho)$, where μ represents a probability measure defined on it. For the rest of this paper, we will assume that ϱ corresponds to the ℓ_∞-norm. The most straightforward measure of the dimension of a probability distribution is the upper Minkowski dimension Falconer (2004) of its support, and it is defined as follows:

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

This concept of dimensionality relies solely on the covering number of the support and does not presume the existence of a smooth mapping to a lower-dimensional Euclidean space. As a result, it encompasses not only smooth Riemannian manifolds but also highly non-smooth sets such as fractals. The statistical convergence properties of various estimators related to the upper Minkowski dimension have been extensively investigated 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. More recently, studies by Nakada and Imaizumi (2020) and Huang et al. (2022) demonstrated how deep learning models can leverage this inherent low-dimensionality in data, which is also reflected in their convergence rates. However, 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 that might not accurately reflect the underlying

To address the aforementioned challenge, in terms of the intrinsic dimension of a measure μ, Chakraborty and Bartlett (2024) introduced the concept of the α-entropic dimension of a measure. Before we proceed, we recall the (ϵ, τ)-cover of a measure (Posner et al., 1967) as:

$$\mathcal{N}_\epsilon(\mu, \tau) = \inf\{\mathcal{N}(\epsilon; S, \varrho) : \mu(S) \geqslant 1 - \tau\},$$

i.e. $\mathcal{N}_\epsilon(\mu, \tau)$ counts the minimum number of ϵ-balls required to cover a set S of probability at least $1 - \tau$, under the probability measure μ.

Definition 6 (Entropic Dimension, Chakraborty and Bartlett, 2024)**.** For any $\alpha > 0$, we define the α-entropic dimension of μ as:

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

The α-entropic dimension extends Dudley's entropic dimension (Dudley, 1969) to characterize the convergence rate for the Bidirectional Generative Adversarial Network (GAN) problem (Donahue et al., 2017). It has been demonstrated that the entropic dimension is no larger than the upper Minkowski dimension and the upper Wasserstein dimension (Weed and Bach, 2019). Moreover, strict inequality holds even for simple examples. For example, for the Pareto distribution, $p_\gamma(x) = \gamma x^{-(\gamma+1)} \mathbb{1}(x \geqslant 1)$, it is easy to check that $\overline{\dim}_M(p_\gamma) = \infty$ and $\bar{d}_\alpha(p_\gamma) = 1 + \alpha/\gamma$. For a more in-depth exploration, we direct the reader to Section 3 of Chakraborty and Bartlett (2024). The study indicated that the entropic dimension serves as a more efficient means of characterizing the intrinsic dimension of data distributions compared to popular measures such as the upper Minkowski dimension or the Wasserstein dimension and enables the derivation of faster rates of convergence for the estimates.

3 Problem Statement

We let $\mathcal{X} = [0,1]^d$ be the data space and $\mathcal{Y} = \mathbb{R}$ be the outcome space. We assume that there are m clients and each client gives rise to n data points. To conceptualize the two-stage sampling framework in a Bayesian setting, we introduce an unobserved hyper-parameter θ, lying in some parameter space Θ, which we assume to be Polish and compact. This θ is used to represent a client's inner state: θ_i represents the state of the i-th participating client and we assume that $\theta_1, \dots, \theta_m$ are independent and identically distributed (i.i.d.) according to the distribution $\pi(\cdot)$ on Θ. $(\boldsymbol{x}_{ij}, y_{ij})$ denotes the j-th sample for the i-th participating client. Conditioned on θ_i, we assume that $\{\boldsymbol{x}_{ij}\}_{i=1}^n$ are i.i.d. $\lambda_{\theta_i}(\cdot)$. Furthermore, we suppose that the true regression function is $f_0(\cdot)$ and $y_{ij} = f_0(\boldsymbol{x}_{ij}) + \epsilon_{ij}$, for zero-mean sub-Gaussian random variables ϵ_{ij}'s which are i.i.d. and are independent of θ_i's and $\boldsymbol{x}_{ij}'s$. To write more succinctly,

$$\theta_1, \dots, \theta_m \overset{i.i.d.}{\sim} \pi(\cdot); \qquad \boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{in} | \theta_i \overset{i.i.d.}{\sim} \lambda_{\theta_i}(\cdot); \qquad y_{ij} = f_0(\boldsymbol{x}_{ij}) + \epsilon_{ij},\, \epsilon_{ij} \overset{i.i.d.}{\sim} \tau. \tag{2}$$

The law of ϵ_{ij}'s are denoted as $\tau(\cdot)$ for notational simplicity. Note that a similar two-level framework was also used by Mohri et al. (2019); Chen et al. (2021); Hu et al. (2023), although under a different model. We posit that this assumption holds practical merit, e.g. cross-device federated learning, where the total number of clients is typically large, and it is reasonable to presume that the m participating clients are selected at random from the pool (Reisizadeh et al., 2020; Wang et al., 2021). In this learning scenario, the training process solely engages with the m distributions $\{\lambda_{\theta_i}\}_{i=1}^m$, where as, the total number of clients and the count of non-participating clients generally far exceeds m (Xu and Wang, 2020; Yang et al., 2020). In practical terms, this two-level sampling framework not only captures the diversity among clients' distributions but also underscores the interdependence among these distributions. A similar framework has been employed in recent literature (Li et al., 2020; Yuan et al., 2021; Wang et al., 2021; Hu et al., 2023).

Throughout the remainder of the analysis, we take the loss function as the squared error loss, which emerges as a natural choice for additive noise models. In practice, one has only access to $\{(\boldsymbol{x}_{ij}, y_{ij})\}_{i\in[m],\, j\in[n]}$ and obtains an estimate for f_0 under the squared error loss as:

$$\hat{f} = \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} - f(\boldsymbol{x}_{ij}))^2. \tag{3}$$

Here, $\mathcal{F}$ is a function class, usually realized through neural networks. In this paper, we take $\mathcal{F} = \mathcal{RN}(L, W, B, R)$ for some choice of the hyper-parameters.

Under model (2), a new data point for a nonparticipating client is generated as $\theta \sim \pi(\cdot)$, $\boldsymbol{x}|\theta \sim \lambda_\theta(\cdot)$ and

$y = f_0(\boldsymbol{x}) + \epsilon$, where, $\epsilon \sim \tau(\cdot)$ and is independent of θ and $\boldsymbol{x}$.

$$\lambda(\cdot) = \int \lambda_\theta(\cdot) d\pi(\theta)$$

denotes the marginal distribution of the explanatory variables for the nonparticipating clients. The excess risk for the nonparticipating clients is denoted as,

$$\begin{aligned}\mathbb{E}_{\theta\sim\pi}\mathbb{E}_{\boldsymbol{x}|\theta\sim\lambda_\theta}\mathbb{E}_{y|\boldsymbol{x},\theta}\left((y-\hat{f}(\boldsymbol{x}))^2-(y-f_0(\boldsymbol{x}))^2\right) =&\mathbb{E}_{\theta\sim\pi}\mathbb{E}_{\boldsymbol{x}|\theta\sim\lambda_\theta}\mathbb{E}_{\epsilon\sim\tau}\left((f_0(\boldsymbol{x})+\epsilon-\hat{f}(\boldsymbol{x}))^2-\epsilon^2\right)\\ =&\mathbb{E}_{\theta\sim\pi}\mathbb{E}_{\boldsymbol{x}|\theta\sim\lambda_\theta}(\hat{f}(\boldsymbol{x})-f_0(\boldsymbol{x}))^2\\ =&\|\hat{f}-f_0\|^2_{\mathbb{L}_2(\lambda)}. \end{aligned} \tag{4}$$

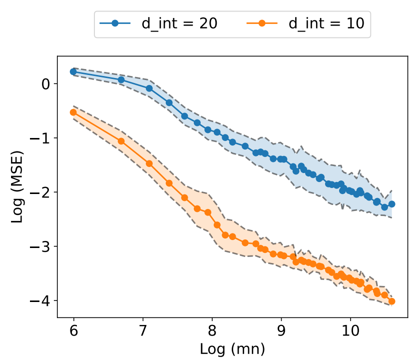
Similarly, the marginal distribution for the explanatory variable for participating clients, selected at random is $\hat{\lambda}^p_m(\cdot) = \frac{1}{m}\sum_{i=1}^m \lambda_{\theta_i}(\cdot)$. Thus, the excess risk for a participating client, selected at random is given by,

$$\begin{aligned}\mathbb{E}_{\theta\sim\hat{\pi}_m}\mathbb{E}_{\boldsymbol{x}|\theta\sim\lambda_\theta}\mathbb{E}_{y|\boldsymbol{x},\theta}\left((y-\hat{f}(\boldsymbol{x}))^2-(y-f_0(\boldsymbol{x}))^2\right) =&\mathbb{E}_{\theta\sim\hat{\pi}_m}\mathbb{E}_{\boldsymbol{x}|\theta\sim\lambda_\theta}\mathbb{E}_{\epsilon\sim\tau}\left((f_0(\boldsymbol{x})+\epsilon-\hat{f}(\boldsymbol{x}))^2-\epsilon^2\right)\\ =&\mathbb{E}_{\theta\sim\hat{\pi}_m}\mathbb{E}_{\boldsymbol{x}|\theta\sim\lambda_\theta}(\hat{f}(\boldsymbol{x})-f_0(\boldsymbol{x}))^2\\ =&\|\hat{f}-f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)}. \end{aligned} \tag{5}$$

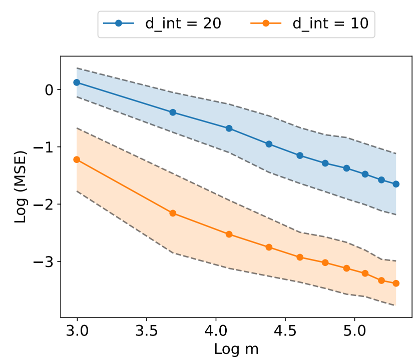
In the above calculations $\hat{\pi}_m \equiv \text{Unif}\left(\{\theta_1,\dots,\theta_m\}\right)$, denotes the empirical distribution on $\{\theta_1,\dots,\theta_m\}$. The goal of this paper is to understand how to choose $\mathcal{F}$ efficiently to obtain tight bounds on the excess risk in (4) and (5).

4 A proof of Concept

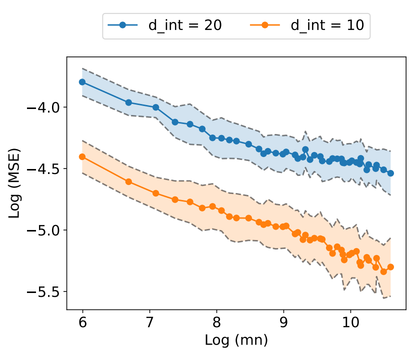
Before delving into the theoretical exploration of the problem, we conduct an experiment aimed at demonstrating that the error rates for deep federated regression are primarily contingent on the intrinsic dimension of the data. We take the true function $f_0^{(1)}(\boldsymbol{x}) = \frac{1}{d-1}\sum_{i=1}^{d-1} x_i x_{i+1} + \frac{2}{d}\sum_{i=1}^{d}\sin(2\pi x_i)\mathbb{1}\{x_i \leqslant 0.5\} + \frac{1}{d}\sum_{i=1}^{d}(4\pi(\sqrt{2}-1)^{-1}(x_i - 2^{-1/2})^2 - \pi(\sqrt{2}-1))\mathbb{1}\{x_i > 0.5\}$. This choice of f_0 was used by Nakada and Imaizumi (2020). Clearly, $f_0 \in \mathcal{H}^2(\mathbb{R}^d, \mathbb{R})$. We take $d = 30$ and the first `d_int` coordinates of $\boldsymbol{x}|\theta$ to be uniformly distributed on the $[\theta, \theta+1]^{\texttt{d\_int}}$. The remaining $d - \texttt{d\_int}$ coordinates of to be 0. θ is varied on the ($\texttt{d\_int}+4$)-dimensional cube $[0,1]^{\texttt{d\_int}+4}$. We generate $y = f_0(\boldsymbol{x}) + \epsilon$, where ϵ are Normal$(0, 0.1)$. For our experiment, we vary $m, n \in \{20, 40, \dots, 200\}$ and $\texttt{d\_int} \in \{10, 20\}$. We train a three-layer network with ReLU activation with hidden layer widths as d and take Adam (Kingma and Ba, 2015) with a learning rate 0.001. We replicate the experiments 20 times and report the logarithm of the test Mean Square Error (MSE) for both participating and non-participating clients in Figures 1(a) and 1(b). We also conduct a similar experiment with $f_0^{(2)}(\boldsymbol{x}) = \frac{1}{d}\sum_{i=1}^{n} x_i^2\mathbb{1}\{x_i \leqslant 0.5\} - \frac{1}{d}\sum_{i=1}^{n}(x_i - 3/4)\mathbb{1}\{x_i > 0.5\}$, which is a member of $\mathcal{H}^1(\mathbb{R}^d, \mathbb{R})$ and report the outcomes in Figures 1(c) and 1(d). It is clear from Figure 1 that the error rates for `d_int` $= 10$ is lower than for the case `d_int` $= 20$, further reinforcing the evidence that the generalization performance of federated learning models are dependent on their intrinsic dimension only and not on the dimension of the representative feature space. The codes pertaining to this section are available at https://github.com/SaptarshiC98/FL.



