Journal of Machine Learning Research 25 (2024) 1-42 Submitted 11/22; Revised 10/23; Published 4/24

Learning with Norm Constrained, Over-parameterized, Two-layer Neural Networks

Fanghui Liu* FANGHUI.LIU@WARWICK.AC.UK
Department of Computer Science, University of Warwick, Coventry, UK
Leello Dadi LEELLO.DADI@EPFL.CH
Lab for Information and Inference Systems, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Volkan Cevher VOLKAN.CEVHER@EPFL.CH
Lab for Information and Inference Systems, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Editor: Mehryar Mohri

Abstract

Recent studies show that a reproducing kernel Hilbert space (RKHS) is not a suitable space to model functions by neural networks as the curse of dimensionality (CoD) cannot be evaded when trying to approximate even a single ReLU neuron (Bach, 2017; Celentano et al., 2021). In this paper, we study a suitable function space for over-parameterized two-layer neural networks with bounded norms (e.g., the path norm, the Barron norm) in the perspective of sample complexity and generalization properties. First, we show that the path norm (as well as the Barron norm) is able to obtain width-independence sample complexity bounds, which allows for uniform convergence guarantees. Based on this result, we derive the improved result of metric entropy for ϵ-covering up to $\mathcal{O}(\epsilon^{-\frac{2d}{d+2}})$ (d is the input dimension and the depending constant is at most polynomial order of d) via the convex hull technique, which demonstrates the separation with kernel methods with $\Omega(\epsilon^{-d})$ to learn the target function in a Barron space. Second, this metric entropy result allows for building a sharper generalization bound under a general moment hypothesis setting, achieving the rate at $\mathcal{O}(n^{-\frac{d+2}{2d+2}})$. Our analysis is novel in that it offers a sharper and refined estimation for metric entropy (with a clear dependence relationship on the dimension d) and unbounded sampling in the estimation of the sample error and the output error.

Keywords: learning theory, Barron space, path-norm, sample complexity, generalization guarantees

1. Introduction

While the number of neurons provide a natural complexity or *capacity* measure for neural networks, recent theoretical approaches focus more on the magnitude of the weights via norm-based constraints (Neyshabur et al., 2015; Savarese et al., 2019; Ongie et al., 2020; Domingo-Enrich and Mroueh, 2022). The norm-based capacity analyses (e.g., even in infinite-width) represent essentially potential functions with small cost. Generally they revolve around minimum-norm *over-parameterized* models and seek to identify correlations between the norm-based capacity and the generalization of neural network models, e.g., min-ℓ_2-norm (Hastie et al., 2022; Mei and Montanari, 2022; Liang et al., 2020), and min-ℓ_1-norm (Liang and Sur, 2022; Chatterji and Long, 2022; Wang et al., 2021).

∗. Part of this work was done when Fanghui was at LIONS, EPFL. Corresponding author: Fanghui Liu.

From a functional perspective, one key issue corresponds to what norm can be defined and controlled on the functions defined by neural networks, and what suitable function space is for learning via norm capacity based neural networks. Indeed, the prototypical two-layer neural networks with the rectified linear unit (ReLU) activation cannot be sufficiently captured by the reproducing kernel Hilbert spaces (RKHSs) of kernel methods such as the random features (RF) model (Rahimi and Recht, 2007) and the neural tangent kernel (NTK) (Jacot et al., 2018). Specifically, to approximate a single ReLU neuron with an ε-approximation error, kernel methods require a number of samples $\Omega(\varepsilon^{-d})$, exponential in feature dimensionality d (Bach, 2017; Yehudai and Shamir, 2019; Ghorbani et al., 2021; Wu and Long, 2022), *a.k.a.*, curse of dimensionality (CoD).

It is possible to characterize over-parameterized two-layer neural networks by a compactly supported Radon probability measure with spectral Barron norm (Barron, 1993), path norm (Neyshabur et al., 2015), bounded total variation norm (Bach, 2017) or other variants (Savarese et al., 2019; Ongie et al., 2020; Parhi and Nowak, 2021). In this context, any function that is well-approximated by a norm-bounded two-layer neural networks lives in a revised Barron space (E et al., 2021), which is the *largest*[1] function space for two-layer neural networks (E and Wojtowytsch, 2020) beyond RKHS. In this case, given n iid training data, the minimax lower bound in excess risk for learning in this non-Hilbertian space by any kernel methods estimators is $\Omega(n^{-1/d})$ (E and Wojtowytsch, 2021; Parhi and Nowak, 2022), which suffers from CoD and unsatisfactory sample complexity bounds.

Interestingly, the Barron space can be closely connected to other typical spaces with various norms. For example, when using the ReLU activation function, the Barron norm is exactly the same as the total variation norm (Bach, 2017), and other total variation norm based versions (Siegel and Xu, 2021; Parhi and Nowak, 2021). Furthermore, the discrete version of the Barron norm is the ℓ_1-path norm (Neyshabur et al., 2015) in the parameter space, and thus is widely used in practice. Based on this, understanding on the Barron space, especially based on the path-norm studied in this paper, for learning such two-layer neural networks is general and required.

1.1 Problem setting

Mathematically, let $X \subseteq \mathbb{R}^d$ be a compact space and the label space $Y \subseteq \mathbb{R}$, we assume that a sample set $\boldsymbol{z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in Z^n$ is independently drawn from a non-degenerate Borel probability measure ρ on $X \times Y$. Our target is to find a hypothesis f over the sample set such that f is a good approximation of the *target function*

$$f_\rho(\boldsymbol{x}) = \int_Y y \mathrm{d}\rho(y|\boldsymbol{x}), \boldsymbol{x} \in X\,,$$

where $\rho(\cdot|\boldsymbol{x})$ is the conditional distribution of ρ at $\boldsymbol{x} \in X$. Here the hypothesis is considered as the class of two-layer neural networks parameterized by $\boldsymbol{\theta} := \{(a_i, \boldsymbol{w}_i)\}_{i=1}^m$ with $\boldsymbol{w}_i \in \mathbb{R}^d$, $a_i \in \mathbb{R}$

$$\mathcal{P}_m = \left\{ f_{\boldsymbol{\theta}}(\cdot) := \frac{1}{m} \sum_{k=1}^m a_k \sigma\big(\langle \boldsymbol{w}_k, \cdot \rangle\big) \right\}, \tag{1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function. It allows for the bias term by adding one extra term but we omit it here for simplicity. The hypothesis is often learned from the sample set via the empirical

1. The terminology "largest" means functions in the Barron space and two-layer neural networks with bounded norm can be efficiently represented without the curse of dimensionality, refer to Section 2 for details.

risk minimization (ERM) with a proper regularization term

$$f_{\boldsymbol{\theta},\boldsymbol{z},\lambda} = \operatorname*{argmin}_{f\in\mathcal{P}_m} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(\boldsymbol{x}_i,\boldsymbol{\theta})) + \lambda\|f\|\,, \tag{2}$$

where $\ell : \mathbb{R}\times\mathbb{R}\to\mathbb{R}^+$ is a *surrogate* loss function, and $\lambda \equiv \lambda(n) > 0$ is a regularization parameter of a certain norm-based regularizer $\|f\|$. In this paper, we employ a certain ℓ_1-path norm regularization, that will be discussed later in Section 3. To facilitate a general analysis on learning with noisy data in practice, we consider the following moment hypothesis concerning unbounded outputs.

Moment hypothesis: There exist constants $M \geqslant 1$, $C > 0$ such that[2]

$$\int_Y |y|^p d\rho(y|\boldsymbol{x}) \leqslant Cp!M^p, \forall p\in\mathbb{N}, \quad \boldsymbol{x}\in X\,. \tag{3}$$

Compared to the standard uniform boundedness assumption with $|y| \leq M$ almost surely, this assumption is general since it covers Gaussian noise, sub-Gaussian noise, sub-exponential noise, etc.

Based on the above problem setting, we are interested in sample complexity for uniform estimates and generalization guarantees (these two questions are connected to each other). The sample complexity issue stems from the following question: A width-independent sample complexity bound can be obtained by Frobenius norm control (Neyshabur et al., 2015). However spectral norm control (Bartlett et al., 2017) is generally insufficient for two-layer neural networks as demonstrated by Vardi et al. (2022). This conclusion is *questionable* when the Barron norm (as well as the path norm) is employed since the Barron space is larger than that of Frobenius norm-constrained two-layer neural networks. In this work, we derive the width-independent sample complexity for uniform estimate in the Barron space, and then study the generalization properties under a general setting. This contributes to provide an in-depth theoretical understanding and refined analysis of learning with over-parameterized two-layer neural networks under norm-based capacity under this general setting.

1.2 Contributions and technical challenges

Our analysis starts with the width-independence sample complexity, study the metric entropy, and provide refined results on generalization bounds under weaker conditions. We make the following contributions:

- based on the Gaussian complexity metric, we prove that the path norm is able to obtain width-independence sample complexity bounds. This result motivates us to derive that the metric entropy for ϵ-covering up to $\mathcal{O}(\epsilon^{-\frac{2d}{d+2}})$ via the convex hull technique using the smoothness structure of $\mathcal{P}_m$. Note that the dependence on the input dimension d (in a polynomial order) can be clearly derived from this technique, while this is unclear (would be exponential order of d) in (Siegel and Xu, 2021) based on the orthogonal function argument.

- when applying our metric entropy result for generalization guarantees, we are faced with how to tackle the output error in Eq. (3). Tackling such unbounded sampling requires the number

2. This assumption is essentially equivalent to Bernstein's condition (Steinwart and Andreas, 2008): $\mathbb{E}\left[|y|^b \mid \boldsymbol{x}\right] \leqslant \frac{1}{2}b!\varsigma^2 B^{b-2}$ when $b \geqslant 2$. For description simplicity, we assume y is of zero mean.

of training data n to be in an exponential order in previous work (Wang and Zhou, 2011; Guo and Zhou, 2013; Liu et al., 2021), i.e., $n \geqslant k^k$ for a sufficiently large k, which implies that a sample complexity suffers from CoD due to $k \geqslant d$. In this work, we introduce new concentration inequalities via sub-Weibull random variables, remove the dependence on the exponential order to bound the output error, and provide non-asymptotic error bounds on finite n. This might be of interest in its own right in the learning theory community.

Combining the improved estimation of metric entropy and the output error, we are ready to present generalization bounds of learning ℓ_1-path norm based, over-parameterized, two-layer ReLU neural networks. To be specific, a sharper rate $\mathcal{O}(n^{-\frac{d+2}{2d+2}})$ of generalization bounds is derived under a more general setting, which is better than previous work (Bach, 2017; E et al., 2019; Wang and Lin, 2021) with $\mathcal{O}(\sqrt{\log n/n})$.[3] This requires a refined analysis in our proof for the sample error and the output error. Besides, optimization over the Barron space is normally NP-hard. We attempt to develop a computational (but not efficient except for the low-rank data) algorithm based on the measure representation and convex duality (Chizat, 2021; Pilanci and Ergen, 2020). The basic over-parameterization condition ($m \geqslant n + 1$) is sufficient.

Our results are immune to CoD even though $d \to \infty$; while kernel methods are not. This demonstrates the separation between over-parameterized neural networks and kernel methods when learning in the Barron space. We hope that our analysis has a better understanding on the improved analysis of neural networks in a norm-based capacity view and function spaces in the machine learning community.

1.3 Organization and notations

The rest of the paper is organized as follows. In Section 2, we give an overview of related work and preliminaries on two-layer neural networks in a norm-based capacity view. In Section 3, we consider a general random design regression problem for two-layer neural networks with bounded norm in the context of statistical learning theory. The assumptions and our main results are stated in Section 4. We outline the proof framework in Section 5. We validate our findings with numerical simulations in Section 6. The conclusion is drawn in Section 7. Some technical lemmas and proofs are deferred to the appendix.

For notations, we use the shorthand $[n] := \{1, 2, \ldots, n\}$ for some positive n. The unit sphere of $\mathbb{R}^d$ is defined as $\mathbb{S}_p^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d \big| \|\boldsymbol{x}\|_p = 1\}$. Let $\delta_{\boldsymbol{w}}(\cdot)$ be the Dirac measure on $\boldsymbol{w} \in \mathbb{R}^d$, i.e., $\int f(\boldsymbol{x})\delta_{\boldsymbol{w}}(\mathrm{d}\boldsymbol{x}) = f(\boldsymbol{w})$. The notation $a(n) \lesssim b(n)$ signifies that there exists a positive constant c independent of n such that $a(n) \leqslant cb(n)$. We use the standard big-O notation with $\mathcal{O}(\cdot)$, $\Omega(\cdot)$ hiding constants and $\widetilde{\mathcal{O}}(\cdot)$ hiding constants and factors polylogarithmic in the problem parameters. Besides, as noted in (E and Wojtowytsch, 2020), if m tends to infinity, $\mathcal{P}_m$ is a closed subspace of a Barron space, and thus is still a Banach space. For any compact $X \subset \mathbb{R}^d$, the $\mathsf{C}(X)$-closure of $\mathcal{F}_\infty := \cup_{m\in\mathbb{N}}\mathcal{F}_m$ is the entire space of $\mathsf{C}(X)$, where $\mathsf{C}(X)$ is the function space of all continuous function on X with $\|\cdot\|_\infty$.

3. We remark that the optimal approximation rate of the Barron space is given by (Siegel and Xu, 2021) and (Wu and Long, 2022) based on different techniques, and the minimax rate at $\mathcal{O}(n^{-\frac{d+3}{2d+3}})$ can be derived under the variation norm based space, e.g., (Parhi and Nowak, 2022) with the skip connection. However, the dependence on d is still unclear. We will detail this discussion in Section 4.3.

2. Related Works and Preliminaries

We give an overview of two-layer neural networks via integral representation in the Barron space (E et al., 2021) and other function spaces with various norms.

Two-layer neural networks and integral representation: We consider a two-layer neural network with m neurons represented as $f(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^{m} a_k \sigma(\boldsymbol{w}_k^\top \boldsymbol{x})$, where $\sigma(\cdot)$ is the activation function and the parameters (weights) of the network are $\{(a_k, \boldsymbol{w}_k)\}_{k=1}^m \subset \mathbb{R}\times\mathbb{R}^d$. This setting allows for neural networks with bias, by taking $(\boldsymbol{x}, 1) \in \mathbb{R}^{d+1}$ as $\boldsymbol{x}$ and $(\boldsymbol{w}, b) \in \mathbb{R}^{d+1}$ as $\boldsymbol{w}$. For notational simplicity, we still use $\boldsymbol{x} \in \mathbb{R}^d$ in this paper. We consider the above two-layer neural network in a general integral representation

$$f(\boldsymbol{x}) = \int_\Omega a\sigma(\boldsymbol{w}^\top \boldsymbol{x})\tilde{\mu}(\mathrm{d}a, \mathrm{d}\boldsymbol{w}), \quad \boldsymbol{x} \in X\,,$$

where $\Omega = \mathbb{R}\times\mathbb{R}^d$ and $\tilde{\mu}$ is a probability measure over $(\Omega, \mathcal{T}(\Omega))$ with $\mathcal{T}(\Omega)$ being a Borel σ-algebra on Ω. The used activation function in this work is ReLU $\sigma(x) = \max(0, x)$.

Due to the scaling invariance of ReLU, we can assume $\boldsymbol{w} \in \mathbb{S}^{d-1}$ such that

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^d} a\sigma\left(\boldsymbol{w}^\top \boldsymbol{x}\right)\tilde{\mu}(\mathrm{d}a, \mathrm{d}\boldsymbol{w}) = \int_{\mathbb{S}^d} a(\boldsymbol{w})\sigma\left(\boldsymbol{w}^\top \boldsymbol{x}\right)\mu(\mathrm{d}\boldsymbol{w})\,, \tag{4}$$

where $a(\boldsymbol{w}) = \frac{\int_{\mathbb{R}} a\tilde{\mu}(a,\boldsymbol{w})\mathrm{d}a}{\pi(\boldsymbol{w})}$ and $\mu(\boldsymbol{w}) = \int_{\mathbb{R}} \tilde{\mu}(a, \boldsymbol{w})\mathrm{d}a$. Accordingly, the representation of f in Eq. (4) can be considered into the associated function space with various functional norms

$$\mathcal{F}_p(R) := \left\{ f(\cdot) = \int_{\mathcal{V}} a(\boldsymbol{w})\sigma(\langle \boldsymbol{w}, \cdot\rangle)\mu(\mathrm{d}\boldsymbol{w}) \Big| \|f\|_{\mathcal{F}_p} := \left(\int_{\mathcal{V}} |a(\boldsymbol{w})|^p \mu(\mathrm{d}\boldsymbol{w})\right)^{1/p} \leqslant R \right\}. \tag{5}$$

If we take $p = 2$, the weights $\boldsymbol{w}$ are defined on a probability space $(\mathcal{V}, \mu)$ with $a \in L^2(\mathcal{V}, \mu)$ and $\mathcal{V} = \mathbb{R}^d$.[4] In this case, $\mathcal{F}_2$ is a RKHS with the associated reproducing kernel $k(\boldsymbol{x}, \boldsymbol{x}') := \int_{\mathcal{V}} \sigma(\boldsymbol{w}^\top \boldsymbol{x})\sigma(\boldsymbol{w}^\top \boldsymbol{x}')\mu(\mathrm{d}\boldsymbol{w})$ (Bach, 2017). Although $\mathcal{F}_2(R)$ is infinite-dimensional, the RKHS-norm regularized empirical risk minimization (ERM) can be solved by a finite dimensional optimization problem by the representer theorem (Schölkopf and Smola, 2018). Note that, for $p \in [1, 2)$ this comprises a rich function class than the original RKHS $\mathcal{F}_2(R)$ due to $\mathcal{F}_2(R) \subset \mathcal{F}_p(R) \subset \mathcal{F}_1(R)$ (Celentano et al., 2021). The special case is $p = 1$ in which μ is a signed Radon measure on $\mathcal{V}$ with finite total variation $|\mu|(\mathcal{V})$ (Bach, 2017), i.e., "convex neural network" (Bengio et al., 2005). This corresponds to a variation norm rather than a RKHS norm, which actually enlarges the space by adding non-smooth functions.

Barron spaces under the ReLU activation function: Apart from the typical $\mathcal{F}_p$ function space, the space of two-layer ReLU neural networks can be built on Fourier transform. By taking $\hat{f}$ as the Fourier transform of an extension of f to $\mathbb{R}^d$, the classical *spectral Barron norm* (Barron, 1993) is given by $\|f\| := \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\||\hat{f}(\boldsymbol{\omega})|\mathrm{d}\boldsymbol{\omega}$. Its variants include $\|f\| := \inf_{\hat{f}} \int_{\mathbb{R}^d} |\hat{f}(\boldsymbol{\omega})|\|\boldsymbol{\omega}\|_1^s \mathrm{d}\boldsymbol{\omega}$ (Klusowski and Barron, 2016) and $\|f\| := \int_{\mathbb{R}^d} |\hat{f}(\boldsymbol{\omega})|[1 + \|\boldsymbol{\omega}\|_1^2]\mathrm{d}\boldsymbol{\omega}$ (Klusowski and Barron, 2018). If we replace $|\hat{f}(\boldsymbol{\omega})|$ with $|\hat{f}(\boldsymbol{\omega})|^2$, we would obtain Sobolev semi-norm.

E et al. (2021) provides a probabilistic interpretation of Barron space as an infinite union of a family of RKHS. The related Barron norm is defined by

$$\|f\|_{\mathcal{B}_p} = \inf_{\tilde{\mu}} \left(\mathbb{E}_{\tilde{\mu}}|a|^p \|\boldsymbol{w}\|_1^p\right)^{1/p}\,,$$

4. This is equivalent to $\mathcal{V} = \mathbb{S}_p^{d-1}$ due to scale invariance of ReLU.

where the infimum is taken over all possible $\tilde{\mu}$. Particularly, for the ReLU, we have $\mathcal{B}_p = \mathcal{B}_\infty$ and $\|f\|_{\mathcal{B}_p} = \|f\|_{\mathcal{B}_\infty}$ for any $1 \leqslant p \leqslant \infty$, and thus we can directly use $\mathcal{B}$ and $\|f\|_{\mathcal{B}}$ to denote the Barron space and the Barron norm. It closely relates to the ℓ_1-path norm $\|\cdot\|_{\mathcal{P}}$

$$\|f_{\boldsymbol{\theta}}\|_{\mathcal{B}} \leqslant \|\boldsymbol{\theta}\|_{\mathcal{P}} := \frac{1}{m}\sum_{k=1}^{m} |a_k| \|\boldsymbol{w}_k\|_1 \leqslant 2\|f\|_{\mathcal{B}}\,, \tag{6}$$

where $\boldsymbol{\theta}$ is sampled from the distribution related to the Barron norm. Note that the path norm is not invariant to the parameterization. Even for the same neural network output, the respective path norm depends on the certain parameter implementation style. Different implementation styles lead to different representation ability, and thus the path norm is a suitable metric to describe the model capacity. It is natural to study two-layer neural networks with bounded ℓ_1-path norm as the discrete version of Barron norm. As suggested by E and Wojtowytsch (2020), Barron space is the largest function space which is well/efficiently approximated by two-layer neural networks with appropriately controlled parameters via the *direct* and *inverse* approximation theorems (E et al., 2021).

