COMMUN. MATH. SCI.
Vol. 18, No. 6, pp. 1755–1774

RADEMACHER COMPLEXITY AND THE GENERALIZATION ERROR OF RESIDUAL NETWORKS*

WEINAN E†, CHAO MA‡, AND QINGCAN WANG§

Abstract. Sharp bounds for the Rademacher complexity and the generalization error are derived for the residual network model. The Rademacher complexity bound has no explicit dependency on the depth of the network, while the generalization bounds are comparable to the Monte Carlo error rates, suggesting that they are nearly optimal in the high dimensional setting. These estimates are achieved by constraining the hypothesis space with an appropriately defined path norm such that the constrained space is large enough for the approximation error rates to be optimal and small enough for the estimation error rates to be optimal at the same time. Comparisons are made with other norm-based bounds.

Keywords. *a priori* estimate; residual network; weighted path norm.

AMS subject classifications. 41A46; 41A63; 62J02; 65D05.

1. Introduction

One of the major theoretical challenges in machine learning is to understand, in a high dimensional setting, the generalization error for deep neural networks, especially residual networks [10] which have become one of the default choices for many machine learning tasks. To this end, one needs to understand both the approximation error, which measures the error of the best approximation of the target function in the hypothesis space, and the estimation error, which measures the additional error due to the fact that we have a finite dataset. Naturally the larger the hypothesis space, the smaller the approximation error. Since the estimation error goes roughly in the opposite direction, one most crucial step in obtaining good estimates for the generalization error is the choice of the hypothesis space in order to achieve the right balance. In this regard, the norm-based bounds use some appropriate norms of the parameters to define the hypothesis space [3,5,9,14]. More sophisticated ways of constraining the hypothesis space can be found in [1] and [12]. As we will show later, these constraints are either too strong in the sense that the hypothesis space is not big enough for the optimal approximation error bounds to hold, or too weak in the sense that the complexity of the hypothesis space is too large.

To achieve the right balance, we define a weighted path norm for the parameters in a residual neural network. Our main result is that the Rademacher complexity scales optimally (i.e. $1/\sqrt{n}$ where n is the size of the dataset) for the hypothesis space constrained by this weighted path norm. By itself this result simply means that we have chosen a strong enough path norm to constrain the hypothesis space. So it is important to complement this result with one in a different direction, namely, we will also show that one can approximate reasonable target functions optimally (i.e. with Monte Carlo type of rates) by residual neural networks with the weighted path norm uniformly bounded by some norm of the target function. Together these results imply

*Received: June 03, 2020; Accepted (in revised form): August 23, 2020. Communicated by Shi Jin.
†Department of Mathematics and Program in Applied and Computational Mathematics, Princeton University, Princeton NJ 08540, USA (weinan@math.princeton.edu).
‡Program in Applied and Computational Mathematics, Princeton University, Princeton NJ 08544, USA (chaom@princeton.edu).
§Program in Applied and Computational Mathematics, Princeton University, Princeton NJ 08544, USA (qingcanw@princeton.edu).

a Monte Carlo kind of bounds for the generalization error. It is possible that the same result can also be established under the usual path norm. But so far we have not succeeded in proving this result.

While existing generalization bounds differ in many ways, they have one thing in common: they depend on information about the final parameters obtained in the training process. Following [7], we call them *a posteriori* bounds. *A posteriori* bounds have the advantage that they can be readily computed at the end of the training process. In practice the numerical values of these bounds are so large that they are often vacuous. In this paper, besides the *a posteriori* bounds, we will also derive *a priori* bounds. These bounds do not depend on the parameters computed. Instead, they depend on some norms of the target function. Although these a priori bounds can not be readily evaluated due to the lack of the required information about the target function, they still provide much-needed insight about the qualitative behavior of different models and different theoretical results. To use these bounds in practice, one can approximately evaluate the required norms of the target function using the output of the model.

We remark that results of similar nature have already been established in [7] for the case of two-layer neural network models.

Our estimates are optimal in the sense that they are comparable to the Monte Carlo rate. Indeed what we succeed here is to show that both the approximation error and the estimation error are controlled by quantities with a Monte Carlo origin. This gives us the $O(1/Lm+1/\sqrt{n})$ type of bounds shown below. It is well-known that one can use standard tricks in Monte Carlo methods (as well as the use of local Rademacher complexity) to improve the exponent by some $O(1/d)$ factors [4, 15], where d is the dimensionality of the problem. In the high dimensional case, this factor does not make a big difference and we will not pursue these possible improvements. For this reason, we will refer to the Monte Carlo-like error rates as being "optimal".

Table 1.1 shows a comparison of our results with other results in the literature. Since existing a posteriori estimates in machine learning are only concerned with the estimation error or the generalization gap (see below for precise definitions), to be able to make a comparison, we develop in Section 3 a standard routine that converts a posteriori estimates to a priori ones and this is how we obtain some of the items in the table.

From the table one can see the following:

(1) The estimates of [13] (l_1 path norm) and [11] (variational norm) contain an exponential and algebraic depth-dependent factor in the bound for the generalization gap respectively.

(2) The spectral norm used in [5] is strong enough to guarantee a Monte Carlo-like bound for the estimation error. But the hypothesis space with a fixed spectral norm is too small for getting a Monte Carlo-like rate for the approximation error. For the latter purpose we have to increase the size of the norm with the depth, and this results in a deterioration of the bound for the total error.

Norm	Weighted path norm	l_1 path norm	Spectral norm	Variational norm
A posteriori	$\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$	$\mathcal{O}\left(\frac{2^L}{\sqrt{n}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$	$\mathcal{O}\left(\frac{L^{3/2}}{\sqrt{n}}\right)$
A priori	$\mathcal{O}\left(\frac{1}{Lm}+\frac{1}{\sqrt{n}}\right)$	$\mathcal{O}\left(\frac{1}{Lm}+\frac{2^L}{\sqrt{n}}\right)$	$\mathcal{O}\left(\frac{1}{Lm}+\frac{(Lm)^{3/2}}{\sqrt{n}}\right)$	$\mathcal{O}\left(\frac{1}{Lm}+\frac{L^{3/2}\sqrt{m}}{\sqrt{n}}\right)$

TABLE 1.1. *Comparison of the a posteriori and a priori estimates for different norms. m is the width of the networks, and L is the depth.*

Other works that control the generalization error using norm-like quantities include [1, 9, 12].

Notations. In this paper, we let $\Omega = [0,1]^d$ be the unit hypercube, and consider target functions with domain Ω. Let π be a probability measure on Ω, for any function $f : \Omega \to \mathbb{R}$, let $\|f\|$ be the L^2 norm of f based on π,

$$\|f\|^2 = \int_\Omega f^2(\boldsymbol{x})\pi(d\boldsymbol{x}). \tag{1.1}$$

Let σ be the ReLU activation function used in the neural network models: $\sigma(x) = \max\{x, 0\}$. For a vector $\boldsymbol{x}$, $\sigma(\boldsymbol{x})$ is a vector of the same size obtained by applying ReLU component-wise. Throughout this paper, we use $\mathbb{S}^{d-1}$ to denote the unit L^1 sphere in $\mathbb{R}^d$.

2. Setup of the problem and the main theorem

2.1. Setup of the problem. We consider the regression problem and residual networks with ReLU activation $\sigma(\cdot)$. Assume that the target function $f^* : \Omega \to [0,1]$. Let the training set be $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where the $\boldsymbol{x}_i$'s are independently sampled from an underlying distribution π and $y_i = f^*(\boldsymbol{x}_i)$. Later we will consider problems with noise.

Consider the following residual network architecture with skip connection in each layer[1]

$$\begin{aligned} \boldsymbol{h}_0 &= \boldsymbol{V}\boldsymbol{x}, \\ \boldsymbol{g}_l &= \sigma(\boldsymbol{W}_l \boldsymbol{h}_{l-1}), \\ \boldsymbol{h}_l &= \boldsymbol{h}_{l-1} + \boldsymbol{U}_l \boldsymbol{g}_l, \ l = 1, \dots, L, \\ f(\boldsymbol{x}; \boldsymbol{\theta}) &= \boldsymbol{u}^T \boldsymbol{h}_L. \end{aligned} \tag{2.1}$$

Here the set of parameters $\boldsymbol{\theta} = \{\boldsymbol{V}, \boldsymbol{W}_l, \boldsymbol{U}_l, \boldsymbol{u}\}$, $\boldsymbol{V} \in \mathbb{R}^{D \times d}$, $\boldsymbol{W}_l \in \mathbb{R}^{m \times D}$, $\boldsymbol{U}_l \in \mathbb{R}^{D \times m}$, $\boldsymbol{u} \in \mathbb{R}^D$, L is the number of layers, m is the width of the residual blocks and D is the width of skip connections. Note that we omit the bias term in the network by assuming that the first element of the input $\boldsymbol{x}$ is always 1.

To simplify the proof we will consider the truncated square loss

$$\ell(\boldsymbol{x}; \boldsymbol{\theta}) = \left| \mathcal{T}_{[0,1]} f(\boldsymbol{x}; \boldsymbol{\theta}) - f^*(\boldsymbol{x}) \right|^2, \tag{2.2}$$

where $\mathcal{T}_{[0,1]}$ is the truncation operator: for any function $h(\cdot)$

$$\mathcal{T}_{[0,1]} h(\boldsymbol{x}) = \min\{\max\{h(\boldsymbol{x}), 0\}, 1\}. \tag{2.3}$$

The truncated population risk and empirical risk functions are

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x} \sim \pi} \ell(\mathbf{x}; \boldsymbol{\theta}), \quad \hat{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{x}_i; \boldsymbol{\theta}), \tag{2.4}$$

REMARK 2.1. The truncation is used in order to simplify the proof for the complexity control (Theorem 2.2). Other truncation methods can also be used. For example, we can truncate the loss function ℓ, instead of f.

