COMMUN. MATH. SCI. © 2019 International Press
Vol. 17, No. 5, pp. 1407–1425

A PRIORI ESTIMATES OF THE POPULATION RISK FOR TWO-LAYER NEURAL NETWORKS*

WEINAN E[†], CHAO MA[‡], AND LEI WU[§]

In memory of Professor David Shenou Cai

Abstract. New estimates for the population risk are established for two-layer neural networks. These estimates are nearly optimal in the sense that the error rates scale in the same way as the Monte Carlo error rates. They are equally effective in the over-parametrized regime when the network size is much larger than the size of the dataset. These new estimates are a priori in nature in the sense that the bounds depend only on some norms of the underlying functions to be fitted, not the parameters in the model, in contrast with most existing results which are a posteriori in nature. Using these a priori estimates, we provide a perspective for understanding why two-layer neural networks perform better than the related kernel methods.

Keywords. Two-layer neural network; Barron space; Population risk; A priori estimate; Rademacher complexity.

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

1. Introduction

One of the main challenges in theoretical machine learning is to understand the errors in neural network models [43]. To this end, it is useful to draw an analogy with classical approximation theory and finite element analysis [13]. There are two kinds of error bounds in finite element analysis depending on whether the target solution (the ground truth) or the numerical solution enters into the bounds. Let f^* and $\hat{f}_n$ be the true solution and the "numerical solution", respectively. "A priori" error estimates usually take the form

$$\|\hat{f}_n - f^*\|_1 \leq Cn^{-\alpha}\|f^*\|_2.$$

where only norms of the true solution enter into the bounds. In "a posteriori" error estimates, the norms of the numerical solution enter into the bounds:

$$\|\hat{f}_n - f^*\|_1 \leq Cn^{-\beta}\|\hat{f}_n\|_3.$$

Here $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_3$ denote various norms. In this language, most recent theoretical results [7,24,32–35] on estimating the generalization error of neural networks should be viewed as "a posteriori" analysis, since the bounds depend on various norms of the neural network model obtained after the training process. As was observed in [4,18,34], the numerical values of these norms are very large, yielding vacuous bounds. For example, [34] calculated the values of various a posteriori bounds for some real two-layer neural networks and it is found that the best bounds are still on the order of $O(10^5)$.

In this paper, we pursue a different line of attack by providing "a priori" analysis. Specifically, we focus on two-layer networks, and we consider models with explicit

[*] Received: April 28, 2019; Accepted (in revised form): July 25, 2019.

[†] Department of Mathematics and Program in Applied and Computational Mathematics, Princeton University, Princeton NJ 08540, USA; Beijing Institute of Big Data Research, Beijing 100871, China (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 (leiwu@princeton.edu).

regularization. We establish estimates for the population risk which are asymptotically sharp with constants depending only on the properties of the target function. Our numerical results suggest that such regularization terms are necessary in order for the model to be “well-posed” (see Section 7 for the precise meaning).

Specifically, our main contributions are:

- We establish a priori estimates of the population risk for learning two-layer neural networks with an explicit regularization. These a priori estimates depend on the Barron norm of the target function. The rates with respect to the number of parameters and number of samples are comparable to the Monte Carlo rate. In addition, our estimates hold for high dimensional and over-parametrized regime.
- We make a comparison between the neural network and kernel methods using these a priori estimates. We show that two-layer neural networks can be understood as kernel methods with the kernel adaptively selected from the data. This understanding partially explains why neural networks perform better than kernel methods in practice.

The present paper is the first in a series of papers in which we analyze neural network models using a classical numerical analysis perspective. Subsequent papers will consider deep neural network models [19, 20], the optimization and implicit regularization problem using gradient descent dynamics [20, 22] and the general function spaces and approximation theory in high dimensions [21].

2. Related work

There are two key problems in learning two-layer neural networks: optimization and generalization. Recent progresses on optimization suggest that over-parametrization is the key factor leading to a nice empirical landscape $\hat{L}_n$ [23, 36, 38], thus facilitating convergence towards global minima of $\hat{L}_n$ for gradient-based optimizers [12, 17, 31]. This leaves the generalization property of learning two-layer neural networks more puzzling, since naive arguments would suggest that more parameters implies worse generalization ability. This contradicts what is observed in practice. In what follows, we survey previous attempts in analyzing the generalization properties of two-layer neural network models.

	rate	over-parametrization
rate of [6]	$\frac{1}{m}+\frac{md\ln(n)}{n}$	No
rate of [27]	$\left(\frac{\ln d}{n}\right)^{1/3}$	No
our rate	$\frac{1}{m}+\ln(n)(\frac{\ln d}{n})^{1/2}$	Yes

Table 2.1. *Comparison of the theoretical bounds. The second column are the bounds and the third column indicates whether the bounds are relevant in the over-parametrized regime, i.e. $m \geq n$.*

2.1. Explicit regularization.

This line of work studies the generalization property of two-layer neural networks with explicit regularization and our work lies in this category. Let n,m denote the number of samples and number of parameters, respectively. For two-layer sigmoidal networks, [6] established a risk bound $O(1/m+md\ln(n)/n)$. By considering smoother activation functions, [27] proved another bound $O((\ln d/n)^{1/3})$ for the case when $m\approx\sqrt{n}$. Both of these results are proved for a regularized estimator. In comparison, the error rate established in this paper,

$O(1/m + \ln n\sqrt{\ln d/n})$ is sharper and in fact nearly optimal, and it is also applicable for the over-parametrized regime. For a better comparison, please refer to Table 2.1.

More recently, [41] considered explicit regularization for classification problems. They proved that for the specific cross-entropy loss, the regularization path converges to the maximum margin solutions. They also proved an a priori bound on how the network size affects the margin. However, their analysis is restricted to the case where the data is well-separated. Our result does not have this restriction.

2.2. Implicit regularization. Another line of work studies how gradient descent (GD) and stochastic gradient descent (SGD) find the generalizable solutions. [9] proved that SGD learns over-parametrized networks that provably generalize for binary classification problem. However, it is not clear how the population risk depends on the number of samples for their compression-based generalization bound. Moreover, their proof highly relies on the strong assumption that the data is linearly separable. The experiments in [34] suggest that increasing the network width can improve the test accuracy of solutions found by SGD. They tried to explain this phenomena by an initialization-dependent (a posterior) generalization bound. However, in their experiments, the largest width $m \approx n$, rather than $m \gg n$. Furthermore their generalization bounds are arbitrarily loose in practice. So their result cannot tell us whether GD can find generalizable solutions for arbitrarily wide networks.

In [15] and [1], it is proved that GD with a particularly chosen initialization, learning rate and early stopping can find generalizable solutions θ_T such that $L(\theta_T) \le \min_\theta L(\theta) + \varepsilon$, as long as $m \ge \text{poly}(n, \frac{1}{\varepsilon})$. These results differ from ours in several aspects. First, both of them assume that the target function $f^* \in \mathcal{H}_{\pi_0}$, where π_0 is the uniform distribution over S^d. Recall that $\mathcal{H}_{\pi_0}$ is the reproducing kernel Hilbert space (RKHS) induced by $k_{\pi_0}(x,x') = \mathbb{E}_{w\sim\pi_0}[\sigma(\langle w,x\rangle)\sigma(\langle w,x'\rangle)]$, which is much smaller than $\mathcal{B}_2(X)$, the space we consider. Secondly, through carefully analyzing the polynomial order in two papers, we can see that the sample complexities they provided scales as $O(1/n^{1/4})$, which is worse than $O(1/\sqrt{n})$ proved here. See also [3, 10] for some even more recent results.

Recent work in [20, 22] has shown clearly that for the kind of initialization schemes considered in these previous works or in the over-parametrized regime, the neural network models do not perform better than the corresponding kernel method with a kernel defined by the initialization. These results do not rule out the possibility that neural network models can still outperform kernel methods in some regimes, but they do show that finding these regimes is quite non-trivial.

3. Preliminaries

We begin by recalling the basics of two-layer neural networks and their approximation properties.

The problem of interest is to learn a function from a training set of n examples $S = \{(x_i, y_i)\}_{i=1}^n$, i.i.d. samples drawn from an underlying distribution $\rho_{x,y}$, which is assumed fixed but known only through the samples. Our target function is $f^*(x) = \mathbb{E}[y|x]$. We assume that the values of y_i are given through the decomposition $y = f^*(x) + \xi$, where ξ denotes the noise. For simplicity, we assume that the data lie in $X = [-1,1]^d$ and $0 \le f^* \le 1$.