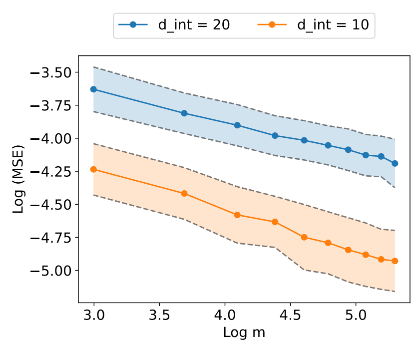
(a) Participating Clients



(b) Nonparticipating Clients



(c) Participating Clients



(d) Nonparticipating Clients

Figure 1: Average test mean squared error (MSE) is presented for both participating and non-participating clients across two distinct intrinsic dimensions, varying training sample sizes on a logarithmic scale. The error bars and bands illustrate the standard deviation over 20 replications. The top row corresponds to experiments with $f_0^{(1)}$, while the bottom row denotes the performance for $f_0^{(2)}$. Notably, the intrinsic dimensions manifest two distinct decay patterns. As anticipated from theoretical analyses, participating clients exhibit a lower error rate compared to non-participating clients.

5 Main Results and Inference

To facilitate the the theoretical analysis, we assume that the problem is smooth in terms of the learning function f_0. As a notion of smoothness, we assume that f_0 is β-Hölder. The assumption of smoothness for the regression function is common in the present literature on the statistical convergence rates for deep learners (Chen et al., 2019; Schmidt-Hieber, 2020; Nakada and Imaizumi, 2020; Chakraborty and Bartlett, 2024) and covers a wide variety of well-behaved functions. Formally,

Assumption 1. $f_0 \in \mathcal{H}^\beta([0,1]^d, \mathbb{R}, C)$, for some positive constant $C > 0$.

For notational simplicity, we define,

$$\tilde{d}_\alpha := \operatorname{ess\,sup}^\pi_{\theta \in \Theta} \bar{d}_\alpha(\lambda_\theta),$$

i.e., the maximum entropic dimension of the explanatory variable for all clients. First, for precipitating clients, the error rate in terms of the total number of samples depends on $\tilde{d}_{2\beta}$. In essence, $\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)}$ scales roughly as $\tilde{\mathcal{O}}((mn)^{-2\beta/(\tilde{d}_{2\beta}+2\beta)})$, barring poly-log factors, with high probability. This result is formally stated in Theorem 7.

Theorem 7 (Error rate for participating clients)**.** Suppose that $\lambda\left([0,1]^d\right) = 1$ and $s > \tilde{d}_{2\beta}$. We can find an $n_0 \in \mathbb{N}$, such that if $m, n \geqslant n_0$, we can choose $\mathcal{F} = \mathcal{RN}(L, W, B, R)$ in such a way that, $L \asymp \log(mn)$, $W \asymp (mn)^{\frac{s}{2\beta+s}} \log(mn)$, $\log B \asymp \log(mn)$ and $R \leqslant 2C$, such that with probability at least $1 - 2\exp\left(-(mn)^{\frac{s}{2\beta+s}}\right)$,

$$\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \precsim (mn)^{-\frac{2\beta}{s+2\beta}} \log^2(mn).$$

Second, for nonparticipating clients, the error rate $\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)}$ exhibits a scaling behavior roughly characterized by

$$\tilde{\mathcal{O}}\left(\Delta(\theta, X) m^{-2\beta/(\bar{d}_{2\beta}(\lambda)+2\beta)} + (mn)^{-2\beta/(\bar{d}_{2\beta}(\lambda)+2\beta)}\right),$$

barring log-factors as shown in Theorem 8. Here the term,

$$\Delta(\theta, \boldsymbol{x}) = \min\left\{\|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1}, 1\right\} := \inf\left\{t > 0 : \mathbb{E}_\theta \exp(|\chi^2(\lambda_\theta, \lambda)|^p / t^p) \leqslant 2\right\} \wedge 1. \tag{6}$$

characterizes the level of dependency among θ and X. It essentially quantifies the "closeness" of the distributions across clients, by evaluating how much the client-specific distribution λ_θ deviates from the mean distribution λ, given that $\theta \sim \pi(\cdot)$. When θ and X are independent, it is straightforward to see that the discrepancy measure $\Delta(\theta; X) = 0$ as well. In this scenario, where the distributions of the explanatory variables across different clients are identical, the overall error rate behaves as though one has access to mn i.i.d. samples, thus reflecting the optimal scenario for error scaling, i.e. $\tilde{\mathcal{O}}\left((mn)^{-2\beta/(\bar{d}_{2\beta}(\lambda)+2\beta)}\right)$. However, when there is some degree of dependency between θ and X, the error rate no longer scales as favorably. Instead, it scales at a rate no faster than $\tilde{\mathcal{O}}\left(m^{-2\beta/(\bar{d}_{2\beta}(\lambda)+2\beta)} \max\left\{\Delta(\theta, \boldsymbol{x}), n^{-2\beta/(\bar{d}_{2\beta}(\lambda)+2\beta)}\right\}\right)$. Therefore, the extent to which the client distributions deviate from one another plays a significant role in determining the overall efficiency and performance for deep federated learners. When one has enough samples from each of the clients, i.e., when $n \geqslant \Delta(\theta, \boldsymbol{x})^{-\frac{\bar{d}_{2\beta}(\lambda)+2\beta}{2\beta}}$, the error rate scales as $\tilde{\mathcal{O}}\left(\Delta(\theta, \boldsymbol{x}) m^{-2\beta/(\bar{d}_{2\beta}(\lambda)+2\beta)}\right)$, depending only on the number of participating clients and the discrepancy of the clients' distributions.

Theorem 8 (Error rate for nonparticipating clients)**.** Suppose that $\lambda\left([0,1]^d\right) = 1$ and $s > \bar{d}_{2\beta}(\lambda)$. We can find an $n_0' \in \mathbb{N}$, such that if $m, n \geqslant n_0'$, we can choose $\mathcal{F} = \mathcal{RN}(L, W, B, R)$ in such a way that, $L \asymp \log\left(\frac{mn}{\Delta(\theta,\boldsymbol{x})n+1}\right)$, $W \asymp \left(\frac{mn}{\Delta(\theta,\boldsymbol{x})n+1}\right)^{\frac{s}{2\beta+s}} \log\left(\frac{mn}{\Delta(\theta,\boldsymbol{x})n+1}\right)$, $\log B \asymp \log\left(\frac{mn}{\Delta(\theta,\boldsymbol{x})n+1}\right)$ and $R \leqslant 2C$, such that with probability at least $1 - 3\exp\left(-(mn)^{\frac{s}{2\beta+s}}\right) - 2\exp\left(-m^{\frac{s}{2\beta+s}}\right)$,

$$\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \lesssim m^{-\frac{2\beta}{s+2\beta}} \left(\Delta(\theta, \boldsymbol{x}) + n^{-\frac{2\beta}{s+2\beta}}\right) \times \log^3 m \left(\log^2 m + \log n\right).$$

Comparison with the Existing Literature Firstly, it is important to note that the current state of the art do not propose any dimension-based error bounds for deep federated learning. The only comparable results in the prior art is in the directions of Gaussian-noise additive regression models (Nakada and Imaizumi, 2020) and GANs (Huang et al., 2022; Dahal et al., 2022; Chakraborty and Bartlett, 2024). The negative exponent for the sample size derived by Nakada and Imaizumi (2020) is roughly, $\frac{2\beta}{2\beta+\overline{\dim}_M(\lambda)}$. Since $\bar{d}_{2\beta}(\lambda) \leqslant \overline{\dim}_M(\lambda)$ (Chakraborty and Bartlett, 2024), the negative exponent of the sample size derived in this paper, i.e. $\frac{2\beta}{2\beta+\bar{d}_{2\beta}(\lambda)} \geqslant \frac{2\beta}{2\beta+\overline{\dim}_M(\lambda)}$, resulting in better rates compared to the existing literature for additive regression models.

Bounds on the Expected Excess Risk Using the high probability bounds in Theorems 7 and 8, we can derive control the expected excess risk for both participating and nonparticipating clients. We state this result as a corollary as follows:

Corollary 9. Suppose that $\lambda\left([0,1]^d\right) = 1$. Then,

(a) if $s > \tilde{d}_{2\beta}$ and if $\mathcal{F}$ is chosen according to Theorem 7, then, $\mathbb{E}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \lesssim (mn)^{-\frac{2\beta}{s+2\beta}} \log^2(mn)$.

(b) if $s > \bar{d}_{2\beta}(\lambda)$ and if $\mathcal{F}$ is chosen according to Theorem 8,

$$\mathbb{E}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\lambda)} \lesssim m^{-\frac{2\beta}{s+2\beta}} \left(\Delta(\theta, \boldsymbol{x}) + n^{-\frac{2\beta}{s+2\beta}}\right) \times \log^3 m \left(\log^2 m + \log(mn)\right).$$

Inference for Data Supported on a Manifold Suppose that the explanatory variables are supported on a $d^\star$-dimensional compact differentiable manifold. From Propositions 8 and 9 of Weed and Bach (2019), we note that the Minkowski and lower Wasserstein dimension of λ is $d^\star$. Since $\bar{d}_\alpha(\lambda)$ lies between these two dimension (Chakraborty and Bartlett, 2024, Proposition 8), we conclude that $\bar{d}_\alpha(\lambda) = d^\star$, for all $\alpha > 0$. Hence the error rates for participating and nonparticipating clients scale as roughly $\mathcal{O}\left((mn)^{-\frac{2\beta}{d^\star+2\beta}}\right)$ and $\mathcal{O}\left(m^{-\frac{2\beta}{d^\star+2\beta}} + (mn)^{-\frac{2\beta}{d^\star+2\beta}}\right)$, respectively, excluding the excess log-factors.

Corollary 10. Suppose that the support of λ is a compact $d^\star$-dimensional differentiable manifold and let $s > d^\star$. Suppose that the assumptions of Theorem 8 hold and and $\mathcal{F}$ is chosen according to Theorem 8, with probability at least $1 - 3\exp\left(-(mn)^{\frac{s}{2\beta+s}}\right) - 2\exp\left(-m^{\frac{s}{2\beta+s}}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \lesssim m^{-\frac{2\beta}{s+2\beta}} \left(\Delta(\theta, \boldsymbol{x}) + n^{-\frac{2\beta}{s+2\beta}}\right) \times \log^3 m \left(\log^2 m + \log(mn)\right).$$

Comparison of error rates for Participating and Non-participating clients One can also infer that the error rate for participating clients decays faster than the error rates for nonparticipating clients. This is because one can show that $\tilde{d}_{2\beta} \leqslant \bar{d}_{2\beta}(\lambda)$, making the upper bound on the error converge faster for participating clients than that of nonparticipating clients. To show this, we state the Lemma 11 which ensures that $\tilde{d}_\alpha$ is at most the α-entropic dimension of λ. Thereafter, we state the result formally in Corollary 12.

Lemma 11. For any $\alpha > 0$, $\tilde{d}_\alpha \leqslant \bar{d}_\alpha(\lambda)$.