[1] In practice, residual networks may use skip connections every several layers. We consider skip connections every layer for the sake of simplicity. It is easy to extend the analysis to the more general cases.

Let $\hat{f}$ be the output of the machine learning model for the target function f^* and f_m be the best approximation to the target function f^* in the hypothesis space $\mathcal{F}_m$. We can decompose the error into

$$f^* - \hat{f} = f^* - f_m + f_m - \hat{f}$$

- $f^* - f_m$ is the approximation error, due entirely to the choice of the hypothesis space.
- $f_m - \hat{f}$ is the estimation error — the additional error due to fact that we only have a finite dataset.

2.2. Function space and norms. In this paper, we consider target functions belonging to the Barron space $\mathcal{B}$. Note that in principle one can and should consider target functions in the compositional spaces defined in [6] since as shown there, they are the natural function spaces associated with residual networks. However, the machineries of the compositional spaces are much more complicated and less intuitive. Therefore to present the main ideas on estimating the Rademacher complexity, we will consider Barron space for simplicity.

The following definitions of the Barron space and the corresponding norm are adopted from [7].

DEFINITION 2.1 (Barron space). *Let $\mathbb{S}^{d-1}$ be the unit L^1 sphere in $\mathbb{R}^d$, and $\mathcal{F}$ be the Borel σ-algebra on $\mathbb{S}^{d-1}$. For any function $f:\Omega\to\mathbb{R}$, define the* Barron norm *of f as*

$$\|f\|_{\mathcal{B}} = \inf\left[\int_{\mathbb{S}^{d-1}} |a(\boldsymbol{\omega})|^2 \pi(d\boldsymbol{\omega})\right]^{1/2}, \tag{2.5}$$

where the infimum is taken over all measurable function $a(\boldsymbol{\omega})$ and probability distribution π on $(\mathbb{S}^{d-1},\mathcal{F})$ that satisfies

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}} a(\boldsymbol{\omega})\sigma(\boldsymbol{\omega}^T\boldsymbol{x})\pi(d\boldsymbol{\omega}), \tag{2.6}$$

for any $\boldsymbol{x}\in\Omega$.

The Barron space *$\mathcal{B}$ is the set of L^2 functions with finite Barron norm,*

$$\mathcal{B} = \{f:\Omega\to\mathbb{R} \mid \|f\|_{\mathcal{B}} < \infty\}. \tag{2.7}$$

The Barron space is large enough to contain many functions of interest. For example, it was shown in [11] that if a function has finite spectral norm, then it belongs to the Barron space.

DEFINITION 2.2 (Spectral norm). *Let $f\in L^2(\Omega)$, and let $F\in L^2(\mathbb{R}^d)$ be an extension of f to $\mathbb{R}^d$, and $\hat{F}$ be the Fourier transform of F. Define the* spectral norm *of f as*

$$\gamma(f) = \inf \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_1^2 |\hat{F}(\boldsymbol{\omega})| d\boldsymbol{\omega}, \tag{2.8}$$

where the infimum is taken over all possible extensions F.

COROLLARY 2.1. *Let $f:\Omega\to\mathbb{R}$ be a function that satisfies $\gamma(f)<\infty$. Then*

$$\|f\|_{\mathcal{B}} \le \gamma(f) < \infty. \tag{2.9}$$

On the other hand, for residual networks, we define the following parameter-based norm to control the estimation error. We call this norm the *weighted path norm* since it is a weighted version of the l_1 path norm studied in [13] and [18].

DEFINITION 2.3 (Weighted path norm). *Given a residual network $f(\cdot;\boldsymbol{\theta})$ with architecture (2.1), define the* weighted path norm *of f as*

$$\|f\|_{\mathrm{P}} = \|\boldsymbol{\theta}\|_{\mathrm{P}} = \left\| |\boldsymbol{u}|^T (\boldsymbol{I} + 2|\boldsymbol{U}_L||\boldsymbol{W}_L|) \cdots (\boldsymbol{I} + 2|\boldsymbol{U}_1||\boldsymbol{W}_1|) 2|\boldsymbol{V}| \right\|_1, \tag{2.10}$$

where $|\boldsymbol{A}|$, with $\boldsymbol{A}$ being a vector or matrix, means taking the absolute values of all the entries of the vector or matrix.

The weighted path norm is a weighted sum over all paths in the neural network flowing from the input to the output, and gives larger weight to the paths that go through more nonlinearities. More precisely, given a path $\mathcal{P}$, let $w_1^{\mathcal{P}}, w_2^{\mathcal{P}}, \ldots, w_L^{\mathcal{P}}$ be the weights on this path, let p be the number of non-linearities that $\mathcal{P}$ goes through. Then, it is straightforward to see that our weighted path norm can also be expressed as

$$\|f\|_{\mathrm{P}} = \sum_{\mathcal{P} \text{ is activated}} 2^{p+1} \prod_{l=1}^{L} |w_l^{\mathcal{P}}|. \tag{2.11}$$

This weighted path norm can also be seen from an "effective depth" viewpoint. It has been observed that although residual networks can be very deep, most information is processed by only a small number of nonlinearities. This has been explored for example in [17], where the authors observed numerically that residual networks behave like ensembles of networks with fewer layers. The weighted path norm naturally takes this into account.

REMARK 2.2. At a first sight, the factor 2 may trigger an alarm for fear that one might end up with some exponential depth-dependent factors in the estimates. This is not the case. In fact, as one can see in the proof, if 2 is changed by a factor of k, the result is that the constant 4 in (2.17) is changed by a factor of k. This does not mean that the factor 2 can be replaced by an arbitrary small number, since we need it to be big enough to control the Rademacher complexity. But it does not introduce any seriously bad term in the estimates. At this point, it is not clear whether the factor 2 can be removed.

2.3. Rademacher complexity. A crucial step in estimating the generalization error is to bound the generalization gap $\mathcal{L}(\boldsymbol{\theta}) - \hat{\mathcal{L}}(\boldsymbol{\theta})$. As usual, this is done by bounding the Rademacher complexity of the hypothesis space. Recall the definition of Rademacher complexity:

DEFINITION 2.4 (Rademacher complexity). *Given a function class $\mathcal{H}$ and sample set $S = \{x_i\}_{i=1}^n$, the* (empirical) Rademacher complexity *of $\mathcal{H}$ with respect to S is defined as*

$$\hat{R}(\mathcal{H}) = \frac{1}{n} \mathbb{E}_\xi \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \xi_i h(x_i) \right], \tag{2.12}$$

where the ξ_i's are independent random variables with $\Pr\{\xi_i = 1\} = \Pr\{\xi_i = -1\} = 1/2$.

By its definition, the Rademacher complexity measures the capability of the function class ability to fit random binary labels (represented by $\{\xi_i\}$). Larger Rademacher

complexity means that the function class can fit noises better, and hence is more vulnerable to overfitting. It is well-known that the Rademacher complexity can be used to control the generalization gap [16].

THEOREM 2.1. *Given a function class $\mathcal{H}$, for any $\delta\in(0,1)$, with probability at least $1-\delta$ over the random samples $\{x_i\}_{i=1}^n$,*

$$\sup_{h\in\mathcal{H}}\left|\mathbb{E}_{\mathrm{x}}[h(\mathrm{x})]-\frac{1}{n}\sum_{i=1}^{n}h(x_i)\right|\le 2\hat{R}(\mathcal{H})+2\sup_{h,h'\in\mathcal{H}}\|h-h'\|_\infty\sqrt{\frac{2\log(8/\delta)}{n}}. \tag{2.13}$$

The following theorem is our first result. It shows that the Rademacher complexity of residual networks can be controlled by the weighted path norm, with the optimal $1/\sqrt{n}$ scaling.

THEOREM 2.2. *Let $\mathcal{F}^Q=\{f(\cdot;\boldsymbol{\theta}):\|\boldsymbol{\theta}\|_{\mathrm{P}}\le Q\}$ where the $f(\cdot,\boldsymbol{\theta})$'s are residual networks defined by (2.1). Assume that the samples $\{\boldsymbol{x}_i\}_{i=1}^n\subset\Omega$. Then we have*

$$\hat{R}(\mathcal{F}^Q)\le 2Q\sqrt{\frac{2\log(2d)}{n}}. \tag{2.14}$$

Note that the definition of $\mathcal{F}^Q$ does not specify the depth or width of the network. Consequently our Rademacher complexity bound does not depend on the depth and width of the network. Hence, the resulted a posteriori estimate has no dependence on L and m either.

