Journal of Machine Learning Research 27 (2026) 1-35 Submitted 11/24; Revised 9/25; Published 1/26

Optimization and Generalization of Gradient Descent for Shallow ReLU Networks with Minimal Width

Yunwen Lei LEIYW@HKU.HK
Department of Mathematics
The University of Hong Kong

Puyu Wang WANGPUYU1026@GMAIL.COM
Department of Computer Science
RPTU Kaiserslautern-Landau

Yiming Ying YIMING.YING@SYDNEY.EDU.AU
School of Mathematics and Statistics
The University of Sydney

Ding-Xuan Zhou DINGXUAN.ZHOU@SYDNEY.EDU.AU
School of Mathematics and Statistics
The University of Sydney

Editor: Mahdi Soltanolkotabi

Abstract

Understanding the generalization and optimization of neural networks is a longstanding problem in modern learning theory. The prior analysis often leads to risk bounds of order $1/\sqrt{n}$ for ReLU networks, where n is the sample size. In this paper, we present a general optimization and generalization analysis for gradient descent applied to shallow ReLU networks. We develop convergence rates of the order $1/T$ for gradient descent with T iterations, and show that the gradient descent iterates fall inside local balls around either an initialization point or a reference point. Then we develop improved Rademacher complexity estimates by using the activation pattern of the ReLU function in these local balls. We apply our general result to NTK-separable data with a margin γ, and develop an almost optimal risk bound of the order $1/(n\gamma^2)$ for the ReLU network with a polylogarithmic width.

Keywords: Learning theory, Shallow neural networks, Generalization analysis, Gradient descent, Rademacher complexity

1 Introduction

Overparameterized neural networks have found impressive success in solving challenging problems in various fields. Surprisingly, these models not only achieve vanishing training errors, but also generalize well for prediction on unseen examples (Zhang et al., 2021). Despite extensive study, this mysterious generalization phenomenon is still not well understood. Furthermore, the associated objective functions are typically non-convex and non-smooth, which also poses significant difficulties for convergence analysis. These challenges motivate the development of modern techniques to study the generalization and optimization for learning with neural networks (Jacot et al., 2018; Allen-Zhu et al., 2019b,a; Bao et al.,

2024; Ji et al., 2020; Dziugaite and Roy, 2017; Bartolucci et al., 2023; Wu et al., 2024; Holzmüller and Steinwart, 2022).

A powerful technique to study the behavior of neural networks with the rectified linear unit (ReLU) activation is the neural tangent kernel (NTK) (Jacot et al., 2018). The intuition is that if the network width is sufficiently large, then the gradient of the neural network would stay almost as a constant in a ball around the initialization point with a small radius (Jacot et al., 2018; Chizat et al., 2019). Consequently, the training trajectory of gradient descent on neural networks can be well approximated by that of a function in the reproducing kernel Hilbert space (RKHS) generated by the NTK. Building on this perspective, linear convergence of gradient methods has been established for both shallow and deep networks in regression tasks (Arora et al., 2019; Du et al., 2019; Zou et al., 2020). However, these results typically require the width m to grow as a polynomial function of the sample size n (Allen-Zhu et al., 2019b; Zou et al., 2020; Arora et al., 2019; Cao and Gu, 2019). This overparameterization requirement is substantially strong as compared to practice, where a much more moderate overparameterization is used and achieves impressive optimization and generalization guarantee.

Recently, there is an increasing interest in studying the generalization behavior of shallow neural networks (SNNs) via another technique called algorithmic stability. The work (Liu et al., 2020) showed that the smallest eigenvalue of the associated Hessian matrix is larger than $-\frac{1}{\sqrt{m}}$ (up to a constant factor), which motivates recent stability analysis under an assumption that the width m increases polynomially with respect to (w.r.t.) the sample size n (Richards and Kuzborskij, 2021; Lei et al., 2022).