Corollary 12. Suppose $\lambda\left([0,1]^d\right)$ and let $s > \bar{d}_{2\beta}(\lambda)$. Then if $\mathcal{F}$ is chosen according to Theorem 7 and the assumptions of Theorem 7 hold, with probability at least $1 - 2\exp\left(-(mn)^{\frac{s}{2\beta+s}}\right)$,

$$\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \lesssim (mn)^{-\frac{2\beta}{s+2\beta}} \log^2(mn).$$

Network Sizes We observe that Theorem 7 and 8 imply that one can select networks with the number of weights as an exponent of m and n, which is smaller than 1. Additionally, this exponent is solely dependent on the intrinsic dimension of the explanatory variables. Furthermore, for smooth models, where β is large, it is feasible to opt for smaller networks that necessitate fewer parameters compared to non-smooth models. This is because the exponent on the number of weights in Theorems 7 and 8 decreases as β increases. Such a trend aligns with practical expectations, where simpler problems often require less complex networks in contrast to more challenging problems.

6 Proof of the Main Results

This section provides a structured overview of proofs the main results, namely Theorems 7 and 8, with comprehensive details available in the appendix. For ease of notation, we denote $\mathbb{P}(\cdot|x,\theta)$ to represent the conditional distribution given $\{\boldsymbol{x}_{i,j}\}_{i\in[m],j\in[n]}$ and $\{\theta_i\}_{i\in[m]}$. Similarly, $\mathbb{P}(\cdot|\theta)$ is used to denote the conditional distribution given $\{\theta_i\}_{i\in[m]}$. As an initial step in establishing bounds on the excess risks for both the participating and nonparticipating clients, we proceed to derive the following oracle inequality. This inequality effectively constrains the excess risk in terms of the approximation error and a generalization gap.

Lemma 13. For any $f \in \mathcal{F}$,

$$\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \leqslant \|f - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \epsilon_{ij}(\hat{f}(\boldsymbol{x}_{ij}) - f(x_{ij})). \tag{7}$$

The first term on the right-hand side (RHS) of the oracle inequality (Lemma 13) bears resemblance to an approximation error, whereas the second term resembles a generalization gap. Importantly, increasing the size of the networks in $\mathcal{F}$ can diminish the approximation error but might concurrently lead to an increased generalization gap, and conversely. The crux lies in choosing a network of optimal size that guarantees the mitigation of both these errors to sufficiently small levels. In the subsequent results, we delve into the individual control of these terms.

6.1 Generalization Gap

To effectively manage the generalization error, we employ localization techniques, as expounded by Wainwright (2019, Chapter 14). These techniques play a pivotal role in achieving fast convergence of the sample estimator to the population estimator under the $\mathbb{L}_2(\hat{\lambda}_m^p)$ and $\mathbb{L}_2(\lambda)$ norms. It is crucial to note that the true function f_0 may not be exactly representable by a ReLU network. In such cases, our alternative approach involves establishing a high-probability bound for the squared $\mathbb{L}_2(\lambda)$ norm difference between our estimated function $\hat{f}$ and f^*, where $f^* \in \mathcal{F}$ is considered sufficiently close to f_0. Our strategy unfolds in a two-step process: firstly, we derive a local complexity bound, detailed in Lemma 14. Subsequently, we leverage this local complexity bound to derive an estimate for $\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_m^p)}$, as expounded in Lemma 15. This result is then utilized to control $\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)}$ in Lemma 19. These results are presented subsequently, with proofs available in the Appendix.

Lemma 14. Suppose $\alpha \in (0, 1/2)$ and $n \geqslant \max\left\{e^{1/\alpha}, \mathrm{Pdim}(\mathcal{F})\right\}$. Then, for any $f^* \in \mathcal{F}$, with probability (under $\mathbb{P}(\cdot|\theta)$) at least, $1 - \exp\left(-n^{1-2\alpha}\right)$,

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

Lemma 15. Suppose that $n \geqslant \mathrm{Pdim}(\mathcal{F})$. Then, with probability (under $\mathbb{P}(\cdot|\theta)$) at least $1 - 3\exp(-(mn)^{1-2\alpha})$,

$$\begin{aligned}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_m^p)} \lesssim & \|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_m^p)} + (mn)^{-2\alpha} + \frac{1}{mn}\left(\mathrm{Pdim}(\mathcal{F})\log(mn) + \log\log(mn)\right) \\ & + \epsilon^2 + \frac{1}{mn}\log\mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right)\end{aligned}$$

Lemma 16. With probability at least $1 - 3\exp\left(-(mn)^{1-\alpha}\right) - 2\exp\left(-m^{1-\alpha'}\right)$,

$$\begin{aligned}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \lesssim & \epsilon^2 + (mn)^{-2\alpha} + \frac{1}{mn}\log\mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right) + \frac{1}{mn}\left(\mathrm{Pdim}(\mathcal{F})\log(mn) + \log\log(mn)\right) \\ & + \frac{\Delta(\theta, \boldsymbol{x})}{m}\left(\log\mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right) + m^{1-2\alpha'} + \log\log m\right).\end{aligned}$$

6.2 Approximation Error

To effectively bound the overall error in Lemma 13, one needs to control the approximation error, denoted by the first term of (7). 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. To control the approximation errors $\|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)}$ and $\|f^* - f_0\|^2_{\mathbb{L}_2(\lambda)}$, we employ the recent approximation results derived by Chakraborty and Bartlett (2024).

Lemma 17 (Chakraborty and Bartlett, 2024)**.** Suppose that $f \in \mathcal{H}^\alpha(\mathbb{R}^d, \mathbb{R}, C)$, for some $C > 0$ and let $s > \bar{d}_{\alpha p}(\mu)$. Then, we can find constants ϵ_0 and a, that might depend on α, d and C, such that, for any $\epsilon \in (0, \epsilon_0]$, there exists a ReLU network, $\hat{f}$ with $\mathcal{L}(\hat{f}) \leqslant a \log(1/\epsilon)$, $\mathcal{W}(\hat{f}) \leqslant a \log(1/\epsilon)\epsilon^{-s/\alpha}$, $\mathcal{B}(\hat{f}) \leqslant a\epsilon^{-1/\alpha}$ and $\mathcal{R}(\hat{f}) \leqslant 2C$, that satisfies, $\|f - \hat{f}\|_{\mathbb{L}_p(\gamma)} \leqslant \epsilon$.

It is noteworthy that when $\text{supp}(\lambda)$ possesses a finite Minkowski dimension, as per Chakraborty and Bartlett (2024, Proposition 8 (c)), we observe that $\bar{d}_{\alpha p} \leqslant \overline{\dim}_M(\mu)$. Consequently, the number of weights needed for an ϵ-approximation, in the $\mathbb{L}_p$ sense, is limited to at most $\mathcal{O}(\epsilon^{-\bar{d}_{\alpha p}/\alpha} \log(1/\epsilon))$. This result improves upon the bounds $\mathcal{O}(\epsilon^{-\overline{\dim}_M(\mu)/\alpha}$, derived by Nakada and Imaizumi (2020) as a special case. It is crucial to highlight that the requisite number of weights for low-dimensional data, specifically when $\bar{d}_{\alpha p}(\gamma) \ll d$, is notably smaller than $\mathcal{O}(\epsilon^{-d/\alpha} \log(1/\epsilon))$. This stands in contrast to the scenario when approximating over the entire space with respect to the ℓ_∞-norm (Yarotsky, 2017; Chen et al., 2019). Using the above results, we are now ready to formally prove Theorem 7 and 8 in Sections 6.3 and 6.4, respectively.

6.3 Proof of Theorem 7

Proof. Suppose that $s > \tilde{d}_{2\beta}$. From Lemma 11, we observe that $\tilde{d}_{2\beta} \geqslant \bar{d}_{2\beta}(\hat{\lambda}^p_m)$, almost surely under $\mathbb{P}(\cdot|\theta)$. Hence, $s > \bar{d}_{2\beta}(\hat{\lambda}^p_m)$ almost surely under $\mathbb{P}(\cdot|\theta)$. From Lemma 17, we can choose $\mathcal{F} = \mathcal{RN}(L, W, B, R)$ with $L \asymp \log(1/\epsilon)$, $W \asymp \epsilon^{-s/\beta}$, $\log B \asymp \log(1/\epsilon)$ and $R \leqslant 2C$ such that $\inf_{f \in \mathcal{F}} \|f - f_0\|_{\mathbb{L}_2(\hat{\lambda}^p_m)} \leqslant \epsilon$. From Lemma 15, we observe that, under $\mathbb{P}(\cdot|\theta)$, with probability at least, $1 - 3\exp\left(-(mn)^{1-2\alpha}\right)$,

$$
\begin{aligned}
&\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \\
\leqslant & 2\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + 2\|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \\
\lesssim & \|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + (mn)^{-2\alpha} + \frac{1}{mn} \text{Pdim}(\mathcal{F}) \log(mn) + \frac{\log\log(mn)}{mn} + \epsilon^2 + \frac{\log \mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right)}{mn} \\
\lesssim & \epsilon^2 + (mn)^{-2\alpha} + \frac{\log(mn)}{mn} WL \log W + \frac{\log\log(mn)}{mn} + \frac{W \log\left(\frac{2LB^L(W+1)^L}{\epsilon}\right)}{mn} \qquad (9)
\end{aligned}
$$

We choose $\epsilon \asymp (mn)^{-\frac{\beta}{2\beta+s}}$. Thus, from (9), we note that with probability at least, $1 - 3\exp\left(-(mn)^{1-2\alpha}\right)$,

$$
\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \lesssim (mn)^{-2\alpha} + (mn)^{-\frac{2\beta}{s+2\beta}} \log^3(mn)
$$

Choosing $\alpha = \frac{\beta}{2\beta+s}$, we note that with probability at least $1 - 3\exp\left(-(mn)^{\frac{s}{2\beta+s}}\right)$,

$$\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \lesssim (mn)^{-\frac{2\beta}{s+2\beta}} \log^3(mn).$$

Thus, for some constant τ,

$$\mathbb{P}\left(\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \leqslant \tau\,(mn)^{-\frac{2\beta}{s+2\beta}} \log^3(mn)\middle|\theta\right) \geqslant 1 - 3\exp\left(-(mn)^{\frac{s}{2\beta+s}}\right).$$

The result now follows from Integrating both sides w.r.t. the distribution of $\theta_1, \dots, \theta_m$. □

6.4 Proof of Theorem 8

Proof. From Lemma 17, we can choose $\mathcal{F} = \mathcal{RN}(L, W, B, R)$ with $L \asymp \log(1/\epsilon)$, $W \asymp \epsilon^{-s/\beta}$, $\log B \asymp \log(1/\epsilon)$ and $R \leqslant 2C$ such that $\inf_{f \in \mathcal{F}} \|f - f_0\|_{\mathbb{L}_2(\lambda)} \leqslant \epsilon$. From Lemma 16, we observe that, with probability at least, $1 - 3\exp\left(-(mn)^{1-2\alpha}\right) - 2\exp\left(-m^{1-2\alpha'}\right)$,

$$\begin{aligned}
&\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \\
\lesssim& \epsilon^2 + (mn)^{-2\alpha} + \frac{1}{mn}\log\mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right) \\
&+ \frac{1}{mn}\left(\mathrm{Pdim}(\mathcal{F})\log(mn) + \log\log(mn)\right) + \frac{\Delta(\theta, \boldsymbol{x})}{m}\left(\log\mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right) + m^{1-2\alpha}\right) \\
\lesssim& \epsilon^2 + (mn)^{-2\alpha} + \Delta(\theta, \boldsymbol{x})m^{-\alpha'} + \frac{\log mn}{mn}\mathrm{Pdim}(\mathcal{F}) + \left(\frac{1}{mn} + \frac{\Delta(\theta, \boldsymbol{x})}{m}\right) W \log\left(\frac{2LB^L(W+1)^L}{\epsilon}\right) + \frac{\log\log mn}{mn}
\end{aligned} \tag{10}$$

We choose $\epsilon \asymp \left(\frac{\Delta(\theta,\boldsymbol{x})}{m} + \frac{1}{mn}\right)^{\frac{\beta}{2\beta+s}}$. Thus, from (10), we note that with probability at least, $1 - 3\exp\left(-(mn)^{1-2\alpha}\right) - 2\exp\left(-m^{1-2\alpha'}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \lesssim \Delta(\theta, \boldsymbol{x})m^{-2\alpha'} + (mn)^{-2\alpha} + \left(\frac{\Delta(\theta, \boldsymbol{x})}{m} + \frac{1}{mn}\right)^{\frac{2\beta}{s+2\beta}} \log^3 m\left(\log^2 m + \log(mn)\right)$$

Choosing $\alpha = \alpha' = \frac{\beta}{2\beta+s}$, we note that with probability at least $1 - 3\exp\left(-(mn)^{\frac{s}{2\beta+s}}\right) - 2\exp\left(-m^{\frac{s}{2\beta+s}}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \lesssim m^{-\frac{2\beta}{s+2\beta}}\left(\Delta(\theta, \boldsymbol{x}) + n^{-\frac{2\beta}{s+2\beta}}\right)\log^3 m\left(\log^2 m + \log(mn)\right).$$

□

We refer the reader to the Appendix for a comprehensive and detailed exposition of the supporting lemmata and essential results.