The two-layer neural network is defined by

$$f(x;\theta) = \sum_{k=1}^m a_k \sigma(w_k^T x), \tag{3.1}$$

where $w_k \in \mathbb{R}^d$, $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is a nonlinear scale-invariant activation function such as

ReLU [30] and Leaky ReLU [25], both satisfy the condition $\sigma(\alpha t) = \alpha\sigma(t)$ for any $\alpha \geq 0, t \in \mathbb{R}$. Without loss of generality, we assume σ is 1-Lipschitz continuous. In the formula (3.1), we omit the bias term for notational simplicity. The effect of bias term can be incorporated if we assume that the first component of x is always 1. We say that a network is over-parametrized if the *network width* $m > n$. We define a truncated form of f through $Tf(x) = \max\{\min\{f(x), 1\}, 0\}$. By an abuse of notation, in the following we still use f to denote Tf. We will use $\theta = \{(a_k, w_k)\}_{k=1}^m$ to denote all the parameters to be learned from the training data,

The ultimate goal is to minimize the population risk

$$L(\theta) = \mathbb{E}_{x,y}[\ell(f(x;\theta), y)].$$

In practice, we have to work with the empirical risk

$$\hat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(f(x_i;\theta), y_i).$$

Here the loss function $\ell(y, y') = \frac{1}{2}(y - y')^2$, unless it is specified otherwise.

Define the path norm [35],

$$\|\theta\|_{\mathcal{P}} := \sum_{k=1}^m |a_k| \|w_k\|_1, \tag{3.2}$$

We will consider the regularized model defined as follows:

Definition 3.1. *For a two-layer neural network $f(\cdot;\theta)$ of width m, we define the regularized risk as*

$$J_\lambda(\theta) := \hat{L}_n(\theta) + \lambda(\|\theta\|_{\mathcal{P}} + 1).$$

The $+1$ term at the right-hand side is included only to simplify the proof. Our result also holds if we do not include this term in the regularized risk. The corresponding regularized estimator is defined as

$$\hat{\theta}_{n,\lambda} = \operatorname{argmin} J_\lambda(\theta).$$

Here $\lambda > 0$ is a tuning parameter that controls the balance between the fitting error and the model complexity. It is worth noting that the minimizer is not necessarily unique, and $\hat{\theta}_{n,\lambda}$ should be understood as any of the minimizers.

In the following, we will call Lipschitz continuous functions with Lipschitz constant C C-Lipschitz continuous. We will use $X \lesssim Y$ to indicate that $X \leq cY$ for some universal constant $c > 0$.

3.1. Barron space.

We begin by defining the natural function space associated with two-layer neural networks, which we will refer to as the Barron space to honor the pioneering work that Barron has done on this subject [5, 27–29]. A more complete discussion can be found in [21].

Let $\mathbb{S}^d := \{w \mid \|w\|_1 = 1\}$, and let $\mathcal{F}$ be the Borel σ-algebra on $\mathbb{S}^d$ and $P(\mathbb{S}^d)$ be the collection of all probability measures on $(\mathbb{S}^d, \mathcal{F})$. Let $\mathcal{B}(X)$ be the collection of functions that admit the following integral representation:

$$f(x) = \int_{\mathbb{S}^d} a(w)\sigma(\langle w, x\rangle) d\pi(w) \quad \forall x \in X, \tag{3.3}$$

where $\pi \in P(\mathbb{S}^d)$, and $a(\cdot)$ is a measurable function with respect to $(\mathbb{S}^d, \mathcal{F})$. For any $f \in \mathcal{B}(X)$ and $p \geq 1$, we define the following norm

$$\gamma_p(f) := \inf_{(a,\pi)\in\Theta_f} \left(\int_{\mathbb{S}^d} |a(w)|^p d\pi(w) \right)^{1/p}, \tag{3.4}$$

where

$$\Theta_f = \big\{(a,\pi) \,|\, f(x) = \int_{\mathbb{S}^d} a(w)\sigma(\langle w, x\rangle) d\pi(w)\big\}.$$

Definition 3.2 (Barron space). *We define Barron space by*

$$\mathcal{B}_p(X) := \{f \in \mathcal{B}(X) \mid \gamma_p(f) < \infty\}.$$

Since $\pi(\cdot)$ is a probability distribution, by Hölder's inequality, for any $q \geq p > 0$ we have $\gamma_p(f) \leq \gamma_q(f)$. Thus, we have $\mathcal{B}_\infty(X) \subset \cdots \subset \mathcal{B}_2(X) \subset \mathcal{B}_1(X)$.

Obviously $\mathcal{B}_p(X)$ is dense in $C(X)$ since all the finite two-layer neural networks belong to Barron space with $\pi(w) = \frac{1}{m}\sum_{k=1}^m \delta(w - \hat{w}_k)$ and the universal approximation theorem [14] tells us that continuous functions can be approximated by two-layer neural networks. Moreover, it is interesting to note that the $\gamma_1(\cdot)$ norm of a two-layer neural network is bounded by the path norm of the parameters.

An important result proved in [8, 27] states that if a function $f : X \mapsto \mathbb{R}$ satisfies $\int_{\mathbb{R}^d} \|\omega\|_1^2 |\hat{f}(\omega)| d\omega < \infty$, where $\hat{f}$ is the Fourier transform of an extension of f, then it can be expressed in the form (3.3) with

$$\gamma_\infty(f) := \sup_{w\in\mathbb{S}^d} |a(w)| \lesssim \int_{\mathbb{R}^d} \|\omega\|_1^2 |\hat{f}(\omega)| d\omega.$$

Thus it lies in $\mathcal{B}_\infty(X)$.

Connection with reproducing kernel Hilbert space. The Barron space has a natural connection with reproducing kernel Hilbert space (RKHS) [2], and as we will show later, this connection will lead to a precise comparison between two-layer neural networks and kernel methods. For a fixed π, we define

$$\mathcal{H}_\pi(X) := \left\{ \int_{\mathbb{S}^d} a(w)\sigma(\langle w, x\rangle) d\pi(w) : \|f\|_{\mathcal{H}_\pi} < \infty \right\},$$

where

$$\|f\|_{\mathcal{H}_\pi}^2 := \mathbb{E}_\pi[|a(w)|^2].$$

Recall that for a symmetric positive definite (PD)[1] function $k : X \times X \mapsto \mathbb{R}$, the induced RKHS $\mathcal{H}_k$ is the completion of $\{\sum_i a_i k(x_i, x)\}$ with respect to the inner product $\langle k(x_i, \cdot), k(x_j, \cdot)\rangle_{\mathcal{H}_k} = k(x_i, x_j)$. It was proved in [37] that $\mathcal{H}_\pi = \mathcal{H}_{k_\pi}$ with the kernel k_π defined by

$$k_\pi(x, x') = \mathbb{E}_\pi[\sigma(\langle w, x\rangle)\sigma(\langle w, x'\rangle)]. \tag{3.5}$$

[1] We say k is PD function, if for any $x_1, \ldots, x_n$, the matrix K^n with $K_{i,j}^n = k(x_i, x_j)$ is positive semidefinite.

Thus Barron space can be viewed as the union of a family of RKHS with kernels defined by π through Equation (3.5), i.e.

$$\mathcal{B}_2(X) = \bigcup_{\pi \in P(\mathbb{S}^d)} \mathcal{H}_\pi(X). \tag{3.6}$$

Note that the family of kernels is only determined by the activation function $\sigma(\cdot)$.

3.2. Approximation property.

THEOREM 3.1. *For any $f \in \mathcal{B}_2(X)$, there exists a two-layer neural network $f(\cdot;\tilde{\theta})$ of width m, such that*

$$\mathbb{E}_x[(f(x) - f(x;\tilde{\theta}))^2] \le \frac{3\gamma_2^2(f)}{m} \tag{3.7}$$

$$\|\tilde{\theta}\|_{\mathcal{P}} \le 2\gamma_2(f) \tag{3.8}$$

This kind of approximation results have been established in many papers, see for example [5,8]. The difference is that we provide the explicit control of the norm of the constructed solution in (3.8), and the bound is independent of the network size. This observation will be useful for what follows.

The proof of Theorem 3.1 can be found in Appendix A. The basic intuition is that the integral representation of f allows us to approximate f by the Monte-Carlo method: $f(x) \approx \frac{1}{m}\sum_{k=1}^m a(w_k)\sigma(\langle w_k, x\rangle)$ where $\{w_k\}_{k=1}^m$ are sampled from the distribution π.