A corollary of this is:

COROLLARY 2.2 (A posteriori estimates). *Let $\|\boldsymbol{\theta}\|_{\mathrm{P}}$ be the weighted path norm of residual network $f(\cdot;\boldsymbol{\theta})$. Let n be the number of training samples. Let $\mathcal{L}(\boldsymbol{\theta})$ and $\hat{\mathcal{L}}(\boldsymbol{\theta})$ be the truncated population risk and empirical risk defined in (2.4). Then for any $\delta\in(0,1)$, with probability at least $1-\delta$ over the random training samples, we have*

$$\left|\mathcal{L}(\boldsymbol{\theta})-\hat{\mathcal{L}}(\boldsymbol{\theta})\right|\le 2(\|\boldsymbol{\theta}\|_{\mathrm{P}}+1)\frac{2\sqrt{2\log(2d)}+1}{\sqrt{n}}+2\sqrt{\frac{2\log(14/\delta)}{n}}. \tag{2.15}$$

2.4. A priori estimates of the generalization error. We adopt a relaxed form of the weighted norm constraint in the form of a regularized model:

$$\min_{\boldsymbol{\theta}}\ \mathcal{J}(\boldsymbol{\theta}):=\hat{\mathcal{L}}(\boldsymbol{\theta})+3\lambda\|\boldsymbol{\theta}\|_{\mathrm{P}}\sqrt{\frac{2\log(2d)}{n}}. \tag{2.16}$$

THEOREM 2.3 (A priori estimate). *Let $f^*:\Omega\to[0,1]$ and assume that the residual network $f(\cdot;\boldsymbol{\theta})$ has architecture (2.1). Let n be the number of training samples, L be the number of layers and m be the width of the residual blocks. Let $\|f\|_{\mathcal{B}}$ be the Barron norm of f^* and $\|\boldsymbol{\theta}\|_{\mathrm{P}}$ be the weighted path norm of $f(\cdot;\boldsymbol{\theta})$ in Definition 2.1 and 2.3.*

For any $\lambda\ge 4+2/[3\sqrt{2\log(2d)}]$, assume that $\hat{\boldsymbol{\theta}}$ is an optimal solution of the regularized model (2.16). Then for any $\delta\in(0,1)$, with probability at least $1-\delta$ over the random training samples, the population risk satisfies

$$\mathcal{L}(\hat{\boldsymbol{\theta}})\le\frac{3\|f\|_{\mathcal{B}}^2}{Lm}+(4\|f\|_{\mathcal{B}}+1)\frac{3(4+\lambda)\sqrt{2\log(2d)}+2}{\sqrt{n}}+4\sqrt{\frac{2\log(28/\delta)}{n}}. \tag{2.17}$$

REMARK 2.3.

(1) The estimate is a priori in nature since the right-hand side of (2.17) depends only on the Barron norm of the target function instead of the norm of $\hat{\boldsymbol{\theta}}$.

(2) We want to emphasize that our estimate is nearly optimal. The first term in (2.17) shows that the convergence rate with respect to the size of the neural network is $\mathcal{O}(1/(Lm))$, which matches the rate in the approximation theory for shallow networks [2]. The last two terms show that the rate with respect to the number of training samples is $\mathcal{O}(1/\sqrt{n})$, which matches the classical estimates of the generalization gap.

(3) The second term depends only on $\|f\|_{\mathcal{B}}$ instead of the network architecture, thus there is no need to increase the sample size n with respect to the network size parameters L and m to ensure that the model generalizes well. This is not the case for existing error bounds (see Section 3).

2.5. Extension to the case with noise. Our a priori estimates can be extended to problems with sub-Gaussian noise.

ASSUMPTION 2.1. *Assume that y_i are given by $y_i = f^*(\boldsymbol{x}_i) + \varepsilon_i$, and $\{\varepsilon_i\}$ are i.i.d. random variables such that $\mathbb{E}\varepsilon_i = 0$ and*

$$\Pr\{|\varepsilon_i| > t\} \leq ce^{-\frac{t^2}{2\sigma^2}}, \quad \forall t \geq \tau, \tag{2.18}$$

for some constants c, σ and τ.

Let $\ell_B(\boldsymbol{x};\boldsymbol{\theta}) = \ell(\boldsymbol{x};\boldsymbol{\theta}) \wedge B^2$ be the square loss truncated by B^2, and define

$$\mathcal{L}_B(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim\pi}\ell_B(\mathbf{x};\boldsymbol{\theta}), \quad \hat{\mathcal{L}}_B(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\ell_B(\boldsymbol{x}_i;\boldsymbol{\theta}). \tag{2.19}$$

Then, we have

THEOREM 2.4 (A priori estimate for noisy problems). *In addition to the conditions in Theorem 2.3, assume Assumption 2.1 holds. Let $\mathcal{L}_B(\boldsymbol{\theta})$ and $\hat{\mathcal{L}}_B(\boldsymbol{\theta})$ be the truncated population risk and empirical risk defined in (2.19). For $B \geq 1 + \max\{\tau, \sigma\sqrt{\log n}\}$ and $\lambda \geq 4 + 2B/[3\sqrt{2\log(2d)}]$, assume that $\hat{\boldsymbol{\theta}}$ is an optimal solution of the regularized model*

$$\min_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}) := \hat{\mathcal{L}}(\boldsymbol{\theta}) + \lambda B\|\boldsymbol{\theta}\|_{\mathrm{P}} \cdot 3\sqrt{\frac{2\log(2d)}{n}}. \tag{2.20}$$

Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the random training sample, the population risk satisfies

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{\theta}}) \leq{}& \frac{16\|f\|_{\mathcal{B}}^2}{Lm} + (12\|f\|_{\mathcal{B}} + 1)\frac{3(4+\lambda)B\sqrt{2\log(2d)} + 2B^2}{\sqrt{n}} \\ &+ 4B^2\sqrt{\frac{2\log(28/\delta)}{n}} + \frac{2c(4\sigma^2+1)}{\sqrt{n}}.\end{aligned} \tag{2.21}$$

We see that the a priori estimates for problems with noise only differ from that for problems without noise by a logarithmic term. In particular, the estimates of the generalization error are still nearly optimal.

2.6. Proof sketch. First, we show that any function f in the Barron space can be approximated by residual networks with increasing depth or width, and with weighted path norm uniformly bounded.

THEOREM 2.5. *For any target function $f^* \in \mathcal{B}$, and any $L, m \geq 1$, there exists a residual network $f(\cdot;\tilde{\boldsymbol{\theta}})$ with depth L and width m, such that*

$$\|f(\boldsymbol{x};\tilde{\boldsymbol{\theta}}) - f^*\|^2 \leq \frac{3\|f^*\|_{\mathcal{B}}^2}{Lm} \tag{2.22}$$

and

$$\|\tilde{\boldsymbol{\theta}}\|_{\mathrm{P}} \leq 4\|f^*\|_{\mathcal{B}}.$$

Next, consider the decomposition

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}}) = \left[\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{J}(\hat{\boldsymbol{\theta}})\right] + \left[\mathcal{J}(\hat{\boldsymbol{\theta}}) - \mathcal{J}(\tilde{\boldsymbol{\theta}})\right] + \left[\mathcal{J}(\tilde{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}})\right]. \tag{2.23}$$

Recall that $\hat{\boldsymbol{\theta}}$ is the optimal solution of the minimization problem (2.16), and $\tilde{\boldsymbol{\theta}}$ corresponds to the approximation in Theorem 2.5.

By the definition of $\mathcal{J}$ (2.16),

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{J}(\hat{\boldsymbol{\theta}}) \leq \left|\mathcal{L}(\hat{\boldsymbol{\theta}}) - \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}})\right| - 3\lambda\|\hat{\boldsymbol{\theta}}\|_{\mathrm{P}}\sqrt{\frac{2\log(2d)}{n}},$$
$$\mathcal{J}(\tilde{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}}) \leq \left|\mathcal{L}(\tilde{\boldsymbol{\theta}}) - \hat{\mathcal{L}}(\tilde{\boldsymbol{\theta}})\right| + 3\lambda\|\tilde{\boldsymbol{\theta}}\|_{\mathrm{P}}\sqrt{\frac{2\log(2d)}{n}}.$$

From the a posteriori estimate (2.15), both $|\mathcal{L}(\hat{\boldsymbol{\theta}}) - \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}})|$ and $|\mathcal{L}(\tilde{\boldsymbol{\theta}}) - \hat{\mathcal{L}}(\tilde{\boldsymbol{\theta}})|$ are bounded with high probability, thus both $\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{J}(\hat{\boldsymbol{\theta}})$ and $\mathcal{J}(\tilde{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}})$ are bounded with high probability. In addition, $\mathcal{J}(\hat{\boldsymbol{\theta}}) - \mathcal{J}(\tilde{\boldsymbol{\theta}}) \leq 0$, and the approximation result (2.22) bounds $\mathcal{L}(\tilde{\boldsymbol{\theta}})$. Plugging all of the above into (2.23) will give us the a priori estimates in Theorem 2.3.

For problems with noise, we can similarly bound $\mathcal{L}_B(\boldsymbol{\theta}) - \mathcal{J}(\boldsymbol{\theta})$ instead of $\mathcal{L}(\boldsymbol{\theta}) - \mathcal{J}(\boldsymbol{\theta})$. Hence, to formulate an a priori estimate, we also need to control $\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}_B(\boldsymbol{\theta})$. This is given by the following lemma:

LEMMA 2.1. *Assume that the noise ε has zero mean and satisfies (2.18), and $B \geq 1 + \max\{\tau, \sigma\sqrt{\log n}\}$. For any $\boldsymbol{\theta}$ we have*

$$|\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}_B(\boldsymbol{\theta})| \leq \frac{c(4\sigma^2+1)}{\sqrt{n}}. \tag{2.24}$$

3. Comparison with norm-based a posteriori estimates

Different norms have been used as a vehicle to bound the generalization error of deep neural networks, including the group norm and path norm given in [14], the spectral norm in [5], and the variational norm in [3]. In these works, the bounds for the generalization gap $\mathcal{L}(\boldsymbol{\theta}) - \hat{\mathcal{L}}(\boldsymbol{\theta})$ is derived from a Rademacher complexity bound of the set $\mathcal{F}^Q = \{f(\boldsymbol{x};\boldsymbol{\theta}) : \|\boldsymbol{\theta}\|_{\mathrm{N}} \leq Q\}$, as in Theorem 2.2, where $\|\boldsymbol{\theta}\|_{\mathrm{N}}$ is some norm or value computed from the parameter $\boldsymbol{\theta}$. These estimates are a posteriori estimates. They are shown to be valid once the complexity of $\mathcal{F}^Q$ is controlled.