7 Discussions and Conclusion

In this paper, we present a comprehensive framework for analyzing error rates in deep federated regression, encompassing both participating and nonparticipating clients, particularly when the data manifests an intrinsically low-dimensional structure within a high-dimensional feature space. We capture this intrinsic low-dimensionality using the entropic dimension of the explanatory variables and establish an error bound on the excess risk, accounting for both misspecification and generalization errors. The derived excess risk bounds are achieved by balancing model misspecification against stochastic errors, enabling the identification of optimal network architectures based on sample size. This framework facilitates a nuanced analysis of model accuracy for both participating and nonparticipating clients, with a focus on the interplay between sample size and intrinsic data dimensionality.

Our contributions extend the existing literature by not only broadening parametric results to encompass more general nonparametric classes but also by incorporating a characterization of the "closeness" of clients' distributions through the Orlicz-1 norm of the corresponding χ^2-divergences in our generalization bounds–a consideration previously overlooked in prior studies. Supported by empirical evidence, we also provide a theoretical comparison of error rates for participating and nonparticipating clients, demonstrating that these rates depend not on the full-data dimensionality (in terms of the number of observations) but rather on the intrinsic dimension, thereby elucidating the effectiveness of federated learning in high-dimensional contexts.

While our findings shed light on the theoretical aspects of deep federated learning, it is crucial to recognize that practical evaluation of total test error must account for an optimization error component. Accurately estimating this component is a significant challenge due to the non-convex and complex nature of the optimization problem. Nevertheless, our error analyses remain independent of the optimization process and can be seamlessly integrated with optimization analyses. Another open question remains whether characterizing the dependence between the two sampling stages using the χ^2-divergence is optimal. It is well-known that alternative divergence measures, such as Kullback-Leibler (KL) or total variation divergences, scale logarithmically compared to the χ^2-divergence. Replacing the χ^2-divergence with KL or TV may further tighten the bounds on excess risk. Additionally, exploring the minimax optimality of the proposed bounds in terms of the sample sizes m and n presents an intriguing direction for future research.

Contents

A Proofs from Section 5

Lemma 18. For any $\alpha > 0$, $\bar{d}_\alpha(\hat{\lambda}^p_m) \leqslant \tilde{d}_\alpha$, almost surely.

Proof. Let, $s > \max_{1 \leqslant i \leqslant m} \bar{d}_\alpha(\lambda_{\theta_i})$. By definition, we can find an $\epsilon_0 \in (0,1)$, such that if $\epsilon \in (0, \epsilon_0]$, $\mathcal{N}_\epsilon(\lambda_{\theta_i}, \epsilon^\alpha) \leqslant \epsilon^{-s}$, for all $i = 1, \dots, m$. By definition, we can find sets A_i's such that, $\lambda_{\theta_i}(A_i) \geqslant 1 - \epsilon^\alpha$ and $\mathcal{N}(\epsilon; A_i, \varrho) \leqslant \epsilon^{-s}$, for all $i = 1, \dots, m$. Let $A = \cup_{i=1}^m A_i$. Then, $\hat{\lambda}^p_m(A) = \frac{1}{m}\sum_{i=1}^m \lambda_{\theta_i}(A) \geqslant \frac{1}{m}\sum_{i=1}^m \lambda_{\theta_i}(A_i) \geqslant 1 - \epsilon^\alpha$. Furthermore, $\mathcal{N}(\epsilon; A, \varrho) \leqslant \sum_{i=1}^m \mathcal{N}(\epsilon; A_i, \varrho) \leqslant m\epsilon^{-s}$. Thus, $\mathcal{N}_\epsilon(\hat{\lambda}^p_m, \epsilon^\alpha) \leqslant m\epsilon^{-s}$. Hence,

$$\bar{d}_\alpha(\hat{\lambda}^p_m) \leqslant \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\hat{\lambda}^p_m, \epsilon^\alpha)}{\log(1/\epsilon)} = \lim_{\epsilon \downarrow 0} \frac{\log m + s \log(1/\epsilon)}{\log(1/\epsilon)} = s.$$

Thus, for any $s > \max_{1 \leqslant i \leqslant m} \bar{d}_\alpha(\lambda_{\theta_i})$, $\bar{d}_\alpha(\hat{\lambda}^p_m) \leqslant s$. Hence, $\bar{d}_\alpha(\hat{\lambda}^p_m) \leqslant \max_{1 \leqslant i \leqslant m} \bar{d}_\alpha(\lambda_{\theta_i}) \leqslant \tilde{d}_\alpha$. ☐

A.1 Proof of Lemma 11

Proof. Suppose that $s > \bar{d}_\alpha(\lambda) = \limsup_{\epsilon\downarrow} \frac{\log \mathcal{N}_\epsilon(\lambda,\epsilon^\alpha)}{\log(1/\epsilon)}$. Thus, we can find $\epsilon_0 \in (0,1)$, such that if $\epsilon \in (0,\epsilon_0]$, $\mathcal{N}_\epsilon(\lambda,\epsilon^\alpha) \leqslant \epsilon^{-s}$. Hence there exists S_ϵ, such that $\lambda(S_\epsilon) \geqslant 1-\epsilon^\alpha$ and $\mathcal{N}(\epsilon; S, \varrho) \leqslant \epsilon^{-s}$. Suppose that

$$\Theta_n = \{\theta \in \Theta : \lambda_\theta(S_{\epsilon/n}) \geqslant 1-\epsilon^\alpha,\ \forall \epsilon \in (0,\epsilon_0]\}.$$

For any $\epsilon \in (0,\epsilon_0]$, we note that,

$$\lambda(S_{\epsilon/n}) = \int \lambda_\theta(S_{\epsilon/n}) d\pi(\theta) = \int_{\Theta_n} \lambda_\theta(S_{\epsilon/n}) d\pi(\theta) + \int_{\Theta_n^\complement} \lambda_\theta(S_{\epsilon/n}) d\pi(\theta) \leqslant \pi(\Theta_n) + (1-\epsilon^\alpha)(1-\pi(\Theta_n))$$

$$\implies 1-(\epsilon/n)^\alpha \leqslant 1-\epsilon_k^\alpha + \epsilon_k^\alpha \pi(\Theta_2)$$

$$\implies \pi(\Theta_2) \geqslant 1-1/n^\alpha.$$

Further, note that if $\theta \in \Theta_n$, for all $\epsilon \in (0,\epsilon_0]$, $\lambda_\theta(S_{\epsilon/n}) \geqslant 1-\epsilon^\alpha$ and

$$\mathcal{N}(\epsilon; S_{\epsilon/n}, \varrho) \leqslant \mathcal{N}(\epsilon/n; S_{\epsilon/n}, \varrho) \leqslant n^s \epsilon^{-s}.$$

Thus, $\mathcal{N}_\epsilon(\lambda_\theta, \epsilon^\alpha) \leqslant n^s \epsilon^{-s}$, for all $\epsilon \leqslant \epsilon_0$. Thus,

$$\bar{d}_\alpha(\lambda_\theta) = \limsup_{\epsilon\downarrow} \frac{\log \mathcal{N}_\epsilon(\lambda\theta, \epsilon^\alpha)}{\log(1/\epsilon)} \leqslant s.$$

Hence, $\Theta_n \subseteq \{\theta \in \Theta : \bar{d}_\alpha(\lambda_\theta) \leqslant s\}$, for all $n \in \mathbb{N}$. Thus, $\pi\left(\{\theta \in \Theta : \bar{d}_\alpha(\lambda_\theta) > s\}\right) \leqslant \pi(\Theta_n^\complement) \leqslant 1/n$, for all $n \in \mathbb{N}$. Taking $n \uparrow \infty$, we get that $\pi\left(\{\theta \in \Theta : \bar{d}_\alpha(\lambda_\theta) > s\}\right) = 0$. Thus, by definition, $s > \operatorname{ess\,sup}^\pi_{\theta\in\Theta} \bar{d}_\alpha(\lambda_\theta)$, which proves the result. □

A.2 Proof of Corollary 9

Proof. We only prove part (b) of the corollary. Part (a) can be proved similarly. Suppose that a be the positive constant that honors the inequality in Theorem 8. Then, with probability at least $1-3\exp\left(-(mn)^{\frac{s}{2\beta+s}}\right) - 2\exp\left(-m^{\frac{s}{2\beta+s}}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \leqslant a m^{-\frac{2\beta}{s+2\beta}} \left(1 + n^{-\frac{2\beta}{s+2\beta}}\right) \log^3 m \left(\log^2 m + \log(mn)\right).$$

We let $\xi = a m^{-\frac{2\beta}{s+2\beta}} \left(1 + n^{-\frac{2\beta}{s+2\beta}}\right) \log^3 m \left(\log^2 m + \log(mn)\right)$. Hence,

$$\begin{aligned}
\mathbb{E}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} =& \mathbb{E}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \mathbb{1}\{\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} > \xi\} + \mathbb{E}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \mathbb{1}\{\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \leqslant \xi\} \\
\precsim& \mathbb{P}(\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} > \xi) + \xi \\
\leqslant& 3\exp\left(-(mn)^{\frac{s}{2\beta+s}}\right) + 2\exp\left(-m^{\frac{s}{2\beta+s}}\right) + a m^{-\frac{2\beta}{s+2\beta}} \left(1 + n^{-\frac{2\beta}{s+2\beta}}\right) \log^3 m \left(\log^2 m + \log(mn)\right) \\
\precsim& m^{-\frac{2\beta}{s+2\beta}} \left(1 + n^{-\frac{2\beta}{s+2\beta}}\right) \log^3 m \left(\log^2 m + \log(mn)\right),
\end{aligned}$$

when m and n are large enough. □

B Proofs from Section 6

B.1 Proof of Lemma 13

Proof. Since $\hat{f}$ is the global minimizer of $\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2$, we note that, for any $f \in \mathcal{F}$,

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \hat{f}(\boldsymbol{x}_{ij}))^2 \leqslant \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - f(\boldsymbol{x}_{ij}))^2 \tag{11}$$

$$\iff \sum_{i=1}^{m}\sum_{j=1}^{n}(f_0(x_{ij}) + \epsilon_{ij} - \hat{f}(\boldsymbol{x}_{ij}))^2 \leqslant \sum_{i=1}^{m}\sum_{j=1}^{n}(f_0(x_{ij}) + \epsilon_{ij} - f(\boldsymbol{x}_{ij}))^2. \tag{12}$$

Taking $f = f_0$, we get,

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(f_0(x_{ij}) - \hat{f}(\boldsymbol{x}_{ij}))^2 \leqslant \sum_{i=1}^{m}\sum_{j=1}^{n}(f_0(x_{ij}) - f(\boldsymbol{x}_{ij}))^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{n}\epsilon_{ij}(\hat{f}(\boldsymbol{x}_{ij}) - f(x_{ij}))$$

$$\iff \|\hat{f} - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \leqslant \|f - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\epsilon_{ij}(\hat{f}(\boldsymbol{x}_{ij}) - f(x_{ij})) \tag{13}$$

□

B.2 Proof of Lemma 14

Proof. We take $\delta = \max\left\{n^{-\alpha}, 2\|\hat{f} - f_0\|_{\mathbb{L}_2(\hat{\lambda}_{m,n})}\right\}$ and let $\eta = e^{-(mn)^{1-2\alpha}}$. We consider two cases as follows.