For classification tasks, a logarithmic dependency has been established for SNNs under NTK-separable data with margin γ (Ji and Telgarsky, 2019; Chen et al., 2021; Taheri and Thrampoulidis, 2024). The work (Ji and Telgarsky, 2019) employs a uniform convergence approach based on Rademacher complexity, focusing on a specific logistic regression problem. However, it remains unclear whether their analysis extends to shallow ReLU networks in more general settings. Moreover, their generalization analysis yields risk bounds of order $\sqrt{\log n}/(\sqrt{n}\gamma^2)$, which is worse than the optimal rate by a factor of $\sqrt{n}$. In contrast, the work (Taheri and Thrampoulidis, 2024) adopts an algorithmic stability framework and achieves the optimal bound of order $1/(n\gamma^2)$, whose key argument is to show that SNNs admit the self-bounded weak-convexity property. Nevertheless, this stability analysis critically relies on the use of smooth activation functions and often imply bounds in expectation. These limitations naturally motivate the following questions:

Can we develop optimization and generalization bounds with high probability for gradient descent when training shallow ReLU networks in a more general problem setup?
Furthermore, can such general analysis yield optimal bounds for NTK-separable data with polylogarithmic network width?

In this paper, we answer the above questions affirmatively by providing both optimization and generalization analyses of gradient descent for training shallow ReLU networks. We summarize our main contributions as follows.

1. We present convergence rates for shallow ReLU networks for a general self-bounding and smooth loss function. Specifically, we show that gradient descent achieves the

convergence rate of order $\widetilde{O}\big(\mathfrak{E}_S(\mathbf{W}^*)/(\eta T)\big)$, where $\mathfrak{E}_S(\mathbf{W}^*) = 3\eta T F_S(\mathbf{W}^*) + \|\mathbf{W}_1 - \mathbf{W}^*\|_2^2$, F_S is the training error, T is the iteration number, $\mathbf{W}^*$ is a reference model and $\mathbf{W}_1$ is the initialization point. Moreover, the iterators stay inside local balls with radius as measured by either the Euclidean norm or the $(2,\infty)$-norm.

2. We give improved Rademacher complexity estimates for a hypothesis space containing the gradient descent iterates. The key idea is to decompose the complexity into two terms: one term associated with relatively few neurons, controlled by a $(2,\infty)$-norm constraint, while the other term contains almost linear functions and can be controlled by a constraint on the Euclidean norm. We then develop optimistic risk bounds of order $\widetilde{O}(\mathfrak{E}(\mathbf{W}^*)/n)$, where $\mathfrak{E}(\mathbf{W}^*)$ is a population counterpart of $\mathfrak{E}_S(\mathbf{W}^*)$. As compared to the bounds stated in expectation in Taheri and Thrampoulidis (2024), we develop high-probability bounds which can shed high-order moment information of the excess risks.

3. To compare our result with the existing NTK analysis, we apply our general result to NTK-separable data with a margin γ. In particular, we show that shallow ReLU networks with a polylogarithmic width can achieve the risk bounds of order $\widetilde{O}\big(1/(n\gamma^2)\big)$, which is almost optimal as illustrated in Shamir (2021); Schliserman and Koren (2024). This improves the existing risk bounds of order $\widetilde{O}\big(1/(\sqrt{n}\gamma^2)\big)$ (Ji and Telgarsky, 2019) by a factor of $\sqrt{n}$.

We organize the paper as follows. We discuss the related work on neural networks in Section 2. The problem is formulated in Section 3. We present our main results on the optimization and generalization in Section 4. We provide the proofs in Section 5. The conclusion is given in Section 6.