However, finding a set of functions with small complexity is not enough to explain the generalization of neural networks. The population risk contains two parts—the approximation error and the estimation error. In general, optimal bounds for the approximation error requires the hypothesis space to be large enough and optimal bounds for the estimation error requires the hypothesis space to be small enough. A posteriori estimates only deal with the estimation error. In a priori estimates, both effects are present and we have to strike a balance between them. In this sense, a priori estimates can better reflect the quality of the norm or the hypothesis space selected. Therefore in order to compare our estimates with previous results, we need to turn the previous a posteriori estimates into a priori estimates by establishing approximation error bounds for the other approaches in the same way as we did for ours. These approximation error bounds allow us to translate existing a posteriori estimates into a priori estimates and thereby put previous results on the same footing as ours.

To start with, based on the analysis in Section 2.6, we provide a general framework for establishing a priori estimates from norm-based a posteriori estimates. This framework holds for both residual networks and deep fully-connected networks:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{W}_L \sigma(\boldsymbol{W}_{L-1}\sigma(\cdots\sigma(\boldsymbol{W}_1\boldsymbol{x}))) \tag{3.1}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{m\times d}$, $\boldsymbol{W}_l \in \mathbb{R}^{m\times m}$, $l=2,\ldots,L-1$ and $\boldsymbol{W}_L \in \mathbb{R}^{1\times m}$, and m is the width of the network.

Let $\|\boldsymbol{\theta}\|_{\mathrm{N}}$ be a general norm of the parameters $\boldsymbol{\theta}$, we make the following assumptions about $\|\boldsymbol{\theta}\|_{\mathrm{N}}$.

ASSUMPTION 3.1. *For any set of parameters $\boldsymbol{\theta}$, let $f(\cdot;\boldsymbol{\theta})$ be a neural network associated with $\boldsymbol{\theta}$. Then, there exists a function $\psi(d,L,m)$, such that the Rademacher complexity of the set $\mathcal{F}^Q_{L,m} = \{f(\cdot;\boldsymbol{\theta})\colon f(\cdot;\boldsymbol{\theta})$ has depth L and width m, and $\|\boldsymbol{\theta}\|_{\mathrm{N}} \le Q\}$ can be bounded by*

$$\hat{R}(\mathcal{F}^Q_{L,m}) \le Q\cdot\frac{\psi(d,L,m)}{\sqrt{n}}, \tag{3.2}$$

where d is the dimension of $\boldsymbol{x}$, L and m are the neural network depth and width respectively.

The above Rademacher complexity bound implies the following a posteriori estimate.

THEOREM 3.1 (A posteriori estimate). *Let n be the number of training samples. Consider parameters $\boldsymbol{\theta}$ of a network with depth L and width m. Let $\mathcal{L}(\boldsymbol{\theta})$ and $\hat{\mathcal{L}}(\boldsymbol{\theta})$ be the truncated population risk and empirical risk defined in (2.4). Then for any $\delta\in(0,1)$, with probability at least $1-\delta$ over the random choice of training samples, we have*

$$\left|\mathcal{L}(\boldsymbol{\theta}) - \hat{\mathcal{L}}(\boldsymbol{\theta})\right| \le 2(\|\boldsymbol{\theta}\|_{\mathrm{N}}+1)\frac{2\psi(d,L,m)+1}{\sqrt{n}} + 2\sqrt{\frac{2\log(14/\delta)}{n}}. \tag{3.3}$$

The proof of Theorem 3.1 follows the same way as the proof of Theorem 2.2. With the a posteriori estimate, we obtain an a priori estimate by formulating a regularized problem, and comparing the solution of the regularized problem to a reference solution with good approximation property.

THEOREM 3.2 (A priori estimate). *Under the same conditions as in Theorem 3.1, for* $\lambda \geq 4+2/\psi(d,L,m)$, *assume that* $\hat{\boldsymbol{\theta}}$ *is a minimizer of the regularized model*

$$\min_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}) := \hat{\mathcal{L}}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_{\mathrm{N}} \cdot \frac{\psi(d,L,m)}{\sqrt{n}}, \tag{3.4}$$

Then, for any $\delta \in (0,1)$, *with probability at least* $1-\delta$ *over the random training samples,*

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) \leq \mathcal{L}(\tilde{\boldsymbol{\theta}}) + \left(\|\tilde{\boldsymbol{\theta}}\|_{\mathrm{N}} + 1\right) \frac{(4+\lambda)\psi(d,L,m)+2}{\sqrt{n}} + 4\sqrt{\frac{2\log(28/\delta)}{n}}. \tag{3.5}$$

where $\tilde{\boldsymbol{\theta}}$ *is an arbitrary set of parameters for the same hypothesis space.*

Next, we apply this general framework to the l_1 path norm [14], spectral complexity norm [5] and variational norm [3]. The definitions of the norms are given below.

l_1 path norm. For a residual network defined by (2.1), the l_1 path norm [14] is defined as

$$\|\boldsymbol{\theta}\| = \left\| |\boldsymbol{u}|^T (\boldsymbol{I} + |\boldsymbol{U}_L||\boldsymbol{W}_L|) \cdots (\boldsymbol{I} + |\boldsymbol{U}_1||\boldsymbol{W}_1|)|\boldsymbol{V}| \right\|_1, \tag{3.6}$$

Spectral complexity norm. For a fully-connected network (3.1), the spectral complexity norm proposed in [5] is given by

$$\|\boldsymbol{\theta}\|_{\mathrm{N}} = \left[\prod_{l=1}^{L} \|\boldsymbol{W}_l\|_\sigma\right] \left[\sum_{l=1}^{L} \frac{\|\boldsymbol{W}_l^T\|_{2,1}^{2/3}}{\|\boldsymbol{W}_l\|_\sigma^{2/3}}\right]^{3/2}, \tag{3.7}$$

where $\|\cdot\|_\sigma$ denotes the matrix spectral norm and $\|\cdot\|_{p,q}$ denotes the (p,q) matrix norm $\|\boldsymbol{W}\|_{p,q} = \|(\|\boldsymbol{W}^{:,1}\|_p, \ldots, \|\boldsymbol{W}^{:,m}\|_p)\|_q$.

Variational norm. For a fully-connected network (3.1), the variational norm proposed in [3] is

$$\|\boldsymbol{\theta}\|_{\mathrm{N}} = \frac{1}{L}\sqrt{V} \sum_{l=1}^{L} \sum_{j_l} \sqrt{V_{j_l}^{\mathrm{in}} V_{j_l}^{\mathrm{out}}}, \tag{3.8}$$

where

$$\begin{aligned} V &= \left\| |\boldsymbol{W}_L| \cdots |\boldsymbol{W}_1| \right\|_1, \\ V_{j_l}^{\mathrm{in}} &= \left\| |\boldsymbol{W}_l^{j_l,:}| |\boldsymbol{W}_{l-1}| \cdots |\boldsymbol{W}_1| \right\|_1, \\ V_{j_l}^{\mathrm{out}} &= \left\| |\boldsymbol{W}_L| \cdots |\boldsymbol{W}_{l+1}| |\boldsymbol{W}_l^{:,j_l}| \right\|_1. \end{aligned}$$

When applying Theorem 3.2, for residual networks, we choose $\tilde{\boldsymbol{\theta}}$ to be the solution given by Theorem 2.5, which is the same solution used in our main theorem. For fully-connected networks, we slightly modify the construction of $\tilde{\boldsymbol{\theta}}$ (see the appendix for details), such that the a priori estimates we obtain for different norms all have the same approximation error. But as $\|\tilde{\boldsymbol{\theta}}\|_{\mathrm{N}}$ and ψ vary for different norms, the estimation error comes out differently. To this end, let us recall the expressions of ψ for the norms mentioned above

$$l_1 \text{ path norm}: \qquad \psi(d,L,m) = 2^L \sqrt{2\log 2m},$$

$$\begin{aligned}\text{Spectral norm}: &\quad \psi(d,L,m)=12\log n\sqrt{2\log 2m},\\ \text{Variational norm}: &\quad \psi(d,L,m)=L\log n\sqrt{(L-2)\log m+\log(8ed)}.\end{aligned}$$

On the other hand, one can derive following bounds for $\|\tilde{\boldsymbol{\theta}}\|_{\mathrm{N}}$ (see the appendix for details):

$$\begin{aligned}l_1 \text{ path norm}: &\quad \|\tilde{\boldsymbol{\theta}}\|_{\mathrm{N}}\le 4\|f^*\|_{\mathcal{B}},\\ \text{Spectral norm}: &\quad \|\tilde{\boldsymbol{\theta}}\|_{\mathrm{N}}\le 16(Lm)^{3/2}\|f^*\|_{\mathcal{B}},\\ \text{Variational norm}: &\quad \|\tilde{\boldsymbol{\theta}}\|_{\mathrm{N}}\le 4\sqrt{m}\|f^*\|_{\mathcal{B}}.\end{aligned}$$

Plugging the results above into Theorem 3.2, we get a priori estimates of the regularized model using different norms. The results are summarized in Table 1.1 in Section 1. They are shown in the order of L, m and n, the logarithmic terms are ignored. The notation $\mathcal{O}(\cdot)$ hides constants that depend only on the target function. We see that the weighted path norm is the only one in which the second term in the a priori error bound scales cleanly as $\mathcal{O}(1/\sqrt{n})$, i.e., it is independent of the depth L.