Case 1: $\|\hat{f} - f^*\|_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \leqslant \delta$

Then, by Lemma 26, with probability at least $1 - \exp\left(-n^{1-2\alpha}\right)$

$$\begin{aligned}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \leqslant & 2\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + 2\|f_0 - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \\ \lesssim & \|f_0 - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \sup_{g \in \mathcal{G}_\delta}\frac{1}{n}\sum_{i=1}^{n}\epsilon_{ij}g(x_{ij}) \qquad (14) \\ \lesssim & \|f_0 - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \delta\sqrt{\frac{\log(1/\eta)}{mn}} + \delta\sqrt{\frac{\text{Pdim}(\mathcal{F})\log(mn/\delta)}{mn}} \qquad (15)\end{aligned}$$

In the above calculations, (14) follows from (13). Inequality (15) follows from Lemma 26. Let $\alpha_1 \geqslant 1$ be the corresponding constant that honors the inequality in (15). Then using the upper bound on δ, we observe that,

$$\begin{aligned}&\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \\ \leqslant & \alpha_1\|f_0 - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \alpha_1\delta\sqrt{\frac{\text{Pdim}(\mathcal{F})\log(mn/\delta)}{mn}} + \alpha_1(mn)^{-2\alpha} \\ \leqslant & \alpha_1\|f_0 - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \frac{\delta^2}{16} + \frac{4\alpha_1^2}{mn}\text{Pdim}(\mathcal{F})\log(mn/\delta) + \alpha_1(mn)^{-2\alpha} \qquad (16) \\ \leqslant & \alpha_1\|f_0 - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + (1/8 + \alpha_1)(mn)^{-2\alpha} + \frac{1}{4}\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \frac{4(1+\alpha)\alpha_1^2}{mn}\text{Pdim}(\mathcal{F})\log(mn)\end{aligned}$$

$$\leqslant \alpha_1 \|f_0 - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + 2\alpha_1 (mn)^{-2\alpha} + \frac{1}{2}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \frac{1}{2}\|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \frac{4(1+\alpha)\alpha_1^2}{mn} \mathrm{Pdim}(\mathcal{F}) \log(mn)$$

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

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

Case 2: $\|\hat{f} - f^*\|_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \geqslant \delta$

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

$$\begin{aligned}
\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} &\leqslant 2\|\hat{f} - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + 2\|f_0 - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \\
&\leqslant \frac{1}{2}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + 2\|f_0 - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \\
\implies \|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} &\lesssim \|f_0 - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})}
\end{aligned}$$

Thus, from the above two cases, with probability at least, $1 - \exp\left(-(mn)^{1-2\alpha}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \lesssim (mn)^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \frac{1}{n} \mathrm{Pdim}(\mathcal{F}) \log(mn). \tag{17}$$

From equation (17), we note that, for some constant B_4,

$$\mathbb{P}\left(\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \leqslant B_4\big((mn)^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \frac{1}{mn} \mathrm{Pdim}(\mathcal{F}) \log(mn)\big) \middle| x, \theta \right) \geqslant 1 - \exp\left(-(mn)^{1-2\alpha}\right)$$

Integrating both sides w.r.t.the joint distribution of $\{\boldsymbol{x}_{ij}\}_{i \in [m], j \in [n]}$, we observe that under $\mathbb{P}(\cdot|\theta)$, with probability at least, $1 - \exp\left(-(mn)^{1-2\alpha}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \lesssim (mn)^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \frac{1}{mn} \mathrm{Pdim}(\mathcal{F}) \log(mn). \tag{18}$$

□

B.3 Proof of Lemma 15

Proof. In the proof all probabilities and expectations are w.r.t. $\mathbb{P}(\cdot|\theta)$. Let $\mathcal{C}\left(\epsilon; \mathcal{RN}(L, W, B, R), \|\cdot\|_{\mathbb{L}_\infty(\lambda)}\right)$ be an ϵ-cover of $\mathcal{RN}(L, W, B, R)$ w.r.t. the $\|\cdot\|_{\mathbb{L}_\infty(\lambda)}$-norm and let, $N = \mathcal{N}\left(\epsilon; \mathcal{RN}(L, W, B, R), \|\cdot\|_{\mathbb{L}_\infty(\lambda)}\right)$. Let, $f \in \mathcal{C}\left(\epsilon; \mathcal{RN}(L, W, B, R), \|\cdot\|_{\mathbb{L}_\infty(\lambda)}\right)$ be such that, $\|f - \hat{f}\|_{\mathbb{L}_\infty(\lambda)} \leqslant \epsilon$. Then

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \leqslant 2\|\hat{f} - f\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + 2\|f - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \leqslant 2\epsilon^2 + 2\|f - f^*\|^2_{\mathbb{L}_2(\lambda)}. \tag{19}$$

For any $g \in \mathcal{C}\left(\epsilon; \mathcal{RN}(L, W, B, R), \|\cdot\|_{\mathbb{L}_\infty(\hat{\lambda}^p_m)}\right)$, we let $Z_{ij} = (g(\boldsymbol{x}_{ij}) - f^*(\boldsymbol{x}_{ij}))^2 - \mathbb{E}(g(\boldsymbol{x}_{ij}) - f^*(\boldsymbol{x}_{ij}))^2$. Let, $u = \max\left\{v, \frac{1}{2}\|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)}\right\}$. Clearly,

$$\mathbb{E}Z_{ij}^2 = \mathrm{Var}\left((g(\boldsymbol{x}_{ij}) - f^*(\boldsymbol{x}_{ij}))^2\right) \leqslant \mathbb{E}(g(\boldsymbol{x}_{ij}) - f^*(\boldsymbol{x}_{ij}))^4 \leqslant 4R^2 \mathbb{E}(g(\boldsymbol{x}_{ij}) - f^*(\boldsymbol{x}_{ij}))^2 \leqslant 8R^2 u.$$

Furthermore, $|Z_{ij}| \leqslant 8R^2$. Thus, from Bernstein's inequality (Lemma 21), we note that,

$$\mathbb{P}\Bigg(\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}(g(\boldsymbol{x}_{ij}) - f^*(\boldsymbol{x}_{ij}))^2 \geqslant \sum_{i=1}^{m}\sum_{j=1}^{n}(g(\boldsymbol{x}_{ij}) - f^*(\boldsymbol{x}_{ij}))^2 + mnu\Bigg) \leqslant \exp\left(-\frac{mnu}{24R^2}\right)$$
$$\implies \mathbb{P}\left(\|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \geqslant \|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + u\right) \leqslant \exp\left(-\frac{mnu}{24R^2}\right) \leqslant \exp\left(-\frac{mnv}{24R^2}\right). \tag{20}$$

Thus, by union bound,

$$\mathbb{P}\Bigg(\|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \leqslant \|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + u,\ \forall\, g \in \mathcal{C}\left(\epsilon; \mathcal{RN}(L, W, B, R), \|\cdot\|_{\mathbb{L}_\infty(\lambda)}\right)\Bigg) \geqslant 1 - N\exp\left(-\frac{mnv}{24R^2}\right).$$

Thus, under $\hat{\lambda}^p_m$, with probability at least, $1 - N\exp\left(-\frac{mnv}{24R^2}\right)$, for all $g \in \mathcal{C}\left(\epsilon; \mathcal{RN}(L, W, B, R), \|\cdot\|_{\mathbb{L}_\infty(\lambda)}\right)$,

$$\begin{aligned}
\|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \leqslant& \|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + u \\
\leqslant& \|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + v + \frac{1}{2}\|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \\
\implies \|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \leqslant& 2\|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + 2v.
\end{aligned}$$

Taking $v = \frac{24R^2}{mn}\left(\log N + (mn)^{1-2\alpha}\right)$, we note that, under $\hat{\lambda}^p_m$ with probability at least, $1 - \exp(-(mn)^{1-2\alpha})$, for all $g \in \mathcal{C}\left(\epsilon; \mathcal{RN}(L, W, B, R), \|\cdot\|_{\mathbb{L}_\infty(\lambda)}\right)$,

$$\|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \leqslant 2\|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \frac{48R^2}{mn}\left(\log N + (mn)^{1-2\alpha}\right)$$

From, (19), we thus observe that under $\hat{\lambda}^p_m$, with probability at least, $1 - \exp(-(mn)^{1-2\alpha})$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \leqslant 2\epsilon^2 + 4\|f - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \frac{96R^2}{mn}\left(\log N + (mn)^{1-2\alpha}\right) \tag{21}$$
$$\leqslant 4\epsilon^2 + 8\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})} + \frac{96R^2}{mn}\left(\log N + (mn)^{1-2\alpha}\right) \tag{22}$$

Applying Lemma 14, we note that under $\hat{\lambda}^p_m$, with probability at least, $1 - 2\exp(-(mn)^{1-2\alpha})$,

$$\begin{aligned}
\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \lesssim& 4\epsilon^2 + (mn)^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{n,m})} + \frac{1}{mn}\operatorname{Pdim}(\mathcal{F})\log(mn) + \frac{\log\log(mn)}{mn} \\
&+ \frac{96R^2}{mn}\left(\log N + (mn)^{1-2\alpha}\right)
\end{aligned} \tag{23}$$

Applying Lemma 32 with $t = (mn)^{-\alpha}$ and $g : \boldsymbol{x} \mapsto (f^*(\boldsymbol{x}) - f_0(\boldsymbol{x}))^2$, we note that, with probability at least, $1 - \exp\left(-(mn)^{1-2\alpha}\right)$,

$$\|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}_{n,m})} \leqslant \|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + (mn)^{-2\alpha} \tag{24}$$

Combining (23) and (24), we note that with probability at least $1 - 3\exp(-(mn)^{1-2\alpha})$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \lesssim 4\epsilon^2 + (mn)^{-2\alpha} + \|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + \frac{1}{mn}\operatorname{Pdim}(\mathcal{F})\log(mn) + \frac{\log\log(mn)}{mn} + \frac{96R^2}{mn}\left(\log N + (mn)^{1-2\alpha}\right) \tag{25}$$

□

B.4 Proof of Lemma 16

Lemma 19. With probability at least, $1 - 3\exp\left(-(mn)^{1-2\alpha}\right) - 2\exp\left(-m^{1-2\alpha'}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \lesssim \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda)} + m^{-2\alpha'} + (mn)^{-2\alpha} + \operatorname{Pdim}(\mathcal{F})\left(\frac{\log^2 m}{m} + \frac{\log(mn)}{mn}\right) + \frac{\log\log m}{m} + \frac{\log\log(mn)}{mn} + \epsilon^2 + \frac{\log \mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right)}{mn}.$$

Proof. Suppose that $\mathcal{H} = \left\{h : \theta \mapsto \int (f - f')^2 d\lambda_\theta : f \in \mathcal{F}\right\}$. We note that if $n \geqslant \operatorname{Pdim}(\mathcal{F})$ and $r \in (0, r_0]$, then, the empirical Rademacher complexity can be bounded as,

$$\begin{aligned}\mathcal{R}_m\left(\mathcal{H}; \{\theta_i\}_{i\in[m]}\right) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\sum_{i=1}^m \sigma_i h(\theta_i) \\ &\lesssim \sqrt{\frac{r\log(1/r)\operatorname{Pdim}(\mathcal{F})\log m}{m}} &(26)\\ &\leqslant \sqrt{\frac{(\operatorname{Pdim}(\mathcal{F}))^2\log m}{m^2} + r\frac{\operatorname{Pdim}(\mathcal{F})\log(m/e\operatorname{Pdim}(\mathcal{F}))\log m}{m}} &(27)\end{aligned}$$