2 Related Work

2.1 Optimization

The concept of the NTK was introduced in a seminal paper (Jacot et al., 2018), which was used to study the global convergence of shallow ReLU networks (Arora et al., 2019). These discussions require $m \gtrsim n^6/\lambda_0^4$, where λ_0 is the smallest eigenvalue of the Gram matrix associated with the ReLU kernel. The work (Cao and Gu, 2019) introduced the neural tangent random feature, and showed the convergence of one-pass SGD in the case $m \gtrsim n^7$. The work (Allen-Zhu et al., 2019b) showed that deep ReLU networks enjoy a semi-smoothness property, and established the global convergence provided that m grows as a polynomial function of n. This result was improved in Zou et al. (2020); Chen et al. (2021) by relaxing the overparameterization requirement. For quadratic loss functions, the works (Liu et al., 2022; Soltanolkotabi et al., 2018) established the PL condition for smooth activations, while the work (Oymak and Soltanolkotabi, 2020) showed the Lipschitzness of the Jacobian matrix for SNNs. These discussions further reduce the overparameterization requirement. For example, it was shown in Oymak and Soltanolkotabi (2020) that gradient descent converges under the quadratic loss if the square-root of the number of parameters exceeds the sample size by a constant factor. Mild overparameterization was also studied for a simpler student-teacher setting (Zhou et al., 2021; Safran et al., 2021). Information theoretical lower

bounds on the overparameterization requirement have also been studied (Bombari et al., 2022).

2.2 Generalization

The uniform convergence of training errors to testing errors for neural networks has been widely studied based on capacity measures such as Rademacher complexities and covering numbers (Neyshabur et al., 2015; Bartlett et al., 2017; Ledent et al., 2021; Golowich et al., 2018; Zhou and Huo, 2024; Liu et al., 2024; Mao and Zhou, 2023; Frei et al., 2023; Yang and Zhou, 2025, 2024; Parhi and Nowak, 2022). For gradient methods, the generalization bounds $\widetilde{O}(1/\sqrt{n})$ have been established based on the uniform convergence approach (Chen et al., 2021; Nitanda et al., 2019; Ji and Telgarsky, 2019; Arora et al., 2019; Chen et al., 2020). Recently, there is growing interest in studying the generalization of SNNs based on algorithmic stability (Richards and Kuzborskij, 2021; Lei et al., 2022; Taheri and Thrampoulidis, 2024; Deora et al., 2024; Lei, 2023; Wang et al., 2025). The work (Liu et al., 2020) showed that empirical risks are weakly convex with the convexity parameter decaying with the order $O(1/\sqrt{m})$. This observation was used to derive the stability bound of $O(T/n)$ for training SNNs with the quadratic loss (Richards and Kuzborskij, 2021; Lei et al., 2022), for which a polynomial width is required. The recent work (Taheri and Thrampoulidis, 2024; Deora et al., 2024) refines the curvature analysis of Richards and Kuzborskij (2021); Lei et al. (2022) by showing that the empirical risks enjoy a self-bound weak convexity, meaning that the smallest eigenvalue decays with the order $O(F_S(\mathbf{W}_t)/\sqrt{m})$ at the t-th iteration. This key observation allows them to derive stability bounds of order $O((\log T)/n)$ for a polylogarithmic width. All these stability analyses focused on the smooth activation function, which cannot apply to the ReLU activation which is widely used in practice. Indeed, their key idea is to control the smallest eigenvalue of the Hessian matrix, which does not exist for the ReLU networks.

3 Shallow ReLU Networks

Let ρ be a probability measure defined on a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is an input space and $\mathcal{Y}$ is an output space. We consider binary classification where $\mathcal{Y} = \{-1, 1\}$. Let $S = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ be a dataset drawn independently from ρ. Let $\sigma(a) = \max\{a, 0\}$ be the ReLU activation function and $\mathbf{w}^\top$ denote the transpose of $\mathbf{w}$. We consider SNNs of the following form

$$\Phi(\mathbf{W}; \mathbf{x}) = \sum_{j=1}^{m} a_j \sigma(\mathbf{w}_j^\top \mathbf{x}),$$