Note that in Table 1.1 the standard l_1 path norm gives an a priori estimate with an exponential dependence on L, different from the case for the weighted path norm. To see why, consider a network $f(\cdot;\boldsymbol{\theta})$ with $\boldsymbol{\theta}=\{\boldsymbol{V},\boldsymbol{W}_l,\boldsymbol{U}_l,\boldsymbol{u}\}$. By the Rademacher complexity bound associated with the weighted path norm (2.14), this function is contained in a set with Rademacher complexity smaller than

$$\frac{C_1}{\sqrt{n}}\big\||\boldsymbol{u}|^T(\boldsymbol{I}+2|\boldsymbol{U}_L||\boldsymbol{W}_L|)\cdots(\boldsymbol{I}+2|\boldsymbol{U}_1||\boldsymbol{W}_1|)|\boldsymbol{V}|\big\|_1. \tag{3.9}$$

On the other hand, if we use the l_1 path norm, this function is contained in a set with Rademacher complexity smaller than

$$\frac{C_2}{\sqrt{n}}\big\||\boldsymbol{u}|^T(2\boldsymbol{I}+2|\boldsymbol{U}_L||\boldsymbol{W}_L|)\cdots(2\boldsymbol{I}+2|\boldsymbol{U}_1||\boldsymbol{W}_1|)|\boldsymbol{V}|\big\|_1, \tag{3.10}$$

where C_1 and C_2 are constants. This gives rise to the exponential dependence. This is not the case in (3.9) as long as the weighted path norm is controlled.

The use of the variational norm eliminates the exponential dependence for the complexity bound, but still retains an algebraic dependence.

The story for the spectral norm is different. It was shown in [5] that the Rademacher complexity of the hypothesis space with bounded spectral norm has an optimal scaling $(1/\sqrt{n})$. However, as the depth of the network goes to infinity, this hypothesis space shrinks to 0 if the bound on the spectral norm is fixed. Therefore, in order to get the desired bound on the approximation error, one has to increase the bound on the spectral norm (the value of Q). This again results in the L dependence in the estimation error.

When deriving the results in Table 1.1, we used a specific construction $\tilde{\boldsymbol{\theta}}$ to control the approximation error. Other constructions may exist. However, they will not change the qualitative dependence of the estimation error, specifically the dependence (or the lack thereof) on L,m in the second term of these bounds, the term that controls the estimation error.

4. Conclusion

By designing proper constraints on the hypothesis space, we have established Monte Carlo-like bounds of the population risk for deep residual networks. This result generalizes the result in [7] for two-layer neural networks. In particular, the error rates established here are dimension-independent. This is the first time that results of this kind

have been achieved. Previous results suffer from either the lack of depth-independent control for the Rademacher complexity, as is the case for the results in [14] and [3], or the fact that the hypothesis space is too small to allow dimension-independent approximation error estimates, as is the case for the results in [5].

The present work still does not explain why vanilla deep residual networks, without regularization, can still perform quite well. This issue of "implicit regularization" still remains quite mysterious.

Appendix A. The full proof in Section 2.6.

A.1. Approximation error. For the approximation error, [7] proved the following result for shallow networks.

THEOREM A.1. *For any target function $f^* \in \mathcal{B}$ and any $M \geq 1$, there exists a two-layer network with width M, such that*

$$\left\| \sum_{j=1}^{M} a_j \sigma(\boldsymbol{b}_j^T \mathbf{x}) - f^*(\mathbf{x}) \right\|^2 \leq \frac{3\|f^*\|_{\mathcal{B}}^2}{M} \tag{A.1}$$

and

$$\sum_{j=1}^{M} |a_j| \|\boldsymbol{b}_j\|_1 \leq 2\|f^*\|_{\mathcal{B}}. \tag{A.2}$$

We have omitted writing out the bias term. This can be accommodated by assuming that the first element of input $\boldsymbol{x}$ is always 1. For residual networks, we prove the approximation result (Theorem 2.5) by splitting the shallow network into several parts and stack them vertically [8]. This is allowed by the special structure of residual networks.

Proof. **(Proof of Theorem 2.5).** We construct a residual network $f(\cdot;\tilde{\boldsymbol{\theta}})$ with input dimension d, depth L, width m, and $D = d+1$ using

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{I}_d & 0 \end{bmatrix}^T, \quad \boldsymbol{u} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T,$$

$$\boldsymbol{W}_l = \begin{bmatrix} \boldsymbol{b}_{(l-1)m+1}^T & 0 \\ \boldsymbol{b}_{(l-1)m+2}^T & 0 \\ \vdots & \vdots \\ \boldsymbol{b}_{lm}^T & 0 \end{bmatrix}, \quad \boldsymbol{U}_l = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \\ a_{(l-1)m+1} & a_{(l-1)m+2} & \cdots & a_{lm} \end{bmatrix}$$

for $l = 1, \ldots, L$. Then it is easy to verify that $f(\boldsymbol{x};\tilde{\boldsymbol{\theta}}) = \sum_{j=1}^{Lm} a_j \sigma(\boldsymbol{b}_j^T \boldsymbol{x})$, and

$$\|\tilde{\boldsymbol{\theta}}\|_{\mathrm{P}} = 2\sum_{j=1}^{Lm} |a_j| \|\boldsymbol{b}_j\|_1 \leq 4\|f^*\|_{\mathcal{B}}.$$

∎

A.2. Rademacher complexity. We use the method of induction to bound the Rademacher complexity of residual networks. We first extend the definition of weighted path norm to hidden neurons in the residual network.

DEFINITION A.1. *Given a residual network defined by (2.1), recall the definition of $\boldsymbol{g}_l$,*

$$\boldsymbol{g}_l(\boldsymbol{x}) = \sigma(\boldsymbol{W}_l \boldsymbol{h}_{l-1}), \quad l = 1, \ldots, L. \tag{A.3}$$

Let g_l^i be the i-th element of $\boldsymbol{g}_l$, and define the weighted path norm

$$\|g_l^i\|_{\mathrm{P}} = \left\| 2|\boldsymbol{W}_l^{i,:}|(\boldsymbol{I}+2|\boldsymbol{U}_{l-1}||\boldsymbol{W}_{l-1}|)\cdots(\boldsymbol{I}+2|\boldsymbol{U}_1||\boldsymbol{W}_1|)2|\boldsymbol{V}| \right\|_1, \tag{A.4}$$

where $\boldsymbol{W}_l^{i,:}$ is the i-th row of $\boldsymbol{W}_l$.

The following lemma establishes the relationship between $\|f\|_{\mathrm{P}}$ and $\|g_l^i\|_{\mathrm{P}}$. Lemma A.2 gives properties of the corresponding function class.

LEMMA A.1. *For the weighted path norm defined in (2.10) and (A.4), we have*

$$\|f\|_{\mathrm{P}} = \sum_{l=1}^{L}\sum_{j=1}^{m}\left(|\boldsymbol{u}|^T|\boldsymbol{U}_l^{:,j}|\right)\|g_l^j\|_{\mathrm{P}} + 2\left\||\boldsymbol{u}|^T|\boldsymbol{V}|\right\|_1, \tag{A.5}$$

and

$$\|g_l^i\|_{\mathrm{P}} = \sum_{k=1}^{l-1}\sum_{j=1}^{m} 2\left(|\boldsymbol{W}_l^{i,:}||\boldsymbol{U}_k^{:,j}|\right)\|g_k^j\|_{\mathrm{P}} + 4\left\||\boldsymbol{W}_l^{i,:}||\boldsymbol{V}|\right\|_1, \tag{A.6}$$

where $\boldsymbol{U}_l^{:,j}$ is the j-th column of $\boldsymbol{U}_l$.

Proof. Recall the definition of $\|f\|_{\mathrm{P}}$, we have

$$\begin{aligned}
\|f\|_{\mathrm{P}} &= \left\||\boldsymbol{u}|^T(\boldsymbol{I}+2|\boldsymbol{U}_L||\boldsymbol{W}_L|)\cdots(\boldsymbol{I}+2|\boldsymbol{U}_1||\boldsymbol{W}_1|)2|\boldsymbol{V}|\right\|_1 \\
&= \left\|\sum_{l=1}^{L}|\boldsymbol{u}|^T|\boldsymbol{U}_l|\cdot 2|\boldsymbol{W}_l|\prod_{j=1}^{l-1}(\boldsymbol{I}+2|\boldsymbol{U}_{l-j}||\boldsymbol{W}_{l-j}|)|\boldsymbol{V}| + 2|\boldsymbol{u}|^T|\boldsymbol{V}|\right\|_1 \\
&= \sum_{l=1}^{L}\sum_{j=1}^{m}\left(|\boldsymbol{u}|^T|\boldsymbol{U}_l^{:,j}|\right)\|g_l^j\|_{\mathrm{P}} + 2\left\||\boldsymbol{u}|^T|\boldsymbol{V}|\right\|_1,
\end{aligned}$$

which gives (A.5). Similarly we obtain (A.6). ∎

LEMMA A.2. *Let $\mathcal{G}_l^Q = \{g_l^i : \|g_l^i\|_{\mathrm{P}} \le Q\}$. Then*

(1) $\mathcal{G}_k^Q \subseteq \mathcal{G}_l^Q$ for $k \le l$;

(2) $\mathcal{G}_l^q \subseteq \mathcal{G}_l^Q$ and $\mathcal{G}_l^q = \frac{q}{Q}\mathcal{G}_l^Q$ for $q \le Q$.

Proof. For any $g_k \in \mathcal{G}_k^Q$, let $\boldsymbol{V}$, $\{\boldsymbol{U}_j, \boldsymbol{W}_j\}_{j=1}^k$ and $\boldsymbol{w}$ be the parameters of g_k, where $\boldsymbol{w}$ is the vector of the parameters in the output layer (the $\boldsymbol{W}_k^{i,:}$ in the definition of g_l^i). Then, for any $l \ge k$, consider g_l generated by parameters $\boldsymbol{V}$, $\{\boldsymbol{U}_j, \boldsymbol{W}_j\}_{j=1}^l$ and $\boldsymbol{w}$, with $\boldsymbol{U}_j = 0$ and $\boldsymbol{W}_j = 0$ for any $k < j \le l$. Now it is easy to verify that $g_l = g_k$ and $\|g_l\|_{\mathrm{P}} = \|g_k\|_{\mathrm{P}} \le Q$. Hence, we have $\mathcal{G}_k^Q \subseteq \mathcal{G}_l^Q$.