- *direct approximation*: For any $f \in \mathcal{B}$, two-layer neural networks with controlled parameters are able to approximate it at a certain convergence rate without the curse of dimensionality.

- *inverse approximation*: Any continuous function that can be efficiently approximated by two-layer neural networks with bounded ℓ_1-path norm belongs to the Barron space.

Hence, it is natural to study two-layer ReLU neural networks with bounded ℓ_1-path norm as the discrete version of Barron norm. In this paper, we mainly focus on the ℓ_1-path norm.

Connection to variation spaces and smoothness: To control the Barron norm of two-layer ReLU networks, it is equivalent to control the total variation of the derivative for univariate functions (Savarese et al., 2019) and the total variation norm in the Radon domain for multivariate cases (Ongie et al., 2020; Domingo-Enrich and Mroueh, 2022). In fact, controlling the function derivative (e.g., total variation) is common in function space theory. For example, Unser (2019) studies the class of (univariate) functions with bounded second total variation norm; Parhi and Nowak (2021) focus on the total variation norm $\|f\|_{(s)} := c_d \left\|\partial_t^s \Lambda^{d-1} \mathcal{R} f\right\|_{\mathcal{M}\left(\mathbb{S}^{d-1}\times\mathbb{R}\right)}$ with $s \geqslant 2$ on on the Randon measure domain $\mathcal{M}$, where $\mathcal{R}$ is the Radon transform, Λ^{d-1} is a ramp filter, and ∂_t^s is the s-th partial derivative with respect to t, the offset variable in the Radon domain, and c_d is a dimension dependent constant. For two-layer ReLU neural networks, this norm is equivalent to $\|f\|_{(s)} = \sum_{k=1}^{m} |a_k| \|\boldsymbol{w}_k\|_2^{s-1}$ (Parhi and Nowak, 2022), which is the ℓ_2-path norm by taking $s = 2$. Note that the function space in a second-order bounded variation sense is a reproducing kernel Banach space (Bartolucci et al., 2023) and the representer theorem for data-fitting variational problem also exists (Parhi and Nowak, 2021).

Apart from the above function spaces in a norm-based capacity view, study on the *smoothness* of function classes also exists in classical approximation theory on various spaces, e.g., Hölder, Sobolev, Besov spaces (Schmidt-Hieber, 2020; Suzuki, 2019). Previous results have shown that the curse of dimensionality in sample complexity can be avoided for (deep) ReLU neural networks if the intrinsic dimensionality of data is small (Chen et al., 2019; Nakada and Imaizumi, 2020; Ghorbani et al., 2020) or the target function is (almost) smooth (Suzuki and Nitanda, 2021).

3. Empirical risk minimization under the ℓ_1 path norm regularization

In this paper, we consider a general random design regression problem for two-layer neural networks with the ℓ_1-path norm in the context of statistical learning theory. Besides, we also introduce some notations needed for our analysis on metric entropy.

For regression, we employ the commonly used squared loss in this paper. The expected risk of such hypothesis $f_{\boldsymbol{\theta}}$ is defined by the mean square error (MSE), i.e., $\mathcal{E}(f) = \int_Z (f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y)^2 \mathrm{d}\rho$, as defined in the introduction. The empirical risk is accordingly defined on the sample $\boldsymbol{z}$, i.e., $\mathcal{E}_{\boldsymbol{z}}(f) = \frac{1}{n}\sum_{i=1}^n \big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\big)^2$, and the excess risk is exactly the distance in $L^2_{\rho_X}$ under the squared loss, i.e., $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}}$ for any $f \in L^2_{\rho_X}$, where $\|f\|_{L^2_{\rho_X}} = \big(\int_X |f(\boldsymbol{x})|^2 \mathrm{d}\rho_X(\boldsymbol{x})\big)^{1/2}$, where ρ_X is the marginal distribution of ρ on X, refer to Cucker and Zhou (2007) for details.

The empirical risk minimization regression problem over two-layer neural networks under the ℓ_1-path norm regularization setting is given by

$$\boldsymbol{\theta}^\star = \underset{f_{\boldsymbol{\theta}} \in \mathcal{P}_m}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n (y_i - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))^2 + \lambda \|\boldsymbol{\theta}\|_{\mathcal{P}}\,, \tag{7}$$

where the regularization parameter $\lambda \equiv \lambda(n) > 0$ is typically assumed to satisfy $\lim_{n\to\infty} \lambda(n) = 0$. It is clear that the solution to Eq. (7) is not unique. For example, under the ReLU setting, if $\{(a_k^*, \boldsymbol{w}_k^*)\}_{k=1}^m$ corresponds to an optimal neural network, any re-scaling scheme $\{(c_k a_k^*, \boldsymbol{w}_k^*/c_k)\}_{k=1}^m$ (with $c_k > 0$) is also optimal, as it does not change the neural network output and the ℓ_1-path norm due to the positive 1-homogeneity of ReLU.

To obtain a tighter bound, we need the following *projection operator* in our analysis.

Definition 1 *(Steinwart and Andreas, 2008, Projection operator) For $B \geqslant 1$, the projection operator $\pi := \pi_B$ on $\mathbb{R}$ is defined as*

$$\pi_B(t) = \begin{cases} B, & \text{if } t > B; \\ t, & \text{if } -B \leqslant t \leqslant B \\ -B, & \text{if } t < -B\,. \end{cases}$$

The projection of a function $f : X \to \mathbb{R}$ is given by $\pi_B(f)(\boldsymbol{x}) = \pi_B(f(\boldsymbol{x}))$, $\forall \boldsymbol{x} \in X$.

The projection operator is commonly used in learning theory, e.g., (Steinwart and Andreas, 2008; Shi et al., 2019; Liu et al., 2021), beneficial to the $\|\cdot\|_\infty$-bounds in the convergence analysis for sharp estimation, i.e., $\|\pi_B(f_{\theta^\star,\boldsymbol{z},\lambda}) - f_\rho\|^2_{L^2_{\rho_X}}$ in this work. This is because, Eq. (3) implies $|f_\rho(\boldsymbol{x}) = \int_Y y \mathrm{d}\rho(y|\boldsymbol{x})| \leqslant CM := M^*$ for any $\boldsymbol{x} \in X$. We assume $M^* \geqslant 1$ for simplicity. It is natural to project $f_{\boldsymbol{\theta}^\star,\boldsymbol{z},\lambda}$ onto the same interval.

Furthermore, to quantitatively understand how the complexity of $\mathcal{P}_m$ affects the learning ability of Eq. (7), we need the capacity (roughly speaking the "size") of $\mathcal{P}_m$ as measured by the ℓ_2-empirical covering number (Koltchinskii and Panchenko, 2005). Formally, define the normalized ℓ_2-metric $\mathscr{D}_2$ on the Euclidean space $\mathbb{R}^l$ as

$$\mathscr{D}_2(\mathbf{a}, \mathbf{b}) = \left(\frac{1}{l}\sum_{i=1}^l |a_i - b_i|^2\right)^{1/2}, \quad \mathbf{a} = (a_i)_{i=1}^l, \mathbf{b} = (b_i)_{i=1}^l \in \mathbb{R}^l\,.$$

Definition 2 *(Koltchinskii and Panchenko, 2005, ℓ_2-empirical covering number) For a subset $\mathcal{S}$ of a pseudo-metric space $(\mathcal{M}, d)$ and $\eta > 0$, the covering number $\mathscr{N}(\mathcal{S}, \eta, d)$ is defined to be the minimal number of balls of radius η whose union covers. For a set $\mathcal{F}$ of functions on X and $\eta > 0$, the ℓ_2-empirical covering number of $\mathcal{F}$ is given by*

$$\mathscr{N}_2(\mathcal{F}, \eta) = \sup_{l \in \mathbb{N}} \sup_{\boldsymbol{u} \in X^l} \mathscr{N}\left(\mathcal{F}|_{\boldsymbol{u}}, \eta, \mathscr{D}_2\right)$$

where for $l \in \mathbb{N}$ and $\boldsymbol{u} = (u_i)_{i=1}^l \subset X^l$, we denote the covering number of the subset $\mathcal{F}|_{\boldsymbol{u}}$ of the metric space $(\mathbb{R}^l, \mathscr{D}_2)$ as $\mathscr{N}\left(\mathcal{F}|_{\boldsymbol{u}}, \eta, \mathscr{D}_2\right)$.

In our analysis, we also need the definition of Gaussian complexity (Bartlett and Mendelson, 2002), that relates better with the metric entropy $\log \mathscr{N}_2(\mathcal{G}_R, \epsilon)$ for ϵ-covering of the index space when compared to the standard Rademacher complexity, where $\mathcal{G}_R$ is defined as $\mathcal{G}_R := \{f_{\boldsymbol{\theta}} \in \mathcal{P}_m : \|\boldsymbol{\theta}\|_{\mathcal{P}} \leqslant R\}$.

Definition 3 *(Bartlett and Mendelson, 2002, Gaussian complexity) The empirical Gaussian complexity of a function class $\mathcal{F}$ over data points $\{\boldsymbol{x}_i\}_{i=1}^n$ is defined as*

$$\mathcal{C}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \xi_i f(\boldsymbol{x}_i)\right],$$

where $\boldsymbol{\xi} = [\xi_i, \xi_2, \cdots, \xi_n]^\top$ is a standard normal random vector.

4. Main Results

We present our main results on sample complexity, metric entropy, and generalization bounds for learning with over-parameterized two-layer neural networks. Before presenting these results, we also need the following assumptions.

4.1 Assumptions

Apart from the moment hypothesis on the label noise in Eq. (3), we consider the following two assumptions that are standard and general.

Assumption 1 *(Bounded data) It holds that* $\operatorname{supp}(\rho_X) \subset \{\boldsymbol{x} \in \mathbb{R}^d | \|\boldsymbol{x}\|_\infty \leqslant S\}$.

For ease of convenience, we take $S = 1$ throughout this paper.

Assumption 2 *(Existence of f_ρ) We assume that the* target function $f_\rho \in \mathcal{B}(R) := \{f \in \mathcal{B} | \|f\|_{\mathcal{B}} \leq R\}$ *exists.*

The target function class is the closure (in $L^2_{\rho_X}$) of any two-layer networks with a finite ℓ_1-path norm, which includes the standard teacher-student model, e.g., (Tian, 2017; Loureiro et al., 2021; Akiyama and Suzuki, 2021).

4.2 Sample complexity and metric entropy

Based on the definition of Gaussian complexity, we are able to estimate the sample complexity of the path-norm based spaces (with proof deferred to Appendix A.1).

Lemma 4 *(sample complexity) Given the data $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ satisfying Assumption 1, denote $\mathcal{G}_R = \{f_{\boldsymbol{\theta}} \in \mathcal{P}_m : \|\boldsymbol{\theta}\|_{\mathcal{P}} \leqslant R\}$, then the empirical Gaussian complexity of $\mathcal{G}_R$ on $\{\boldsymbol{x}_i\}_{i=1}^n$ is at most ϵ, if the number of training data satisfies $n \geqslant \frac{8R^2 \log d}{\epsilon^2}$.*

Remark: The similar result via Rademacher complexity can be given by (E et al., 2021) but we still put it here as a good starting point: a width-free sample complexity bound under the ℓ_1-path norm can be achieved. This result enlarges the application scope of (Vardi et al., 2022) on the bounded Frobenius norm.

Similarly, our results can be directly extended to neural networks with the general ℓ_p-path norm for $p \geq 1$, e.g., the ℓ_2-path norm $\sum_{k=1}^m |a_k| \|\boldsymbol{w}_k\|_2$ used in (Wang and Lin, 2021; Parhi and Nowak, 2021). We give an example of the commonly used ℓ_2-path norm, demonstrating the sample complexity is $n \geq \frac{4R^2}{\epsilon^2}$ independent of d. See Appendix A.1 for details.

Based on Lemma 4, the Gaussian complexity leads to a basic estimation of the metric entropy $\log \mathscr{N}_2(\mathcal{G}_1, \epsilon) \lesssim \log d \left(\frac{1}{\epsilon}\right)^2$, see the proof in Appendix A.2. Note that this metric entropy based on the ℓ_2-empirical covering number can be improved in the following proposition by the convex hull technique (Van Der Vaart et al., 1996) (with proof deferred to Appendix A.3).

Proposition 5 *(metric entropy) Under Assumption 1, denote $\mathcal{G}_R = \{f_{\boldsymbol{\theta}} \in \mathcal{P}_m : \|\boldsymbol{\theta}\|_{\mathcal{P}} \leqslant R\}$, the metric entropy of $\mathcal{G}_1$ can be bounded by*

$$\log \mathscr{N}_2(\mathcal{G}_1, \epsilon) \leqslant 6144 d^5 \epsilon^{-\frac{2d}{d+2}}, \quad \forall \epsilon > 0 \quad \textit{and} \quad d > 5\,. \tag{8}$$

Remark: We make the following remarks.
i) The exact constant depends on the input dimension d *polynomially*, at most d^5. In fact, if d is large, e.g., $d > 100$, this polynomial order can be further improved to d^4. Nevertheless, the derived sample complexity is still better than Lemma 4 as well as (Vardi et al., 2022).
ii) If we directly apply previous results on (deep) ReLU neural networks, e.g., (Schmidt et al., 2011; Suzuki, 2019; Bartlett et al., 2019), we have $\log \mathscr{N}_2(\mathcal{G}_1, \epsilon) \lesssim s \log\left(\frac{md}{\epsilon}\right)$, where s is the number of *free* parameters in our two-layer neural network or $\|\boldsymbol{a}\|_0 + \|\boldsymbol{W}\|_0 \leqslant s$. The sparsity $s = o(m)$ holds true for deep ReLU neural networks as they have few activations (Hanin and Rolnick, 2019) but normally this is not valid for our over-parameterized two-layer setting. In this case, we only have $s = \Omega(m)$, leading to a vacuous convergence rate $\mathcal{O}(m/n)$.

4.3 Convergence rates of the excess risk

Based on the above results, now we can state our main results on over-parameterized two-layer neural networks in the Barron space. We follow Schmidt-Hieber (2020); Chen et al. (2019); Suzuki and Nitanda (2021) on generalization guarantees for neural networks that directly assume the attainable property of the global minimum. In the next subsection, we will develop a computational algorithm to obtain a global minimum of the original non-convex optimization problem.

Our result on the generalization guarantees under the ReLU activation function is given as below (with proofs deferred to Appendix C).

Theorem 6 *Considering problem (7) in $\mathcal{G}_R = \{f_{\boldsymbol{\theta}} \in \mathcal{P}_m : \|\boldsymbol{\theta}\|_{\mathcal{P}} \leqslant R\}$ for over-parameterized two-layer ReLU neural networks with $d \geq 5$, under Assumptions 1, 2, and moment hypothesis in Eq.* (3)*, taking $R \geqslant CM \geqslant 1$, a large B and any $0 < \delta < 1$, with probability $1 - \delta$, there exists one global minimum $\boldsymbol{\theta}^\star$ of problem (7) such that*

$$\left\|\pi_B(f_{\boldsymbol{\theta}^\star}) - f_\rho\right\|^2_{L^2_{\rho_X}} \lesssim \lambda + \frac{R^2}{m} + R^2 d^2 n^{-\frac{d+2}{2d+2}} \log\frac{4}{\delta} + BCM^2 \exp\Big(-\frac{B}{4CM^2}\Big).$$

Remark: We make the following remarks:
i) The first two term of RHS involves the regularization error which relates to the approximation ability. The third term of RHS is the estimation of the sample error, depending on generalization error; and the last term of RHS involves the output error. For example, taking $\lambda := 1/n$ (faster than $\mathcal{O}(n^{-\frac{d+2}{2d+2}})$ is enough), we can obtain a certain convergence rate at $\mathcal{O}(n^{-\frac{d+2}{2d+2}})$ of the excess risk under our over-parameterized setting. Note that this rate depends on d but tends to $\mathcal{O}(1/\sqrt{n})$ when $d \to \infty$, and thus is immune to CoD. Instead, kernel methods estimators can not evade CoD due to the lower bound $\Omega(n^{-\frac{1}{d}})$. Accordingly, the separation between kernel methods and neural networks is built in the perspective of function space for approximation. This separation can be also studied via sample complexity (Yehudai and Shamir, 2019), Kolmogorov width (Wu and Long, 2022).
ii) The pre-given radius R in $\mathcal{G}_R$ can be properly chosen by the standard iteration technique (Wu et al., 2006) in approximation theory, leading to a certain rate at $\mathcal{O}(n^{\epsilon-\frac{d+2}{2d+2}})$ for some sufficiently small ϵ. We do not include this result in our paper as it is classical and standard.
Discussion on the improved approximation rate: There are some literature building the improved approximation rate as well as metric entropy. To be specific, given some constants $c_1, c_2 \in \mathbb{R}$, considering the following function space in (Siegel and Xu, 2021)

$$\mathbb{T} := \overline{\left\{\sum_{i=1}^m a_i \sigma(\boldsymbol{w}_i^\top \boldsymbol{x} + b_i), \boldsymbol{w}_i \in \mathbb{S}^{d-1}, b \in [c_1, c_2], \sum_{i=1}^m |a_i| \leq 1\right\}},$$

which is the closure of the convex, symmetric hull of two-layer ReLU neural networks with the bounded ℓ_1 path norm. The estimation of the metric entropy is given by Siegel and Xu (2021)

$$\epsilon^{-\frac{2d+3}{2d}d} \lesssim \log \mathscr{N}_2(\mathbb{T}, \epsilon) \lesssim_d \epsilon^{-\frac{2d+3}{2d}}, \tag{9}$$

where $\lesssim_d$ denotes that some constant depending on d is omitted. Though both our result and (Siegel and Xu, 2021) use the smoothness structure of the symmetric convex hull, the techniques are different. (Siegel and Xu, 2021) employ the orthogonal argument for ridge functions in Hilbert space and then the metric entropy can be lower bounded by the minimum Hilbert norm of these nearly orthogonal ridge functions. Instead, our results exploit the relationship between the symmetric convex hull and VC-hull class, and then the metric entropy of the convex hull of any polynomial class is of lower order than ϵ^{-r} with $r < 2$. It appears that an improved estimation is provided by (Siegel and Xu, 2021), however, the dependence on d is unclear and difficult to be checked. If the derived result depends on c^d for some $c > 1$, the metric entropy as well as the generalization analysis still suffers from CoD. The claim for approximation and generalization cannot be well supported in this case. Besides, the dependence on d can be clearly calculated at most the polynomial order (Wu and Long, 2022) via a different spectral decomposition approach based on spherical harmonics. Nevertheless, this result is applied to Kolmogorov width but is still unanswered for metric entropy.

Based on this optimal metric entropy in Eq. (9) as well as the optimal approximation in (Siegel and Xu, 2021), it can be applied to the variation norm based space equipped with two-layer ReLU neural networks, e.g., (Parhi and Nowak, 2022, Theorem 8), (Yang and Zhou, 2024, Theorem 4.2). This leads to the minimax rate at $\mathcal{O}(n^{-\frac{d+3}{2d+3}})$ for the excess risk. However, the dependence on d is still unclear due to the use of (Siegel and Xu, 2021). Apart from this main difference, our results also differ from them in the label noise type as well as a computational algorithm that will be described as below.

4.4 A computational algorithm

As mentioned before, we directly assume that one global optimal solution can be obtained. In this subsection, we devise a certain algorithm based on the measure representation (Pilanci and Ergen, 2020; Akiyama and Suzuki, 2021; Zweig and Bruna, 2021). Note that, the developed algorithm is not the main contribution of this work, but bridges the gap between theoretical and practical use of learning in the Barron space.

Normally, optimization in the Barron space is difficult, or even NP-hard, e.g., the conditional gradient algorithm (a.k.a. Frank-Wolfe algorithm) developed in (Bach, 2017). If the standard gradient descent (or gradient flow) is employed, the CoD can not be avoided (Wojtowytsch and Weinan, 2020, Theorem 1) or an exponentially large number of widths in terms of n and d is required (Akiyama and Suzuki, 2021; Takakura and Suzuki, 2024) under mean field analysis. Besides, Akiyama and Suzuki (2022) develop a two-phase noisy gradient descent algorithm, which provably reaches the near-optimal solution. Nevertheless, the considered function space in their work is smaller than the Barron space as kernel methods do not suffer from CoD in that space; besides the convergence rate $\mathcal{O}(m^5/n)$ of the excess risk in their work is not applicable under our *over-parameterized* setting.