where $\mathbf{x} \in \mathcal{X}$ is the input data, m is the number of nodes in the hidden layer, $a_j \in \{-1/\sqrt{m}, 1/\sqrt{m}\}$ indicates the connection weight between the j-th node in the hidden layer to the node in the output layer, and $\mathbf{w}_j \in \mathbb{R}^d$ denotes the connection weight between the j-th hidden node and the nodes in the input layer. We collect all $\mathbf{w}_j$ into a column vector $\mathbf{W} = (\mathbf{w}_1^\top, \ldots, \mathbf{w}_m^\top)^\top \in \mathbb{R}^{dm}$. Let $\ell : \mathbb{R} \mapsto \mathbb{R}_+$ and use $f(\mathbf{W}; \mathbf{z})$ to denote the loss suffered by using $\Phi(\mathbf{W}; \cdot)$ to do prediction at $\mathbf{z} = (\mathbf{x}, y)$, i.e.,

$$f(\mathbf{W}; \mathbf{z}) = \ell(y\Phi(\mathbf{W}; \mathbf{x})).$$

The empirical and testing behaviors of a model $\mathbf{W}$ are then quantified by the empirical risk $F_S(\mathbf{W})$ and the population risk $F(\mathbf{W})$, respectively:

$$F_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{W}; \mathbf{z}_i) \quad \text{and} \quad F(\mathbf{W}) = \mathbb{E}_{\mathbf{z}}[f(\mathbf{W}; \mathbf{z})],$$

where $\mathbb{E}_{\mathbf{z}}[\cdot]$ denotes the expectation w.r.t. $\mathbf{z} \sim \rho$. In this paper, we fix the output weights $a_j \in \{-1/\sqrt{m}, 1/\sqrt{m}\}$, and only train the weights $\mathbf{W}$ by gradient descent (Ji and Telgarsky, 2019; Kuzborskij and Szepesvári, 2022; Arora et al., 2019). For simplicity, we assume m is an even number and consider the symmetric initialization. Specifically, we initialize $a_j \in \{-1, +1\}$ for $j \leq m/2$ and set $a_j = -a_{j-m/2}$ for $j > m/2$. We set $\mathbf{W}_1 = (\mathbf{w}_{1,1}^\top, \ldots, \mathbf{w}_{1,m}^\top)^\top \in \mathbb{R}^{dm}$ with $\mathbf{w}_{1,j} \sim N(0, I_d)$ for $j \leq m/2$ and $\mathbf{w}_{1,j} = \mathbf{w}_{1,j-m/2}$ for $j > m/2$, where $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix.

Definition 1 (Gradient descent) *Let $\eta > 0$. At the t-th iteration, we update*

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \nabla F_S(\mathbf{W}_t), \quad t \in [T] := \{1, \ldots, T\}. \tag{3.1}$$

Note F_S is not differentiable. We choose a subgradient in gradient descent as follows

$$\mathbf{w}_{t+1,j} = \mathbf{w}_{t,j} - \frac{\eta a_j}{n}\sum_{i=1}^{n} y_i \ell'(y_i \Phi(\mathbf{W}_t; \mathbf{x}_i)) \mathbb{I}_{[\mathbf{w}_{t,j}^\top \mathbf{x}_i \geq 0]} \mathbf{x}_i, \quad \forall j \in [m], \tag{3.2}$$

where $\ell'(a)$ denotes the derivative of ℓ at a and $\mathbb{I}_{[\cdot]}$ denotes the indicator function (i.e., taking the value 1 if the argument holds true, and 0 otherwise). For convenience, we denote

$$\nabla \Phi(\mathbf{W}; \mathbf{x}) = \begin{pmatrix} a_1 \mathbf{x} \mathbb{I}_{[\mathbf{w}_1^\top \mathbf{x} \geq 0]} \\ a_2 \mathbf{x} \mathbb{I}_{[\mathbf{w}_2^\top \mathbf{x} \geq 0]} \\ \vdots \\ a_m \mathbf{x} \mathbb{I}_{[\mathbf{w}_m^\top \mathbf{x} \geq 0]} \end{pmatrix} \in \mathbb{R}^{dm}. \tag{3.3}$$

Then, Eq. (3.2) can be written as

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \frac{\eta}{n}\sum_{i=1}^{n} y_i \ell'(y_i \Phi(\mathbf{W}_t; \mathbf{x}_i)) \nabla \Phi(\mathbf{W}_t; \mathbf{x}_i).$$

Our analysis requires some standard assumptions on the loss ℓ and the input $\mathbf{x}$. We say ℓ is convex if $\ell(a) \geq \ell(a') + (a - a')\ell'(a')$ for any $a, a' \in \mathbb{R}$.