4. Main results

For simplicity we first discuss the case without noise, i.e. $\xi = 0$. In the next section, we deal with the noise. We also assume $\ln(2d) \ge 1$, and let $\hat{\gamma}_p(f) = \max\{1, \gamma_p(f)\}, \lambda_n = 4\sqrt{2\ln(2d)/n}$. Here d is the dimension of input and the definition of $\gamma_p(\cdot)$ is given in Equation (3.4).

THEOREM 4.1 (Noiseless case). *Assume that the target function $f^* \in \mathcal{B}_2(X)$ and $\lambda \ge \lambda_n$. Then for any $\delta > 0$, with probability at least $1-\delta$ over the choice of the training set S, we have*

$$\mathbb{E}_x|f(x;\hat{\theta}_{n,\lambda}) - f^*(x)|^2 \lesssim \frac{\gamma_2^2(f^*)}{m} + \lambda\hat{\gamma}_2(f^*) \tag{4.1}$$

$$+ \frac{1}{\sqrt{n}}\big(\hat{\gamma}_2(f^*) + \sqrt{\ln(n/\delta)}\big). \tag{4.2}$$

The above theorem provides an a priori estimate for the population risk. The a priori nature is reflected by dependence of the $\gamma_2(\cdot)$ norm of the target function. The first term at the right-hand side controls the approximation error. The second term bounds the estimation error. Surprisingly, the bound for the estimation error is independent of the network width m. Hence the bound also makes sense in the over-parametrization regime.

In particular, if we take $\lambda \asymp \lambda_n$ and $m \ge \sqrt{n}$, the bound becomes $O(1/\sqrt{n})$ up to some logarithmic terms. This bound is nearly optimal in a minimax sense [28, 42].

4.1. Comparison with kernel methods.

Consider $f^* \in \mathcal{B}_2(X)$, and without loss of generality, we assume that $(a^*, \pi^*) \in \Theta_{f^*}$ is one of the best representations of f^* (it is easy to prove that such a representation exists), i.e. $\gamma_2^2(f^*) = \mathbb{E}_{\pi^*}[|a^*(w)|^2]$. For a fixed π_0, we have,

$$f^*(x) = \int_{\mathbb{S}^d} a^*(w)\sigma(\langle w, x\rangle)d\pi^*(w) = \int_{\mathbb{S}^d} a^*(w)\frac{d\pi^*}{d\pi_0}(w)\sigma(\langle w, x\rangle)d\pi_0(w) \tag{4.3}$$

as long as π is absolutely continuous with respect to π_0. In this sense, we can view f^* from the perspective of $\mathcal{H}_{\pi_0}$. Note that $\mathcal{H}_{\pi_0}$ is induced by PD function $k_{\pi_0}(x,x')=\mathbb{E}_{w\sim\pi_0}[\sigma(\langle w,x\rangle)\sigma(\langle w,x'\rangle)]$, and the norm of f^* in $\mathcal{H}_{\pi_0}$ is given by

$$\|f^*\|^2_{\mathcal{H}_{\pi_0}}=\mathbb{E}_{w\sim\pi_0}[|a^*(w)\frac{d\pi^*}{d\pi_0}(w)|^2].$$

Let $\hat{h}_{n,\lambda}$ be the solution of the kernel ridge regression (KRR) problem defined by:

$$\min_{h\in\mathcal{H}_{\pi_0}}\frac{1}{2n}\sum_{i=1}^{n}(h(x_i)-y_i)^2+\lambda\|h\|_{\mathcal{H}_{\pi_0}}. \tag{4.4}$$

We are interested in the comparison between the two population risks $L(\hat{\theta}_{n,\lambda})$ and $L(\hat{h}_{n,\lambda})=\mathbb{E}[\ell(\hat{h}_{n,\lambda}(x),y)]$.

If $\|f^*\|_{\mathcal{H}_{\pi_0}}<\infty$, then we have $f^*\in\mathcal{H}_{\pi_0}$ and $\inf_{h\in\mathcal{H}_{\pi_0}}L(h)=0$. In this case, it was proved in [11] that the optimal learning rate is

$$L(\hat{h}_{n,\lambda})\sim\frac{\|f^*\|_{\mathcal{H}_{\pi_0}}}{\sqrt{n}}. \tag{4.5}$$

Compared to Theorem 4.1, we can see that both rates have the same scaling with respect to n, the number of samples. The only difference appears in the two norms: $\gamma_2(f^*)$ and $\|f^*\|_{\mathcal{H}_{\pi_0}}$. From the definition (3.4), we always have $\gamma_2(f^*)\le\|f^*\|_{\mathcal{H}_{\pi_0}}$, since $(a^*\frac{d\pi^*}{d\pi_0},\pi_0)\in\Theta_{f^*}$. If π^* is nearly singular with respect to π_0, then $\|f^*\|_{\mathcal{H}_{\pi_0}}\gg\gamma_2(f^*)$. In this case, the population risk for the kernel methods should be much larger than the population risk for the neural network model.

Example. Take π_0 to be the uniform distribution over $\mathbb{S}^d$ and $f^*(x)=\sigma(\langle w^*,x\rangle)$, for which $\pi^*(w)=\delta(w-w^*)$ and $a^*(w)=1$. In this case $\gamma_2(f^*)=1$, but $\|f^*\|_{\mathcal{H}_{\pi_0}}=+\infty$. Thus the rate (4.5) becomes trivial. Assume that the population risk scales as $O(n^{-\beta})$, and it is interesting to see how β depends on the dimension d. We numerically estimate β's for two methods, and report the results in Table 4.1. It does show that the higher the dimensionality, the slower the rate of the kernel method. In contrast, the rates for the two-layer neural networks are independent of the dimensionality, which confirms the the prediction of Theorem 4.1. For this particular target function, the value of $\beta\ge1$ is bigger than the lower bound (1/2) proved in Theorem 4.1. This is not a contradiction since the latter holds for any $f\in\mathcal{B}_2(X)$.

d	10	100	1000
β_{nn}	1.18	1.23	1.02
β_{ker}	0.50	0.35	0.14

TABLE 4.1. *The error rates of learning the one-neuron function in different dimensions. The second and third lines correspond to the two-layer neural network and the kernel ridge regression method, respectively.*

The two-layer neural network model as an adaptive kernel method. Recall that $\mathcal{B}_2(X)=\cup_\pi\mathcal{H}_\pi(X)$. The norm $\gamma_2(\cdot)$ characterizes the complexity of the target function by selecting the best kernel among a family of kernels $\{k_\pi(\cdot,\cdot)\}_{\pi\in P(\mathbb{S}^d)}$. The kernel method works with a specific RKHS with a particular choice of the kernel or the probability distribution π. In contrast, the neural network models work with the union of all these RKHS and select the kernel or the probability distribution adapted to the data. From this perspective, we can view the two-layer neural network model as an adaptive kernel method.

4.2. Tackling the noise. We first make the following sub-Gaussian assumption on the noise.

Assumption 4.1. *We assume that the noise satisfies*

$$\mathbb{P}[|\xi| > t] \leq c_0 e^{-\frac{t^2}{\sigma}} \quad \forall t \geq \tau_0. \tag{4.6}$$

Here c_0, τ_0 *and* σ *are constants.*

In the presence of noise, the population risk can be decomposed into

$$L(\theta) = \mathbb{E}_x (f(x;\theta) - f^*(x))^2 + \mathbb{E}[\xi^2]. \tag{4.7}$$

This suggests that, in spite of the noise, we still have $\operatorname{argmin}_\theta L(\theta) = \operatorname{argmin}_\theta \mathbb{E}_x |f(x;\theta) - f^*(x)|^2$, and the latter is what we really want to minimize. However due to the noise, $\ell(f(x_i), y_i)$ might be unbounded. We cannot directly use the generalization bound in Theorem 5.2. To address this issue, we consider the truncated risk defined as follows,

$$L_B(\theta) = \mathbb{E}_{x,y}[\ell(f(x;\theta), y) \wedge \frac{B^2}{2}]$$
$$\hat{L}_B(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(x_i;\theta), y_i) \wedge \frac{B^2}{2}.$$

Let $B_n = 1 + \max\{\tau_0, \sigma^2 \ln n\}$. For the noisy case, we consider the following regularized risk:

$$J_\lambda(\theta) := \hat{L}_{B_n}(\theta) + \lambda B_n (\|\theta\|_{\mathcal{P}} + 1). \tag{4.8}$$

The corresponding regularized estimator is given by $\hat{\theta}_{n,\lambda} = \operatorname{argmin} J_\lambda(\theta)$. Here for simplicity we slightly abused the notation.