Here, $\sigma_i' s$ are independent Rademacher random variables. In the above calculations, (26) follows from Lemma 28 and (27) follows from Lemma 31. Applying Lemma 22, we note that, the RHS of (27) has a fixed point of r^* and $r^* \lesssim \frac{\operatorname{Pdim}(\mathcal{F})\log^2 m}{m}$. Then, by Theorem 6.1 of Bousquet (2002), we note that with probability at least $1 - e^{-x}$, for all $h \in \mathcal{H}$,

$$\int h d\pi \lesssim B_3\left(\int h d\hat{\pi}_m + \frac{\operatorname{Pdim}(\mathcal{F})\log^2 m}{m} + \frac{x}{m} + \frac{\log\log m}{m}\right), \tag{28}$$

for some absolute constant B_3. Now, taking $x = (mn)^{1-2\alpha}$ in (28), we note that, with probability at least $1 - \exp\left(-m^{1-2\alpha'}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \lesssim m^{-2\alpha'} + \|\hat{f} - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + \frac{1}{m}\operatorname{Pdim}(\mathcal{F})\log^2 m + \frac{\log\log m}{m} \tag{29}$$

Combining (29) with Lemma 15, we observe that with probability at least $1-3\exp\left(-(mn)^{1-2\alpha}\right)-\exp\left(-m^{1-2\alpha'}\right)$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \lesssim \|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + m^{-2\alpha'} + (mn)^{-2\alpha} + \frac{1}{m}\operatorname{Pdim}(\mathcal{F})\log^2 m + \frac{1}{mn}\operatorname{Pdim}(\mathcal{F})\log(mn) + \frac{\log\log m}{m} + \frac{\log\log(mn)}{mn} + \epsilon^2 + \frac{\log \mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right)}{mn} \tag{30}$$

Applying Lemma 32 with $t = m^{-\alpha'}$, $Z_i = \theta_i$ and $g : \theta \mapsto \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda_\theta)}$, we note that, with probability at least, $1 - 3\exp\left(-(mn)^{1-2\alpha}\right) - 2\exp\left(-m^{1-2\alpha'}\right)$,

$$\begin{aligned}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \lesssim & \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda)} + m^{-2\alpha'} + (mn)^{-2\alpha} + \frac{1}{m}\operatorname{Pdim}(\mathcal{F})\log^2 m \\ & + \frac{1}{mn}\operatorname{Pdim}(\mathcal{F})\log(mn) + \frac{\log\log m}{m} + \frac{\log\log(mn)}{mn} + \epsilon^2 + \frac{\log \mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right)}{mn} \end{aligned} \tag{31}$$

□

Lemma 20. Suppose that $\lambda_\theta \ll \lambda$, almost surely under π. Then, with probability at least $1-3\exp\left(-(mn)^{1-\alpha}\right) - 2\exp\left(-m^{1-\alpha'}\right)$,

$$\begin{aligned}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \lesssim & \epsilon^2 + (mn)^{-2\alpha} + \frac{1}{mn}\log \mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right) + \frac{1}{mn}\left(\operatorname{Pdim}(\mathcal{F})\log(mn) + \log\log(mn)\right) \\ & + \frac{\|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1}}{m}\left(\log \mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right) + m^{1-2\alpha'}\right).\end{aligned}$$

Proof. Let $\mathcal{C}\left(\epsilon; \mathcal{RN}(L, W, B, R), \|\cdot\|_{\mathbb{L}_\infty(\lambda)}\right)$ be an ϵ-cover of $\mathcal{RN}(L, W, B, R)$ w.r.t. the $\|\cdot\|_{\mathbb{L}_\infty(\lambda)}$-norm and let,

$$N = \mathcal{N}\left(\epsilon; \mathcal{RN}(L, W, B, R), \|\cdot\|_{\mathbb{L}_\infty(\lambda)}\right).$$

Let, $f \in \mathcal{C}\left(\epsilon; \mathcal{RN}(L, W, B, R), \|\cdot\|_{\mathbb{L}_\infty(\lambda)}\right)$ be such that, $\|f - \hat{f}\|_{\mathbb{L}_\infty(\lambda)} \leqslant \epsilon$. Then

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \leqslant 2\|\hat{f} - f\|^2_{\mathbb{L}_2(\lambda)} + 2\|f - f^*\|^2_{\mathbb{L}_2(\lambda)} \leqslant 2\epsilon^2 + 2\|f - f^*\|^2_{\mathbb{L}_2(\lambda)}. \tag{32}$$

Suppose that $\lambda_\theta \ll \lambda$, almost surely under π. Then, by Radon-Nykodym theorem, the density of λ_θ w.r.t. λ exists and is denoted by $p_\theta = \frac{d\lambda_\theta}{d\lambda}$. By Lemma 29, with probability at least, $1 - 2N\exp\left(-\frac{c_3 m v}{\|\chi^2(\lambda_\theta,\lambda)\|_{\psi_1}}\right)$, for all $g \in \mathcal{C}\left(\epsilon; \mathcal{RN}(L, W, B, R), \|\cdot\|_{\mathbb{L}_\infty(\lambda)}\right)$,

$$\|g - f^*\|^2_{\mathbb{L}_2(\lambda)} \leqslant 2\|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + 2v.$$

Taking $v = \frac{\|\chi^2(\lambda_\theta,\lambda)\|_{\psi_1}}{c_3 m}\left(\log(2N) + m^{1-2\alpha'}\right)$, we note that with probability at least, $1 - \exp(-m^{1-2\alpha'})$, for all $g \in \mathcal{C}\left(\epsilon; \mathcal{RN}(L, W, B, R), \|\cdot\|_{\mathbb{L}_\infty(\lambda)}\right)$,

$$\|g - f^*\|^2_{\mathbb{L}_2(\lambda)} \leqslant 2\|g - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + \frac{2\|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1}}{c_3 m}\left(\log(2N) + m^{1-2\alpha'}\right)$$

From, (32), we thus observe that, with probability at least $1 - \exp(-m^{1-2\alpha'})$,

$$\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \lesssim \epsilon^2 + \|f - f^*\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + \frac{\|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1}}{m}\left(\log(2N) + m^{1-2\alpha'}\right) \tag{33}$$

Applying Lemma 15, we note that, with probability at least $1 - 2\exp(-m^{1-2\alpha'})$,

$$\begin{aligned}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \lesssim & \epsilon^2 + (mn)^{-2\alpha} + \frac{1}{mn}\log \mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right) + \frac{1}{mn}\left(\operatorname{Pdim}(\mathcal{F})\log(mn) + \log\log(mn)\right) \\ & + \frac{\|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1}}{m}\left(\log \mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right) + m^{1-2\alpha'}\right)\end{aligned} \tag{34}$$

Similarly, using Lemma 29, we note that, with probability at least, $1 - \exp\left(-m^{1-2\alpha'}\right)$,

$$\|f^* - f_0\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \lesssim \|f^* - f_0\|^2_{\mathbb{L}_2(\lambda)} + \sqrt{\|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1}} m^{-2\alpha} \tag{35}$$

Combining (34) and (35), we note that with probability at least $1 - 3\exp(-(mn)^{1-2\alpha}) - 2\exp(-m^{1-2\alpha'})$,

$$\begin{aligned}\|\hat{f} - f^*\|^2_{\mathbb{L}_2(\lambda)} \lesssim & \epsilon^2 + (mn)^{-2\alpha} + \frac{1}{mn} \log \mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right) + \frac{1}{mn}\left(\mathrm{Pdim}(\mathcal{F}) \log(mn) + \log\log(mn)\right) \\ & + \frac{\|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1}}{m}\left(\log \mathcal{N}\left(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)}\right) + m^{1-2\alpha'}\right).\end{aligned}$$

□

Combining Lemmata 19 and 20, we get Lemma 16.

C Auxiliary Results

C.1 Supporting Results From the Literature

This section outlines, without proof, a selection of relevant theoretical underpinnings from the literature that are employed in this paper.

Lemma 21 (Bernstein's Inequality for Bounded Distributions, Theorem 2.8.4 of Vershynin (2018))**.** Let $X_1, \ldots, X_N$ be independent, mean zero random variables such that $|X_i| \leqslant K$ for all $i \in [N]$. Then, for every $t \geqslant 0$, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^N X_i\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{Kt}{3}}\right).$$

Here, $\sigma^2 = \sum_{i=1}^N \mathbb{E}(X_i^2)$ is the variance of the sum.

Lemma 22 (Lemma B.1 of Yousefi et al. (2018))**.** Let $c_1, c_2 > 0$ and $s > q > 0$. Then the equation $x^s - c_1 x^q - c_2 = 0$ has a unique positive solution x_0 satisfying

$$x_0 \leqslant \left(c_1^{\frac{s}{s-1}} + \frac{sc_2}{s-q}\right)^{\frac{1}{s}}.$$

Moreover, for any $x \geqslant x_0$, we have $x^s \geqslant c_1 x^q + c_2$.

Lemma 23 (Lemma 21 of Nakada and Imaizumi (2020))**.** Let $\mathcal{F} = \mathcal{RN}(W, L, B)$ be a space of ReLU networks with the number of weights, the number of layers, and the maximum absolute value of weights bounded by W, L, and B respectively. Then,

$$\log \mathcal{N}(\epsilon; \mathcal{F}, \ell_\infty) \leqslant W \log\left(\frac{2LB^L(W+1)^L}{\epsilon}\right).$$

Lemma 24 (Theorem 12.2 of Anthony and Bartlett, 2009)**.** Assume for all $f \in \mathcal{F}$, $\|f\|_\infty \leqslant M$. Denote the pseudo-dimension of $\mathcal{F}$ as $\text{Pdim}(\mathcal{F})$, then for $n \geqslant \text{Pdim}(\mathcal{F})$, we have for any ϵ and any $X_1, \ldots, X_n$,

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

C.2 Additional Lemmata

Lemma 25. Suppose that $Z_1, \ldots, Z_n$ are independent and identically distributed sub-Gaussian random variables with variance proxy σ^2 and suppose that $\|f\|_\infty \leqslant b$ for all $f \in \mathcal{F}$. Then with probability at least $1 - \delta$,

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

Proof. Let $g(Z) = \frac{1}{n} \sup_{f \in \mathcal{F}} \sum_{i=1}^n Z_i f(x_i)$. Using the notations of Maurer and Pontil (2021), we note that

$$\begin{aligned}
\|g_k(Z)\|_{\psi_2} =& \frac{1}{n} \left\| \sup_{f \in \mathcal{F}} \left(\sum_{i \neq k} z_i f(x_i) + Z_k f(x_k) \right) - \mathbb{E}_{Z'_k} \sup_{f \in \mathcal{F}} \left(\sum_{i \neq k} z_i f(x_i) + Z'_k f(x_k) \right) \right\|_{\psi_2} \\
\leqslant& \frac{1}{n} \left\| \mathbb{E}_{Z'_k} |Z_k - Z'_k f(x_k)| \right\|_{\psi_2} \\
\leqslant& \frac{b}{n} \left\| \mathbb{E}_{Z'_k} |Z_k - Z'_k| \right\|_{\psi_2} \qquad (36) \\
\leqslant& \frac{b}{n} \left\| Z_k - Z'_k \right\|_{\psi_2} \\
\leqslant& \frac{2b}{n} \|Z_k\|_{\psi_2} \\
\lesssim& \frac{b\sigma}{n}.
\end{aligned}$$

Here, (36) follows from (Maurer and Pontil, 2021, Lemma 6). Thus, $\left\| \sum_{k=1}^n \|g_k(Z)\|_{\psi_2}^2 \right\|_\infty \lesssim b^2\sigma^2/n$. Hence applying (Maurer and Pontil, 2021, Theorem 3), we note that with probability at least $1 - \delta$,

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

□

Lemma 26. Suppose that $\mathcal{G}_\delta = \left\{ f - f' : \|f - f'\|_{\mathbb{L}_\infty(\hat{\lambda}_{m,n})} \leqslant \delta \text{ and } f, f' \in \mathcal{F} \right\}$, with $\delta \leqslant 1/e$. Also let, $n \geqslant \text{Pdim}(\mathcal{F})$. Then, for any $t > 0$, under $\mathbb{P}(\cdot|\boldsymbol{x}_{1:n})$, with probability at least $1 - \eta$,

$$\sup_{g \in \mathcal{G}_\delta} \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \epsilon_{ij} g(x_{ij}) \lesssim \delta \sqrt{\frac{\log(1/\eta)}{mn}} + \delta \sqrt{\frac{\text{Pdim}(\mathcal{F}) \log(mn/\delta)}{mn}}$$