On the other hand, obviously we have $\mathcal{G}_l^q \subseteq \mathcal{G}_l^Q$ for any $q \le Q$. For any $g_l \in \mathcal{G}_l^q$, define $\tilde{g}_l$ by replacing the output parameters $\boldsymbol{w}$ by $\frac{Q}{q}\boldsymbol{w}$, then we have $\|\tilde{g}_l\|_{\mathrm{P}} = \frac{Q}{q}\|g_l\|_{\mathrm{P}} \le Q$, and hence $\tilde{g}_l \in \mathcal{G}_l^Q$. Therefore, we have $\frac{Q}{q}\mathcal{G}_l^q \subseteq \mathcal{G}^Q$. Similarly we can obtain $\frac{q}{Q}\mathcal{G}_l^Q \subseteq \mathcal{G}^q$. Consequently, we have $\mathcal{G}_l^q = \frac{q}{Q}\mathcal{G}_l^Q$. ∎

We will also use the following two lemmas about Rademacher complexity [16]. Lemma A.3 bounds the Rademacher complexity of linear functions, and Lemma A.4 gives the contraction property of the Rademacher complexity.

LEMMA A.3. *Let* $\mathcal{H}=\{h(\boldsymbol{x})=\boldsymbol{u}^T\boldsymbol{x}:\|\boldsymbol{u}\|_1\le 1\}$. *Assume that the samples* $\{\boldsymbol{x}_i\}_{i=1}^n\subset\mathbb{R}^d$. *Then*

$$\hat{R}(\mathcal{H})\le\max_i\|\boldsymbol{x}_i\|_\infty\sqrt{\frac{2\log(2d)}{n}}. \tag{A.7}$$

LEMMA A.4. *Assume that* $\phi_i, i=1,\dots,n$ *are Lipschitz continuous functions with uniform Lipschitz constant* L_ϕ, *i.e.,* $|\phi_i(x)-\phi_i(x')|\le L_\phi|x-x'|$ *for* $i=1,\dots,n$. *Then*

$$\mathbb{E}_\xi\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^n\xi_i\phi_i(h(x_i))\right]\le L_\phi\mathbb{E}_\xi\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^n\xi_i h(x_i)\right]. \tag{A.8}$$

With Lemma A.1–A.4, we can come to prove Theorem 2.2.

Proof. **(Proof of Theorem 2.2).** We first estimate the Rademacher complexity of $\mathcal{G}_l^Q$,

$$\hat{R}(\mathcal{G}_l^Q)\le Q\sqrt{\frac{2\log(2d)}{n}}. \tag{A.9}$$

This is done by induction. By definition, $g_1^i(\boldsymbol{x})=\sigma(\boldsymbol{W}_1^{i,:}\boldsymbol{V}\boldsymbol{x})$. Hence, using Lemma A.3 and A.4, we conclude that the statement (A.9) holds for $l=1$. Now, assume that the result holds for $1,2,\dots,l$. Then, for $l+1$ we have

$$\begin{aligned}
n\hat{R}(\mathcal{G}_{l+1}^Q)&=\mathbb{E}_\xi\sup_{g_{l+1}\in\mathcal{G}_{l+1}^Q}\sum_{i=1}^n\xi_i g_{l+1}(\boldsymbol{x}_i)\\
&=\mathbb{E}_\xi\sup_{(1)}\sum_{i=1}^n\xi_i\sigma(\boldsymbol{w}_{l+1}^T(\boldsymbol{U}_l\boldsymbol{g}_l+\boldsymbol{U}_{l-1}\boldsymbol{g}_{l-1}+\cdots+\boldsymbol{U}_1\boldsymbol{g}_1+\boldsymbol{h}_0))\\
&\le\mathbb{E}_\xi\sup_{(1)}\sum_{i=1}^n\xi_i(\boldsymbol{w}_{l+1}^T(\boldsymbol{U}_l\boldsymbol{g}_l+\boldsymbol{U}_{l-1}\boldsymbol{g}_{l-1}+\cdots+\boldsymbol{U}_1\boldsymbol{g}_1+\boldsymbol{h}_0))\\
&=\mathbb{E}_\xi\sup_{(1)}\sum_{i=1}^n\xi_i(\boldsymbol{w}_{l+1}^T(\boldsymbol{U}_l\boldsymbol{g}_l+\boldsymbol{U}_{l-1}\boldsymbol{g}_{l-1}+\cdots+\boldsymbol{U}_1\boldsymbol{g}_1+\boldsymbol{V}\sigma(\boldsymbol{x})-\boldsymbol{V}\sigma(-\boldsymbol{x})))\\
&\le\mathbb{E}_\xi\sup_{(2)}\left\{\sum_{k=1}^l a_k\sup_{g\in\mathcal{G}_k^1}\left|\sum_{i=1}^n\xi_i g(\boldsymbol{x}_i)\right|+2b\sup_{g\in\mathcal{G}_1^1}\left|\sum_{i=1}^n\xi_i g(\boldsymbol{x}_i)\right|\right\}\\
&\le\mathbb{E}_\xi\sup_{\substack{2a+4b\le Q\\ a,b\ge 0}}(a+2b)\sup_{g\in\mathcal{G}_l^1}\left|\sum_{i=1}^n\xi_i g(\boldsymbol{x}_i)\right|\\
&\le\frac{Q}{2}\mathbb{E}_\xi\sup_{g\in\mathcal{G}_l^1}\left|\sum_{i=1}^n\xi_i g(\boldsymbol{x}_i)\right|
\end{aligned}$$

where condition (1) is $\sum_{k=1}^l\sum_{j=1}^m 2\left(|\boldsymbol{w}_{l+1}|^T|U_k^{:,j}|\right)\|g_k^j\|_{\mathrm{P}}+4\left\||\boldsymbol{w}_{l+1}|^T|\boldsymbol{V}|\right\|_1\le Q$, and condition (2) is $2\sum_{k=1}^l a_k+4b\le Q$. The first inequality is due to the contraction lemma, while the third inequality is due to Lemma A.2. On the one hand, we have

$$\mathbb{E}_\xi\sup_{\|\boldsymbol{u}\|_1\le 1}\left|\sum_{i=1}^n\xi_i\boldsymbol{u}^T\boldsymbol{x}_i\right|=\mathbb{E}_\xi\sup_{\|\boldsymbol{u}\|_1\le 1}\sum_{i=1}^n\xi_i\boldsymbol{u}^T\boldsymbol{x}_i\le n\sqrt{\frac{2\log(2d)}{n}}.$$

On the other hand, since $0 \in \mathcal{G}_l^1$, for any $\{\xi_1, \ldots, \xi_n\}$, we have

$$\sup_{g \in \mathcal{G}_l^1} \sum_{i=1}^n \xi_i g(\boldsymbol{x}_i) \geq 0.$$

Hence, we have

$$\begin{aligned}\sup_{g \in \mathcal{G}_l^1} \left| \sum_{i=1}^n \xi_i g(\boldsymbol{x}_i) \right| &\leq \max \left\{ \sup_{g \in \mathcal{G}_l^1} \sum_{i=1}^n \xi_i g(\boldsymbol{x}_i), \ \sup_{g \in \mathcal{G}_l^1} \sum_{i=1}^n -\xi_i g(\boldsymbol{x}_i) \right\} \\ &\leq \sup_{g \in \mathcal{G}_l^1} \sum_{i=1}^n \xi_i g(\boldsymbol{x}_i) + \sup_{g \in \mathcal{G}_l^1} \sum_{i=1}^n -\xi_i g(\boldsymbol{x}_i),\end{aligned}$$

which gives

$$\mathbb{E}_\xi \sup_{g \in \mathcal{G}_l^1} \left| \sum_{i=1}^n \xi_i g(\boldsymbol{x}_i) \right| \leq 2 \mathbb{E}_\xi \sup_{g \in \mathcal{G}_l^1} \sum_{i=1}^n \xi_i g(\boldsymbol{x}_i) = 2n \hat{R}(\mathcal{G}_l^1).$$

Therefore, we have

$$\hat{R}(\mathcal{G}_{l+1}^Q) \leq \frac{Q}{2} 2 \sqrt{\frac{2 \log(2d)}{n}} \leq Q \sqrt{\frac{2 \log(2d)}{n}}.$$

Similarly, based on the control for the Rademacher complexity of $\mathcal{G}_1^Q, \ldots, \mathcal{G}_L^Q$, we get

$$\hat{R}(\mathcal{F}^Q) \leq 2Q \sqrt{\frac{2 \log(2d)}{n}}.$$

∎

A.3. A posteriori estimates.