Theorem 4.2 (Main result, noisy case). *Assume that the target function* $f^* \in \mathcal{B}_2(X)$ *and* $\lambda \geq \lambda_n$. *Then for any* $\delta > 0$, *with probability at least* $1 - \delta$ *over the choice of the training set* S, *we have*

$$\mathbb{E}_x |f(x;\hat{\theta}_{n,\lambda}) - f^*(x)|^2 \lesssim \frac{\gamma_2^2(f^*)}{m} + \lambda B_n \hat{\gamma}_2(f^*) + \frac{B_n^2}{\sqrt{n}}\left(\hat{\gamma}_2(f^*) + \sqrt{\ln(n/\delta)}\right)$$
$$+ \frac{B_n^2}{\sqrt{n}}\left(c_0\sigma^2 + \sqrt{\frac{\mathbb{E}[\xi^2]}{n^{1/2}\lambda}}\right).$$

Compared to Theorem 4.1, the noise introduces at most several logarithmic terms. The case with no noise corresponds to the situation with $B_n = 1$.

4.3. Extension to classification problems. Let us consider the simplest setting: binary classification problem, where $y \in \{0,1\}$. In this case, $f^*(x) = \mathbb{P}\{y = 1|x\}$ denotes the probability of $y = 1$ given x. Given $f^*(\cdot)$ and $f(\cdot;\theta_{n,\lambda})$, the corresponding plug-in classifiers are defined by $\eta^*(x) = 1[f^*(x) \geq \frac{1}{2}]$ and $\hat{\eta}(x) = 1[f(x;\hat{\theta}_{n,\lambda}) \geq \frac{1}{2}]$, respectively. η^* is the optimal Bayes classifier.

For a classifier η, we measure its performance by the 0-1 loss defined by $\mathcal{E}(\eta) = \mathbb{P}\{\eta(x) \neq y\}$.

Corollary 4.1. *Under the same assumption as in Theorem 4.2 and taking $\lambda = \lambda_n$, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\mathcal{E}(\hat{\eta}) \lesssim \mathcal{E}(\eta^*) + \frac{\gamma_2(f^*)}{\sqrt{m}} + \hat{\gamma}_2^{1/2}(f^*) \frac{\ln^{1/4}(d) + \ln^{1/4}(n/\delta)}{n^{1/4}}.$$

Proof. According to the Theorem 2.2 of [16], we have

$$\begin{aligned} \mathcal{E}(\hat{\eta}) - \mathcal{E}(\eta^*) &\leq 2\mathbb{E}[|f(x;\hat{\theta}_{n,\lambda}) - f^*(x)|] \\ &\leq 2\mathbb{E}[|f(x;\hat{\theta}_{n,\lambda}) - f^*(x)|^2]. \end{aligned} \tag{4.9}$$

In this case, $\varepsilon_i = y_i - f^*(x_i)$ is bounded by 1, thus $\tau_0 = 1, c = \sigma = 0$. Applying Theorem 4.2 yields the result. □

The above theorem suggests that our a priori estimates also hold for classification problems, although the error rate only scales as $O(n^{-1/4})$. It is possible to improve the rate with more a delicate analysis. One potential way is to specifically develop a better estimate for L_1 loss, as can be seen from inequality (4.9). Another way is to make a stronger assumption on the data. For example, we can assume that there exists $f^* \in \mathcal{B}_2(X)$ such that $\mathbb{P}_{x,y}(yf^*(x) \geq 1) = 1$, for which the Bayes error $\mathcal{E}(\eta^*) = 0$. We leave these to future work.

5. Proofs

5.1. Bounding the generalization gap.

Definition 5.1 (Rademacher complexity). *Let $\mathcal{F}$ be a hypothesis space, i.e. a set of functions. The Rademacher complexity of $\mathcal{F}$ with respect to samples $S = (z_1, z_2, \ldots, z_n)$ is defined as $\hat{\mathcal{R}}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon}}[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \varepsilon_i f(z_i)]$, where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. random variables with $\mathbb{P}(\varepsilon_i = +1) = \mathbb{P}(\varepsilon_i = -1) = \frac{1}{2}$.*

The generalization gap can be estimated via the Rademacher complexity by the following theorem [39].

Theorem 5.1. *Fix a hypothesis space $\mathcal{F}$. Assume that for any $f \in \mathcal{F}$ and z, $|f(z)| \leq B$. Then for any $\delta > 0$, with probability at least $1-\delta$ over the choice of $S = (z_1, z_2, \ldots, z_n)$, we have,*

$$\left|\frac{1}{n}\sum_{i=1}^n f(z_i) - \mathbb{E}_z[f(z)]\right| \leq 2\mathbb{E}_S[\hat{\mathcal{R}}_n(\mathcal{F})] + B\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

Let $\mathcal{F}_Q = \{f(x;\theta) \,|\, \|\theta\|_{\mathcal{P}} \leq Q\}$ denote all the two-layer networks with path norm bounded by Q. It was proved in [35] that

$$\hat{\mathcal{R}}_n(\mathcal{F}_Q) \leq 2Q\sqrt{\frac{2\ln(2d)}{n}}. \tag{5.1}$$

By combining the above result with Theorem 5.1, we obtain the following a posterior bound of the generalization gap for two-layer neural networks. The proof is deferred to Appendix B.

Theorem 5.2 (A posterior generalization bound). *Assume that the loss function $\ell(\cdot, y)$ is $A-$Lipschitz continuous and bounded by B. Then for any $\delta > 0$, with probability* *at least* $1-\delta$ *over the choice of the training set* S*, we have, for any two-layer network* $f(\cdot;\theta)$*,*

$$|L(\theta)-\hat{L}_n(\theta)| \le 4A\sqrt{\frac{2\ln(2d)}{n}}(\|\theta\|_{\mathcal{P}}+1) \tag{5.2}$$

$$+B\sqrt{\frac{2\ln(2c(\|\theta\|_{\mathcal{P}}+1)^2/\delta)}{n}}, \tag{5.3}$$

where $c=\sum_{k=1}^{\infty} 1/k^2$*.*

We see that the generalization gap is bounded roughly by $\|\theta\|_{\mathcal{P}}/\sqrt{n}$ up to some logarithmic terms.

5.2. Proof for the noiseless case. The intuition is as follows. The path norm of the special solution $\tilde{\theta}$ which achieves the optimal approximation error is independent of the network width, and this norm can also be used to bound the generalization gap (Theorem 5.2). Therefore, if the path norm is suitably penalized during training, we should be able to control the generalization gap without harming the approximation accuracy.

We first have the estimate for the regularized risk of $\tilde{\theta}$.

PROPOSITION 5.1. *Let* $\tilde{\theta}$ *be the network constructed in Theorem 3.1, and* $\lambda \ge \lambda_n$*. Then with probability at least* $1-\delta$*, we have*

$$J_\lambda(\tilde{\theta}) \le L(\tilde{\theta}) + 8\lambda\hat{\gamma}_2(f^*) + 2\sqrt{\frac{2\ln(2c/\delta)}{n}}. \tag{5.4}$$

Proof. First $\ell(y,y_i)=\frac{1}{2}(y-y_i)^2$ is 1-Lipschitz continuous and bounded by 2. According to Definition 3.1 and the property that $\|\tilde{\theta}\|_{\mathcal{P}} \le 2\gamma_2(f^*)$, the regularized risk of $\tilde{\theta}$ satisfies

$$\begin{aligned} J_\lambda(\tilde{\theta}) &= \hat{L}_n(\tilde{\theta}) + \lambda(\|\tilde{\theta}\|_{\mathcal{P}}+1) \\ &\le L(\tilde{\theta}) + (\lambda_n+\lambda)(\|\tilde{\theta}\|_{\mathcal{P}}+1) + 2\sqrt{\frac{2\ln(2c(\|\tilde{\theta}\|_{\mathcal{P}}+1)^2/\delta)}{n}} \\ &\le L(\tilde{\theta}) + 6\lambda\hat{\gamma}_2(f^*) + 2\sqrt{\frac{2\ln(2c(1+2\gamma_2(f^*))^2/\delta)}{n}}. \end{aligned} \tag{5.5}$$