Proof. From the definition of $\mathcal{G}_\delta$, it is clear that $\log \mathcal{N}(\epsilon; \mathcal{G}_\delta, \|\cdot\|_{\mathbb{L}_\infty(\hat{\lambda}_{m,n})}) \leqslant 2 \log \mathcal{N}(\epsilon/2; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty(\hat{\lambda}_{m,n})})$. Let $Z_f = \frac{1}{\sqrt{mn}} \sum_{i=1}^m \sum_{j=1}^n \epsilon_{ik} f(x_{ij})$. Clearly, $\mathbb{E}_\epsilon Z_f = 0$. Furthermore, we observe that,

$$\mathbb{E}_\epsilon \exp(\lambda(Z_f - Z_g)) = \mathbb{E}_\epsilon \exp\left(\frac{\lambda}{\sqrt{mn}} \sum_{i=1}^m \sum_{j=1}^n \epsilon_{ij} (f(x_{ij}) - g(x_{ij})) \right)$$

$$
\begin{aligned}
&= \prod_{i=1}^{m}\prod_{j=1}^{n} \mathbb{E}_\epsilon \exp\left(\frac{\lambda}{\sqrt{mn}}\epsilon_{ij}(f(x_{ij}) - g(x_{ij}))\right) \\
&\leqslant \prod_{i=1}^{m}\prod_{j=1}^{n} \mathbb{E}_\epsilon \exp\left(\frac{\lambda^2\sigma^2}{2n}(f(x_{ij}) - g(x_{ij}))^2\right) \\
&\leqslant \exp\left(\frac{\lambda^2\sigma^2}{2mn}\sum_{i=1}^{m}\sum_{j=1}^{m}(f(x_{ij}) - g(x_{ij}))^2\right) \\
&= \exp\left(\frac{\lambda^2\sigma^2}{2}\|f-g\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})}\right).
\end{aligned}
$$

Thus, $(Z_f - Z_g)$ is $\|f-g\|^2_{\mathbb{L}_2(\hat{\lambda}_{m,n})}\sigma^2$-subGaussian. Furthermore,

$$
\sup_{f,g\in\mathcal{G}_\delta}\|f-g\|_{\mathbb{L}_2(\hat{\lambda}_{m,n})} = \sup_{f,f'\in\mathcal{F}:\|f-f'\|_{\mathbb{L}_\infty(\hat{\lambda}_{m,n})}\leqslant\delta}\|f-f'\|_{\mathbb{L}_2(\hat{\lambda}_{m,n})} \leqslant \sup_{f,f'\in\mathcal{F}:\|f-f'\|_{\mathbb{L}_\infty(\hat{\lambda}_{m,n})}\leqslant\delta}\|f-f'\|_{\mathbb{L}_\infty(\hat{\lambda}_{m,n})} \leqslant \delta.
$$

From (Wainwright, 2019, Proposition 5.22),

$$
\begin{aligned}
\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}_\delta}\frac{1}{\sqrt{mn}}\sum_{i=1}^{m}\sum_{j=1}^{n}\epsilon_{ij}g(x_{ij}) =& \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}_\delta} Z_g \qquad (37) \\
=& \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}_\delta}(Z_g - Z_{g'}) \\
\leqslant& \mathbb{E}_\epsilon \sup_{g,g'\in\mathcal{G}_\delta}(Z_g - Z_{g'}) \\
\leqslant& 32\int_0^\delta \sqrt{\log\mathcal{N}(\epsilon;\mathcal{G}_\delta,\mathbb{L}_2(\hat{\lambda}_{m,n}))}d\epsilon \\
\lesssim& \int_0^\delta \sqrt{\log\mathcal{N}(\epsilon/2;\mathcal{F},\mathbb{L}_\infty(\hat{\lambda}_{m,n}))}d\epsilon \\
\lesssim& \int_0^\delta \sqrt{\mathrm{Pdim}(\mathcal{F})\log(mn/\epsilon)}d\epsilon \\
\leqslant& \delta\sqrt{\mathrm{Pdim}(\mathcal{F})\log mn} + \sqrt{\mathrm{Pdim}(\mathcal{F})}\int_0^\delta \sqrt{\log(1/\epsilon)}d\epsilon \\
\leqslant& \delta\sqrt{\mathrm{Pdim}(\mathcal{F})\log mn} + 2\sqrt{\mathrm{Pdim}(\mathcal{F})}\delta\sqrt{\log(1/\delta)} \qquad (38) \\
\lesssim& \delta\sqrt{\mathrm{Pdim}(\mathcal{F})\log(mn/\delta)} \qquad (39)
\end{aligned}
$$

(38) follows from Lemma 30. Thus,

$$
\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}_\delta}\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\epsilon_{ij}g(x_{ij}) \lesssim \delta\sqrt{\frac{\mathrm{Pdim}(\mathcal{F})\log(mn/\delta)}{mn}} \qquad (40)
$$

Applying Lemma 25, we note that, with probability at least $1-\eta$,

$$
\sup_{g\in\mathcal{G}_\delta}\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\epsilon_{ij}g(x_{ij}) \lesssim \delta\sqrt{\frac{\log(1/\eta)}{mn}} + \delta\sqrt{\frac{\mathrm{Pdim}(\mathcal{F})\log(mn/\delta)}{mn}}. \qquad (41)
$$

□

Lemma 27. Let $\mathcal{H}_r = \{h = (f - f')^2 : f, f' \in \mathcal{F} \text{ and } \lambda_n h \leqslant r\}$ with $\sup_{f \in \mathcal{F}} \|f\|_{\mathbb{L}_\infty(\lambda_n)} < \infty$. Then, we can find $r_0 > 0$, such that if $0 < r \leqslant r_0$ and $n \geqslant \operatorname{Pdim}(\mathcal{F})$,

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

Proof. Let $B = 4 \sup_{f \in \mathcal{F}} \|f\|^2_{\mathbb{L}_\infty(\lambda_n)}$. We first fix $\epsilon \leqslant \sqrt{2Br}$ and let $h = f - f'$ be a member of $\mathcal{H}_r$ with $f, f' \in \mathcal{F}$. We use the notation $\mathcal{F}_{|x_{1:n}} = \{(f(x_1), \dots, f(x_n))^\top : f \in \mathcal{F}\}$. Suppose that $\boldsymbol{v}^f, \boldsymbol{v}^{f'} \in \mathcal{C}(\epsilon; \mathcal{F}_{|x_{1:n}}, \|\cdot\|_\infty)$ be such that $|\boldsymbol{v}^f_i - f(x_i)|, |\boldsymbol{v}^{f'}_i - f'(x_i)| \leqslant \epsilon$, for all i. Here $\mathcal{C}(\epsilon; \mathcal{F}_{|x_{1:n}}, \|\cdot\|_\infty)$ denotes the ϵ cover of $\mathcal{F}_{|x_{1:n}}$ w.r.t. the ℓ_∞-norm. Let $\boldsymbol{v} = \boldsymbol{v}^f - \boldsymbol{v}^{f'}$ Then

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (h(x_i) - v_i^2)^2 =& \frac{1}{n}\sum_{i=1}^n ((f(x_i) - f'(x_i))^2 - (v_i^f - v_i^{f'})^2)^2 \\
\leqslant& \frac{2}{n}\sum_{i=1}^n ((f(x_i) - f'(x_i))^2 + (v_i^f - v_i^{f'})^2) \times ((f(x_i) - f'(x_i)) - (v_i^f - v_i^{f'}))^2 \qquad (42)\\
\precsim& \epsilon^2. \qquad (43)
\end{aligned}$$

Here (42) follows from the fact that $(t^2 - r^2)^2 = (t+r)^2(t-r)^2 \leqslant 2(t^2 + r^2)(t-r)^2$, for any $t, r \in \mathbb{R}$. Hence, from the above calculations, $\mathcal{N}(\epsilon; \mathcal{H}_r, \mathbb{L}_2(\lambda_n)) \leqslant (\mathcal{N}(a_1\epsilon; \mathcal{F}, \mathbb{L}_\infty(\lambda_n)))^2$, for some absolute constant a_1.

$$\begin{aligned}
\operatorname{diam}^2(\mathcal{H}_r, \mathbb{L}_2(\lambda_n)) = \sup_{h, h' \in \mathcal{H}_r} \|h - h'\|^2_{\mathbb{L}_2(\lambda_n)} \leqslant& \sup_{h, h' \in \mathcal{H}_r} \frac{1}{n}\sum_{i=1}^n (h(x_i) - h'(x_i))^2 \\
\leqslant& 2 \sup_{h \in \mathcal{H}_r} \frac{1}{n}\sum_{i=1}^n h^2(x_i) \\
\leqslant& 2B \sup_{h \in \mathcal{H}_r} \frac{1}{n}\sum_{i=1}^n h(x_i) \\
\leqslant& 2Br.
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_r} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i) \precsim& \int_0^{\sqrt{2Br}} \sqrt{\frac{1}{n} \log \mathcal{N}(\epsilon; \mathcal{H}_r, \mathbb{L}_2(\lambda_n))} d\epsilon \\
\leqslant& \int_0^{\sqrt{2Br}} \sqrt{\frac{2 \operatorname{Pdim}(\mathcal{F})}{n} \log\left(\frac{a_2 n}{\epsilon}\right)} d\epsilon \\
\precsim& \sqrt{2Br}\sqrt{\frac{\operatorname{Pdim}(\mathcal{F}) \log n}{n}} + \int_0^{\sqrt{2Br}} \sqrt{\frac{\operatorname{Pdim}(\mathcal{F})}{n} \log(a_2/\epsilon)} d\epsilon \\
\precsim& \sqrt{\frac{r \log(1/r) \operatorname{Pdim}(\mathcal{F}) \log n}{n}}. \qquad (44)\\
\leqslant& \sqrt{\frac{(\operatorname{Pdim}(\mathcal{F}))^2 \log n}{n^2} + r \frac{\operatorname{Pdim}(\mathcal{F}) \log(n/e \operatorname{Pdim}(\mathcal{F})) \log n}{n}}. \qquad (45)
\end{aligned}$$

Here, (44) follows from Lemma 30. Here, (45) follows from Lemma 31 with $x = r$ and $y = \operatorname{Pdim}(\mathcal{F})/n$. ☐

Lemma 28. Let $\mathcal{H}_r = \{h : \theta \mapsto \int (f - f')^2 d\lambda_\theta : f, f' \in \mathcal{F} \text{ and } \hat{\pi}_m h \leqslant r\}$ with $\sup_{f \in \mathcal{F}} \|f\|_{\mathbb{L}_\infty(\lambda_n)} < \infty$. Then, we can find $r_0 > 0$, such that if $0 < r \leqslant r_0$ and $n \geqslant \operatorname{Pdim}(\mathcal{F})$,

$$\mathbb{E}_\sigma \sup_{h \in \mathcal{H}_r} \frac{1}{m} \sum_{i=1}^m \sigma_i h(\theta_i) \precsim \sqrt{\frac{(\operatorname{Pdim}(\mathcal{F}))^2 \log m}{m^2} + r \frac{\operatorname{Pdim}(\mathcal{F}) \log(m/e \operatorname{Pdim}(\mathcal{F})) \log m}{m}}.$$

Proof. Let $B = 4 \sup_{f \in \mathcal{F}} \|f\|_{\mathbb{L}_\infty([0,1]^d)}$. Let $\hat{f}, \hat{f}' \in \mathcal{C}(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbb{L}_\infty([0,1]^d)})$ be such that, $\|f - \hat{f}\|_{\mathbb{L}_\infty([0,1]^d)}, \|f' - \hat{f}'\|_{\mathbb{L}_\infty([0,1]^d)} \leqslant \epsilon$. Let $\hat{h}(\theta) = \int \left(\hat{f} - \hat{f}'\right)^2 d\lambda_\theta$. Then, for any θ,

$$\begin{aligned}
\frac{1}{m} \sum_{i=1}^m |h(\theta_i) - \hat{h}(\theta_i)|^2 \leqslant & \frac{1}{m} \sum_{i=1}^m \int \left| (f - f')^2 - \left(\hat{f} - \hat{f}'\right)^2 \right|^2 d\lambda_{\theta_i} \\
\leqslant & \frac{2}{m} \sum_{i=1}^m \int \left((f - f')^2 + \left(\hat{f} - \hat{f}'\right)^2 \right) \left(f - f' - \left(\hat{f} - \hat{f}'\right) \right)^2 d\lambda_{\theta_i} \\
\precsim & \epsilon^2
\end{aligned}$$