Here we directly employ one computational algorithm developed by (Pilanci and Ergen, 2020) that uses the measure representation of over-parameterized, two-layer ReLU neural networks and convex duality. This type algorithm based on the strong duality works in a high dimensional convex program but actually admits approximations that are successful in practice (Mishkin et al., 2022). To be specific, let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$ be the data matrix, and $\{\boldsymbol{D}_i\}_{i=1}^P$ be the set of diagonal matrices whose diagonal is given by $[1_{\{\boldsymbol{x}_1^\top \boldsymbol{w}\geqslant 0\}}, \ldots, 1_{\{\boldsymbol{x}_n^\top \boldsymbol{w}\geqslant 0\}}]$ for all possible $\boldsymbol{w} \in \mathbb{R}^d$ through the scaling-invariance of ReLU. There is a finite number P of such matrices. Under the basic over-parameterization condition with $m \geqslant n+1$, strong duality holds (Rosset et al., 2007) and then a globally optimal solution for the non-convex optimization problem (7) can be solved by a high dimensional convex program (see Appendix B for details)

$$\{\boldsymbol{u}_i^*\}_{i=1}^{2P} = \underset{\boldsymbol{u}_i\in\mathcal{C}_i}{\operatorname{argmin}} \ \frac{1}{n}\Big\| \sum_{i=1}^{2P} \boldsymbol{D}_i \boldsymbol{X}\boldsymbol{u}_i - \boldsymbol{y}\Big\|_2^2 + \lambda \sum_{i=1}^{2P} \|\boldsymbol{u}_i\|_1 \,, \tag{10}$$

where $\mathcal{C}_i = \{\boldsymbol{u} \in \mathbb{R}^d | (2\boldsymbol{D}_i - \boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{u} \geqslant 0\}$, $\mathcal{C}_{i+p} = \mathcal{C}_i$, $\forall i \in [P]$ by setting $\boldsymbol{D}_{i+p} = -\boldsymbol{D}_i$. We need to remark that this convex program has $2dP$ variables and $2nP$ linear inequalities, where $P = 2r[e(n-1)/r]^r$ and $r = \operatorname{rank}(\boldsymbol{X})$. If the data matrix is low rank, this problem can be solved efficiently. Accordingly, we are ready to present our generalization results for two-layer ReLU neural networks via a convex program (with proofs deferred to Appendix C).

Proposition 7 *[optimization] Given an over-parameterized ($m \geqslant n+1$), two-layer, ReLU neural network with $d > 5$ in Eq.* (7) *endowed by $\mathcal{G}_R = \{f_{\boldsymbol{\theta}} \in \mathcal{P}_m : \|\boldsymbol{\theta}\|_{\mathcal{P}} \leqslant R\}$ with a global minima*

$f_{\boldsymbol{\theta}}^{(T)}(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^{m} a_k^{(T)}\sigma(\langle \boldsymbol{w}_k^{(T)}, \boldsymbol{x}\rangle)$ *by solving a high dimensional convex problem (10) with* $\mathcal{O}(dr(n/r)^r)$ *variables and* $\mathcal{O}(nr(n/r)^r)$ *linear inequalities with* $r = rank(\boldsymbol{X})$ *after* T *iterations. Under Assumptions 1, 2, and moment hypothesis in Eq.* (3)*, and taking* $R \geqslant CM \geqslant 1$*, then for a large* B *and any* $0 < \delta < 1$*, the following result holds with probability* $1-\delta$

$$\left\|\pi_B(f_{\boldsymbol{\theta}^{(T)}}) - f_\rho\right\|_{L^2_{\rho_X}}^2 \lesssim \lambda + \frac{R^2}{m} + CMRd^2 n^{-\frac{d+2}{2d+2}} \log\frac{4}{\delta} + BCM^2 \exp\left(-\frac{B}{4CM^2}\right) + \mathcal{O}\left(\frac{1}{T^2}\right).$$

Remark: We make the following remarks:
i) Our results are still valid if stochastic approximation algorithms are used for solving problem (10), see Appendix C.2 for details. Though the convergence rate $\mathcal{O}(1/T^2)$ with T iterations can be achieved for optimization, the computational complexity would be high as we mentioned before. Optimization over the Barron space is difficult and such convex program admits $\mathcal{O}(dr(n/r)^r)$ variables and $\mathcal{O}(nr(n/r)^r)$ linear inequalities with $r = \text{rank}(\boldsymbol{X})$. Such problem can be efficiently solved if the data are low-rank.
iii) Note that, different from the mean field analysis, the training dynamics at each iteration cannot be well posed under the convex program, which could be a drawback. Nevertheless, all of globally optimal solutions could be obtained by a similar convex program by permutation and splitting/merging of the neurons (Wang et al., 2020), which provides a possible direction to study the scaling law of certain datasets. We hope it would facilitate the future research on the optimization over the Barron space in terms of statistical-computational gap.
iii) If we consider the classical ℓ_2 regularization (i.e., weight decay) in the original non-convex objective function, the related high dimensional optimization problem only differs in its regularizer, i.e., $\lambda\sum_{i=1}^{2P}\|\boldsymbol{u}_i\|_2$. Such differences in fact relate to the classical ℓ_1 *vs.* ℓ_2 regularization in terms of sample complexity (Gopi et al., 2013; Wei et al., 2019).

5. Proof framework

In this section, we establish the framework of proofs for Theorem 6. Since Theorem 6 is the special case of Proposition 7 without employing a certain optimization algorithm, we consider the proofs of Proposition 7 but leave the optimization error to Appendix C.2.

1) Error decomposition: The excess risk in the Barron space can be upper bounded by four terms via the following decomposition scheme (with proofs deferred to Appendix C.1).

Proposition 8 *The excess risk* $\mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^{(T)}})] - \mathcal{E}(f_\rho)$ *can be decomposed as*

$$\mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^{(T)}})] - \mathcal{E}(f_\rho) \leqslant \mathtt{Opt}(\boldsymbol{z},\lambda) + \mathtt{Out}(y) + \mathtt{D}(\lambda) + \mathtt{S}(\boldsymbol{z},\lambda,\boldsymbol{\theta}),$$

where $\mathtt{Opt}(\boldsymbol{z},\lambda) := \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^{(T)}})] - \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^\star})]$ *is the optimization error for a global optimal solution* $f_{\boldsymbol{\theta}^\star}$*;* $\mathtt{Out}(y) := \frac{1}{n}\sum_{i=1}^{n}|\pi_B(y_i) - y_i|^2$ *is the output error;* $\mathtt{D}(\lambda) := \inf_{f\in\mathcal{P}_m}\left\{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\|\boldsymbol{\theta}\|_{\mathcal{P}}\right\}$ *is the regularization error; and the sample error is* $\mathtt{S}(\boldsymbol{z},\lambda,\boldsymbol{\theta}) := \mathcal{E}\big[\pi_B(f_{\boldsymbol{\theta}^\star})\big] - \mathcal{E}_{\boldsymbol{z}}\big[\pi_B(f_{\boldsymbol{\theta}^\star})\big] + \mathcal{E}_{\boldsymbol{z}}\big(f_{\boldsymbol{\theta}}^\lambda\big) - \mathcal{E}\big(f_{\boldsymbol{\theta}}^\lambda\big)$ *with* $f_{\boldsymbol{\theta}}^\lambda = \text{argmin}_{f_{\boldsymbol{\theta}}\in\mathcal{P}_m}\left\{\mathcal{E}(f_{\boldsymbol{\theta}}) - \mathcal{E}(f_\rho) + \lambda\|\boldsymbol{\theta}\|_{\mathcal{P}}\right\}$.

In the following, we provide the key ideas needed to estimate these error terms.

2) Output error: We focus on the output error $\mathtt{Out}(y) := \frac{1}{n}\sum_{i=1}^{n}|\pi_B(y_i) - y_i|^2$, which is more intractable due to the squared order. In this case, the random variable $(|y_i| - B)^2\mathbb{I}_{\{|y_i|\geqslant B\}}$ is

no longer sub-exponential but still admits the exponential-type tail decay. We introduce sub-Weibull random variables (Vladimirova et al., 2020; Zhang and Chen, 2020) to tackle this issue, with the proof deferred to Appendix C.3.

Proposition 9 (Output error) *Let $B \geqslant CM \geqslant 1$. Under moment hypothesis in Eq.* (3)*, there exists a subset Z_1 of Z^n with probability at least $1-\delta/4$ such that*

$$\frac{1}{n}\sum_{i=1}^{n}|\pi_B(y_i)-y_i|^2 \lesssim \frac{1}{n}\log^2\frac{4}{\delta}+\exp\left(-\frac{B}{CM^2}\right)(CM^2)^2\,,\ \ \forall \boldsymbol{z}:=(\boldsymbol{x},y)\in Z_1\,.$$

Remark: The derived error bound is $\mathcal{O}(1/n)+\mathrm{Err}$, where the residual term Err admits an exponential decay w.r.p to B, which is better than previous results on unbounded outputs in (Wang and Zhou, 2011; Guo and Zhou, 2013; Liu et al., 2021) with $\mathrm{Err}:=2^kB^{-k}k^kM^k$. They require $B:=n^\epsilon$ (increasing slowly) and large k such that B^{-k} would behave like $1/n$. However, this needs $n\geqslant(2kM)^k$, and hence Err cannot easily tend to zero in prior results as the sample complexity suffers from CoD when $k\geqslant d$.

3) Regularization error: This term can be estimated by the approximation properties in the Barron space: denote $\boldsymbol{\theta}_\rho$ as the parameter of f_ρ, we have $\mathtt{D}(\lambda)\leqslant\lambda\|f_\rho\|_{\mathcal{P}}+\frac{3\|\boldsymbol{\theta}_\rho\|_{\mathcal{P}}^2}{m}$ due to $f_\rho\in\mathcal{B}(R)$ in Assumption 2 and the approximation error in (E et al., 2021, Theorem 1).

4) Sample error: Estimation of the sample error is also one key part in our proof (see Appendix C.4). The techniques differ from previous learning theory literature in terms of the function space, the estimation for truncated outputs, and the complexities of the target function.

Proposition 10 *Under Assumptions 1, 2 and moment hypothesis in Eq.* (3)*, let $R\geqslant B\geqslant M\geqslant 1$ and $CM\geqslant 1$. Then, there exists a subset of Z' of Z^n with confidence at least $1-3\delta/4$ with $0<\delta<1$ such that for any $\boldsymbol{z}:=(\boldsymbol{x},y)\in Z'$ and $f_{\boldsymbol{\theta}^\star}\in\mathcal{G}_R$*

$$\mathtt{S}(\boldsymbol{z},\lambda,\boldsymbol{\theta})\lesssim(CM)^2\left(\frac{S}{n}\log\frac{4}{\delta}+\lambda\right)+Rn^{-\frac{d+2}{2d+2}}\log\frac{4}{\delta}+BCM^2\exp\left(-\frac{B}{4CM^2}\right).$$

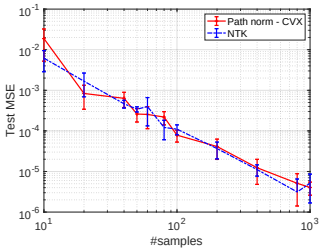
Combining the above results, we conclude the proof of Theorem 6 (as well as Proposition 7).

6. Numerical Validation

In this section, we conduct numerical experiments to validate our theoretical results in the perspective of the convergence rate of the excess risk.

To validate whether the derived (sharper) convergence rate is attainable or not, we construct a simple synthetic dataset under a known f_ρ in the over-parameterized regime. To be specific, we assume that the data are sampled from a normal Gaussian distribution, i.e., $\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)$ and normalized with $\|\boldsymbol{x}\|_2=1$. The feature dimension is $d=3$, a low dimension setting to ensure P in Theorem 14 is not large as mentioned before. We set the number of training points to range from 10 to 1000 while the number of test points is held fixed at 20. Albeit simple, such an experimental setting still works in the over-parameterized regime, see Table 1 (Left). We consider the noiseless case, where the target function is generated by a single ReLU, i.e., $y=f_\rho(\boldsymbol{x})=\sigma(\langle\boldsymbol{w}^*,\boldsymbol{x}\rangle)$ with $\boldsymbol{w}^*\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)$. The regularization parameter is set to $\lambda=10^{-8}$ for both two methods, kernel ridge regression via the NTK and the path norm based algorithm. We solve the convex program in Eq. (17)

n (#training data)	10	20	30	40	50
m (#parameters)	32	116	192	250	325



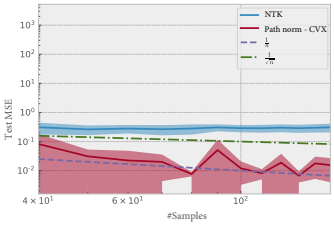


Table 1: Left: the number of (activated) parameters *v.s.* the number of training data in the synthetic dataset. Convergence rates of problem (7) under the path norm *v.s.* NTK on a synthetic dataset (Middle) and the UCI ML Breast Cancer dataset (Right), respectively.

using CVX (Grant and Boyd, 2014) to obtain the exact global minima and then compute the test MSE for regression over 5 runs.

The (middle) figure of Table 1 shows that, when learning a single ReLU beyond RKHS, our algorithm still achieves the same convergence rate as the NTK in RKHS regime. This is because, the input dimension $d = 3$ is not large, so there is no significant difference on the convergence rate.

Besides, we also conduct this experiment on a real-world dataset, i.e., the UCI ML Breast Cancer dataset with 569 samples and the dimension $d = 30$. We set 80% of samples used for training and 20% of samples for test. Here the number of training data ranges from 40 to 300, and the number of test data ranges from 10 to 75, accordingly. The remaining experimental setting is the same as that of the synthetic dataset.

The (right) figure of Table 1 shows that, when increasing the number of training data, the test MSE of NTK slightly decreases. Instead, the path norm based algorithm achieves a significant lower test MSE, which demonstrates the attainability of our theoretical results. Nevertheless, we also need to point out that, the path norm based algorithm is quite inefficient and unstable when compared to NTK. The performance is based on an extreme accurate solution by CVX, which restricts the utility of this convex program algorithm in practice. Additionally, we remark here that we do not claim this algorithm is better than SGD.

7. Conclusion and discussion

This work provides a theoretical understanding on the separation between kernel methods and neural networks from the perspective of function space. Our work sheds light on the theoretical guarantees of learning with over-parameterized two-layer neural networks under a general norm based capacity, and demonstrates the possibility of achieving sharper convergence rates with a clear dependence on d via a computational high-dimensional convex algorithm.

Our results have several findings, 1) learning with the ℓ_1-path norm is able to achieve faster convergence rate than $\mathcal{O}(\frac{1}{\sqrt{n}})$ on the excess risk under the general setting. 2) while kernel methods suffers from CoD, including the popular random features model, the NTK approach and other kernel estimators, neural networks can avoid this CoD and outperform kernel estimators from the perspective of function space theory. Hence, we hope that our analysis opens the door to improved analysis of neural networks in a norm-based capacity view and function spaces in the machine learning community.

We admit that transforming from the parameter space to the measure space leads to a very high dimensional convex problem, but the developed computational algorithm nonetheless provides a possible way to obtain the global minima. Optimization over the Barron space is quite difficult. Maybe the Barron space is still a bit large from the perspective of optimization. Identifying a suitable function space that is *data-adaptive* than RKHS as well as computationally efficient to learn a high-dimensional function, e.g., (Spek et al., 2022; Chen et al., 2023) has its own interest in approximation theory and deep learning theory and still is unanswered well, see more references (Steinwart, 2024; Schölpple and Steinwart, 2023). This is always our target to understand approximation-optimization trade-off and statistical-computational gap from kernel methods to neural networks.

Acknowledgement

This work was supported by the Hasler Foundation Program: Hasler Responsible AI (project number 21043), by the Army Research Office and was accomplished under Grant Number W911NF-24-1-0048, by the Swiss National Science Foundation (SNSF) under grant number 200021_205011.

Appendix A. Estimation of sample complexity and covering number

A.1 Proof of Lemma 4

Here we provide the upper bound of the Gaussian complexity and then transform this bound to sample complexity.

Proof [Proof of Lemma 4] According to (Barron and Klusowski, 2019, Lemma 1), due to 1-Lipschitz of the ReLU activation function σ, we have

$$\mathcal{C}_n(\sigma \circ \mathcal{G}_R) \leqslant \mathcal{C}_n(\mathcal{G}_R)\,. \tag{11}$$

Based on the definition of $\mathcal{C}_n(\mathcal{G}_R)$ in Definition 3, we have

$$\mathcal{C}_n(\mathcal{G}_R) = \mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{f_{\boldsymbol{\theta}}\in\mathcal{G}_R}\frac{1}{n}\sum_{i=1}^{n}\xi_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right] = \mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{f_{\boldsymbol{\theta}}\in\mathcal{G}_R}\frac{1}{m}\sum_{k=1}^{m}a_k\frac{1}{n}\sum_{i=1}^{n}\xi_i\sigma(\boldsymbol{w}_k^\top\boldsymbol{x}_i)\right],$$

which can be further upper bounded by

$$
\begin{aligned}
\mathcal{C}_n(\mathcal{G}_R) &\leqslant \mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{f_{\boldsymbol{\theta}}\in\mathcal{G}_R}\frac{1}{m}\sum_{k=1}^{m}|a_k|\|\boldsymbol{w}_k\|_1\frac{1}{n}\left|\sum_{i=1}^{n}\xi_i\sigma(\frac{\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|_1}^{\top}\boldsymbol{x}_i)\right|\right] \\
&\leqslant R\,\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\|\boldsymbol{w}\|_1\leqslant 1}\frac{1}{n}\left|\sum_{i=1}^{n}\xi_i\sigma(\boldsymbol{w}^{\top}\boldsymbol{x}_i)\right|\right] \\
&\leqslant 2R\,\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\|\boldsymbol{w}\|_1\leqslant 1}\frac{1}{n}\sum_{i=1}^{n}\xi_i\sigma(\boldsymbol{w}^{\top}\boldsymbol{x}_i)\right] \text{ [using symmetry of Gaussians]} \\
&\leqslant 2R\,\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\|\boldsymbol{w}\|_1\leqslant 1}\frac{1}{n}\sum_{i=1}^{n}\xi_i\boldsymbol{w}^{\top}\boldsymbol{x}_i\right] \text{ [using Eq. (11)]} \\
&\leqslant 2R\,\mathbb{E}_{\boldsymbol{\xi}}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\boldsymbol{x}_i\right\|_{\infty}\right] \text{ [using Hölder's inequality]} \\
&= 2R\,\mathbb{E}_{\boldsymbol{\xi}}\left[\max_{k=1,\dots d}\frac{1}{n}\sum_{i=1}^{n}\xi_i x_i^{(k)}\right] \\
&\leqslant 2R\sqrt{\frac{2\log d}{n}}\,, \quad \text{[using Assumption 1]}
\end{aligned}
$$

where the last inequality uses the upper bounds for d sub-Gaussian maxima, see (Wainwright, 2019, Exercise 2.12). Taking the RHS of the above equation as ϵ, we conclude the proof. ■

Remark: If we consider other types of the path norm for the two-layer ReLU neural networks, e.g., ℓ_p-path norm, we can obtain the similar result by

$$
\mathbb{E}_{\boldsymbol{\xi}}\sup_{\|\boldsymbol{w}\|_p\leq 1}\left\langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^{n}\xi_i\boldsymbol{x}_i\right\rangle \leq \mathbb{E}_{\boldsymbol{\xi}}\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\boldsymbol{x}_i\|_q = \frac{1}{n}\mathbb{E}_{\boldsymbol{\xi}}\|\boldsymbol{X}^{\top}\boldsymbol{\xi}\|_q\,,
$$

where we use Hölder's inequality. Then we can estimate $\mathbb{E}_{\boldsymbol{\xi}}\|\boldsymbol{X}^{\top}\boldsymbol{\xi}\|_q$ under different q to conclude the proof. For the commonly used ℓ_2-path norm, we have $\frac{1}{n}\mathbb{E}_{\boldsymbol{\xi}}\|\boldsymbol{X}^{\top}\boldsymbol{\xi}\|_2 \leq \frac{1}{n}\sqrt{\mathrm{tr}(\boldsymbol{X}^{\top}\boldsymbol{X})} \leq \frac{1}{\sqrt{n}}$ due to Assumption 1. Accordingly, we have $\mathcal{C}_n(\sigma\circ\mathcal{G}_R) \leq \frac{2R}{\sqrt{n}}$ under the ℓ_2-path norm. Finally we conclude the proof.

A.2 Relation between Gaussian complexity and metric entropy

Here we build the connection between Gaussian complexity and metric entropy by the following proposition, which would be beneficial to our proof.

Proposition 11 *Under Assumption 1, denote $\mathcal{G}_R = \{f_{\boldsymbol{\theta}} \in \mathcal{P}_m : \|\boldsymbol{\theta}\|_{\mathcal{P}} \leqslant R\}$, then the metric entropy of $\mathcal{G}_1$ is estimated by*

$$
\log \mathscr{N}_2(\mathcal{G}_1,\epsilon) \leqslant c\log d\left(\frac{1}{\epsilon}\right)^2, \quad \forall\epsilon>0\,, \quad \textit{with } c := 32\pi\log 2\,.
$$

To prove Theorem 11, we need lower bound the empirical Gaussian complexity using Gaussian comparison theorems. To this end, we need the following metric on packing number (Wainwright, 2019, Definition 5.4).