The last term can be simplified by using $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$ and $\ln(1+a) \le a$ for $a \ge 0, b \ge 0$. So we have

$$\sqrt{2\ln(2c(1+2\gamma_2(f^*))^2/\delta)} \le \sqrt{2\ln(2c/\delta)} + 3\hat{\gamma}_2(f^*).$$

Plugging it into Equation (5.5) completes the proof. □

PROPOSITION 5.2 (Properties of regularized solutions). *The regularized estimator* $\hat{\theta}_{n,\lambda}$ *satisfies:*

$$J_\lambda(\hat{\theta}_{n,\lambda}) \le J_\lambda(\tilde{\theta})$$

$$\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}} \le \frac{L(\tilde{\theta})}{\lambda} + 8\hat{\gamma}_2(f^*) + \frac{1}{2}\sqrt{\ln(2c/\delta)}$$

Proof. The first claim follows from the definition of $\hat{\theta}_n$. For the second claim, note that

$$\lambda(\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}+1)\leq J_\lambda(\hat{\theta}_n)\leq J_\lambda(\tilde{\theta}),$$

Applying Proposition 5.1 completes the proof. □

Remark 5.1. The above proposition establishes the connection between the regularized solution and the special solution $\tilde{\theta}$ constructed in Theorem 3.1. In particular, by taking $\lambda=t\lambda_n$ with $t\geq 1$ the generalization gap of the regularized solution is bounded by $\frac{\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}}{\sqrt{n}}\to L(\tilde{\theta})/(t\sqrt{\ln 2d})$ as $n\to\infty$, up to some constant. This suggests that our regularization term is appropriate, and it forces the generalization gap to be roughly in the same order as the approximation error.

Proof. **(Proof of Theorem 4.1.)** Now we are ready to prove the main result. Following the a posteriori generalization bound given in Theorem 5.2, we have with probability at least $1-\delta$,

$$\begin{aligned} L(\hat{\theta}_{n,\lambda}) &\leq \hat{L}_n(\hat{\theta}_{n,\lambda})+\lambda_n(\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}+1)+3Q_n \\ &\overset{(1)}{\leq} J_\lambda(\hat{\theta}_{n,\lambda})+3Q_n, \end{aligned}$$

where $Q_n=\sqrt{\ln(2c(1+\|\hat{\theta}_{n,\lambda}\|)^2/\delta)/n}$. The inequality (1) is due to the choice $\lambda\geq\lambda_n$. The first term can be bounded by $J_\lambda(\hat{\theta}_{n,\lambda})\leq J_\lambda(\tilde{\theta})$, which is given by Proposition 5.1. It remains to bound Q_n,

$$\begin{aligned} \sqrt{n}Q_n &\leq \sqrt{\ln(2nc/\delta)}+\sqrt{2\ln(1+n^{-1/2}\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}})} \\ &\leq \sqrt{\ln(2nc/\delta)}+\sqrt{2\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}/\sqrt{n}}. \end{aligned}$$

By Proposition 5.2, we have

$$\begin{aligned} \sqrt{\frac{2\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}}{\sqrt{n}}} &\leq \sqrt{\frac{2(L(\tilde{\theta})/\lambda+8\hat{\gamma}_2(f^*)+0.5\sqrt{\ln(2c/\delta)})}{\sqrt{n}}} \\ &\leq \sqrt{\frac{2L(\tilde{\theta})}{\lambda n^{1/2}}}+\frac{3\hat{\gamma}_2(f^*)}{n^{1/4}}+\left(\frac{\ln(1/\delta)}{n}\right)^{1/4}. \end{aligned}$$

Thus after some simplification, we obtain

$$Q_n\leq 2\sqrt{\frac{\ln(n/\delta)}{n}}+\sqrt{\frac{2L(\tilde{\theta})}{\lambda n^{3/2}}}+\frac{3\hat{\gamma}_2(f^*)}{\sqrt{n}}. \tag{5.6}$$

By combining Equation (5.4) and (5.6), we obtain

$$L(\hat{\theta}_n)\lesssim L(\tilde{\theta})+8\lambda\hat{\gamma}_2(f^*)+\frac{3}{\sqrt{n}}\Big(\sqrt{\frac{L(\tilde{\theta})}{n^{1/2}\lambda}}+\hat{\gamma}_2(f^*)+\sqrt{\ln(n/\delta)}\Big).$$

By applying $L(\tilde{\theta})\leq 3\gamma_2^2(f^*)/m$, we complete the proof. □

5.3. Proof for the noisy case.

We need the following lemma. The proof is deferred to Appendix C.

LEMMA 5.1. *Under Assumption 4.1, we have*

$$\sup_{\theta} |L(\theta) - L_{B_n}(\theta)| \leq \frac{2c_0\sigma^2}{\sqrt{n}},$$

Therefore we have,

$$L(\theta) = L(\theta) - L_{B_n}(\theta) + L_{B_n}(\theta) \leq \frac{2c_0\sigma^2}{\sqrt{n}} + L_{B_n}(\theta).$$

This suggests that as long as we can bound the truncated population risk, the original risk will be bounded accordingly.

Proof. **(Proof of Theorem 4.2.)** The proof is almost the same as the noiseless case. The loss function $\ell(y,y_i)\wedge B^2/2$ is B-Lipschitz continuous and bounded by $B^2/2$. By analogy with the proof of Proposition 5.1, we obtain that with probability at least $1-\delta$ the following inequality holds,

$$J_\lambda(\tilde{\theta}) \leq L_{B_n}(\tilde{\theta}) + 8B_n\lambda\hat{\gamma}_2(f^*) + B_n^2\sqrt{\frac{\ln(2c/\delta)}{n}}. \tag{5.7}$$

Following the proof in Proposition 5.2, we similarly obtain $J_\lambda(\hat{\theta}_{n,\lambda}) \leq J_\lambda(\tilde{\theta})$ and

$$\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}} \leq \frac{L_{B_n}(\tilde{\theta})}{B_n\lambda} + 8\hat{\gamma}(f^*) + \frac{B_n}{2}\sqrt{\ln(2c/\delta)}. \tag{5.8}$$

Following the proof of Theorem 4.1, we have

$$L_{B_n}(\hat{\theta}_{n,\lambda}) \leq J_\lambda(\tilde{\theta}) + \frac{B_n^2}{2}\sqrt{2\ln(2c(1+\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}})^2/\delta)/n}. \tag{5.9}$$

Plugging (5.7) and (5.8) into (5.9), we get

$$\begin{aligned} L_{B_n}(\hat{\theta}_{n,\lambda}) \leq\; & L_{B_n}(\tilde{\theta}) + 8B_n\hat{\gamma}_2(f^*)\lambda \\ & + \frac{3B_n^2}{\sqrt{n}}\left(\sqrt{\frac{L_{B_n}(\tilde{\theta})}{n^{1/2}\lambda}} + \hat{\gamma}_2(f^*) + \sqrt{\ln(n/\delta)}\right). \end{aligned}$$

Using Lemma 5.1 and the decomposition (4.7), we complete the proof. □

6. Numerical experiments

In this section, we evaluate the regularized model using numerical experiments. We consider two datasets, MNIST[2] and CIFAR-10[3]. Each example in MNIST is a 28 × 28 grayscale image, while each example in CIFAR-10 is a 32 × 32 × 3 color image. For MNIST, we map numbers $\{0,1,2,3,4\}$ to label 0 and $\{5,6,7,8,9\}$ to 1. For CIFAR-10, we select the examples with labels 0 and 1 to construct our new training and validation sets. Thus, our new MNIST has 60,000 training examples, and CIFAR-10 has 10,000 training examples.

[2] http://yann.lecun.com/exdb/mnist/

[3] https://www.cs.toronto.edu/~kriz/cifar.html

The two-layer ReLU network is initialized using $a_i \sim \mathcal{N}(0, \frac{2\kappa}{m})$, $w_{i,j} \sim \mathcal{N}(0, 2\kappa/d)$. We use $\kappa=1$ and train the regularized models using the Adam optimizer [26] for $T=10{,}000$ steps, unless it is specified otherwise. The initial learning rate is set to be 0.001, and it is then multiplied by a decay factor of 0.1 at $0.7T$ and again at $0.9T$. We set the trade-off parameter $\lambda = 0.1\lambda_n$[4] .