Hence, from the above calculations, $\mathcal{N}(\epsilon; \mathcal{H}_r, \mathbb{L}_2(\hat{\pi}_m)) \leqslant \left(\mathcal{N}\left(a_3 \epsilon; \mathcal{F}, \mathbb{L}_\infty(\hat{\pi}_m)\right)\right)^2$, for some absolute constant a_3. We also note that,

$$\begin{aligned}
\operatorname{diam}^2(\mathcal{H}_r, \mathbb{L}_2(\lambda_n)) = \sup_{h, h' \in \mathcal{H}_r} \|h - h'\|^2_{\mathbb{L}_2(\hat{\pi}_m)} \leqslant & \sup_{h, h' \in \mathcal{H}_r} \frac{1}{m} \sum_{i=1}^m (h(\theta_i) - h'(\theta_i))^2 \\
\leqslant & 2 \sup_{h \in \mathcal{H}_r} \frac{1}{m} \sum_{i=1}^m h^2(\theta_i) \\
\leqslant & 2B \sup_{h \in \mathcal{H}_r} \frac{1}{m} \sum_{i=1}^m h(\theta_i) \\
\leqslant & 2Br.
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_r} \frac{1}{m} \sum_{i=1}^m \sigma_i h(\theta_i) \precsim & \int_0^{\sqrt{2Br}} \sqrt{\frac{1}{n} \log \mathcal{N}(\epsilon; \mathcal{H}_r, \mathbb{L}_2(\hat{\pi}_m))} d\epsilon \\
\leqslant & \int_0^{\sqrt{2Br}} \sqrt{\frac{2 \operatorname{Pdim}(\mathcal{F})}{m} \log\left(\frac{a_4 m}{\epsilon}\right)} d\epsilon \\
\precsim & \sqrt{2Br} \sqrt{\frac{\operatorname{Pdim}(\mathcal{F}) \log m}{m}} + \int_0^{\sqrt{2Br}} \sqrt{\frac{\operatorname{Pdim}(\mathcal{F})}{m} \log(a_2/\epsilon)} d\epsilon \\
\precsim & \sqrt{\frac{r \log(1/r) \operatorname{Pdim}(\mathcal{F}) \log m}{m}}. \qquad (46) \\
\leqslant & \sqrt{\frac{(\operatorname{Pdim}(\mathcal{F}))^2 \log m}{m^2} + r \frac{\operatorname{Pdim}(\mathcal{F}) \log(m/e \operatorname{Pdim}(\mathcal{F})) \log m}{m}}. \qquad (47)
\end{aligned}$$

Here, (44) follows from Lemma 30. □

Lemma 29. Suppose that $v > 0$ and $\|f\|_{\mathbb{L}_\infty([0,1]^d)} \leqslant 2R$. Then with probability at least $1 - 2\exp\left(-\frac{c_3 m v}{\|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1}}\right)$,

$$\|f\|^2_{\mathbb{L}_2(\hat{\lambda}_m^p)} \leqslant \frac{3}{2} \|f\|^2_{\mathbb{L}_2(\lambda)} + v \quad \text{and} \quad \|f\|^2_{\mathbb{L}_2(\lambda)} \leqslant 2\|f\|^2_{\mathbb{L}_2(\hat{\lambda}_m^p)} + 2v.$$

Here c_3 is a constant that depends on R.

Proof. Suppose that $\lambda_\theta \ll \lambda$, almost surely under π. Then, by Radon-Nykodym theorem, the density of λ_θ w.r.t. λ exists and is denoted by $p_\theta = \frac{d\lambda_\theta}{d\lambda}$. We let,

$$Z_i = \int f^2(\boldsymbol{x})d\lambda_{\theta_i}(\boldsymbol{x}) - \int f^2(\boldsymbol{x})d\lambda(\boldsymbol{x}) = \int f^2(\boldsymbol{x})(p_{\theta_i}(\boldsymbol{x}) - 1)d\lambda$$

Suppose $h(\boldsymbol{x}) = f^2(\boldsymbol{x})$ and $u = \max\left\{v, \frac{1}{2}\|f\|^2_{\mathbb{L}_2(\lambda)}\right\}$. Then,

$$|Z_i|^2 \leqslant \int h^2(\boldsymbol{x})d\lambda \cdot \int (p_{\theta_i}(\boldsymbol{x}) - 1)^2 d\lambda = \|h\|^2_{\mathbb{L}_2(\lambda)} \cdot \chi^2(\lambda_{\theta_i}, \lambda)$$

$$\implies \|Z_i\|^2_{\psi_2} = \|Z_i^2\|_{\psi_1} \leqslant \|h\|^2_{\mathbb{L}_2(\lambda)} \cdot \|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1} \leqslant 8R^2 u \|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1} \tag{48}$$

In (48), the first equality follows from Vershynin (2018, Lemma 2.7.6). By Hoeffding's inequality,

$$\mathbb{P}\left(\left|\sum_{i=1}^m Z_i\right| > mu\right) \leqslant 2\exp\left(-\frac{cm^2u^2}{8R^2 mu \|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1}}\right) = 2\exp\left(-\frac{cmu}{8R^2\|\chi^2(\lambda_\theta, \lambda)\|_{\psi_1}}\right). \tag{49}$$

Here, $c_3 = \frac{c}{8R^2}$. Hence, with probability at least, $1 - 2\exp\left(-\frac{c_3 mv}{\|\chi^2(\lambda_\theta,\lambda)\|_{\psi_1}}\right)$,

$$\|f\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} \leqslant \|f\|^2_{\mathbb{L}_2(\lambda)} + u \leqslant \|f\|^2_{\mathbb{L}_2(\lambda)} + v + \frac{1}{2}\|f\|^2_{\mathbb{L}_2(\lambda)} = \frac{3}{2}\|f\|^2_{\mathbb{L}_2(\lambda)} + v.$$

Furthermore,

$$\|f\|^2_{\mathbb{L}_2(\lambda)} \leqslant \|f\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + u \leqslant \|f\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + v + \frac{1}{2}\|f\|^2_{\mathbb{L}_2(\lambda)} \implies \|f\|^2_{\mathbb{L}_2(\lambda)} \leqslant 2\|f\|^2_{\mathbb{L}_2(\hat{\lambda}^p_m)} + 2v.$$

□

Lemma 30. For any $\delta \leqslant 1/e$, $\int_0^\delta \sqrt{\log(1/\epsilon)}d\epsilon \leqslant 2\delta\sqrt{\log(1/\delta)}$.

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

$$\begin{aligned}\int_0^\delta \sqrt{\log(1/\epsilon)}d\epsilon = \int_{\log(1/\delta)}^\infty \sqrt{x}e^{-x}dx = & \int_{\log(1/\delta)}^\infty \sqrt{x}e^{-x/2}e^{-x/2}dx \\ \leqslant & \sqrt{\log(1/\delta)}e^{-\frac{1}{2}\log(1/\delta)} \int_{\log(1/\delta)}^\infty e^{-x/2}dx \qquad (50) \\ = & 2\delta\sqrt{\log(1/\delta)}.\end{aligned}$$

In the above calculations, (50) follows from the fact that the function $\sqrt{x}e^{-x/2}$ is decreasing when $x \geqslant 1$. □

Lemma 31. For any $x, y > 0$, $x\log x \leqslant y + x\log(1/ye)$.

Proof. Let $f(r) = r\log(1/r)$. Then, $f'(r) = -\log(re)$ and $f''(r) = -1/r$. Thus, $f(\cdot)$ is concave and thus, for any $x, y > 0$,

$$f(x) \leqslant f(y) + f'(y)(x - y) = -y\log y - (\log y + 1)(x - y) = y - x\log(ye) = y + x\log(1/ye).$$

□

Lemma 32. Suppose that $g(\cdot)$ be a non-negative real-valued function such that $B = \|g\|_\infty < \infty$. Let $Z_1, \dots, Z_n$ be independent random variables. Then, with probability at least $1 - e^{-nt}$,

$$\frac{1}{n}\sum_{i=1}^n g(Z_i) \leqslant \frac{2}{n}\sum_{i=1}^n \mathbb{E}g(Z_i) + \frac{7Bt}{3}.$$

Proof. Let $Y_i = g(Z_i) - \mathbb{E}g(Z_i)$. Also let $v > 0$ and $u = \max\left\{v, \frac{1}{n}\sum_{i=1}^n \mathbb{E}g(Z_i)\right\}$. Clearly,

$$\mathbb{E}Y_i^2 = \mathrm{Var}(Y_i) = \mathrm{Var}(g(Z_i)) \leqslant \mathbb{E}g^2(Z_i) \leqslant B\mathbb{E}g(Z_i). \tag{51}$$

Thus, $\sigma^2 = \sum_{i=1}^n \mathbb{E}Y_i^2 \leqslant B\sum_{i=1}^n \mathbb{E}g(Z_i) \leqslant nBu$. From Lemma 21, we observe that,

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^n g(Z_i) - \mathbb{E}g(Z_i)\right| > nu\right) &\leqslant \exp\left(-\frac{n^2u^2}{2\sigma^2 + nBu/3}\right)\\ &\leqslant \exp\left(-\frac{3nu}{7B}\right) \leqslant \exp\left(-\frac{3nv}{7B}\right)\end{aligned}$$

Thus, with probability at least $1 - \exp\left(-\frac{3nv}{7B}\right)$,

$$\frac{1}{n}\sum_{i=1}^n g(Z_i) \leqslant \frac{1}{n}\sum_{i=1}^n \mathbb{E}g(Z_i) + u \leqslant \frac{2}{n}\sum_{i=1}^n \mathbb{E}g(Z_i) + v.$$

Taking $v = 7Bt/3$, we get the desired result. □

Acknowledgments

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.

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. L. (2024). 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.

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

Chen, S., Zheng, Q., Long, Q., and Su, W. J. (2021). A theorem of the alternative for personalized federated learning. *arXiv preprint arXiv:2103.01901*.

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

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

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.

Falconer, K. (2004). *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons.

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

Hu, X., Li, S., and Liu, Y. (2023). Generalization bounds for federated learning: Fast rates, unparticipating clients and unbounded losses. In *International Conference on Learning Representations*.

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*.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR).*

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

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.

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.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Mishchenko, K., Malinovsky, G., Stich, S., and Richtárik, P. (2022). Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pages 15750–15769. PMLR.

Mitra, A., Jaafar, R., Pappas, G. J., and Hassani, H. (2021). Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619.

Mohri, M., Sivek, G., and Suresh, A. T. (2019). Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR.

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

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

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.

Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., and Lu, Z. (2022). Generalized federated learning via sharpness aware minimization. In *International Conference on Machine Learning*, pages 18250–18280. PMLR.

Reisizadeh, A., Farnia, F., Pedarsani, R., and Jadbabaie, A. (2020). Robust federated learning: The case of affine distribution shifts. *Advances in Neural Information Processing Systems*, 33:21554–21565.

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.

Suzuki, T. (2019). Adaptivity of deep reLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*.

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

Villani, C. (2021). *Topics in optimal transportation*, volume 58. American Mathematical Society.

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

Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al. (2021). A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*.

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.

Xu, J. and Wang, H. (2020). Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective. *IEEE Transactions on Wireless Communications*, 20(2):1188–1200.

Yang, H. H., Arafa, A., Quek, T. Q., and Poor, H. V. (2020). Age-based scheduling policy for federated learning in mobile edge networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8743–8747. IEEE.

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

Yousefi, N., Lei, Y., Kloft, M., Mollaghasemi, M., and Anagnostopoulos, G. C. (2018). Local rademacher complexity-based learning guarantees for multi-task learning. *The Journal of Machine Learning Research*, 19(1):1385–1431.

Yuan, H., Morningstar, W. R., Ning, L., and Singhal, K. (2021). What do we mean by generalization in federated learning? In *International Conference on Learning Representations*.

Yun, C., Rajput, S., and Sra, S. (2022). Minibatch vs local sgd with shuffling: Tight convergence bounds and beyond. In *International Conference on Learning Representations (ICLR)*. PMLR.

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., and Gao, Y. (2021). A survey on federated learning. *Knowledge-Based Systems*, 216:106775.