For $\epsilon > 0$, denote $\mathscr{D}_2(\mathcal{G}_R, \epsilon)$ as the cardinality of the largest ϵ-packing of $\mathcal{G}_R$, it admits

$$\mathscr{N}_2(\mathcal{G}_R, \epsilon) \leqslant \mathscr{D}(\mathcal{G}_R, \epsilon)\,.$$

Lemma 12 *Let $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ satisfying Assumption 1, denote $\mathcal{G}_R = \{f_{\boldsymbol{\theta}} \in \mathcal{P}_m : \|\boldsymbol{\theta}\|_{\mathcal{P}} \leqslant R\}$, then for any $\epsilon > 0$, we have*

$$\mathcal{C}_n(\mathcal{G}_1) \geqslant \frac{1}{2\sqrt{\pi \log 2}} \frac{\epsilon}{\sqrt{n}} \sqrt{\log \mathscr{N}_2(\mathcal{G}_1, \epsilon)}\,.$$

Proof Let $\epsilon > 0$. Let $\mathscr{D}$ be the largest ϵ-packing of $\mathcal{G}_1$. Due to $\mathscr{D} \subseteq \mathcal{G}_1$, we have that

$$\mathcal{C}_n(\mathcal{G}_1) \geqslant \mathbb{E}\left[\sup_{f_{\boldsymbol{\theta}} \in \mathscr{D}} \frac{1}{n} \sum_{i=1}^n Z_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right].$$

Denote $X_f := \frac{1}{n}\sum_{i=1}^n Z_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$, we can observe that for any $f, g \in \mathscr{D}$,

$$\mathbb{E}[(X_f - X_g)^2] \geqslant \frac{1}{n}\epsilon^2.$$

Define iid random variables $(Y_f)_{f \in \mathscr{D}}$ such that

$$Y_f = \frac{\epsilon}{2\sqrt{n}} W_f\,, \quad \text{with } W_f \sim \mathcal{N}(0, 1)\,,$$

then we have

$$\mathbb{E}\left[\sup_{f \in \mathscr{D}} X_f\right] \geqslant \mathbb{E}\left[\sup_{f \in \mathscr{D}} Y_f\right].$$

Since the condition for Sudakov-Fernique's comparison in Lemma 23 is verified, we have the following result by Lemma 22

$$\mathbb{E}\left[\sup_{f \in \mathscr{D}} Y_f\right] = \frac{\epsilon}{2\sqrt{n}} \mathbb{E}\left[\sup_{f \in \mathscr{D}} Z_f\right] \geqslant \frac{1}{2\sqrt{\pi \log 2}} \frac{\epsilon}{\sqrt{n}} \sqrt{\log \mathscr{D}(\mathcal{G}_1, \epsilon)}\,,$$

which concludes the proof by recalling that $\mathscr{D}(\mathcal{G}_1, \epsilon) \geqslant \mathscr{N}_2(\mathcal{G}_1, \epsilon)$. ■

Combing Lemmas 4 and 12, we can directly finish the proof of Proposition 11.

A.3 Proof of Proposition 5

To improve the estimation of the metric entropy in Proposition 11, we need a useful lemma (Van Der Vaart et al., 1996) on the convex hull technique. The considered convex hull is sequentially closed and its envelope has a weak second moment, and thus this function class is Donsker. Accordingly, the metric entropy of such convex hull of any polynomial class is of lower order than ϵ^{-r} with $r < 2$, which is enough to ensure that Dudley's entropy integral is finite.

Lemma 13 *(Van Der Vaart et al., 1996, Theorem 2.6.9) Let ρ_X be a probability measure over the input space $\boldsymbol{x} \in X$, and $\mathcal{F}$ be a class of measurable functions with a measurable squared integrable envelope F, i.e., $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \geqslant |f_{\boldsymbol{\theta}}(\boldsymbol{x})|$ for any $\boldsymbol{x} \in X$ such that we have for some $V > 0$,*

$$\mathscr{N}(\mathcal{F}, \epsilon \|F\|_{L^2_{\rho_X}}, L^2_{\rho_X}) \leqslant C \left(\frac{1}{\epsilon}\right)^V, \quad 0 < \epsilon < 1 .$$

Then there exists a constant C_V depending on C and V such that

$$\log \mathscr{N}(\overline{\mathrm{conv}}(\mathcal{F}), \epsilon \|F\|_{L^2_{\rho_X}}, L^2_{\rho_X}) \leqslant C_V \left(\frac{1}{\epsilon}\right)^{\frac{2V}{V+2}} . \tag{12}$$

Here we rewrite it by clearly indicating the constant. Let $W := 1/2 + 1/V$ and $L := C^{1/V} \|F\|_{L^2_{\rho_X}}$, the function space $\mathcal{F}$ can be covered by n balls of radius at most $Ln^{-1/V}$. Form the set $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}$ such that $\mathcal{F}_n$ is a maximal, $Ln^{-1/V}$-separated net over $\mathcal{F}$. Then Eq. (12) *is equivalent to prove*

$$\log \mathscr{N}(C_k L n^{-W}, \overline{\mathrm{conv}}(\mathcal{F}_{nk^q}), L^2_{\rho_X}) \leqslant D_k n , \quad n, k \geq 1 ,$$

where $\{C_k\}_{k=1}^\infty$ and $\{D_k\}_{k=1}^\infty$ are two sequences admitting

$$C_k = C_{k-1} + \frac{1}{k^2} ,$$
$$D_k = D_{k-1} + 2^{2q/V+1} \frac{1 + \log(1 + k^{2q/V-4+q})}{k^{2q/C-4}} ,$$

where q is some constant satisfying $q \geq 3 + V$ and we need to choose proper parameters to ensure the convergence of $\{C_k\}_{k=1}^\infty$ and $\{D_k\}_{k=1}^\infty$.

Now we are ready to prove Proposition 5.

Proof [Proof of Proposition 5] Recall the considered function space $\mathcal{P}_m$

$$\mathcal{P}_m = \left\{ f_{\boldsymbol{\theta}}(\cdot) = \frac{1}{m} \sum_{k=1}^m a_k \sigma\big(\langle \boldsymbol{w}_k, \cdot \rangle\big) = \frac{1}{m} \sum_{k=1}^m \widetilde{a}_k \sigma\left(\langle \widetilde{\boldsymbol{w}}_k, \cdot \rangle\right) \right\} , \tag{13}$$

where $\widetilde{a}_k = \|\boldsymbol{w}_k\|_1 a_k / m$ and $\widetilde{\boldsymbol{w}}_k = \boldsymbol{w}_k / \|\boldsymbol{w}_k\|_1$ by the scaling variance property of ReLU for any $\boldsymbol{w}_k \in \mathbb{R}^d / \{\mathbf{0}\}$, $\forall k \in [m]$, so we have $\|\boldsymbol{\theta}\|_{\mathcal{P}} = \sum_{k=1}^m |\widetilde{a}_k|$. In this case, we consider $\mathcal{G}_R = \{f \in \mathcal{P}_m : \|\boldsymbol{\theta}\|_{\mathcal{P}} \leqslant R\}$, and the related parameter space $\widetilde{\boldsymbol{w}} \in \mathcal{W} = \mathbb{S}_1^{d-1}$ is the ℓ_1-norm ball. [5]

We consider the following function space

$$\mathcal{F} = \{\sigma(\langle \widetilde{\boldsymbol{w}}, \cdot \rangle) : \widetilde{\boldsymbol{w}} \in \mathcal{W}\} \cup \{0\} \cup \{-\sigma(\langle \widetilde{\boldsymbol{w}}, \cdot \rangle) : \widetilde{\boldsymbol{w}} \in \mathcal{W}\} ,$$

and the convex hull of $\mathcal{F}$ is

$$\overline{\mathrm{conv}}\mathcal{F} = \left\{ \sum_{i=1}^m \alpha_i f_i \,\middle|\, f_i \in \mathcal{F}, \sum_{i=1}^m \alpha_i = 1, \alpha_i \geqslant 0, m \in \mathbb{N} \right\} . \tag{14}$$

5. If $\|\boldsymbol{w}_k\|_1 = 0$ with $k \in [m]$, we can directly set $\widetilde{a}_k = 0$ and $\widetilde{\boldsymbol{w}}_k = \mathbf{0}$. In this case, we only consider non-zero $\widetilde{\boldsymbol{w}}_k$ for convenience.

It can be found that $\mathcal{G}_1 \subset \overline{\mathrm{conv}}\mathcal{F}$, and hence we focus on bounding the covering number of $\overline{\mathrm{conv}}\mathcal{F}$. One can see that the convex hull of $\mathcal{F}$ in Eq. (14) is actually the same as the data-dependent hypothesis space with the ℓ_1 coefficient regularization (Shi et al., 2011) and we employ the result here.

Recall that in our setting the data are bounded in Assumption 1, which implies that the used ReLU in our work is 1-Lipschitz continuous with respect to the weights. Consequently, if $\{\widetilde{\boldsymbol{w}}_j\}_{j=1}^m$ is an ϵ-net of $\mathcal{W}$, then $\{\sigma(\langle\widetilde{\boldsymbol{w}}_j, \cdot\rangle)\}_{j=1}^m$ is an ϵ-net of $\{\sigma(\langle\widetilde{\boldsymbol{w}}, \cdot\rangle) : \widetilde{\boldsymbol{w}} \in \mathcal{W}\}$ in $\mathtt{C}(X)$. Accordingly, $\mathcal{F}_m$, defined as

$$\mathcal{F}_m = \{\sigma(\langle\widetilde{\boldsymbol{w}}_j, \cdot\rangle)\}_{j=1}^m \cup \{0\} \cup \{-\sigma(\langle\widetilde{\boldsymbol{w}}_j, \cdot\rangle)\}_{j=1}^m ,$$

is an ϵ-net of $\mathcal{F}$ in $\mathtt{C}(X)$. That means,

$$\mathscr{N}(\mathcal{F}, \epsilon, \|\cdot\|_\infty) \leqslant 2\mathscr{N}(\mathcal{W}, \epsilon) + 1 ,$$

where $\mathscr{N}(\mathcal{W}, \epsilon)$ is the covering number of $\mathcal{W}$ with respect to the Euclidean distance. Due to $\mathcal{W} = \mathbb{S}_1^{d-1}$ in our problem, we have $\mathscr{N}(\mathbb{S}_1^{d-1}, \epsilon) \leqslant (3/\epsilon)^d$ (Wainwright, 2019, Chapter 5).

The class $\mathcal{F}$ satisfies the conditions of Lemma 13 with F being the constant 1 as both weights and data are bounded. Accordingly, we have

$$\mathscr{N}(\mathcal{F}, \epsilon\|F\|_{L^2_{\rho_X}}, L^2_{\rho_X}) \leqslant 2^{d+1}\left(\frac{1}{\epsilon}\right)^d + 1 .$$

By taking $V := d$ and $C := 2^{d+1} + 1$ in Lemma 13 and the empirical measure $\mu := \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_i}$, then we have

$$\log \mathscr{N}_2(\mathcal{G}_1, \epsilon) \leqslant \log \mathscr{N}_2(\overline{\mathrm{conv}}\mathcal{F}, \epsilon, \mu) \leqslant c\left(\frac{1}{\epsilon}\right)^{\frac{2d}{d+2}} , \tag{15}$$

where c is some constant depending on d.

To ensure our estimate in Eq. (15) non-vacuous, we need to carefully check the proof of (Van Der Vaart et al., 1996, Theorem 2.6.9) to ensure that c cannot depend on the exponential order of d. To be specific, we have

$$\log \mathscr{N}_2(\overline{\mathrm{conv}}\mathcal{F}, \epsilon, \mu) \leqslant D_k[C_k(2^{d+1}+1)^{\frac{1}{d}}]^{\frac{2d}{d+2}}\left(\frac{1}{\epsilon}\right)^{\frac{2d}{d+2}} ,$$

where $\{C_k\}_{k=1}^\infty$ and $\{D_k\}_{k=1}^\infty$ are two converged sequences under some proper parameters admitting

$$C_k = C_{k-1} + \frac{1}{k^2} ,$$
$$D_k = D_{k-1} + 64\frac{1 + \log(1 + k^{2+3d})}{k^2} .$$

By some calculation, we have $C_k \leq C_1 + 2$ for any k, and D_k admits

$$D_k \leq D_{k-1} + \frac{64}{k^2} + \frac{192(1+d)\log k}{k^2} \leq D_{k-1} + \frac{64}{k^2} + \frac{384(1+d)}{k^{3/2}} \leq D_1 + 896 + 768d .$$

Therefore, we only need to check C_1 and D_1, respectively.

For C_1, the proof of (Van Der Vaart et al., 1996, Theorem 2.6.9) requires that the following inequality holds

$$e^{n/d}(6d^{\frac{1}{2}+\frac{1}{d}})\left(e + \frac{eC_1^2}{d^{\frac{2}{d}}}\right)^{8d^{\frac{2}{d}}C_1^{-2}n} \leq e^n , \tag{16}$$

where $\mathcal{F}$ is covered by n balls with the radius at most $(2^{d+1}+1)^{\frac{1}{d}} n^{-\frac{1}{d}}$. Then Eq. (16) is equivalent to

$$8d^{\frac{2}{d}} C_1^{-2} \log\left(e + \frac{eC_1^2}{d^{\frac{2}{d}}}\right) \leq 1 - \frac{1}{d}\left[1 + \log 6 + (\frac{1}{2} + \frac{1}{d}) \log d\right] .$$

Clearly, taking $C_1 = d^2$ is sufficient to ensure this inequality for any $d > 5$. For D_1, Eq. (16) implies that taking $D_1 = 1$ is enough, precisely, using Eq. (2.6.10) in (Van Der Vaart et al., 1996, Theorem 2.6.9).

Based on the above discussion, we have for any $d > 5$

$$\begin{aligned}
\log \mathcal{N}_2(\overline{\text{conv}}\mathcal{F}, \epsilon, \mu) &\leq D_k[C_k(2^{d+1}+1)^{\frac{1}{d}}]^{\frac{2d}{d+2}} \left(\frac{1}{\epsilon}\right)^{\frac{2d}{d+2}} \\
&\leq (897 + 768d)[(d^2+2)(2^{d+1}+1)^{\frac{1}{d}}]^{\frac{2d}{d+2}} \left(\frac{1}{\epsilon}\right)^{\frac{2d}{d+2}} \\
&\leq 6144 d^5 \left(\frac{1}{\epsilon}\right)^{\frac{2d}{d+2}} .
\end{aligned}$$

Finally we conclude the proof. ■

Appendix B. Optimization via a high dimensional convex program

In this section, we give the exact equivalence between the original non-convex optimization problem and a high dimensional convex problem by adapting the work of (Pilanci and Ergen, 2020) to our setting.

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times d}$ be the data matrix, and $\{\boldsymbol{D}_i\}_{i=1}^P$ be the set of diagonal matrices whose diagonal is given by $[1_{\{\boldsymbol{x}_1^\top \boldsymbol{w} \geqslant 0\}}, \ldots, 1_{\{\boldsymbol{x}_n^\top \boldsymbol{w} \geqslant 0\}}]$ for all possible $\boldsymbol{w} \in \mathbb{R}^d$. There is a finite number P of such matrices. We then have the following proposition on the globally optimal solutions for the non-convex optimization problem (7).

Proposition 14 *In the over-parameterized regime with $m \geqslant n+1$, the non-convex regularized problem* (7) *admits a global minimum that is recovered from the solution of the convex program*

$$\{\boldsymbol{u}_i^\star\}_{i=1}^{2P} = \underset{\boldsymbol{u}_i \in \mathcal{C}_i}{\text{argmin}} \ \frac{1}{n}\Big\| \sum_{i=1}^{2P} \boldsymbol{D}_i \boldsymbol{X} \boldsymbol{u}_i - \boldsymbol{y}\Big\|_2^2 + \lambda \sum_{i=1}^{2P} \|\boldsymbol{u}_i\|_1 , \tag{17}$$

where $\mathcal{C}_i = \{\boldsymbol{u} \in \mathbb{R}^d | (2\boldsymbol{D}_i - \boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{u} \geqslant 0\}$, $\mathcal{C}_{i+p} = \mathcal{C}_i$, $\forall i \in [P]$ by setting $\boldsymbol{D}_{i+p} = -\boldsymbol{D}_i$. The solution $\{(a_k^\star, \boldsymbol{w}_k^\star)\}_{k=1}^m$ can be constructed by non-zero elements of $\{\boldsymbol{u}_i^*\}_{i=1}^{2P}$ in Eq. (17) *such that $m := \sum_{i:\boldsymbol{u}_i^\star \neq \boldsymbol{0}}^{2P} 1$. To be specific, denote the index set $\mathcal{J} = [j_1, j_2, \cdots, j_m]$ for non-zero elements of $\{\boldsymbol{u}_i^\star\}_{i=1}^{2P}$ with $j_1 \geqslant 1$ and $j_m \leqslant 2P$, we have for any $k \in [m]$, the weights are $\boldsymbol{w}_k^\star = \boldsymbol{u}_{j_k}^\star / \|\boldsymbol{u}_{j_k}^\star\|_1$; and $a_k^\star = m\|\boldsymbol{u}_{j_k}^\star\|_1$ if $j_k \leqslant P$ and $a_k^\star = -m\|\boldsymbol{u}_i^\star\|_1$ if $j_k \geqslant P+1$.*

Remark: *i*) The number of neurons m is not fixed but larger than $\sum_{i:\boldsymbol{u}_i^\star \neq 0}^{2P} 1$ (also larger than $n+1$) to satisfy the strong duality. When transformed to a convex program in measure spaces (Pilanci and Ergen, 2020; Chizat, 2021), the problem has $2dP$ variables and $2nP$ linear inequalities, where

$P = 2r[e(n-1)/r]^r$ and $r = \mathrm{rank}(\boldsymbol{X})$. If the data are low-dimensional or low-rank, then P is not very large. We remark that, compared to most previous work going beyond RKHSs assuming that a global minimum is given, we provide this computational approach for analysis.
ii) Solving problem Eq. (17) can find an optimal neural network $f_{\boldsymbol{\theta}^\star}(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^m a_k^\star \sigma(\langle \boldsymbol{w}_k^\star, \boldsymbol{x}\rangle)$. In fact, all globally optimal neural networks (i.e., all of globally optimal solutions) could be obtained by a similar convex program by permutation and splitting/merging of the neurons (Wang et al., 2020). Our analysis just focuses on one global minimum obtained by solving Eq. (17) for description simplicity.

In the next, we first give some explanations on the additional ℓ_2 regularization in Eq. (17) in Section B.1 and then present proofs of Theorem 14 for the regularization problem in Section B.2.

B.1 Additional ℓ_2 regularization in Eq. (17)

The optimal solutions of Eq. (17) may not be unique, which would increase the difficulty for the generalization analysis. To overcome this issue, we add an extra ℓ_2 regularization term, i.e., *elastic net penalty* (Zou and Hastie, 2005), into the objective function problem (17) to make it strongly convex. This scheme is common in optimization (Bruer et al., 2014) and learning theory (De Mol et al., 2009; Rosasco et al., 2019).

Eq. (17) can be equivalently transformed to the following composite convex program

$$\min_{\{\boldsymbol{u}_i\}_{i=1}^{2P}} \frac{1}{n}\Big\| \sum_{i=1}^{2P} \boldsymbol{D}_i \boldsymbol{X} \boldsymbol{u}_i - \boldsymbol{y}\Big\|_2^2 + \lambda \sum_{i=1}^{2P} \|\boldsymbol{u}_i\|_1 + \sum_{i=1}^{2P} \iota_{\mathcal{C}_i}(\boldsymbol{u}_i)\,, \tag{18}$$

where the first term is convex, smooth, gradient Lipschitz continuous; the second term is convex but nonsmooth; and the third term is related to the indicator function $\iota_{\mathcal{C}_i} : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ defined as

$$\iota_{\mathcal{C}_i}(\boldsymbol{u}_i) := \begin{cases} 0 & \text{if } \boldsymbol{u}_i \in \mathcal{C}_i; \\ +\infty & \text{otherwise}\,, \end{cases}$$

where $\mathcal{C}_i = \{\boldsymbol{u} \in \mathbb{R}^d | (2\boldsymbol{D}_i - \boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{u} \geqslant 0\}$, $\mathcal{C}_{i+p} = \mathcal{C}_i$, $\forall i \in [P]$ by setting $\boldsymbol{D}_{i+p} = -\boldsymbol{D}_i$.