6.1. Shaper bounds for the generalization gap.

Theorem 5.2 shows that the generalization gap is bounded by $\frac{\|\theta\|_{\mathcal{P}}}{\sqrt{n}}$ up to some logarithmic terms. Previous works [18, 34] showed that (stochastic) gradient descent tends to find solutions with huge norms, causing the a posterior bound to be vacuous. In contrast, our theory suggests there exist good solutions (i.e. solutions with small generalization error) with small norms, and these solutions can be found by the explicit regularization.

To see how this works in practice, we trained both the regularized models and unregularized models ($\lambda=0$) for fixed network width $m=$10,000. To cover the over-parametrized regime, we also consider the case $n=100$ where $m/n=100\gg 1$. The results are summarized in Table 6.1.

dataset	λ	n	training accuracy	testing accuracy	$\frac{\|\theta\|_{\mathcal{P}}}{\sqrt{n}}$
CIFAR-10	0	10^4	100%	84.5%	58
		100	100%	70.5%	507
	0.1	10^4	87.4%	86.9%	**0.14**
		100	91.0%	72.0%	**0.43**
MNIST	0	6×10^4	100%	98.8%	58
		100	100%	78.7%	162
	0.1	6×10^4	98.1%	97.8%	**0.27**
		100	100%	74.9%	**0.41**

TABLE 6.1. *Comparison of regularized* ($\lambda=0.1$) *and unregularized* ($\lambda=0$) *models. For each case, the experiments are repeated for* 5 *times, and the mean values are reported.*

As we can see, the test accuracies of the regularized and unregularized solutions are generally comparable, but the values of $\frac{\|\theta\|_{\mathcal{P}}}{\sqrt{n}}$, which serve as an upper bound for the generalization gap, are drastically different. The bounds for the unregularized models are always vacuous, as was observed in [4,18,34]. In contrast, the bounds for the regularized models are always several orders of magnitude smaller than that for the unregularized models. This is consistent with the theoretical prediction in Proposition 5.2.

To further explore the impact of over-parametrization, we trained various models with different widths. For both datasets, all the training examples are used. In Figure 6.1, we display how the value of $\frac{\|\theta\|_{\mathcal{P}}}{\sqrt{n}}$ of the learned solution varies with the network width. We find that for the unregularized model this quantity increases with network width, whereas for the regularized model it is almost constant. This is consistent with our theoretical result.

6.2. Dependence on the Initialization.

Since the neural network model is non-convex, it is interesting to see how initialization affects the performance of the different models, regularized and unregularized, especially in the over-parametrized regime. To this end, we fix $m=10000, n=100$ and vary the variance of random initialization

[4] Our proof of theoretical results requires $\lambda \geq \lambda_n$. However, this condition is not necessarily optimal.

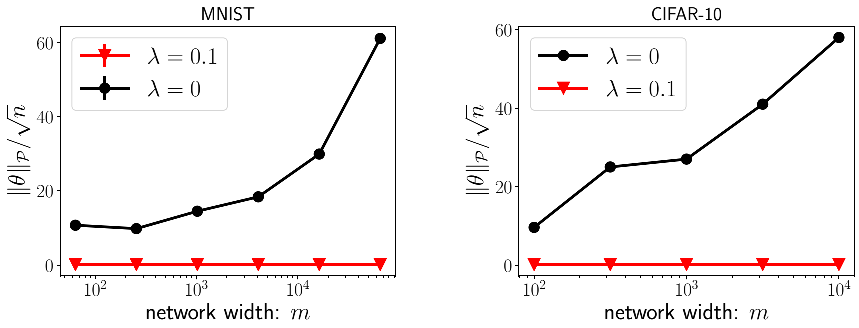

FIG. 6.1. *Comparison of the path norms between the regularized and unregularized solutions for varying widths.*

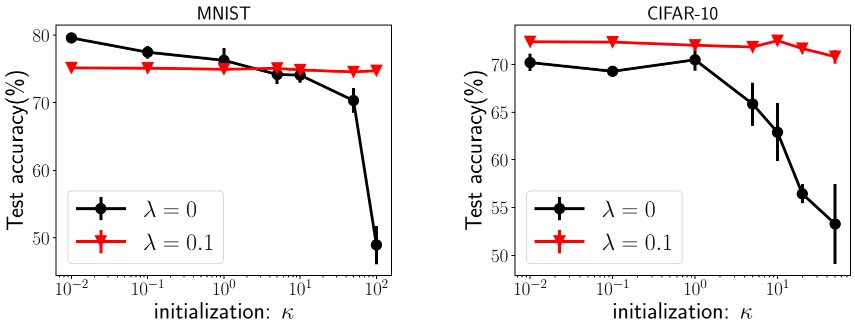

FIG. 6.2. *Test accuracies of solutions obtained from different initializations. Each experiment is repeated for* 5 *times, and we report the mean and standard deviation.*

κ. The results are reported in Figure 6.2. In general, we find that regularized models are much more stable than the unregularized models. For large initialization, the regularized model always performs significantly better.

7. Conclusion

In this paper, we proved nearly optimal a priori estimates of the population risk for learning two-layer neural networks. Our results also give some insight regarding the advantage of neural network models over the kernel method. We should also mention that the main result of this paper has also been extended to deep residual network models in [19].

The most unsatisfactory aspect of our result is that it is proved for the regularized model since practitioners rely on the so-called implicit regularization. At the moment it is unclear where the "implicit regularization" comes from and how it actually works. Existing works consider special initialization schemes and require strong assumptions on the target function [1, 9, 15, 20, 22]. In particular, the work in [20, 22] demonstrates clearly that in the regimes considered the neural network models are no better than the kernel method in terms of implicit regularization. This is quite unsatisfactory.

There is overwhelming evidence that by tuning the optimization procedure, including the algorithm, the initialization, the hyper-parameters, etc., one can find solutions with superior performance on the test data. The problem is that excessive tuning and serious experience is required to find good solutions. Until we have a good understanding about the mysteries surrounding implicit regularization, the business of parameter tuning for unregularized models will remain an art. In contrast, the regularized model proposed here is rather robust and much more fool-proof. Borrowing the terminology from mathematical physics, one is tempted to say that the regularized model considered here is "well-posed" whereas the unregularized model is "ill-posed" [40].

Acknowledgement. The work presented here is supported in part by a gift to Princeton University from iFlytek and the ONR grant N00014-13-1-0338.

Appendix A. Proof of Theorem 3.1. Without loss of generality, let (a,π) be the best representation of f, i.e. $\gamma_2^2(f)=\mathbb{E}_\pi[|a(w)|^2]$. Let $U=\{w_j\}_{j=1}^m$ be *i.i.d.* random variables sampled from $\pi(\cdot)$, and define

$$\hat{f}_U(x)=\frac{1}{m}\sum_{j=1}^m a(w_j)\sigma(\langle w_j,x\rangle).$$

Let $L_U=\mathbb{E}_x|\hat{f}_U(x)-f(x)|^2$ denote the population risk, we have

$$\begin{aligned}\mathbb{E}_U[L_U]&=\mathbb{E}_x\mathbb{E}_U|\hat{f}_U(x)-f(x)|^2\\&=\frac{1}{m^2}\mathbb{E}_x\sum_{j,l=1}^m\mathbb{E}_{w_j,w_l}[(a(w_j)\sigma(\langle w_j,x\rangle)-f(x))(a(w_l)\sigma(\langle w_l,x\rangle)-f(x))]\\&\leq\frac{\gamma_2^2(f)}{m}.\end{aligned}$$

On the other hand, denote the path norm of $\hat{f}_U(x)$ by A_U, we have $\mathbb{E}_U[A_U]=\gamma_1(f)\leq\gamma_2(f)$.

Define the event $E_1=\{L_U<\frac{3\gamma_2^2(f)}{m}\}$, and $E_2=\{A_U<2\gamma_1(f)\}$. By Markov's inequality, we have

$$\begin{aligned}\mathbb{P}\{E_1\}&=1-\mathbb{P}\{L_U\geq\frac{3\gamma_2^2(f)}{m}\}\geq1-\frac{\mathbb{E}_U[L(U)]}{3\gamma_2^2(f)/m}\geq\frac{2}{3}\\\mathbb{P}\{E_2\}&=1-\mathbb{P}\{A_U\geq2\gamma_2(f)\}\geq1-\frac{\mathbb{E}[A_U]}{2\gamma_2(f)}\geq\frac{1}{2}.\end{aligned}$$

Therefore, we have the probability of two events happening together,

$$\mathbb{P}\{E_1\cap E_2\}=\mathbb{P}\{E_1\}+\mathbb{P}\{E_2\}-1\geq\frac{2}{3}+\frac{1}{2}-1>0.$$

This completes the proof.