Assumption 1 *We assume the function $a \mapsto \ell(a)$ is convex, nonnegative and L-smooth, i.e., $|\ell'(a) - \ell'(a')| \leq L|a - a'|$ for all $a, a' \in \mathbb{R}$. Furthermore, assume for any a, $|\ell'(a)| \leq \widetilde{L}\ell(a)$ for some $\widetilde{L} > 0$.*

Assumption 1 is satisfied by the logistic loss $\ell(a) = \log(1+\exp(-a))$ (Ji and Telgarsky, 2019; Schliserman and Koren, 2022). Other examples include the polynomially-tailed losses and sub-exponential tailed losses (Schliserman and Koren, 2022). Hence, our analysis applies to a broad class of loss functions, whereas Ji and Telgarsky (2019) focused exclusively on the logistic loss. For simplicity, we always assume $\widetilde{L} = 1$ throughout the paper. We let $\|\cdot\|_2$ denote the Euclidean norm. The following assumption is standard in learning with SNNs (Ji and Telgarsky, 2019; Chen et al., 2021; Arora et al., 2019).

Assumption 2 *We assume* $\|\mathbf{x}\|_2 = 1$ *for all* $\mathbf{x} \in \mathcal{X}$.

Under these assumptions, we can build the following self-bounding property of f

$$\|\nabla f(\mathbf{W}; \mathbf{z})\|_2 \leq \|\nabla \Phi(\mathbf{W}; \mathbf{x})\|_2 |\ell'(y\Phi(\mathbf{W}; \mathbf{x}))| \leq \ell(y\Phi(\mathbf{W}; \mathbf{x})) = f(\mathbf{W}; \mathbf{z}), \tag{3.4}$$

where we have used the fact $\|\nabla \Phi(\mathbf{W}; \mathbf{x})\|_2 \leq 1$ and $|\ell'(a)| \leq \ell(a)$.

Error decomposition. In this paper, we are interested in the performance of gradient descent iterators as measured by the population risk. Our basic idea is to decompose $F(\mathbf{W}_T)$ as

$$F(\mathbf{W}_T) = \big(F(\mathbf{W}_T) - F_S(\mathbf{W}_T)\big) + F_S(\mathbf{W}_T).$$

We call the first term $F(\mathbf{W}_T) - F_S(\mathbf{W}_T)$ the generalization gap, which measures the difference between training and testing. We call the second term $F_S(\mathbf{W}_T)$ the optimization error since for SNNs we often encounter overparameterization where the best SNN achieves a zero training error. We will use tools in optimization theory to control the optimization error (Orabona, 2019; Jin et al., 2021), and use the Rademacher complexity to control the generalization gap (Bartlett and Mendelson, 2002; Zhang, 2023; Steinwart and Christmann, 2008).

4 Main Results

4.1 Optimization Analysis

In this section, we present results on optimization. The following lemma shows the behavior of the network initialization.

Lemma 2 (Ji and Telgarsky 2019) *Let* $R > 0$. *For any* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$ *w.r.t. the randomization of* $\mathbf{W}_1$, *it holds simultaneously for all* $i \in [n]$ *that*

$$\big|\big\{j \in [m] : |\mathbf{w}_{1,j}^\top \mathbf{x}_i| \leq R\big\}\big| \leq 2Rm/\sqrt{\pi} + \big(m \log(n/\delta)\big)^{\frac{1}{2}} =: C_R, \tag{4.1}$$