Note that, since Eq. (18) is not strongly convex as $\boldsymbol{D}_i\boldsymbol{X}$ might be not full rank, it could be difficult to obtain convergence on sequence for optimization error estimation, which increases the difficulty on generalization analysis. Thankfully, the strongly convexity property can be obtained considering an *elastic net penalty* (Zou and Hastie, 2005), that is adding a small strongly convex term to the sparsity inducing penalty. This will lead to a nice sequence convergence in optimization. Thereby, Eq. (18) can be transformed to

$$\min_{\{\boldsymbol{u}_i\}_{i=1}^{2P}} \underbrace{\frac{1}{n}\Big\| \sum_{i=1}^{2P} \boldsymbol{D}_i \boldsymbol{X} \boldsymbol{u}_i - \boldsymbol{y}\Big\|_2^2 + \widetilde{\lambda} \sum_{i=1}^{2P} \|\boldsymbol{u}_i\|_2^2}_{\triangleq g(\{\boldsymbol{u}_i\}_{i=1}^{2P})} + \lambda \sum_{i=1}^{2P} \|\boldsymbol{u}_i\|_1 + \sum_{i=1}^{2P} \iota_{\mathcal{C}_i}(\boldsymbol{u}_i)\,, \tag{19}$$

where $\widetilde{\lambda}$ is an additional regularization parameter, which can be sufficiently small, and the function g is hence strongly convex. Interestingly, in some cases, e.g., the exact recovery problem (Bruer et al., 2014), adding *elastic net penalty* does not change the optimal solution. In fact, this penalty is

not only used in optimization, but also common in learning theory as the strongly convex property on the empirical risk does not always hold. This is because, such requirement depends on the probability measure ρ and is typically not satisfied in high (possibly infinite) dimensional settings, see (De Mol et al., 2009; Rosasco et al., 2019) for details. Accordingly, we do not strictly distinguish the difference between Eq. (17) and Eq. (19) in this paper.

B.2 Proof of Proposition 14

In this subsection, we use the measure representation of two-layer neural networks via convex duality (Pilanci and Ergen, 2020; Akiyama and Suzuki, 2021). The proof technique follows (Pilanci and Ergen, 2020) except the ℓ_1-path norm regularization. For self-completeness, we provide a brief proof here.

Proof We re-write Eq. (7) as

$$p^\star = \min_{\{a_k\}_{k=1}^m, \{\boldsymbol{w}_k\}_{k=1}^m} \frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{k=1}^m a_k \sigma(\boldsymbol{w}_k^\top \boldsymbol{x}_i) \right)^2 + \frac{\lambda}{m}\sum_{k=1}^m |a_k| \|\boldsymbol{w}_k\|_1 \,, \tag{20}$$

which is equivalent to the following optimization problem

$$p^\star = \min_{\|\boldsymbol{w}_k\|_1 \leqslant 1} \max_{\substack{\boldsymbol{v}\in\mathbb{R}^n \text{ s.t.} \\ |\boldsymbol{v}^\top (\boldsymbol{X}\boldsymbol{w}_k)_+| \leqslant \lambda,\ \forall k \in [m]}} -\frac{n}{4}\left\| \boldsymbol{v} - \frac{2\boldsymbol{y}}{n} \right\|_2^2 + \frac{1}{n}\|\boldsymbol{y}\|_2^2 \,. \tag{21}$$

Interchanging the order of min and max in Eq. (21), we obtain the lower bound $d^\star$ via weak duality

$$p^\star \geqslant d^\star = \max_{\substack{\boldsymbol{v}\in\mathbb{R}^n \text{ s.t.} \\ |\boldsymbol{v}^\top (\boldsymbol{X}\boldsymbol{w})_+| \leqslant \lambda, \forall \boldsymbol{w}, \|\boldsymbol{w}\|_1 \leqslant 1}} -\frac{n}{4}\left\| \boldsymbol{v} - \frac{2\boldsymbol{y}}{n} \right\|_2^2 + \frac{1}{n}\|\boldsymbol{y}\|_2^2 \,. \tag{22}$$

If $m \geqslant n+1$, the strong duality holds (Rosset et al., 2007), i.e., $d^\star = p^\star$.

Similar derivation from (Pilanci and Ergen, 2020), we have

$$\begin{aligned} \min_{\substack{\boldsymbol{\zeta},\boldsymbol{\zeta}'\in\mathbb{R}^P \\ \boldsymbol{\zeta},\boldsymbol{\zeta}'\geqslant 0}} \max_{\substack{\boldsymbol{v}\in\mathbb{R}^n \\ \boldsymbol{\alpha}_i,\boldsymbol{\beta}_i\in\mathbb{R}^n \\ \boldsymbol{\alpha}_i,\boldsymbol{\beta}_i\geqslant 0,\forall i\in[M] \\ \boldsymbol{\alpha}_i',\boldsymbol{\beta}_i'\in\mathbb{R}^n \\ \boldsymbol{\alpha}_i',\boldsymbol{\beta}_i'\geqslant 0,\forall i\in[P]}} \min_{\substack{\boldsymbol{r}_i\in\mathbb{R}^d,\|\boldsymbol{r}_i\|_1\leqslant 1 \\ \boldsymbol{r}_i'\in\mathbb{R}^d,\|\boldsymbol{r}_i'\|_1\leqslant 1 \\ \forall i\in[P]}} & -\frac{n}{4}\left\| \boldsymbol{v} - \frac{2\boldsymbol{y}}{n} \right\|_2^2 + \frac{1}{n}\|\boldsymbol{y}\|_2^2 \\ & + \sum_{i=1}^P \zeta_i \big(\lambda + \boldsymbol{r}_i^\top \boldsymbol{X}^\top \boldsymbol{D}(\mathcal{S}_i)\big(\boldsymbol{v} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_i\big) - \boldsymbol{r}_i^\top \boldsymbol{X}^\top \boldsymbol{\beta}_i\big) \\ & + \sum_{i=1}^P \zeta_i' \big(\lambda + \boldsymbol{r}_i'^\top \boldsymbol{X}^\top \boldsymbol{D}(\mathcal{S}_i)\big(-\boldsymbol{v} + \boldsymbol{\alpha}_i' + \boldsymbol{\beta}_i'\big) - \boldsymbol{r}_i'^\top \boldsymbol{X}^\top \boldsymbol{\beta}_i'\big) \,, \end{aligned}$$

where we use the dual norm of ℓ_∞: $\|\boldsymbol{x}\|_\infty = \sup_{\boldsymbol{z}}\{|\boldsymbol{z}^\top \boldsymbol{x}| \big| \|\boldsymbol{x}\|_1 \leqslant 1\}$. The constraints $\|\boldsymbol{r}_i\|_1 \leqslant 1$ and $\|\boldsymbol{r}_i'\|_1 \leqslant 1$ are convex and compact, and optimization over $\boldsymbol{v}, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{r}_i, \boldsymbol{r}_i', \forall i \in [M]$ are convex. Accordingly, we can exchange the order of the inner $\max\min$ problem as a $\min\max$ problem. By maximizing over $\boldsymbol{v}, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\alpha}_i', \boldsymbol{\beta}_i'$, the dual optimization problem in Eq. (22) can be formulated as

$$\min_{\substack{\boldsymbol{\zeta},\boldsymbol{\zeta}'\in\mathbb{R}^P\\ \boldsymbol{\zeta},\boldsymbol{\zeta}'\geqslant 0}} \min_{\substack{\boldsymbol{r}_i\in\mathbb{R}^d,\|\boldsymbol{r}_i\|_1\leqslant 1\\ \boldsymbol{r}_i'\in\mathbb{R}^d,\|\boldsymbol{r}_i'\|_1\leqslant 1\\ (2\boldsymbol{D}(\mathcal{S}_i)-\boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{r}_i\geqslant 0\\ (2\boldsymbol{D}(\mathcal{S}_i)-\boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{r}_i'\geqslant 0}} \frac{1}{n}\Big\|\sum_{i=1}^P \zeta_i\boldsymbol{D}(\mathcal{S}_i)\boldsymbol{X}\boldsymbol{r}_i-\zeta_i'\boldsymbol{D}(\mathcal{S}_i)\boldsymbol{X}\boldsymbol{r}_i'-\boldsymbol{y}\Big\|_2^2+\lambda\sum_{i=1}^P(\zeta_i+\zeta_i'),$$

where we use the fact that, the constraint $(2\boldsymbol{D}(\mathcal{S}_i)-\boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{r}\leqslant 0$ is equivalent to $(\boldsymbol{D}(\mathcal{S}_i)-\boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{r}\leqslant 0$ and $\boldsymbol{D}(\mathcal{S}_i)\boldsymbol{X}\boldsymbol{r}\leqslant 0$ for any $\boldsymbol{r}$ since $\boldsymbol{D}(\mathcal{S}_i)-\boldsymbol{I}_n$ and $\boldsymbol{D}(\mathcal{S}_i)$ have no overlap. Then we have the constraint $(2\boldsymbol{D}(\mathcal{S}_i)-\boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{r}\geqslant 0$ by flipping the sign of $\boldsymbol{r}$. Likewise, we have

$$\min_{\boldsymbol{u}_i,\boldsymbol{u}_i'\in\mathcal{P}_{\mathcal{S}_i}} \frac{1}{n}\Big\|\sum_{i=1}^P \boldsymbol{D}(\mathcal{S}_i)\boldsymbol{X}(\boldsymbol{u}_i'-\boldsymbol{u}_i)-\boldsymbol{y}\Big\|_2^2+\lambda\sum_{i=1}^P(\|\boldsymbol{u}_i\|_1+\|\boldsymbol{u}_i'\|_1),$$

which concludes the proof by taking the compact form. ■

Appendix C. Proofs of Theorem 6

In this section, we give the proofs for Theorem 6. In fact, this theorem is a special case of Proposition 7, and accordingly we present the proofs here for the error decomposition in Section C.1, the error bounds for the optimization error in Section C.2, the output error in Section C.3, and the sample error in Section C.4, respectively.

Before presenting our error bound, we give a formal definition of regularization error for notational simplicity.

Definition 15 *(regularization error) The regularization error of $\mathcal{P}_m$ is defined as*

$$\mathtt{D}(\lambda):=\inf_{f_{\boldsymbol{\theta}}\in\mathcal{P}_m}\Big\{\mathcal{E}(f_{\boldsymbol{\theta}})-\mathcal{E}(f_\rho)+\lambda\|\boldsymbol{\theta}\|_{\mathcal{P}}\Big\}. \tag{23}$$

For any $\lambda>0$, the regularizing function *is defined as*

$$f_{\boldsymbol{\theta}}^{\lambda}=\operatorname*{argmin}_{f_{\boldsymbol{\theta}}\in\mathcal{P}_m}\Big\{\mathcal{E}(f_{\boldsymbol{\theta}})-\mathcal{E}(f_\rho)+\lambda\|\boldsymbol{\theta}\|_{\mathcal{P}}\Big\}. \tag{24}$$

The decay of $\mathtt{D}(\lambda)$ as $\lambda\to 0$ measures the approximation ability of the function space $\mathcal{P}_m$. We have a natural result for the regularization error from the approximation ability of the Barron space

$$\mathtt{D}(\lambda)=\inf_{f_{\boldsymbol{\theta}}\in\mathcal{P}_m}\Big\{\|f_{\boldsymbol{\theta}}-f_\rho\|_{L^2_{\rho_X}}^2+\lambda\|f_{\boldsymbol{\theta}}\|_{\mathcal{P}}\Big\}\leqslant\lambda\|f_\rho\|_{\mathcal{P}}+\frac{3\|\boldsymbol{\theta}_\rho\|_{\mathcal{P}}^2}{m},\ \ \forall\lambda>0, \tag{25}$$

where $\boldsymbol{\theta}_\rho$ denotes the parameter of the target function f_ρ. This is because $f_\rho\in\mathcal{B}(R)$ in Assumption 2 and the basic approximation performance of the Barron space (E et al., 2021, Theorem 1). In the literature of learning theory for regularized kernel based methods (Cucker and Zhou, 2007; Steinwart and Andreas, 2008), the regularization error is assumed to be satisfied $\mathtt{D}(\lambda):=\inf_{f\in\mathcal{H}}\Big\{\mathcal{E}(f)-\mathcal{E}(f_\rho)+\lambda\|f\|_{\mathcal{H}}^2\Big\}\leqslant\mathcal{O}(\lambda^\beta)$ with $0<\beta\leqslant 1$, where $\mathcal{H}$ is a RKHS associated with a positive definite kernel and $\|\cdot\|_{\mathcal{H}}$ is a Hilbert norm. In our work, this assumption naturally holds because of the off-the-shelf approximation result in the Barron space.

C.1 Proof of the error decomposition

Here we give the proof of the error decomposition presented in Proposition 8. Before presenting the results, we decompose the sample error $\mathtt{S}(\boldsymbol{z}, \lambda, \boldsymbol{\theta})$ into the following two parts $\mathtt{S}(\boldsymbol{z}, \lambda, \boldsymbol{\theta}) = \mathtt{S}_1(\boldsymbol{z}, \lambda, \boldsymbol{\theta}) + \mathtt{S}_2(\boldsymbol{z}, \lambda, \boldsymbol{\theta})$ with

$$\begin{aligned}\mathtt{S}_1(\boldsymbol{z}, \lambda, \boldsymbol{\theta}) &= \mathcal{E}\left[\pi_B(f_{\boldsymbol{\theta}^\star})\right] - \mathcal{E}(f_\rho) - \mathcal{E}_{\boldsymbol{z}}\left[\pi_B(f_{\boldsymbol{\theta}^\star})\right] + \mathcal{E}_{\boldsymbol{z}}(f_\rho)\,,\\ \mathtt{S}_2(\boldsymbol{z}, \lambda, \boldsymbol{\theta}) &= \left\{\mathcal{E}_{\boldsymbol{z}}\big(f_{\boldsymbol{\theta}}^\lambda\big) - \mathcal{E}_{\boldsymbol{z}}(f_\rho)\right\} - \left\{\mathcal{E}(f_{\boldsymbol{\theta}}^\lambda) - \mathcal{E}(f_\rho)\right\}.\end{aligned}$$

Now we are ready to present our proof of Proposition 8.

Proof [Proof of Proposition 8] According to the project operator π_B in Definition 1, we have $0 \leqslant \pi_B(a) - \pi_B(b) \leqslant a - b$ if $a \geqslant b$; and $a - b \leqslant \pi_B(a) - \pi_B(b) \leqslant 0$ if $a \leqslant b$. That means, for the squared loss, there holds

$$\left(\pi_B(f_{\boldsymbol{\theta}}(\boldsymbol{x})) - \pi_B(y)\right)^2 \leqslant \left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y\right)^2,$$

which implies

$$\begin{aligned}\mathcal{E}_{\boldsymbol{z}}[\pi_B(f_{\boldsymbol{\theta}})] &= \frac{1}{n}\sum_{i=1}^n \left(\pi_B(f_{\boldsymbol{\theta}}(\boldsymbol{x})) - y\right)^2 \leqslant \frac{1}{n}\sum_{i=1}^n \left(\pi_B(f_{\boldsymbol{\theta}}(\boldsymbol{x})) - \pi_B(y)\right)^2 + \frac{1}{n}\sum_{i=1}^n \left|\pi_B(y_i) - y_i\right|^2\\ &\leqslant \mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{\theta}}) + \frac{1}{n}\sum_{i=1}^n \left|\pi_B(y_i) - y_i\right|^2,\end{aligned} \tag{26}$$

where the second term is termed as the output error. Accordingly, we write the excess risk as

$$\begin{aligned}\mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^{(T)}})] - \mathcal{E}(f_\rho) &\leqslant \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^{(T)}})] - \mathcal{E}(f_\rho) + \lambda\|f_{\boldsymbol{\theta}^\star}\|_{\mathcal{P}}\\ &= \mathtt{Opt}(\boldsymbol{z}, \lambda) + \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^\star})] - \mathcal{E}(f_\rho) + \lambda\|f_{\boldsymbol{\theta}^\star}\|_{\mathcal{P}}\\ &= \mathtt{Opt}(\boldsymbol{z}, \lambda) + \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^\star})] - \mathcal{E}_{\boldsymbol{z}}[\pi_B(f_{\boldsymbol{\theta}^\star})] - \mathcal{E}(f_\rho) + \mathcal{E}_{\boldsymbol{z}}[\pi_B(f_{\boldsymbol{\theta}^\star})] + \lambda\|f_{\boldsymbol{\theta}^\star}\|_{\mathcal{P}}\\ &\leqslant \mathtt{Opt}(\boldsymbol{z}, \lambda) + \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^\star})] - \mathcal{E}_{\boldsymbol{z}}[\pi_B(f_{\boldsymbol{\theta}^\star})] - \mathcal{E}(f_\rho) + \mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{\theta}^\star}) + \lambda\|f_{\boldsymbol{\theta}^\star}\|_{\mathcal{P}} + \mathtt{Out}(y)\\ &\leqslant \mathtt{Opt}(\boldsymbol{z}, \lambda) + \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^\star})] - \mathcal{E}_{\boldsymbol{z}}[\pi_B(f_{\boldsymbol{\theta}^\star})] - \mathcal{E}(f_\rho) + \mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{\theta}}^\lambda) + \lambda\|f_{\boldsymbol{\theta}}^\lambda\|_{\mathcal{P}} + \mathtt{Out}(y)\\ &= \mathtt{Opt}(\boldsymbol{z}, \lambda) + \mathtt{D}(\lambda) + \mathtt{Out}(y) + \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^\star})] - \mathcal{E}_{\boldsymbol{z}}[\pi_B(f_{\boldsymbol{\theta}^\star})] + \mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{\theta}}^\lambda) - \mathcal{E}(f_{\boldsymbol{\theta}}^\lambda)\\ &= \mathtt{Opt}(\boldsymbol{z}, \lambda) + \mathtt{Out}(y) + \mathtt{D}(\lambda) + \mathtt{S}_1(\boldsymbol{z}, \lambda, \boldsymbol{\theta}) + \mathtt{S}_2(\boldsymbol{z}, \lambda, \boldsymbol{\theta}),\end{aligned}$$

where the second inequality uses Eq. (26) for the output error, and the third inequality holds by the condition that $f_{\boldsymbol{\theta}^\star}$ is a global minimizer. ■

C.2 Proof of the optimization error

There are several off-the-shelf algorithms to solve the convex optimization problem in Eq. (17), e.g., proximal-proximal gradient algorithm (Ryu and Yin, 2019), operator splitting (Mishchenko and Richtárik, 2019; Salim et al., 2020; Davis and Yin, 2017), which is able to enjoy $\|\boldsymbol{u}_i^{(T)} - \boldsymbol{u}_i^\star\|_2^2 \lesssim \mathcal{O}(1/T^2)$ for strongly convex problem in Eq. (17) (with additional ℓ_2 regularization), where $\boldsymbol{u}_i^{(T)}$ is the solution after T-th iterations. Accordingly, we can transform convergence on sequence in optimization to the expected risk of empirical functional.