Appendix B. Proof of Theorem 5.2. Before we provide the upper bound for the Rademacher complexity of two-layer networks, we first need the following two lemmas.

Lemma B.1 (Lemma 26.11 of [39]). *Let $S=(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ be n vectors in $\mathbb{R}^d$. Then the Rademacher complexity of $\mathcal{H}_1=\{\mathbf{x}\mapsto\boldsymbol{u}\cdot\mathbf{x}\mid\|\boldsymbol{u}\|_1\leq1\}$ has the following upper bound,*

$$\hat{\mathcal{R}}_n(\mathcal{H}_1)\leq\max_i\|\mathbf{x}_i\|_\infty\sqrt{\frac{2\ln(2d)}{n}}.$$

The above lemma characterizes the Rademacher complexity of a linear predictor with ℓ_1 norm bounded by 1. To handle the influence of nonlinear activation function, we need the following contraction lemma.

Lemma B.2 (Lemma 26.9 of [39]). *Let $\phi_i : \mathbb{R} \mapsto \mathbb{R}$ be a ρ−Lipschitz function, i.e. for all $\alpha, \beta \in \mathbb{R}$ we have $|\phi_i(\alpha) - \phi_i(\beta)| \leq \rho|\alpha - \beta|$. For any $\boldsymbol{a} \in \mathbb{R}^n$, let $\boldsymbol{\phi}(\boldsymbol{a}) = (\phi_1(a_1), \ldots, \phi_n(a_n))$, then we have*

$$\hat{\mathcal{R}}_n(\boldsymbol{\phi} \circ \mathcal{H}) \leq \rho \hat{\mathcal{R}}_n(\mathcal{H}).$$

We are now ready to estimate the Rademacher complexity of two-layer networks.

Lemma B.3. *Let $\mathcal{F}_Q = \{f_m(x;\theta) \,|\, \|\theta\|_{\mathcal{P}} \leq Q\}$ be the set of two-layer networks with path norm bounded by Q, then we have*

$$\hat{\mathcal{R}}_n(\mathcal{F}_Q) \leq 2Q\sqrt{\frac{2\ln(2d)}{n}}.$$

Proof. To simplify the proof, we let $c_k = 0$, otherwise we can define $w_k = (w_k^T, c_k)^T$ and $\mathbf{x} = (\mathbf{x}^T, 1)^T$.

$$\begin{aligned}
n\hat{\mathcal{R}}_n(\mathcal{F}_Q) &= \mathbb{E}_\xi\big[\sup_{\|\theta\|_{\mathcal{P}} \leq Q} \sum_{i=1}^n \xi_i \sum_{k=1}^m a_k \|w_k\|_1 \sigma(\hat{w}_k^T \mathbf{x}_i)\big] \\
&\leq \mathbb{E}_\xi\big[\sup_{\|\theta\|_{\mathcal{P}} \leq Q, \|\boldsymbol{u}_k\|_1 = 1} \sum_{i=1}^n \xi_i \sum_{k=1}^m a_k \|w_k\|_1 \sigma(\boldsymbol{u}_k^T \mathbf{x}_i)\big] \\
&= \mathbb{E}_\xi\big[\sup_{\|\theta\|_{\mathcal{P}} \leq Q, \|\boldsymbol{u}_k\|_1 = 1} \sum_{k=1}^m a_k \|w_k\|_1 \sum_{i=1}^n \xi_i \sigma(\boldsymbol{u}_k^T \mathbf{x}_i)\big] \\
&\leq \mathbb{E}_\xi\big[\sup_{\|\theta\|_{\mathcal{P}} \leq Q} \sum_{k=1}^m |a_k \|w_k\|_1| \sup_{\|\boldsymbol{u}\|_1 = 1} |\sum_{i=1}^n \xi_i \sigma(\boldsymbol{u}^T \mathbf{x}_i)|\big] \\
&\leq Q\mathbb{E}_\xi\big[\sup_{\|\boldsymbol{u}\|_1 = 1} |\sum_{i=1}^n \xi_i \sigma(\boldsymbol{u}^T \mathbf{x}_i)|\big] \\
&\leq Q\mathbb{E}_\xi\big[\sup_{\|\boldsymbol{u}\|_1 \leq 1} |\sum_{i=1}^n \xi_i \sigma(\boldsymbol{u}^T \mathbf{x}_i)|\big].
\end{aligned}$$

Due to the symmetry, we have that

$$\begin{aligned}
\mathbb{E}_\xi\big[\sup_{\|\boldsymbol{u}\|_1 \leq 1} |\sum_{i=1}^n \xi_i \sigma(\boldsymbol{u}^T \mathbf{x}_i)|\big] &\leq \mathbb{E}_\xi\big[\sup_{\|\boldsymbol{u}\|_1 \leq 1} \sum_{i=1}^n \xi_i \sigma(\boldsymbol{u}^T \mathbf{x}_i) + \sup_{\|\boldsymbol{u}\|_1 \leq 1} \sum_{i=1}^n -\xi_i \sigma(\boldsymbol{u}^T \mathbf{x}_i)\big] \\
&= 2\mathbb{E}_\xi\big[\sup_{\|\boldsymbol{u}\|_1 \leq 1} \sum_{i=1}^n \xi_i \sigma(\boldsymbol{u}^T \mathbf{x}_i)\big].
\end{aligned}$$

Since σ is Lipschitz continuous with Lipschitz constant 1, by applying Lemma B.2 and Lemma B.1, we obtain

$$\hat{\mathcal{R}}_n(\mathcal{F}_Q) \leq 2Q\sqrt{\frac{2\ln(2d)}{n}}.$$

□

PROPOSITION B.1. *Assume the loss function $\ell(\cdot,y)$ is $A-$Lipschitz continuous and bounded by B, then with probability at least $1-\delta$ we have,*

$$\sup_{\|\theta\|_{\mathcal{P}}\leq Q} |L(\theta)-\hat{L}_n(\theta)| \leq 4AQ\sqrt{\frac{2\ln(2d)}{n}} + B\sqrt{\frac{2\ln(2/\delta)}{n}}. \tag{B.1}$$

Proof. Define $\mathcal{H}_Q = \{\ell\circ f \,|\, f\in\mathcal{F}_Q\}$, then we have $\hat{\mathcal{R}}_n(\mathcal{H}_Q)\leq 2BQ\sqrt{\frac{2\ln(2d)}{n}}$, which follows from Lemma B.2 and B.3. Then directly applying Theorem 5.1 yields the result. ∎

Proof. **(Proof of Theorem 5.2.)** Consider the decomposition $\mathcal{F}=\cup_{l=1}^{\infty}\mathcal{F}_l$, where $\mathcal{F}_l=\{f_m(\mathbf{x};\theta)\,|\,\|\theta\|_{\mathcal{P}}\leq l\}$. Let $\delta_l=\frac{\delta}{cl^2}$ where $c=\sum_{l=1}^{\infty}\frac{1}{l^2}$. According to Proposition B.1, if we fix l in advance, then with probability at least $1-\delta_l$ over the choice of S, we have

$$\sup_{\|\theta\|_{\mathcal{P}}\leq l} |L(\theta)-\hat{L}_n(\theta)| \leq 4Al\sqrt{\frac{2\ln(2d)}{n}} + B\sqrt{\frac{2\ln(2/\delta_l)}{n}}. \tag{B.2}$$

So the probability that there exists at least one l such that (B.2) fails is at most $\sum_{l=1}^{\infty}\delta_l = \delta$. In other words, with probability at least $1-\delta$, the inequality (B.2) holds for all l.

Given an arbitrary set of parameters θ, denote $l_0=\min\{l\,|\,\|\theta\|_{\mathcal{P}}\leq l\}$, then $l_0\leq \|\theta\|_{\mathcal{P}}+1$. Equation (B.2) implies that

$$\begin{aligned}|L(\theta)-\hat{L}_n(\theta)| &\leq 4Al_0\sqrt{\frac{2\ln(2d)}{n}} + B\sqrt{\frac{2\ln(2cl_0^2/\delta)}{n}}\\ &\leq 4A(\|\theta\|_{\mathcal{P}}+1)\sqrt{\frac{2\ln(2d)}{n}} + B\sqrt{\frac{2\ln(2c(1+\|\theta\|_{\mathcal{P}})^2/\delta)}{n}}.\end{aligned}$$

∎

Appendix C. Proof of Lemma 5.1.