Proof. **(Proof of Corollary 2.2).** Let $\mathcal{H} = \{\ell(\cdot; \boldsymbol{\theta}) : \|\boldsymbol{\theta}\|_{\mathrm{P}} \leq Q\}$. Notice that for all $\boldsymbol{x}$,

$$|\ell(\boldsymbol{x}; \boldsymbol{\theta}) - \ell(\boldsymbol{x}; \boldsymbol{\theta}')| \leq 2 |f(\boldsymbol{x}; \boldsymbol{\theta}) - f(\boldsymbol{x}; \boldsymbol{\theta}')|.$$

By Lemma A.4,

$$\hat{R}(\mathcal{H}) = \frac{1}{n} \mathbb{E}_\xi \left[\sup_{\|\boldsymbol{\theta}\|_{\mathrm{P}} \leq Q} \sum_{i=1}^n \xi_i \ell(\boldsymbol{x}_i; \boldsymbol{\theta}) \right] \leq \frac{2}{n} \mathbb{E}_\xi \left[\sup_{\|\boldsymbol{\theta}\|_{\mathrm{P}} \leq Q} \sum_{i=1}^n \xi_i f(\boldsymbol{x}_i; \boldsymbol{\theta}) \right] = 2 \hat{R}(\mathcal{F}^Q).$$

From Theorem 2.1, with probability at least $1 - \delta$,

$$\begin{aligned}\sup_{\|\boldsymbol{\theta}\|_{\mathrm{P}} \leq Q} \left| \mathcal{L}(\boldsymbol{\theta}) - \hat{\mathcal{L}}(\boldsymbol{\theta}) \right| &\leq 2 \hat{R}(\mathcal{H}) + 2 \sup_{h, h' \in \mathcal{H}} \|h - h'\|_\infty \sqrt{\frac{2 \log(8/\delta)}{n}} \\ &\leq 4Q \sqrt{\frac{2 \log(2d)}{n}} + 2 \sqrt{\frac{2 \log(8/\delta)}{n}}. \qquad \text{(A.10)}\end{aligned}$$

Now take $Q = 1, 2, 3, \ldots$ and $\delta_Q = \frac{6\delta}{(\pi Q)^2}$, then with probability at least $1 - \sum_{Q=1}^\infty \delta_Q = 1 - \delta$, the bound

$$\sup_{\|\boldsymbol{\theta}\|_{\mathrm{P}} \leq Q} \left| \mathcal{L}(\boldsymbol{\theta}) - \hat{\mathcal{L}}(\boldsymbol{\theta}) \right| \leq 4Q \sqrt{\frac{2 \log(2d)}{n}} + 2 \sqrt{\frac{2}{n} \log \frac{4(\pi Q)^2}{3\delta}}$$

holds for all $Q \in \mathbb{N}^*$. In particular, for given $\boldsymbol{\theta}$, the inequality holds for $Q = \lceil \|\boldsymbol{\theta}\| \rceil < \|\boldsymbol{\theta}\|_{\mathrm{P}} + 1$, thus

$$\begin{aligned}\left|\mathcal{L}(\boldsymbol{\theta}) - \hat{\mathcal{L}}(\boldsymbol{\theta})\right| &\leq 4(\|\boldsymbol{\theta}\|_{\mathrm{P}}+1)\sqrt{\frac{2\log(2d)}{n}} + 2\sqrt{\frac{2}{n}\log\frac{14(\|\boldsymbol{\theta}\|_{\mathrm{P}}+1)^2}{\delta}} \\ &\leq 4(\|\boldsymbol{\theta}\|_{\mathrm{P}}+1)\sqrt{\frac{2\log(2d)}{n}} + 2\left[\frac{\|\boldsymbol{\theta}\|_{\mathrm{P}}+1}{\sqrt{n}} + \sqrt{\frac{2\log(14/\delta)}{n}}\right] \\ &= 2(\|\boldsymbol{\theta}\|_{\mathrm{P}}+1)\frac{2\sqrt{2\log(2d)}+1}{\sqrt{n}} + 2\sqrt{\frac{2\log(14/\delta)}{n}}.\end{aligned}$$

∎

A.4. A priori estimates. Now we are ready to prove the main Theorem 2.3.

Proof. **(Proof of Theorem 2.3).** Let $\hat{\boldsymbol{\theta}}$ be the optimal solution of the regularized model (2.16), and $\tilde{\boldsymbol{\theta}}$ be the approximation in Theorem 2.5. Consider

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) = \mathcal{L}(\tilde{\boldsymbol{\theta}}) + \left[\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{J}(\hat{\boldsymbol{\theta}})\right] + \left[\mathcal{J}(\hat{\boldsymbol{\theta}}) - \mathcal{J}(\tilde{\boldsymbol{\theta}})\right] + \left[\mathcal{J}(\tilde{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}})\right]. \tag{A.11}$$

From (2.22) in Theorem 2.5, we have

$$\mathcal{L}(\tilde{\boldsymbol{\theta}}) \leq \frac{3\|f^*\|_{\mathcal{B}}^2}{Lm}. \tag{A.12}$$

Compare the definition of $\mathcal{J}$ in (2.16) and the gap $\mathcal{L} - \hat{\mathcal{L}}$ in (2.15), with probability at least $1 - \delta/2$,

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{J}(\hat{\boldsymbol{\theta}}) &\leq \left(\|\hat{\boldsymbol{\theta}}\|_{\mathrm{P}}+1\right)\frac{3(4-\lambda)\sqrt{2\log(2d)}+2}{\sqrt{n}} + 3\lambda\sqrt{\frac{2\log(2d)}{n}} + 2\sqrt{\frac{2\log(14/\delta)}{n}} \\ &\leq 3\lambda\sqrt{\frac{2\log(2d)}{n}} + 2\sqrt{\frac{2\log(14/\delta)}{n}}\end{aligned} \tag{A.13}$$

since $\lambda \geq 4 + 2/[3\sqrt{2\log(2d)}]$; with probability at least $1 - \delta/2$, we have

$$\mathcal{J}(\tilde{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}}) \leq \left(\|\tilde{\boldsymbol{\theta}}\|_{\mathrm{P}}+1\right)\frac{3(4+\lambda)\sqrt{2\log(2d)}+2}{\sqrt{n}} - 3\lambda\sqrt{\frac{2\log(2d)}{n}} + 2\sqrt{\frac{2\log(14/\delta)}{n}} \tag{A.14}$$

Thus with probability at least $1 - \delta$, (A.13) and (A.14) hold simultaneously. In addition, we have

$$\mathcal{J}(\hat{\boldsymbol{\theta}}) - \mathcal{J}(\tilde{\boldsymbol{\theta}}) \leq 0 \tag{A.15}$$

since $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta})$.

Now plugging (A.12–A.15) into (A.11), and noticing that $\|\tilde{\boldsymbol{\theta}}\|_{\mathrm{P}} \leq 4\|f^*\|_{\mathcal{B}}$ from Theorem 2.5, we see that the main theorem (2.17) holds with probability at least $1 - \delta$.

∎

Finally, we deal with the case with noise and prove Theorem 2.4. For problems with noise, we decompose $\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}})$ as

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}}) = \left[\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}_B(\hat{\boldsymbol{\theta}})\right] + \left[\mathcal{L}_B(\hat{\boldsymbol{\theta}}) - \mathcal{J}_B(\hat{\boldsymbol{\theta}})\right] + \left[\mathcal{J}_B(\hat{\boldsymbol{\theta}}) - \mathcal{J}_B(\tilde{\boldsymbol{\theta}})\right]$$

$$+\left[\mathcal{J}_B(\tilde{\boldsymbol{\theta}})-\mathcal{L}_B(\tilde{\boldsymbol{\theta}})\right]+\left[\mathcal{L}_B(\tilde{\boldsymbol{\theta}})-\mathcal{L}(\tilde{\boldsymbol{\theta}})\right]. \tag{A.16}$$

Based on the results we had for the case without noise, in (A.16) we only have to estimate the first and the last terms. This is given by Lemma 2.1. Finally, we prove Lemma 2.1.

Proof. **(Proof of Lemma 2.1).** Let $Z=f(\boldsymbol{x};\boldsymbol{\theta})-f^*(\boldsymbol{x})-\varepsilon$. Then we have

$$\begin{aligned}|\mathcal{L}(\boldsymbol{\theta})-\mathcal{L}_B(\boldsymbol{\theta})| &= \mathbb{E}\left[(Z^2-B^2)\mathbf{1}_{|Z|\geq B}\right]\\ &= \int_0^\infty \Pr\left\{Z^2-B^2\geq t^2\right\}dt^2\\ &= \int_0^\infty \Pr\left\{|Z|\geq\sqrt{B^2+t^2}\right\}dt^2.\end{aligned}$$

As $0\leq f(\boldsymbol{x};\boldsymbol{\theta})\leq 1$ and $0\leq f^*(\boldsymbol{x};\boldsymbol{\theta})\leq 1$, we have

$$\int_0^\infty \Pr\left\{|Z|\geq\sqrt{B^2+t^2}\right\}dt^2\leq\int_0^\infty \Pr\left\{|\varepsilon|\geq\sqrt{B^2+t^2}-1\right\}dt^2.$$

Let $s=\sqrt{B^2+t^2}$, then

$$\begin{aligned}&\int_0^\infty \Pr\left\{|\varepsilon|\geq\sqrt{B^2+t^2}-1\right\}dt^2\leq\int_B^\infty ce^{-\frac{(s-1)^2}{2\sigma^2}}ds^2\\ &=\int_{B-1}^\infty 2ce^{-\frac{s^2}{2\sigma^2}}ds^2+\int_{B-1}^\infty 4ce^{-\frac{s^2}{2\sigma^2}}ds\\ &\leq 4c\sigma^2e^{-\frac{(B-1)^2}{2\sigma^2}}+\sqrt{\frac{2}{\pi}}ce^{-\frac{(B-1)^2}{2\sigma^2}}\\ &\leq\frac{c(4\sigma^2+1)}{\sqrt{n}}.\end{aligned}$$

∎

Appendix B. The missing details in Section 3.