Proposition 16 *[Optimization error] Denote the over-parameterized, two-layer ReLU neural network* $f_{\boldsymbol{\theta}^{(T)}}(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^{m} a_k^{(T)}\sigma(\langle \boldsymbol{w}_k^{(T)}, \boldsymbol{x}\rangle)$ *corresponding to Eq.* (17) *with* $\mathcal{O}(dr(n/r)^r)$ *variables and* $\mathcal{O}(nr(n/r)^r)$ *linear inequalities with* $r = rank(\boldsymbol{X})$ *solved by some convex optimization algorithms after* T *iterations. Under Assumptions 1, the expected risk of the optimal solution* $f^*_{\boldsymbol{z},\lambda}(\boldsymbol{x})$ *and the numerical solution* $f_{\boldsymbol{\theta}^{(T)}}(\boldsymbol{x})$ *can be upper bounded by*

$$\mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^{(T)}})] - \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^\star})] \lesssim \mathcal{O}\left(\frac{1}{T^2}\right).$$

Proof The optimization error under deterministic optimization algorithms can be estimated by

$$\begin{aligned}\mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^{(T)}})] - \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^\star})] &= \|\pi_B(f_{\boldsymbol{\theta}^{(T)}}) - \pi_B(f_{\boldsymbol{\theta}^\star})\|^2_{L^2_{\rho_X}} \\ &= \mathbb{E}_{\boldsymbol{x}}\left|\pi_B(f_{\boldsymbol{\theta}^{(T)}}(\boldsymbol{x})) - \pi_B(f_{\boldsymbol{\theta}^\star}(\boldsymbol{x}))\right|^2 \leqslant \mathbb{E}_{\boldsymbol{x}}\left|f_{\boldsymbol{\theta}^{(T)}}(\boldsymbol{x}) - f_{\boldsymbol{\theta}^\star}(\boldsymbol{x})\right|^2 \\ &= \mathbb{E}_{\boldsymbol{x}}\left|\frac{1}{m}\sum_{k=1}^{m}\sigma(\langle \boldsymbol{u}_k^{(T)}, \boldsymbol{x}\rangle) - \frac{1}{m}\sum_{k=1}^{m}\sigma(\langle \boldsymbol{u}_k^\star, \boldsymbol{x}\rangle)\right|^2,\end{aligned}$$

where the inequality holds by $\pi_B(a) - \pi_B(b) \leqslant |a-b|$ and the last equality benefits from the positive homogeneity of ReLU. Furthermore, the above equation can be upper bounded by

$$\begin{aligned}\mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^{(T)}})] - \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^\star})] &\leqslant \mathbb{E}_{\boldsymbol{x}}\left|\frac{1}{m}\sum_{k=1}^{m}(\boldsymbol{u}_k^{(T)} - \boldsymbol{u}_k^\star)^\top \boldsymbol{x}\right|^2 \leqslant \mathbb{E}_{\boldsymbol{x}}\left|\frac{1}{m}\sum_{k=1}^{m}\|\boldsymbol{u}_k^{(T)} - \boldsymbol{u}_k^\star\|_2\|\boldsymbol{x}\|_2\right|^2 \\ &\leqslant \frac{1}{m}\sum_{k=1}^{m}\|\boldsymbol{u}_k^{(T)} - \boldsymbol{u}_k^\star\|_2^2\mathbb{E}_{\boldsymbol{x}}\|\boldsymbol{x}\|_2^2 \\ &\lesssim \mathcal{O}\left(\frac{1}{T^2}\right),\end{aligned}$$

where the first inequality holds by 1-Lipschitz continuous of ReLU and the last inequality uses the convergence on sequence from optimization $\|\boldsymbol{u}^{(T)} - \boldsymbol{u}^\star\|_2^2 \lesssim \mathcal{O}(1/T^2)$, e.g., (Mishchenko and Richtárik, 2019; Davis and Yin, 2017). Accordingly, we finish the proof. ■

When employing stochastic approximation algorithms, e.g., stochastic decoupling algorithm with time-varying stepsize (Mishchenko and Richtárik, 2019) to solve problem (18), we have $\mathbb{E}_{\mathcal{J}}\|\boldsymbol{u}_i^{(T)} - \boldsymbol{u}_i^\star\|_2^2 \lesssim \mathcal{O}\left(\frac{1}{T^2}\right)$, where the randomness stems from sampling i from the set $\mathcal{J} = \{1, 2, \ldots, 2P\}$ to efficiently conduct the proximal operation. In this case, the optimization error under stochastic approximation algorithms can be still estimated in the similar way as below.

$$\begin{aligned}\mathbb{E}_{\mathcal{J}}\big(\mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^{(T)}})]\big) - \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^\star})] &= \mathbb{E}_{\mathcal{J}}\Big\{\mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^{(T)}})] - \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^\star})]\Big\} = \mathbb{E}_{\mathcal{J}}\|\pi_B(f_{\boldsymbol{\theta}^{(T)}}) - \pi_B(f_{\boldsymbol{\theta}^\star})\|^2_{L^2_{\rho_X}} \\ &\leqslant \mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{x}}\left|f_{\boldsymbol{\theta}^{(T)}}(\boldsymbol{x}) - f_{\boldsymbol{\theta}^\star}(\boldsymbol{x})\right|^2 \\ &= \mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{x}}\left|\frac{1}{m}\sum_{k=1}^{m}\sigma(\langle \boldsymbol{u}_k^{(T)}, \boldsymbol{x}\rangle) - \frac{1}{m}\sum_{k=1}^{m}\sigma(\langle \boldsymbol{u}_k^\star, \boldsymbol{x}\rangle)\right|^2.\end{aligned}$$

Further, by virtue of 1-Lipschitz continuous of ReLU and Assumption 1, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{J}}\big(\mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^{(T)}})]\big) - \mathcal{E}[\pi_B(f_{\boldsymbol{\theta}^\star})] &\leqslant \mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{x}}\left|\frac{1}{m}\sum_{k=1}^{m}(\boldsymbol{u}_k^{(T)} - \boldsymbol{u}_k^\star)^\top \boldsymbol{x}\right|^2 \\ &\leqslant \frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{J}}\|\boldsymbol{u}_k^{(T)} - \boldsymbol{u}_k^\star\|_2^2\mathbb{E}_{\boldsymbol{x}}\|\boldsymbol{x}\|_2^2 \\ &\lesssim \mathcal{O}\left(\frac{1}{T^2}\right).\end{aligned}$$

C.3 Proof of the output error

Estimation on the output error $\texttt{Out}(y) := \frac{1}{n}\sum_{i=1}^{n}|\pi_B(y_i) - y_i|^2$ is one key result in our proof, which is significantly different from (Guo and Zhou, 2013; Shi et al., 2014; Liu et al., 2021) in formulation and techniques. They focus on $\frac{1}{n}\sum_{i=1}^{n}|\pi_B(y_i) - y_i|$ based on moment hypothesis for high moments, and thus leave an extra term difficult to converge to zero. Instead, we focus on the output error $\texttt{Out}(y) := \frac{1}{n}\sum_{i=1}^{n}|\pi_B(y_i) - y_i|^2$, which is more intractable due to the squared order. In this case, the random variable $(|y_i| - B)^2\mathbb{I}_{\{|y_i|\geqslant B\}}$ is no longer sub-exponential but still admits the exponential-type decay. We introduce sub-Weibull random variables (Vladimirova et al., 2020; Zhang and Chen, 2020) to tackle this issue. The extra term in our result is $B\exp(-B)$ in an exponential decaying order. Accordingly, our result is able to work in a non-asymptotic regime as it does not require the exponential sample complexity.

To bound the output error, we need the following lemma. Note that, the results presented here are also needed for the sample error to tackle the unbounded outputs.

Lemma 17 *Let $B \geqslant 1$ and $CM \geqslant 1$, under the moment hypothesis in Eq.* (3)*, the error bound on truncated outputs can be estimated by*

$$\int_{|y|\geqslant B}|y|\mathrm{d}\rho_Y \leqslant 2(B + 4CM^2)\exp\left(-\frac{B}{4CM^2}\right).$$

Proof Using the integral expectation formula $\mathbb{E}(X) = \int_0^\infty \Pr(X \geqslant t)\mathrm{d}t$ for any non-negative random variable X, we have

$$\begin{aligned}\mathbb{E}\left[|y|\mathbb{I}_{\{|y|\geqslant B\}}\right] &= \mathbb{E}\left[\int_0^{|y|} 1\mathrm{d}t\mathbb{I}_{\{|y|\geqslant B\}}\right] = \mathbb{E}\left[\int_0^\infty \mathbb{I}_{\{|y|\geqslant t\}}\mathbb{I}_{\{|y|\geqslant B\}}\mathrm{d}t\right] \\ &= \int_0^\infty \Pr\left[|y| \geqslant t,\ |Y| \geqslant B\right]\mathrm{d}t = \int_0^\infty \Pr\left[|y| \geqslant \max\{B, t\}\right]\mathrm{d}t \\ &= \int_0^B \Pr\left[|y| \geqslant B\right]\mathrm{d}t + \int_B^\infty \Pr\left[|y| \geqslant t\right]\mathrm{d}t \\ &= B\Pr\left[|y| \geqslant B\right] + \int_B^\infty \Pr\left[|y| \geqslant t\right]\mathrm{d}t \\ &\leqslant 2B\exp\left(-\frac{B}{4CM^2}\right) + 8CM^2\exp\left(-\frac{B}{4CM^2}\right),\end{aligned} \tag{27}$$

where the inequality holds for the moment hypothesis in Eq. (3), which implies the sub-exponential property of $|y|$ when y is zero-mean, i.e., $\Pr(|y| \geqslant t) \leqslant 2\exp\left(-\frac{t^2}{2(CM^2+Mt)}\right)$ and we use the fact

$\frac{t^2}{2(CM^2+Mt)} \geqslant \frac{t}{4CM^2}$ when $t \geqslant B \geqslant 1$ and $CM \geqslant 1$. ■

Lemma 17 shows that the truncated outputs admit an exponential decay with the threshold B. Now we are ready to prove Proposition 9 for the output error.

Proof [Proof of Proposition 9] It is clear that

$$|y - \pi_B(y)| = (|y| - B)\mathbb{I}_{\{|y| \geqslant B\}}\,, \quad \text{and } |y - \pi_B(y)|^2 = (|y| - B)^2\mathbb{I}_{\{|y| \geqslant B\}}\,,$$

and thus we set a random variable $\zeta := (|y| - B)\mathbb{I}_{\{|y| \geqslant B\}}$ on (Z, ρ). Similar to Eq. (27), we have

$$\begin{aligned}
\int_{|y| \geqslant B} (|y| - B)^p d\rho &= \mathbb{E}\left[\int_0^{(|y|-B)^p \mathbb{I}_{\{|y| \geqslant B\}}} 1 \mathrm{d}t\right] = \mathbb{E}\left[\int_0^{\infty} \mathbb{I}_{\{(|y|-B)^p \mathbb{I}_{\{|y| \geqslant B\}}\}} \mathrm{d}t\right] \\
&= \int_0^{\infty} \Pr\left[(|y| - B)^p \mathbb{I}_{\{|y| \geqslant B\}} \geqslant t\right] \mathrm{d}t \\
&= \int_0^{\infty} \Pr\left[(|y| - B)^p \mathbb{I}_{\{|y| \geqslant B\}} \geqslant u\right] p u^{p-1} \mathrm{d}u \\
&= \int_0^{\infty} \Pr\left[|y| \geqslant B + u\right] p u^{p-1} \mathrm{d}u \\
&\leqslant 2\int_0^{\infty} \exp\left(-\frac{B+u}{4CM^2}\right) p u^{p-1} \mathrm{d}u \\
&= 2\exp\left(-\frac{B}{4CM^2}\right)(4CM^2)^p p \int_0^{\infty} \exp(-t) t^{p-1} \mathrm{d}t \\
&= 2\exp\left(-\frac{B}{4CM^2}\right)(4CM^2)^p p!\,,
\end{aligned} \tag{28}$$

where the last equality follows with the expression of the Gamma function. Accordingly, we deduce that the random variable ζ is sub-exponential as

$$(\mathbb{E}\zeta^p)^{1/p} \leqslant \left[2\exp\Big(-\frac{B}{4CM^2}\Big)\right]^{1/p} (4CM^2)(p!)^{1/p} \leqslant \widetilde{C} p^{1+\frac{1}{2p}} \to \mathcal{O}(p) \quad \text{as } p \to \infty\,,$$

where $\widetilde{C}$ is some constant depending on B, C, M and we use Stirling's approximation for factorials.

Based on the above discussion, denote the random variable $\nu_i := (|y_i| - B)^2 \mathbb{I}_{\{|y_i| \geqslant B\}}$, it is clear that ν_i is not sub-exponential but still admits the exponential-type decay. This is in fact a sub-Weibull random variable (Vladimirova et al., 2020; Zhang and Chen, 2020). Precisely, a random variable X satisfies $\Pr(|X| \geqslant t) \leqslant a\exp(-bt^\theta)$ for given a, b, θ, denoted as $X \sim \mathrm{subW}(\theta)$. We can also use Orlicz-type norms for definition. To be specific, the sub-Weibull norm is defined as

$$\|X\|_{\psi_\theta} := \inf\left\{c \in (0, \infty) : \mathbb{E}[\exp(|X|^\theta / c^\theta)] \leqslant 2\right\}\,.$$

In this case, X is a sub-Weibull random variable if it has a bounded ψ_θ-norm. In particular, if we take $\theta = 1$, we get the sub-exponential norm. Obviously, the random variable ν_i is sub-Weibull and $\theta = 1/2$. Using concentration for sub-Weilbull summation in (Zhang and Wei, 2021, Theorem 1),

we have

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\nu_i-\mathbb{E}\nu\right|\geqslant s\right)\leqslant 2\exp\left\{-\left(\frac{s^\theta}{[4eC(\theta)\|\boldsymbol{b}\|_2 L_n(\theta,\boldsymbol{b})]^\theta}\wedge\frac{s^2}{16e^2C^2(\theta)\|\boldsymbol{b}\|_2^2}\right)\right\}$$
$$=2e^{-s^\theta/[4eC(\theta)\|\boldsymbol{b}\|_2 L_n(\theta,\boldsymbol{b})]^\theta},\ \text{if } s>4eC(\theta)\|\boldsymbol{b}\|_2 L_n^{\theta/(\theta-2)}(\theta,\boldsymbol{b})\,, \tag{29}$$

where $\boldsymbol{b}=\frac{1}{n}[\|\nu_1\|_{\psi_\theta},\|\nu_2\|_{\psi_\theta},\cdots,\|\nu_n\|_{\psi_\theta}]^\top=\frac{\|\boldsymbol{\nu}\|_{\psi_\theta}}{n}\mathbf{1}$, and $C(\theta)$ is defined as

$$C(\theta):=2\left[\log^{1/\theta}2+e^3\left(\Gamma^{1/2}\left(\frac{2}{\theta}+1\right)+3^{\frac{2-\theta}{3\theta}}\sup_{p\geqslant 2}p^{-\frac{1}{\theta}}\Gamma^{1/p}\left(\frac{p}{\theta}+1\right)\right)\right],$$

and $L_n(\theta,\boldsymbol{b})=\gamma^{2/\theta}A(\theta)\frac{\|\boldsymbol{b}\|_\infty}{\|\boldsymbol{b}\|_2}$ with

$$A(\theta)=:\inf_{p\geqslant 2}\frac{e^3 3^{\frac{2-\theta}{3\theta}}p^{-1/\theta}\Gamma^{1/p}\left(\frac{p}{\theta}+1\right)}{2\left[\log^{1/\theta}2+e^3\left(\Gamma^{1/2}\left(\frac{2}{\theta}+1\right)+3^{\frac{2-\theta}{3\theta}}\sup_{p\geqslant 2}p^{-\frac{1}{\theta}}\Gamma^{1/p}\left(\frac{p}{\theta}+1\right)\right)\right]},$$

and γ is the smallest solution of the equality $\{k>1: e^{2k^{-2}}-1+\frac{e^{2\left(1-k^2\right)/k^2}}{k^2-1}\leqslant 1\}$. An approximate solution is $\gamma\approx 1.78$.

By elementary calculation, we have $C(\theta)=2[\log^2 2+2(e^3+3/4)\sqrt{6}]$ and $L_n(\theta,\boldsymbol{b})=\gamma^4\frac{4e^3\sqrt{6}}{\log^2 2+2(e^3+3/4)\sqrt{6}}n^{-1/2}$ by taking $\theta=1/2$. If $s\geqslant 8e[\log^2 2+2(e^3+3/4)\sqrt{6}]^{\frac{4}{3}}\|\nu\|_{\psi_{\frac{1}{2}}}n^{-1/3}=\mathcal{O}(n^{-1/3})$, we have

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\nu_i-\mathbb{E}\nu\right|\geqslant s\right)\leqslant\exp\left(-\frac{(ns)^{1/2}}{\widetilde{C}}\right),$$

where $\widetilde{C}$ is some constant independent of n. Note that the condition $s>cn^{-1/3}$ for some cosntant c is fair as n is large in practice. Setting the right-hand side to be $\delta/4$, we deduce that with probability at least $1-\delta/4$, there holds

$$\frac{1}{n}\sum_{i=1}^{n}\nu_i\lesssim\mathbb{E}(\nu)+\frac{1}{n}\log^2\frac{4}{\delta},$$

which implies

$$\frac{1}{n}\sum_{i=1}^{n}(|y_i|-B)^2\mathbb{I}_{\{|y_i|\geqslant B\}}\lesssim\exp\left(-\frac{B}{4CM^2}\right)(4CM^2)^2+\frac{1}{n}\log^2\frac{4}{\delta}.$$

Finally, we conclude the proof. ■

C.4 Proof of the sample error

In this subsection, we give the proofs for estimating the sample error.

C.4.1 PROOF OF $\mathtt{S}_2$ IN THE SAMPLE ERROR

This part for estimation on $\mathtt{S}_2$ is similar to (Wang and Zhou, 2011) apart from the studied function space is different. For completeness of proof, we include the proof here for self-completeness.

In our problem, we set $\{\xi_i\}_{i=1}^n$ to be independent random variables on (Z,ρ) with $\boldsymbol{z}=(\boldsymbol{x},y)$ defined as

$$\xi(\boldsymbol{x},y) := \left(y - f_{\boldsymbol{\theta}}^{\lambda}(\boldsymbol{x})\right)^2 - \left(y - f_{\rho}(\boldsymbol{x})\right)^2,$$

such that $A_i = \mathbb{E}\xi - \xi(\boldsymbol{z}_i)$ for the function ξ and $\mathbb{E}\xi = \int_Z \xi(\boldsymbol{z})\mathrm{d}\rho$.

Proposition 18 *Under Assumptions 1, 2 and moment hypothesis in Eq.* (3)*, there exists a subset of Z_2 of Z^n with confidence at least $1-\delta/4$ with $0<\delta<1$, such that $\forall \boldsymbol{z}:=(\boldsymbol{x},y)\in Z_2$*

$$\mathtt{S}_2(\boldsymbol{z},\lambda,\boldsymbol{\theta}) \lesssim (CM)^2\left(\frac{S}{n}\log\frac{4}{\delta}+\lambda\right).$$

Proof According to the definition of the considered function space $\mathcal{P}_m$, we have

$$|f_{\boldsymbol{\theta}}(\boldsymbol{x})| = \frac{1}{m}\sum_{k=1}^m |a_k|\max(\boldsymbol{\omega}_k^\top\boldsymbol{x},0) \leqslant \frac{1}{m}\sum_{k=1}^m |a_k||\boldsymbol{w}_k^\top\boldsymbol{x}| \leqslant \frac{1}{m}\sum_{k=1}^m |a_k||\boldsymbol{w}_k\|_1\|\boldsymbol{x}\|_\infty = \|\boldsymbol{x}\|_\infty\|\boldsymbol{\theta}\|_{\mathcal{P}}, \tag{30}$$

where the last inequality holds by Hölder inequality. Accordingly, we have

$$\|f_{\boldsymbol{\theta}}^{\lambda}\|_\infty \leqslant \|\boldsymbol{x}\|_\infty\|\boldsymbol{\theta}^{\lambda}\|_{\mathcal{P}} \leqslant \|\boldsymbol{x}\|_\infty\frac{\mathtt{D}(\lambda)}{\lambda} \leqslant CM\|\boldsymbol{x}\|_\infty, \tag{31}$$

where $\boldsymbol{\theta}^{\lambda}$ denotes the parameter of $f_{\boldsymbol{\theta}}^{\lambda}$. The last two inequalities hold by the definition of $f_{\boldsymbol{\theta}}^{\lambda}$ in Eq. (24) and $\mathtt{D}(\lambda)$ in Eq. (23), respectively.

By elementary inequalities, we have

$$\begin{aligned}
|\xi(\boldsymbol{z})|^p &= \left|f_\rho(\boldsymbol{x}) - f_{\boldsymbol{\theta}}^{\lambda}(\boldsymbol{x})\right|^p\left|f_\rho(\boldsymbol{x}) + f_{\boldsymbol{\theta}}^{\lambda}(\boldsymbol{x}) - 2y\right|^p \\
&\leqslant 3^p\left(\left|f_{\boldsymbol{\theta}}^{\lambda}(\boldsymbol{x})\right|^p + |f_\rho(\boldsymbol{x})|^p + 2^\ell|y|^p\right)2^{\ell-2}\left(\left|f_{\boldsymbol{\theta}}^{\lambda}(\boldsymbol{x})\right|^{p-2} + |f_\rho(\boldsymbol{x})|^{p-2}\right)\left|f_{\boldsymbol{\theta}}^{\lambda}(\boldsymbol{x}) - f_\rho(\boldsymbol{x})\right|^2,
\end{aligned}$$

which implies

$$\begin{aligned}
\mathbb{E}|\xi(\boldsymbol{z})|^p &= \int_X\int_Y |\xi(\boldsymbol{z})|^p\mathrm{d}\rho(y|\boldsymbol{x})\mathrm{d}\rho_X(\boldsymbol{x}) \\
&\leqslant 3^p\left(\left|f_{\boldsymbol{\theta}}^{\lambda}(\boldsymbol{x})\right|^p + |f_\rho(\boldsymbol{x})|^p + 2^\ell|y|^p\right)2^{\ell-2}\left(\left|f_{\boldsymbol{\theta}}^{\lambda}(\boldsymbol{x})\right|^{p-2} + |f_\rho(\boldsymbol{x})|^{p-2}\right)\int_X [f_{\boldsymbol{\theta}}^{\lambda}(\boldsymbol{x}) - f_\rho(\boldsymbol{x})]^2\rho_X(\boldsymbol{x}) \\
&\leqslant p!6^p\left((CM+1)^2\|\boldsymbol{x}\|_\infty + 2(C+1)^2M^2\right)^{p-1}\mathtt{D}(\lambda),
\end{aligned}$$

where the last inequality holds by Eq. (31). Accordingly, we have $\mathbb{E}|\xi-\mathbb{E}\xi|^p \leqslant 2^p\mathbb{E}|\xi|^p$.