Proof. Let $Z=f(x;\theta)-f^*(x)-\varepsilon$, then for any $B\geq 2+\tau_0$, we have

$$\begin{aligned}|L(\theta)-L_B(\theta)| &= \mathbb{E}\left[(Z^2-B^2)\mathbf{1}_{|Z|\geq B}\right]\\ &= \int_0^{\infty}\mathbb{P}\{Z^2-B^2\geq t^2\}dt^2 \leq \int_0^{\infty}\mathbb{P}\{|Z|\geq\sqrt{B^2+t^2}\}dt^2\\ &\leq \int_0^{\infty}\mathbb{P}\{|\varepsilon|\geq\sqrt{B^2+t^2}-2\}dt^2\\ &= c_0\int_B^{\infty} e^{-\frac{s^2}{2\sigma^2}}ds^2 = 2c_0\sigma^2 e^{-B^2/2\sigma^2}\end{aligned}$$

Since $B_n\geq\sigma^2\ln n$, we have $2c_0\sigma^2 e^{-\frac{B_n^2}{2\sigma^2}}\leq 2c_0\sigma^2 n^{-1/2}$. We thus complete the proof. ∎

REFERENCES

[1] Z. Allen-Zhu, Y. Li, and Y. Liang, *Learning and generalization in overparameterized neural networks, going beyond two layers*, ArXiv preprint, arXiv:1811.04918, 2018. 2.2, 7

[2] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68(3):337–404, 1950. 3.1

[3] S. Arora, S.S. Du, W. Hu, Z. Li, and R. Wang, *Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks*, ArXiv preprint arXiv:1901.08584, 2019. 2.2

[4] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, *Stronger generalization bounds for deep nets via a compression approach*, Proc. Int. Conf. Mach. Learn., 80:254–263, 2018. 1, 6.1

[5] A.R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inf. Theory, 39(3):930–945, 1993. 3.1, 3.2

[6] A.R. Barron, *Approximation and estimation bounds for artificial neural networks*, Mach. Learn., 14(1):115–133, 1994. 2, 2.1

[7] P.L. Bartlett, D.J. Foster, and M.J. Telgarsky, *Spectrally-normalized margin bounds for neural networks*, Adv. Neural. Inf. Process. Syst., 30:6240–6249, 2017. 1

[8] L. Breiman, *Hinging hyperplanes for regression, classification, and function approximation*, IEEE Trans. Inf. Theory, 39(3):999–1013, 1993. 3.1, 3.2

[9] A. Brutzkus, A. Globerson, E. Malach, and S. Shalev-Shwartz, *SGD learns over-parameterized networks that provably generalize on linearly separable data*, in International Conference on Learning Representations, 2018. 2.2, 7

[10] Y. Cao and Q. Gu, *A generalization theory of gradient descent for learning over-parameterized deep ReLu networks*, ArXiv preprint, arXiv:1902.01384, 2019. 2.2

[11] A. Caponnetto and E. De Vito, *Optimal rates for the regularized least-squares algorithm*, Found. Comput. Math., 7(3):331–368, 2007. 4.1

[12] L. Chizat and F. Bach, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, Adv. Neural Inf. Process. Syst., 31:3040–3050, 2018. 2

[13] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, SIAM, Philadelphia, USA 2002. 1

[14] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Math. Control Signal. Systems, 2(4):303–314, 1989. 3.1

[15] A. Daniely, *SGD learns the conjugate kernel class of the network*, Adv. Neural Inf. Process. Syst., 30:2422–2430, 2017. 2.2, 7

[16] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer Science & Business Media, 31, 2013. 4.3

[17] S.S. Du, X. Zhai, B. Poczos, and A. Singh, *Gradient descent provably optimizes over-parameterized neural networks*, in International Conference on Learning Representations, 2019. 2

[18] G.K. Dziugaite and D.M. Roy, *Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data*, Proc. Conf. Uncertain. Artif. Intell., 2017. 1, 6.1, 6.1

[19] W. E., C. Ma, and Q. Wang, *A priori estimates of the population risk for residual networks*, ArXiv preprint, arXiv:1903.02154, 2019. 1, 7

[20] W. E., C. Ma, Q. Wang, and L. Wu, *Analysis of the gradient descent algorithm for a deep neural network model with skip-connections*, ArXiv preprint, arXiv:1904.05263, 2019. 1, 2.2, 7

[21] W. E., C. Ma, and L. Wu, *Barron spaces and compositional function spaces for neural network models*, ArXiv preprint, arXiv:1906.08039, 2019. 1, 3.1

[22] W. E., C. Ma, and L. Wu, *A comparative analysis of the optimization and generalization property of two-layer neural network and random feature models under gradient descent dynamics*, ArXiv preprint, arXiv:1904.04326, 2019. 1, 2.2, 7

[23] C.D. Freeman and J. Bruna, *Topology and geometry of half-rectified network optimization*, in International Conference on Learning Representations, 2017. 2

[24] N. Golowich, A. Rakhlin, and O. Shamir, *Size-independent sample complexity of neural networks*, in Proceedings of the 31st Conference on Learning Theory, 75:297–299, 2018. 1

[25] K. He, X. Zhang, S. Ren, and J. Sun, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, Proc. IEEE Int. Conf. Comput. Vis., 1026–1034, 2015. 3

[26] D.P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, in International Conference on Learning Representations, 2015. 6

[27] J.M. Klusowski and A.R. Barron, *Risk bounds for high-dimensional ridge function combinations including neural networks*, arXiv preprint, arXiv:1607.01434, 2016. 2, 2.1, 3.1, 3.1

[28] J.M. Klusowski and A.R. Barron, *Minimax lower bounds for ridge combinations including neural nets*, Proc. IEEE Int. Symp. Info. Theory, 1376–1380, 2017. 3.1, 4

[29] J.M. Klusowski and A.R. Barron, *Approximation by combinations of ReLu and squared ReLu ridge functions with l^1 and l^0 controls*, IEEE Trans. Inf. Theory, 64(12):7649–7656, 2018. 3.1

[30] A. Krizhevsky, I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks*, Adv. Neural Inf. Process. Syst., 25:1097–1105, 2012. 3

[31] S. Mei, A. Montanari, and P.-M. Nguyen, *A mean field view of the landscape of two-layers neural networks*, Proc. Natl. Acad. Sci. USA, 115:E7665–E7671, 2018. 2

[32] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro, *Exploring generalization in deep learning*, Adv. Neural Inf. Process. Syst., 30:5949–5958, 2017. 1

[33] B. Neyshabur, S. Bhojanapalli, and N. Srebro, *A PAC-Bayesian approach to spectrally-normalized*

margin bounds for neural networks, in International Conference on Learning Representations, 2018. 1

[34] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro, *The role of over-parametrization in generalization of neural networks*, in International Conference on Learning Representations, 2019. 1, 2.2, 6.1, 6.1

[35] B. Neyshabur, R. Tomioka, and N. Srebro, *Norm-based capacity control in neural networks*, in Conference on Learning Theory, 1376–1401, 2015. 1, 3, 5.1

[36] Q. Nguyen, M.C. Mukkamala, and M. Hein, *On the loss landscape of a class of deep neural networks with no bad local valleys*, in International Conference on Learning Representations, 2019. 2

[37] A. Rahimi and B. Recht, *Uniform approximation of functions with random bases*, in 46th Annual Allerton Conference on Communication, Control, and Computing, IEEE, 555–561, 2008. 3.1

[38] I. Safran and O. Shamir, *On the quality of the initial basin in overspecified neural networks*, in International Conference on Machine Learning, 774–782, 2016. 2

[39] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory To Algorithms*, Cambridge University Press, 2014. 5.1, B.1, B.2

[40] A.N. Tikhonov and V.Y. Arsenin, *Solutions of Ill-Posed Problems*, V. H. Winston & Sons, Washington, 1977. 7

[41] C. Wei, J.D. Lee, Q. Liu, and T. Ma, *On the margin theory of feedforward neural networks*, ArXiv preprint, arXiv:1810.05369, 2018. 2.1

[42] Y. Yang and A. Barron, *Information-theoretic determination of minimax rates of convergence*, Ann. Stat., 27(5):1564–1599, 1999. 4

[43] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning requires rethinking generalization*, in International Conference on Learning Representations, 2017. 1