B.1. Approximation properties of deep fully-connected networks. Consider a deep fully-connected network with depth L and width m (3.1) in the form:

$$f(\boldsymbol{x};\boldsymbol{\theta})=\boldsymbol{W}_L\sigma(\boldsymbol{W}_{L-1}\sigma(\cdots\sigma(\boldsymbol{W}_1\boldsymbol{x})))$$

where $\boldsymbol{W}_1\in\mathbb{R}^{m\times d}$, $\boldsymbol{W}_l\in\mathbb{R}^{m\times m}$, $l=2,\ldots,L-1$ and $\boldsymbol{W}_L\in\mathbb{R}^{1\times m}$. Taking the same approach as in Theorem 2.5 and [8], we construct the deep fully-connected network from a two-layer network. From Theorem A.1, there exists a two-layer network with width M, such that

$$\left\|\sum_{j=1}^M a_j\sigma(\boldsymbol{b}_j^T\mathbf{x})-f^*(\mathbf{x})\right\|^2\leq\frac{3\|f^*\|_{\mathcal{B}}^2}{M}$$

and

$$\sum_{j=1}^M|a_j|\|\boldsymbol{b}_j\|_1\leq 4\|f^*\|_{\mathcal{B}}.$$

Since the ReLU activation $\sigma(\cdot)$ is positively homogeneous, we can assume without loss of generality that $a_1 = a_2 = \cdots = a_M = a \le 4\|f^*\|_{\mathcal{B}}$ and $\|\boldsymbol{b}_1\|_1 + \|\boldsymbol{b}_2\|_1 + \cdots + \|\boldsymbol{b}_M\|_1 = 1$. Now let $M = (m-d)(L-1)$, and rewrite the subscripts as $\boldsymbol{b}_{l,j} = \boldsymbol{b}_{(m-d)(l-1)+j}$, $l = 1,\ldots,L-1$, $j = 1,\ldots,m-d$. Define a fully-connected network $f(\cdot;\tilde{\boldsymbol{\theta}})$ by

$$\boldsymbol{W}_1 = \begin{bmatrix} \boldsymbol{I}_d \\ \boldsymbol{b}_{1,1}^T \\ \vdots \\ \boldsymbol{b}_{1,m-d}^T \end{bmatrix}, \quad \boldsymbol{W}_l = \begin{bmatrix} \boldsymbol{I}_d & 0 \\ \boldsymbol{b}_{l,1}^T & \\ \vdots & \boldsymbol{I}_{m-d} \\ \boldsymbol{b}_{l,m-d}^T & \end{bmatrix}, \ l = 2,\ldots,L-1,$$
$$\boldsymbol{W}_L = \begin{bmatrix} 0 & 0 & \cdots & 0 & a & a & \cdots & a \end{bmatrix},$$

then it is easy to verify that $f(\boldsymbol{x};\tilde{\boldsymbol{\theta}}) = a\sum_{j=1}^{M} \sigma(\boldsymbol{b}_j^T \boldsymbol{x})$. This ensures that the approximation property of fully-connected multi-layer neural network is at least as good as the two-layer network.

B.2. Calculation of the spectral complexity norm. Recall the spectral complexity norm (3.7) proposed in [5]

$$\|\boldsymbol{\theta}\|_{\mathrm{N}} = \left[\prod_{l=1}^{L} \|\boldsymbol{W}_l\|_\sigma\right] \left[\sum_{l=1}^{L} \frac{\|\boldsymbol{W}_l^T\|_{2,1}^{2/3}}{\|\boldsymbol{W}_l\|_\sigma^{2/3}}\right]^{3/2}.$$

For $l = 1,\ldots,L-1$, the matrix spectral norm satisfies $\|\boldsymbol{W}_l\|_\sigma \ge 1$, and

$$\|\boldsymbol{W}_l\|_\sigma - 1 \le \|\boldsymbol{W}_l - \boldsymbol{I}\|_\sigma \le \|\boldsymbol{W}_l - \boldsymbol{I}\|_F = \left[\sum_{j=1}^{m-d} \|\boldsymbol{b}_{l,j}\|_2^2\right]^{1/2} \le \sum_{j=1}^{m-d} \|\boldsymbol{b}_{l,j}\|_1,$$

thus

$$\prod_{l=1}^{L-1} \|\boldsymbol{W}_l\|_\sigma \le \prod_{l=1}^{L-1} \left[1 + \sum_{j=1}^{m-d} \|\boldsymbol{b}_{l,j}\|_1\right] < e$$

since $\sum_{l=1}^{L-1}\sum_{j=1}^{m-d} \|\boldsymbol{b}_{l,j}\|_1 = 1$. The $(p,q) = (2,1)$ matrix norm satisfies

$$\|\boldsymbol{W}_l^T\|_{2,1} = \left\|(\|\boldsymbol{W}_l^{1,:}\|_2, \ldots, \|\boldsymbol{W}_l^{:,m}\|_2)\right\|_1 = d + \sum_{j=1}^{m-d} \sqrt{1 + \|\boldsymbol{b}_{l,j}\|_2^2} < \sqrt{2}m.$$

In addition,

$$\|\boldsymbol{W}_L\|_\sigma = \|\boldsymbol{W}_L\|_{2,1} = \|\boldsymbol{W}_L\|_2 = a\sqrt{m-d} \le 4\|f^*\|_{\mathcal{B}}\sqrt{m-d}.$$

Therefore, the spectral complexity norm satifies

$$\|\tilde{\boldsymbol{\theta}}\|_{\mathrm{N}} \le e \cdot 4\|f^*\|_{\mathcal{B}}\sqrt{m-d} \cdot L^{3/2} \cdot \sqrt{2}m \le 16(Lm)^{3/2}\|f^*\|_{\mathcal{B}}.$$

B.3. Calculation of the the variational norm. Recall the variational norm (3.8) proposed in [3]

$$\|\boldsymbol{\theta}\|_{\mathrm{N}} = \frac{1}{L}\sqrt{V}\sum_{l=1}^{L}\sum_{j_l} \sqrt{V_{j_l}^{\mathrm{in}} V_{j_l}^{\mathrm{out}}},$$

where

$$V = \big\| |\boldsymbol{W}_L| \cdots |\boldsymbol{W}_1| \big\|_1,$$
$$V_{j_l}^{\text{in}} = \big\| |\boldsymbol{W}_l^{j_l,:}| |\boldsymbol{W}_{l-1}| \cdots |\boldsymbol{W}_1| \big\|_1,$$
$$V_{j_l}^{\text{out}} = \big\| |\boldsymbol{W}_L| \cdots |\boldsymbol{W}_{l+1}| |\boldsymbol{W}_l^{:,j_l}| \big\|_1.$$

Notice that for any l,

$$\sum_{j_l=1}^{m} V_{j_l}^{\text{in}} V_{j_l}^{\text{out}} = V.$$

Therefore

$$\|\boldsymbol{\theta}\|_{\text{N}} \le \frac{1}{L} \sqrt{V} \cdot L \cdot \sqrt{mV} = \sqrt{m} V.$$

Now it is easy to verify that

$$V = a \sum_{l=1}^{L-1} \sum_{j=1}^{m-d} \|b_{l,j}\|_1 = a \le 4 \|f^*\|_{\mathcal{B}}.$$

REFERENCES

[1] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, *Stronger generalization bounds for deep nets via a compression approach*, Proceedings of the Thirty-Fifth International Conference on Machine Learning, Stockholm, Sweden, 80, 2018. 1, 1

[2] A.R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inf. theory, 39(3):930–945, 1993. 2

[3] A.R. Barron and J.M. Klusowski, *Approximation and estimation for high-dimensional deep learning networks*, arXiv preprint, arXiv:1809.03090, 2018. 1, 3, 3, 3, 4, B.3

[4] P.L. Bartlett, O. Bousquet, and S. Mendelson, *Local Rademacher complexities*, Ann. Stat., 33(4):1497–1537, 2005. 1

[5] P.L. Bartlett, D.J. Foster, and M.J. Telgarsky, *Spectrally-normalized margin bounds for neural networks*, Adv. Neural Inf. Process. Syst., 6241–6250, 2017. 1, 2, 3, 3, 3, 3, 4, B.2

[6] W. E, C. Ma, and L. Wu, *Barron spaces and the compositional function spaces for neural network models*, arXiv preprint, arXiv:1906.08039, 2019. 2.2

[7] W. E, C. Ma, and L. Wu, *A priori estimates of the population risk for two-layer neural networks*, Commun. Math. Sci., 17(5):1407–1425, 2019. 1, 2.2, 4, A.1

[8] W. E. and Q. Wang, *Exponential convergence of the deep neural network approximation for analytic functions*, Sci. China Math., 61(10):1733–1740, 2018. A.1, B.1

[9] N. Golowich, A. Rakhlin, and O. Shamir, *Size-independent sample complexity of neural networks*, Inf. Inference, 9(2):473–504, 2017. 1, 1

[10] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778, 2016. 1

[11] J.M. Klusowski and A.R. Barron, *Risk bounds for high-dimensional ridge function combinations including neural networks*, arXiv preprint, arXiv:1607.01434, 2016. 1, 2.2

[12] T. Liang, T. Poggio, A. Rakhlin, and J. Stokes, *Fisher-Rao metric, geometry, and complexity of neural networks*, arXiv preprint, arXiv:1711.01530, 2017. 1, 1

[13] B. Neyshabur, R.R. Salakhutdinov, and N. Srebro, *Path-SGD: Path-normalized optimization in deep neural networks*, Adv. Neural Inf. Process. Syst., 2422–2430, 2015. 1, 2.2

[14] B. Neyshabur, R. Tomioka, and N. Srebro, *Norm-based capacity control in neural networks*, Conference on Learning Theory, 1376–1401, 2015. 1, 3, 3, 4

[15] A. Pinkus, *Approximation theory of the MLP model in neural networks*, Acta Numer., 8:143–195, 1999. 1

[16] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014. 2.3, A.2

[17] A. Veit, M.J. Wilber, and S. Belongie, *Residual networks behave like ensembles of relatively shallow networks*, Adv. Neural Inf. Process. Syst., 550–558, 2016. 2.2

[18] S. Zheng, Q. Meng, H. Zhang, W. Chen, N. Yu, and T.-Y. Liu, *Capacity control of ReLU neural networks by basis-path norm*, arXiv preprint, arXiv:1809.07122, 2018. 2.2