Denoting $\widetilde{M} := 12(CM+1)^2\|\boldsymbol{x}\|_\infty + 2(C+1)^2M^2$ and $\widetilde{v} := 576\widetilde{M}\mathtt{D}(\lambda)$ such that $\mathbb{E}|\xi-\mathbb{E}\xi|^p \leqslant \frac{1}{2}p\widetilde{M}^{p-2}\widetilde{v}$, then by Lemma 24, we have

$$\mathrm{Prob}_{\boldsymbol{z}\in Z^n}\left\{\mathbb{E}\xi - \mathbb{E}_{\boldsymbol{z}}\xi \geqslant \epsilon\right\} \leqslant \exp\left\{-\frac{n\epsilon^2}{2\left(\widetilde{v}+\widetilde{M}\epsilon\right)}\right\}.$$

By setting the right-hand side to be $\delta/4$ such that

$$\epsilon = \frac{1}{n}\left\{\widetilde{M}\log\frac{4}{\delta} + \sqrt{\widetilde{M}^2\log^2\frac{4}{\delta} + 2n\widetilde{v}\log\frac{4}{\delta}}\right\} \leqslant \frac{20\widetilde{M}}{n}\log\frac{4}{\delta} + 18\mathtt{D}(\lambda)\,. \tag{32}$$

Then there exists a subset Z_2 of Z^n with confidence $1-\delta/4$ such that

$$\mathbb{E}\xi - \mathbb{E}_{\boldsymbol{z}}\xi \leqslant \frac{20\widetilde{M}}{n}\log\frac{4}{\delta} + 18\mathtt{D}(\lambda) \lesssim (CM)^2\left(\frac{\|\boldsymbol{x}\|_\infty}{n}\log\frac{4}{\delta} + \lambda\right),$$

and thus we conclude the proof. ■

C.4.2 PROOF OF $\mathtt{S}_1$ IN THE SAMPLE ERROR

Estimation for $\mathtt{S}_1$ is more challenging than $\mathtt{S}_2$ as the random variable $(f_{\boldsymbol{\theta}}(\boldsymbol{x})-y)^2 - (f_\rho(\boldsymbol{x})-y)^2$ is also depends on the sample $\boldsymbol{z}$ itself, in which the result on the metric entropy is needed. This is also one key technical result in our proof framework, which needs the results of the truncated output error in Appendix C.3 as well.

Proposition 19 *Under Assumptions 1, 2, and moment hypothesis in Eq.* (3) *let $R \geqslant B \geqslant M \geqslant 1$ and $CM \geqslant 1$, there exists a subset of Z' of Z^n with confidence at least $1-\delta/2$ with $0<\delta<1$ such that for any $\boldsymbol{z} := (\boldsymbol{x}, y) \in Z'$ and $f_{\boldsymbol{\theta}^\star} \in \mathcal{G}_R$*

$$\mathtt{S}_1 \leqslant \frac{1}{2}\Big\{\mathcal{E}\big[\pi_B(f_{\boldsymbol{\theta}^\star})\big] - \mathcal{E}(f_\rho)\Big\} + \widetilde{C}Rd^5n^{-\frac{d+2}{2d+2}}\log\frac{4}{\delta} + 4(B+CM)(B+72CM^2)\exp\left(-\frac{B}{4CM^2}\right) + \frac{\widetilde{C}}{n}\log\frac{4}{\delta}\,,$$

where $\widetilde{C}$ is some constant depending on B, C, M, c_q, c_q' but independent of n, δ, and d.

To prove Proposition 19, we need to the following lemma on concentration of truncated outputs as below.

Lemma 20 *Let $B \geqslant 1$ and $CM \geqslant 1$, under the moment hypothesis in Eq.* (3)*, there exists a subset Z_3 of Z^n with probability at least $1-\frac{\delta}{4}$ such that $\forall \boldsymbol{z} := (\boldsymbol{x}, y) \in Z_3$*

$$\frac{1}{n}\sum_{i=1}^n (|y_i|-B)\,\mathbb{I}_{\{|y_i|\geqslant B\}} - \int_Z (|y_i|-B)\,\mathbb{I}_{\{|y_i|\geqslant B\}}\mathrm{d}\rho \leqslant \frac{24CM^2}{n}\log\frac{4}{\delta} + 128CM^2\exp\left(-\frac{B}{CM^2}\right).$$

Proof Define a random variable $\zeta := (|y|-B)\mathbb{I}_{\{|y|\geqslant B\}}$ on (Z,ρ), similar to Eq. (28), we have

$$\begin{aligned}\mathbb{E}|\zeta - \mathbb{E}\zeta|^p \leqslant 2^{p+1}\mathbb{E}|\zeta|^p &= 2^{p+1}\int_{|y|\geqslant B}(|y|-B)^p\mathbb{I}_{\{|y|>B\}}d\rho = 2^{p+1}\int_{|y|\geqslant B}(|y|-B)^p d\rho \\ &\leqslant 2^{p+2}\exp\left(-\frac{B}{4CM^2}\right)(4CM^2)^p p! = \frac{1}{2}p!\widetilde{M}^{p-2}\widetilde{v}\,,\end{aligned}$$

where $\widetilde{M} := 8CM^2$ and $\widetilde{v} := 8\widetilde{M}^2\exp\left(-\frac{B}{4CM^2}\right)$.

By Lemma 24, setting the right-hand side of Eq. (42) to be $\delta/4$ such that

$$\varepsilon = \widetilde{M}\log\frac{4}{\delta} + \sqrt{\widetilde{M}^2\log^2\frac{4}{\delta} + 2n\tilde{v}\log\frac{4}{\delta}} \leqslant 3\widetilde{M}\log\frac{4}{\delta} + 128CM^2 n\exp\left(-\frac{B}{CM^2}\right),$$

which concludes the proof. ■

Our error analysis on $\mathtt{S}_1$ replies on the following concentration inequality which can be found in (Blanchard et al., 2008). Before introducing this, we need the definition of *sub-root* function.

Definition 21 *A function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is sub-root if it is non-negative, non-decreasing, and if $\psi(x)/\sqrt{x}$ is non-increasing.*

It is easy to see for a sub-root function and any $D > 0$, the equation $\psi(r) = r/D$ has unique positive solution.

Now we are ready to prove Proposition 19 for estimation of $\mathtt{S}_1$.

Proof [Proof of Proposition 19] Denote the empirical and expected risk on truncated output as

$$\widetilde{\mathcal{E}}_{\boldsymbol{z}}(f_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\left[\pi_B(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) - \pi_B(y_i)\right]^2, \qquad \widetilde{\mathcal{E}}(f_{\boldsymbol{\theta}}) = \mathbb{E}[\pi_B(f_{\boldsymbol{\theta}}(\boldsymbol{x})) - \pi_B(y)]^2,$$

and recall the definition of $\mathtt{S}_1(\boldsymbol{z}, \lambda, \boldsymbol{\theta})$, we can further decompose it as

$$\begin{aligned}
\mathtt{S}_1(\boldsymbol{z}, \lambda, \boldsymbol{\theta}) &= \left[\mathcal{E}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \mathcal{E}(f_\rho)\right] - \left[\mathcal{E}_{\boldsymbol{z}}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \mathcal{E}_{\boldsymbol{z}}(f_\rho)\right] && (33)\\
&= \left[\mathcal{E}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \mathcal{E}(f_\rho)\right] - \left[\widetilde{\mathcal{E}}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \widetilde{\mathcal{E}}(f_\rho)\right] && (33a)\\
&\quad + \left[\widetilde{\mathcal{E}}_{\boldsymbol{z}}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \widetilde{\mathcal{E}}_{\boldsymbol{z}}(f_\rho)\right] - \left[\mathcal{E}_{\boldsymbol{z}}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \mathcal{E}_{\boldsymbol{z}}(f_\rho)\right] && (33b)\\
&\quad + \left[\widetilde{\mathcal{E}}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \widetilde{\mathcal{E}}(f_\rho)\right] - \left[\widetilde{\mathcal{E}}_{\boldsymbol{z}}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \widetilde{\mathcal{E}}_{\boldsymbol{z}}(f_\rho)\right]. && (33c)
\end{aligned}$$

We aim to estimate the above three parts, respectively.

Estimation of Eq. (33a): $\mathtt{S}_{11}$ for short, it can be bounded by

$$\begin{aligned}
\mathtt{S}_{11} &:= \left[\mathcal{E}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \mathcal{E}(f_\rho)\right] - \left[\widetilde{\mathcal{E}}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \widetilde{\mathcal{E}}(f_\rho)\right]\\
&= \int_Z \left\{[\pi_B(f_{\boldsymbol{\theta}^\star}(\boldsymbol{x})) - y]^2 - [f_\rho(\boldsymbol{x}) - y]^2 - [\pi_B(f_{\boldsymbol{\theta}^\star}(\boldsymbol{x})) - \pi_B(y)]^2 + [f_\rho(\boldsymbol{x}) - \pi_B(y)]^2\right\}\mathrm{d}\rho\\
&= 2\int_Z [\pi_B(f_{\boldsymbol{\theta}^\star}(\boldsymbol{x})) - f_\rho(\boldsymbol{x})][y - \pi_B(y)]\mathrm{d}\rho\\
&\leqslant 2(B + CM)\int_{|y|\geqslant B} |y|\mathrm{d}\rho_Y\\
&\leqslant 4(B + CM)(B + 4CM^2)\exp\left(-\frac{B}{4CM^2}\right) := \Omega,
\end{aligned} \tag{34}$$

where the last inequality holds by Lemma 17.

Estimation of Eq. (33b): $\mathtt{S}_{12}$ for short, then for each $\boldsymbol{z} \in Z_3$, with confidence $1-\delta/4$, it can be bounded by

$$\begin{aligned}
\mathtt{S}_{12} &:= \left[\widetilde{\mathcal{E}}_{\boldsymbol{z}}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \widetilde{\mathcal{E}}_{\boldsymbol{z}}(f_\rho)\right] - \left[\mathcal{E}_{\boldsymbol{z}}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \mathcal{E}_{\boldsymbol{z}}(f_\rho)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{[\pi_B(f_{\boldsymbol{\theta}^\star}(\boldsymbol{x}_i)) - \pi_B(y_i)]^2 - [f_\rho(\boldsymbol{x}_i) - \pi_B(y_i)]^2 - [\pi_B(f_{\boldsymbol{\theta}^\star}(\boldsymbol{x}_i)) - y_i]^2 + [f_\rho(\boldsymbol{x}_i) - y_i]^2\right\} \\
&= \frac{2}{n}\sum_{i=1}^{n}[\pi_B(f_{\boldsymbol{\theta}^\star}(\boldsymbol{x}_i)) - f_\rho(\boldsymbol{x}_i)][y_i - \pi_B(y_i)] \\
&\leqslant \frac{2(B+CM)}{n}\sum_{i=1}^{n}(|y_i| - B)\mathbb{I}_{\{|y_i|\geqslant B\}} \\
&\leqslant 2(B+CM)\left[\int_Y (|y|-B)\mathbb{I}_{\{|y_i|\geqslant B\}}\mathrm{d}\rho + \frac{24CM^2}{n}\log\frac{4}{\delta} + 128CM^2\exp\Big(-\frac{B}{4CM^2}\Big)\right] \\
&\leqslant 2(B+CM)\left[\frac{24CM^2}{n}\log\frac{4}{\delta} + 136CM^2\exp\Big(-\frac{B}{4CM^2}\Big)\right],
\end{aligned} \tag{35}$$

where the second inequality holds by Lemma 20 and the last inequality uses Eq. (28).

Estimation of Eq. (33c): $\mathtt{S}_{13}$ for short, it is in fact the gap between empirical and expected version of a function $g_\pi(\boldsymbol{z}) := \left(\pi_B(f_{\boldsymbol{\theta}}(\boldsymbol{x})) - \pi_B(y)\right)^2 - \left(f_\rho(\boldsymbol{x}) - \pi_B(y)\right)^2$, which can be bounded by Lemma 25 and the metric entropy result. The proof is relatively complex and we also split it into the following four steps.

Step 1: the metric entropy

Define the set $\mathcal{G}_{R,\pi}$ of measurable functions given by $\mathcal{G}_{R,\pi} := \{g_\pi(\boldsymbol{z}) + \Omega : f_{\boldsymbol{\theta}} \in \mathcal{G}_R\}$. The metric entropy of this function class can be estimated as follows. For any two functions $g_{1,\pi} + \Omega \in \mathcal{G}_{R,\pi}$ and $g_{2,\pi} + \Omega \in \mathcal{G}_{R,\pi}$, and $\boldsymbol{z} := (\boldsymbol{x}, y) \in Z$, we have

$$\begin{aligned}
|(g_{1,\pi}+\Omega) - (g_{2,\pi}+\Omega)| &= \left|\left(\pi_B[f_1(\boldsymbol{x})] - \pi_B(y)\right)^2 - \left(\pi_B[f_2(\boldsymbol{x})] - \pi_B(y)\right)^2\right| \\
&= \left|\pi_B[f_1(\boldsymbol{x})] - \pi_B[f_2(\boldsymbol{x})]\right|\left|\pi_B[f_1(\boldsymbol{x})] + \pi_B[f_1(\boldsymbol{x})] - 2\pi_B(y)\right| \\
&\leqslant 4B\left|f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})\right|,
\end{aligned}$$

which implies

$$\mathscr{N}_2\left(\mathcal{G}_{R,\pi}, \varepsilon\right) \leq \mathscr{N}_2\left(\mathcal{G}_R, \frac{\varepsilon}{4B}\right) = \mathscr{N}_2\left(\mathcal{G}_1, \frac{\varepsilon}{4BR}\right). \tag{36}$$

Accordingly, denoting $q := \frac{2d}{d+2}$, the metric entropy result in Proposition 5 yields

$$\log \mathscr{N}_2\left(\mathcal{G}_{R,\pi}, \varepsilon\right) \leqslant c_q d^5 R^q (4B)^q \varepsilon^{-q}, \tag{37}$$

where c_q is a universal constant independent of n and d.

Step 2: Finite second moment

In the next, we use Lemma 25 and the metric entropy in Eq. (37) on $\mathcal{G}_{R,\pi}$ to find the sub-root function ψ in our setting. To this end, let $\{\xi_i\}_{i=1}^n$ with $\Pr(\xi_i = 1) = \Pr(\xi_i = -1) = 1/2$ be iid Rademancher

sequence, the second moment of g_π exists, we have (Van Der Vaart et al., 1996, Lemma 2.3.1)

$$\mathbb{E}\left[\sup_{g_\pi\in\mathcal{G}_{R,\pi},\,\mathbb{E}g_\pi^2\leqslant r}\left|\mathbb{E}g-\frac{1}{n}\sum_{i=1}^n g_\pi\left(\boldsymbol{z}_i\right)\right|\right]\leqslant 2\mathbb{E}\left[\sup_{g_\pi\in\mathcal{G}_{R,\pi},\mathbb{E}g_\pi^2\leqslant r}\left|\frac{1}{n}\sum_{i=1}^n \xi_i g_\pi\left(\boldsymbol{z}_i\right)\right|\right]. \tag{38}$$

To use this result, we need check the condition on uniform boundedness of g_π and its second moment. For $\|g_\pi\|_\infty$, we have

$$\begin{aligned}\|g_\pi\|_\infty&=\sup_{\boldsymbol{z}\in Z}\left|\left[\pi_B(f_{\boldsymbol{\theta}}(\boldsymbol{x}))-f_\rho(\boldsymbol{x})\right]\left[\pi_B(f_{\boldsymbol{\theta}}(\boldsymbol{x}))+f_\rho(\boldsymbol{x})-2\pi_B(y)\right]\right|\\&\leqslant(B+CM)(3B+CM),\end{aligned}$$

which implies

$$|g_\pi(\boldsymbol{z})+\Omega|\leqslant(B+CM)\left[4(B+4CM^2)\exp\left(-\frac{B}{4CM^2}\right)+3B+CM\right],$$

where we use Eq. (34). For the second-order moment of $g_\pi(\boldsymbol{z})$, we have

$$\begin{aligned}\mathbb{E}[g_\pi]^2&=\int_X\int_Y[\pi_B(f_{\boldsymbol{\theta}}(\boldsymbol{x}))-f_\rho(\boldsymbol{x})]^2[\pi_B(f_{\boldsymbol{\theta}}(\boldsymbol{x}))+f_\rho(\boldsymbol{x})-2\pi_B(y)]^2\mathrm{d}\rho(y|\boldsymbol{x})\mathrm{d}\rho_X(\boldsymbol{x})\\&\leqslant\left[\mathcal{E}(\pi_B(f))-\mathcal{E}(f_\rho)\right](3B+CM)^2\\&\leqslant\mathbb{E}[g_\pi+\Omega](3B+CM)^2,\end{aligned}$$

where we use $\int_X[\pi_B(f_{\boldsymbol{\theta}}(\boldsymbol{x}))-f_\rho(\boldsymbol{x})]^2\mathrm{d}\rho_X=\mathcal{E}(\pi_B(f))-\mathcal{E}(f_\rho)=\mathbb{E}g$, with $g(\boldsymbol{z}):=(\pi_B(f_{\boldsymbol{\theta}}(\boldsymbol{x}))-y)^2-(f_\rho(\boldsymbol{x})-y)^2$. That means

$$\begin{aligned}\mathbb{E}[g_\pi+\Omega]^2&=\mathbb{E}[g_\pi]^2+2\Omega\mathbb{E}[g_\pi]+\Omega^2\\&\leqslant\mathbb{E}[g_\pi+\Omega](3B+CM)\left[3B+CM+8(B+4CM^2)\exp\Big(-\frac{B}{4CM^2}\Big)\right].\end{aligned}$$

Step 3: Bound Eq. (38)

Since we have already verified the finite second moment of g_π, in the next, we aim to bound the right-hand of Eq. (38) by the ℓ_2-empirical covering number. Since $f\mapsto\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i f(\boldsymbol{z}_i)$ is a sub-Gaussian process, according to the chain argument (Giné and Nickl, 2021), there exists a universal constant C such that

$$\begin{aligned}\frac{1}{\sqrt{n}}\mathbb{E}_\xi\sup_{g_\pi\in\mathcal{G}_{R,\pi},\mathbb{E}g_\pi^2\leqslant r}\left|\sum_{i=1}^n\xi_i g_\pi\left(\boldsymbol{z}_i\right)\right|&\leqslant C\int_0^{\sqrt{V}}\sqrt{\log_2\mathscr{N}_2\left(\mathcal{G}_{R,\pi},\nu\right)}\mathrm{d}\nu\\&\leqslant C\int_0^{\sqrt{V}}\sqrt{\log_2\mathscr{N}_2\Big(\mathcal{G}_1,\frac{\nu}{4BR}\Big)}\mathrm{d}\nu\\&\leqslant Cd^{\frac{5}{2}}c_q^{1/2}(4BR)^{\frac{q}{2}}\int_0^{\sqrt{V}}\nu^{-\frac{q}{2}}\mathrm{d}\nu\\&=c_q'd^{\frac{5}{2}}(4BR)^{\frac{q}{2}}\left[\sup_{g_\pi\in\mathcal{G}_{R,\pi},\mathbb{E}g_\pi^2\leqslant r}\frac{1}{n}\sum_{i=1}^n g_\pi^2(\boldsymbol{z}_i)\right]^{\frac{1}{2}-\frac{q}{4}},\end{aligned} \tag{39}$$

where $V := \sup_{g_\pi \in \mathcal{G}_{R,\pi}, \mathbb{E}g_\pi^2 \leqslant r} \frac{1}{n}\sum_{i=1}^n g_\pi^2(\boldsymbol{z}_i)$ and we use the result in Eq. (36) and Eq. (37) in **Step 1**. By virtue of Talagrand's concentration inequality (Talagrand, 1996)

$$\mathbb{E} \sup_{g_\pi \in \mathcal{G}_{R,\pi}, \mathbb{E}g_\pi^2 \leqslant r} \frac{1}{n}\sum_{i=1}^n g_\pi^2(\boldsymbol{z}_i) \leqslant \frac{8B}{n}\mathbb{E}\mathbb{E}_\xi \sup_{g_\pi \in \mathcal{G}_{R,\pi}, \mathbb{E}g_\pi^2 \leqslant r} \left|\sum_{i=1}^n \xi_i g_\pi(\boldsymbol{z}_i)\right| + r\,,$$

taking it back to Eq. (39), we have

$$\mathscr{A}_n := \frac{1}{\sqrt{n}}\mathbb{E}\mathbb{E}_\xi \sup_{g_\pi \in \mathcal{G}_{R,\pi}, \mathbb{E}g_\pi^2 \leqslant r} \left|\sum_{i=1}^n \xi_i g_\pi\left(\boldsymbol{z}_i\right)\right| \leqslant \tilde{c}_q \left(\frac{B\mathscr{A}_n}{\sqrt{n}} + r\right)^{\frac{1}{2}-\frac{q}{4}} (4BR)^{\frac{q}{2}} d^{\frac{5}{2}}\,,$$

where we use the same notation on the constant c_q' that may change in the following derivations for notational simplicity. After solving this inequality, and taking it back to Eq. (38), we have

$$\begin{aligned}\mathbb{E}\left[\sup_{g_\pi \in \mathcal{G}_{R,\pi}, \mathbb{E}g_\pi^2 \leqslant r} \left|\mathbb{E}g_\pi - \frac{1}{n}\sum_{i=1}^n g_\pi\left(\boldsymbol{z}_i\right)\right|\right] &\leqslant c_q' d^{\frac{5}{2}} \max\left\{B^{\frac{2-q}{2+q}} n^{-\frac{2}{2+q}} (4BR)^{\frac{2q}{2+q}}, r^{\frac{1}{2}-\frac{q}{4}} n^{-\frac{1}{2}} (4BR)^{\frac{q}{2}}\right\} \\ &\leqslant c_q' d^{\frac{5}{2}} (4BR)^{\frac{2q}{2+q}} \max\left\{B^{\frac{2-q}{2+q}} n^{-\frac{2}{2+q}}, r^{\frac{1}{2}-\frac{q}{4}} n^{-\frac{1}{2}}\right\}\,.\end{aligned} \tag{40}$$

Step 4: Bound $\mathtt{S}_{13}$

According to Lemma 25, we take the right-hand of the above inequality (40) as the sub-root function $\psi(r)$. Then the solution r^* to the equation $\psi(r) = r/D$ satisfies

$$r^* \leq c_q' \max\left\{D^{\frac{4}{2+q}}, DB^{\frac{2-q}{2+q}}\right\} n^{-\frac{2}{2+q}} (4BR)^{\frac{2q}{2+q}} d^{\frac{5}{2}}\,.$$

Then all the conditions in Lemma 25 are satisfied for the function set $\mathcal{G}_{R,\pi}$ with $\alpha = 1$, $Q := (B+CM)\left[4(B+4CM^2)\exp\left(-\frac{B}{4CM^2}\right) + 3B + CM\right]$, $a := c_q' R^q (4B)^q d^{\frac{5}{2}}$, and $\tau := (3B + CM)\left[3B + CM + 8(B+4CM^2)\exp\left(-\frac{B}{4CM^2}\right)\right]$. Therefore, by Lemma 25, there exist a subset Z_4 of Z^n with probability $1-\delta/4$ such that

$$\begin{aligned}\mathtt{S}_{13} &:= \left[\widetilde{\mathcal{E}}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \widetilde{\mathcal{E}}(f_\rho)\right] - \left[\widetilde{\mathcal{E}}_{\boldsymbol{z}}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \widetilde{\mathcal{E}}_{\boldsymbol{z}}(f_\rho)\right] \\ &\leqslant \frac{1}{2}\mathbb{E}[g_\pi + \Omega] + c_q'\eta + \frac{18Q + 2\tau}{n}\log\frac{4}{\delta}, \quad \forall \boldsymbol{z} \in Z_4, f \in \mathcal{G}_R\,,\end{aligned} \tag{41}$$

where

$$\eta = \max\left\{Q^{\frac{2-q}{2+q}}, \tau^{\frac{2-q}{2+q}}\right\}\left(\frac{a}{n}\right)^{\frac{2}{2+q}} \leqslant 16B^2 R^{\frac{2q}{2+q}} \tau \left[c_q^{\frac{2}{2+q}} n^{-\frac{2}{2+q}}\right] d^{\frac{5}{2+q}}\,.$$

Combining the above three equations (34), (35), (41) into Eq. (33) to estimate $\mathtt{S}_1$, for any $\boldsymbol{z} \in Z' := Z_3 \cap Z_4$ and $f \in \mathcal{G}_R$, the following result holds with probability at least $1 - \delta/2$

$$
\begin{aligned}
&\Big\{\mathcal{E}\big[\pi_B(f_{\boldsymbol{\theta}^\star})\big] - \mathcal{E}(f_\rho)\Big\} - \Big\{\mathcal{E}_{\boldsymbol{z}}\big[\pi_B(f_{\boldsymbol{\theta}^\star})\big] - \mathcal{E}_{\boldsymbol{z}}(f_\rho)\Big\} \leqslant \frac{1}{2}\Big\{\mathcal{E}\big[\pi_B(f_{\boldsymbol{\theta}^\star})\big] - \mathcal{E}(f_\rho)\Big\} \\
&\quad + 4(B + CM)(B + 4CM^2)\exp\left(-\frac{B}{4CM^2}\right) + \tau\left(16B^2R^{\frac{2q}{2+q}}c_q'c_q^{\frac{2}{2+q}}d^{\frac{5}{2+q}}n^{-\frac{2}{2+q}} + \frac{20}{n}\log\frac{4}{\delta}\right) \\
&\quad + 2(B + CM)\left[\frac{24CM^2}{n}\log\frac{4}{\delta} + 136CM^2\exp\Big(-\frac{B}{4CM^2}\Big)\right] \\
&\leqslant \frac{1}{2}\left[\mathcal{E}\big(\pi_B(f_{\boldsymbol{\theta}^\star})\big) - \mathcal{E}(f_\rho)\right] + 4(B + CM)(B + 72CM^2)\exp\Big(-\frac{B}{4CM^2}\Big) \\
&\quad + 68(3B + CM)\left[3B + 49CM^2 + 8(B + 4CM^2)\exp\Big(-\frac{B}{4CM^2}\Big)\right]\frac{1}{n}\log\frac{4}{\delta} \\
&+ 16B(3B + CM)\left[3B + CM + 8(B + 4CM^2)\exp\Big(-\frac{B}{4CM^2}\Big)\right]R^{\frac{2q}{q+2}}c_q'c_q^{\frac{2}{2+q}}d^{\frac{5}{2+q}}n^{-\frac{2}{2+q}}\log\frac{4}{\delta},
\end{aligned}
$$

where we conclude the proof by taking $\widetilde{C} := B(B + CM^2)\left[\exp\Big(-\frac{B}{4CM^2}\Big) + 2CM^2 + \|\boldsymbol{x}\|_\infty\right]$ independent of n, δ.

Finally, combining the bounds for the output error in Proposition 9, and the sample error including $\mathtt{S}_2$ in Proposition 18 and $\mathtt{S}_1$ in Proposition 19, and the regularization error, for any $0 < \delta < 1$, with probability at least $1 - \delta$, we have (for $R \geqslant M \geqslant 1$)

$$
\begin{aligned}
\mathcal{E}(\pi_B(f_{\boldsymbol{\theta}^{(T)}})) - \mathcal{E}(f_\rho) \leqslant C_1R^{\frac{2q}{q+2}}d^{\frac{5}{2+q}}n^{-\frac{2}{2+q}}\log\frac{4}{\delta} + C_2\exp\Big(-\frac{B}{4CM^2}\Big) \\
+ \frac{C_3}{n}\log^2\frac{4}{\delta} + 38CM\lambda + \frac{114R^2}{m},
\end{aligned}
$$

where $C_1 = \mathtt{poly}(B, C, M, \|\boldsymbol{x}\|_\infty, c_q, c_q')$, $C_2 = \mathtt{poly}(B, C, M)$, and $C_3 = \mathtt{poly}(C, M, \|\boldsymbol{x}\|_\infty)$ are some constants independent of n, δ, and R. The $\mathtt{poly}$ order on B is at most 3, which means that $C_2\exp\Big(-\frac{B}{4CM^2}\Big)$ can converge to zero in an exponential order for a large B. Finally we finish the proof by taking $q = \frac{2d}{d+2}$.

■

Appendix D. Auxiliary lemmas

Here we present some useful lemmas that our proof is needed. Let us recall some useful properties of Gaussian processes by the following two lemmas.

Lemma 22 *(Kamath, 2015) Let $\{X_i\}_{i=1}^n$ be i.i.d $\mathcal{N}(0, 1)$ random variables, then consider the random variable $Z := \max_{i=1,2,\cdots,n} X_i$, we have*

$$
\mathbb{E}[Z] \geqslant \frac{1}{\sqrt{\pi\log 2}}\sqrt{\log n}\,.
$$

Lemma 23 *(Wainwright, 2019, Theorem 5.27) (Sudakov-Fernique) Let $\{X_t\}_{t=1}^T$ and $\{Y_t\}_{t=1}^T$ be a pair of zero-mean Gaussian vectors, if $\mathbb{E}(X_t - X_s)^2 \geqslant \mathbb{E}(Y_t - Y_s)^2$ for all $t, s \in T$, then, we have*

$$\mathbb{E}\left[\max_{t=1,2,\cdots,T} X_t\right] \geqslant \mathbb{E}\left[\max_{t=1,2,\cdots,T} Y_t\right] .$$

We also need the following two concentration inequalities for random variables.

Lemma 24 *(Bennett, 1962) Let $A_1, A_2, \cdots, A_n$ be independent random variables with $\mathbb{E}(A_i) = 0$. If there exists some constants $M, v > 0$ such that $\mathbb{E}|A_i|^p \leqslant \frac{1}{2}p!M^{p-2}v$ holds for $2 \leqslant p \in \mathbb{N}$, then*

$$\text{Prob}\left\{\sum_{i=1}^m A_i \geqslant \varepsilon\right\} \leqslant \exp\left\{\frac{-\varepsilon^2}{2(mv + M\varepsilon)}\right\}, \forall \varepsilon > 0 . \tag{42}$$

Lemma 25 *(Wu et al., 2007, Proposition 6) Let $\mathcal{F}$ be a set of measurable functions on Z, assume that there exists two positive constants Q, τ and $\alpha \in [0, 1]$ such that $\|f\|_\infty \leqslant Q$ and $\mathbb{E}[f^2] \leqslant \tau\mathbb{E}[f^\alpha]$ for every $f \in \mathcal{F}$. If for some $a > 0$ and $0 < q < 2$,*

$$\sup_{n\in\mathbb{N}} \sup_{\boldsymbol{z}\in Z^m} \log \mathcal{N}_2(\mathcal{F}, \varepsilon) \leqslant a\varepsilon^{-q}, \quad \forall \varepsilon > 0 . \tag{43}$$

Then there exists a constant c_q' only depending on q such that for any $t > 0$, the following proposition holds with probability at least $1 - e^{-t}$

$$\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{z}_i) \leqslant \frac{1}{2}\eta^{1-\alpha}(\mathbb{E}f)^\alpha + c_q'\eta + 2\left(\frac{\tau t}{n}\right)^{\frac{1}{2-\alpha}} + \frac{18Qt}{n} \quad \forall f \in \mathcal{F},$$

where

$$\eta := \max\left\{\tau^{\frac{2-q}{4-2\alpha+q\alpha}}\left(\frac{a}{n}\right)^{\frac{2}{4-2\alpha+q\alpha}}, \quad Q^{\frac{2-q}{2+q}}\left(\frac{a}{n}\right)^{\frac{2}{2+q}}\right\} . \tag{44}$$

References

Shunta Akiyama and Taiji Suzuki. On learnability via gradient method for two-layer relu neural networks in teacher-student setting. In *International Conference on Machine Learning*, pages 152–162, 2021.

Shunta Akiyama and Taiji Suzuki. Excess risk of two-layer relu neural networks in teacher-student settings and its superiority to kernel methods. *arXiv preprint arXiv:2205.14818*, 2022.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1):629–681, 2017.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Andrew R Barron and Jason M Klusowski. Complexity, statistical risk, and metric entropy of deep nets using total path variation. *arXiv preprint arXiv:1902.00800*, 2019.

Peter Bartlett, Dylan Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6241–6250, 2017.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Francesca Bartolucci, Ernesto De Vito, Lorenzo Rosasco, and Stefano Vigogna. Understanding neural networks with reproducing kernel Banach spaces. *Applied and Computational Harmonic Analysis*, 2023.

Yoshua Bengio, Nicolas Le Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. In *Advances in Neural Information Processing Systems*, pages 123–130, 2005.

George Bennett. Probability inequalities for the sum of independent random variables. *Publications of the American Statistical Association*, 57(297):33–45, 1962.

Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. *Annals of Statistics*, 36(2):489–531, 2008.

John J Bruer, Joel A Tropp, Volkan Cevher, and Stephen Becker. Time–data tradeoffs by aggressive smoothing. *Advances in Neural Information Processing Systems*, 27:1664–1672, 2014.

Michael Celentano, Theodor Misiakiewicz, and Andrea Montanari. Minimum complexity interpolation in random features models. *arXiv preprint arXiv:2103.15996*, 2021.

Niladri S Chatterji and Philip M Long. Foolish crowds support benign overfitting. *Journal of Machine Learning Research*, 23(125):1–12, 2022.

Hongrui Chen, Jihao Long, and Lei Wu. A duality framework for generalization analysis of random feature models and two-layer neural networks. *arXiv preprint arXiv:2305.05642*, 2023.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. *Advances in neural information processing systems*, 32:8174–8184, 2019.

Lénaïc Chizat. Convergence rates of gradient methods for convex optimization in the space of measures. *arXiv preprint arXiv:2105.08368*, 2021.

Felipe Cucker and Dingxuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-valued and variational analysis*, 25(4):829–858, 2017.

Christine De Mol, Ernesto De Vito, and Lorenzo Rosasco. Elastic-net regularization in learning theory. *Journal of Complexity*, 25(2):201–230, 2009.

Carles Domingo-Enrich and Youssef Mroueh. Tighter sparse approximation bounds for relu neural networks. In *International Conference on Learning Representations*, 2022.

Weinan E and Stephan Wojtowytsch. Representation formulas and pointwise properties for barron functions. *arXiv preprint arXiv:2006.05982*, 2020.

Weinan E and Stephan Wojtowytsch. Kolmogorov width decay and poor approximators in machine learning: Shallow neural networks, random feature models and neural tangent kernels. *Research in the Mathematical Sciences*, 8(1):1–28, 2021.

Weinan E, Chao Ma, and Lei Wu. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019.

Weinan E, Chao Ma, and Lei Wu. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, pages 1–38, 2021.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? In *Advances in Neural Information Processing Systems*, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *Annals of Statistics*, 49(2):1029–1054, 2021.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, 2021.

Sivakant Gopi, Praneeth Netrapalli, Prateek Jain, and Aditya Nori. One-bit compressed sensing: Provable support and vector recovery. In *International Conference on Machine Learning*, pages 154–162. PMLR, 2013.

Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, 2014.

Zheng-Chu Guo and Ding-Xuan Zhou. Concentration estimates for learning with unbounded sampling. *Advances in Computational Mathematics*, 38(1):207–223, 2013.

Boris Hanin and David Rolnick. Deep relu networks have surprisingly few activation patterns. *Advances in Neural Information Processing Systems*, 32, 2019.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

Gautam Kamath. Bounds on the expectation of the maximum of samples from a gaussian. *URL http://www. gautamkamath. com/writings/gaussian max. pdf*, 10:20–30, 2015.

Jason M Klusowski and Andrew R Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016.

Jason M Klusowski and Andrew R Barron. Approximation by combinations of relu and squared relu ridge functions with ℓ_1 and ℓ_0 controls. *IEEE Transactions on Information Theory*, 64(12): 7649–7656, 2018.

Vladimir Koltchinskii and Dmitry Panchenko. Complexities of convex combinations and bounding the generalization error in classification. *The Annals of Statistics*, 33(4):1455–1496, 2005.

Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and min-ℓ_l-norm interpolated classifiers. *Annals of Statistics*, 2022.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711, 2020.

Fanghui Liu, Lei Shi, Xiaolin Huang, Jie Yang, and Johan A.K. Suykens. Generalization properties of hyper-rkhs and its applications. *Journal of Machine Learning Research*, 22(140):1–38, 2021.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. In *Advances in Neural Information Processing Systems*, 2021.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022.

Konstantin Mishchenko and Peter Richtárik. A stochastic decoupling method for minimizing the sum of smooth and non-smooth functions. *arXiv preprint arXiv:1905.11535*, 2019.

Aaron Mishkin, Arda Sahiner, and Mert Pilanci. Fast convex optimization for two-layer relu networks: Equivalent model classes and cone decompositions. In *International Conference on Machine Learning*, 2022.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21:1–38, 2020.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401. PMLR, 2015.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. In *International Conference on Learning Representations*, 2020.

Rahul Parhi and Robert D Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22:1–43, 2021.

Rahul Parhi and Robert D Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Transactions on Information Theory*, 2022.

Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *International Conference on Machine Learning*, pages 7695–7705. PMLR, 2020.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2007.

Lorenzo Rosasco, Silvia Villa, and Băng Công Vũ. Convergence of stochastic proximal gradient algorithm. *Applied Mathematics & Optimization*, pages 1–27, 2019.

Saharon Rosset, Grzegorz Swirszcz, Nathan Srebro, and Ji Zhu. ℓ_1 regularization in infinite dimensional feature spaces. In *International Conference on Computational Learning Theory*, pages 544–558. Springer, 2007.

Ernest K Ryu and Wotao Yin. Proximal-proximal-gradient method. *Journal of Computational Mathematics*, 37(6):778–812, 2019.

Adil Salim, Laurent Condat, Konstantin Mishchenko, and Peter Richtárik. Dualize, split, randomize: Fast nonsmooth optimization algorithms. *arXiv preprint arXiv:2004.02635*, 2020.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pages 2667–2690. PMLR, 2019.

Mark Schmidt, Nicolas Roux, and Francis Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, pages 1458–1466, 2011.

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

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018.

Max Schölpple and Ingo Steinwart. Which spaces can be embedded in reproducing kernel hilbert spaces? *arXiv preprint arXiv:2312.14711*, 2023.

Lei Shi, Yun-Long Feng, and Ding-Xuan Zhou. Concentration estimates for learning with ℓ_1-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, 31(2):286–302, 2011.

Lei Shi, Xiaolin Huang, Zheng Tian, and Johan A.K. Suykens. Quantile regression with ℓ_1-regularization and Gaussian kernels. *Advances in Computational Mathematics*, 40(2):517–551, 2014.

Lei Shi, Xiaolin Huang, Yunlong Feng, and Johan AK Suykens. Sparse kernel regression with coefficient-based ℓ_q- regularization. *Journal of Machine Learning Research*, 20(161):1–44, 2019.

Jonathan W Siegel and Jinchao Xu. Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *arXiv preprint arXiv:2101.12365*, 2021.

Len Spek, Tjeerd Jan Heeringa, and Christoph Brune. Duality for neural networks through reproducing kernel banach spaces. *arXiv preprint arXiv:2211.05020*, 2022.

Ingo Steinwart. Reproducing kernel hilbert spaces cannot contain all continuous functions on a compact metric space. *Archiv der Mathematik*, pages 1–5, 2024.

Ingo Steinwart and Christmann Andreas. *Support Vector Machines*. Springer Science and Business Media, 2008.

Taiji Suzuki. 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*, 2019.

Taiji Suzuki and Atsushi Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. In *Advances in Neural Information Processing Systems*, 2021.

Shokichi Takakura and Taiji Suzuki. Mean-field analysis on two-layer neural networks from a kernel perspective. *arXiv preprint arXiv:2403.14917*, 2024.

Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126 (3):505–563, 1996.

Yuandong Tian. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. In *International Conference on Machine Learning*, pages 3404–3413. PMLR, 2017.

Michael Unser. A representer theorem for deep neural networks. *Journal of Machine Learning Research*, 20(110):1–30, 2019.

Aad W Van Der Vaart, Adrianus Willem van der Vaart, Aad van der Vaart, and Jon Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

Gal Vardi, Ohad Shamir, and Nathan Srebro. The sample complexity of one-hidden-layer neural networks. *arXiv preprint arXiv:2202.06233*, 2022.

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9 (1):e318, 2020.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Cheng Wang and Ding-Xuan Zhou. Optimal learning rates for least squares regularized regression with unbounded sampling. *Journal of Complexity*, 27(1):55–67, 2011.

Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum l1-norm interpolation of noisy data. *arXiv preprint arXiv:2111.05987*, 2021.

Huiyuan Wang and Wei Lin. Harmless overparametrization in two-layer neural networks. *arXiv preprint arXiv:2106.04795*, 2021.

Yifei Wang, Jonathan Lacotte, and Mert Pilanci. The hidden convex optimization landscape of two-layer relu neural networks: an exact characterization of the optimal solutions. *arXiv e-prints*, pages arXiv–2006, 2020.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pages 9712–9724, 2019.

Stephan Wojtowytsch and E Weinan. Can shallow neural networks beat the curse of dimensionality? a mean field training perspective. *IEEE Transactions on Artificial Intelligence*, 1(2):121–129, 2020.

Lei Wu and Jihao Long. A spectral-based analysis of the separation between two-layer neural networks and linear methods. *Journal of Machine Learning Research*, 119:1–34, 2022.

Qiang Wu, Yiming Ying, and Dingxuan Zhou. Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6(2):171–192, 2006.

Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23(1):108–134, 2007.

Yunfei Yang and Ding-Xuan Zhou. Optimal rates of approximation by shallow relu k neural networks and applications to nonparametric regression. *Constructive Approximation*, pages 1–32, 2024.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2019.

Huiming Zhang and Song Xi Chen. Concentration inequalities for statistical inference. *arXiv preprint arXiv:2011.02258*, 2020.

Huiming Zhang and Haoyu Wei. Sharper sub-weibull concentrations: Non-asymptotic bai-yin theorem. *arXiv preprint arXiv:2102.02450*, 2021.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Aaron Zweig and Joan Bruna. A functional perspective on learning symmetric functions with neural networks. In *International Conference on Machine Learning*, pages 13023–13032. PMLR, 2021.