where $|\cdot|$ *denotes the cardinality of a set.*

We denote $A \lesssim B$ if there exists a universal constant C such that $A \leq CB$, and $A \gtrsim B$ if there exists a universal constant C such that $A \geq CB$. We denote $A \asymp B$ if $A \lesssim B$ and $A \gtrsim B$. For any $\mathbf{W}, \mathbf{W}'$, we denote $\|\mathbf{W} - \mathbf{W}'\|_{2,\infty} = \max_{j \in [m]} \|\mathbf{w}_j - \mathbf{w}_j'\|_2$. We consider a reference model $\mathbf{W}^*$ and introduce the notation

$$\mathfrak{C}_S(\mathbf{W}^*) := 3\eta T F_S(\mathbf{W}^*) + \|\mathbf{W}_1 - \mathbf{W}^*\|_2^2, \quad \widetilde{\mathfrak{C}}_S(\mathbf{W}^*) = \frac{1}{n}\sum_{i=1}^n |\ell'(y_i \Phi(\mathbf{W}^*; \mathbf{x}_i))|. \tag{4.2}$$

Let R_1 be a number independent of $\mathbf{W}_1$ and E_1 be the event (w.r.t. $\mathbf{W}_1$) such that Eq. (4.1) holds with $R = R_1$ and

$$\max\big\{\mathfrak{C}_S(\mathbf{W}^*)/\sqrt{m}, \|\mathbf{W}^* - \mathbf{W}_1\|_{2,\infty}\big\} \leq R_1 \leq \frac{\sqrt{\pi}}{16\mathfrak{C}_S(\mathbf{W}^*)}. \tag{4.3}$$

We will show the event E_1 indeed happens for an appropriate R_1 independent of $\mathbf{W}_1$ with high probability under an NTK separability assumption. In particular, we can choose R_1 dependent on m and then Eq. (4.3) will hold under an overparameterization assumption.

Theorem 3, to be proved in Section 5.1, illustrates the convergence behavior of gradient descent. In particular, Eq. (4.4) gives convergence rates for the averaged empirical risk.

Theorem 3 *Let Assumptions 1, 2, and the event E_1 hold. Suppose $m \geq 64\log(n/\delta)\mathfrak{C}_S^2(\mathbf{W}^*)$. If $\eta \leq \min\{4/(5L), 1/(5\widetilde{\mathfrak{C}}_S(\mathbf{W}^*))\}$, then for all $t \in [T+1]$ it holds (we denote $\sum_{k=1}^0 F_S(\mathbf{W}_k) = 0$)*

$$\|\mathbf{W}_t - \mathbf{W}^*\|_2^2 \leq \mathfrak{C}_S(\mathbf{W}^*) \quad \text{and} \quad \eta \sum_{k=1}^{t-1} F_S(\mathbf{W}_k) \leq \mathfrak{C}_S(\mathbf{W}^*). \tag{4.4}$$

Furthermore, for all $t \in [T+1]$, we have

$$\|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty} \leq \frac{\mathfrak{C}_S(\mathbf{W}^*)}{\sqrt{m}}. \tag{4.5}$$

Remark 4 Other than convergence rates, Theorem 3 also shows that the gradient descent iterates would stay inside a local ball. Eq. (4.4) controls the distance between the iterates and the reference model as measured by the Euclidean norm, while Eq. (4.5) gives such estimates based on the $(2,\infty)$ norm. This shows that the SNNs traversed by gradient descent have a controlled complexity. As we will see in the generalization analysis, Eq. (4.4) and Eq. (4.5) are used to construct a hypothesis space, to which the gradient descent iterates would belong with high probability (over the randomness of $\mathbf{W}_1$).

Remark 5 (Comparison) The work (Taheri and Thrampoulidis, 2024) considered SNNs with a smooth activation function satisfying $|\sigma'(u)| \leq G_\sigma$ and $|\sigma''(u)| \leq L_\sigma$ for all $u \in \mathbb{R}$. If there is a reference model $\mathbf{W}^*$ satisfying $\|\mathbf{W}^* - \mathbf{W}_1\|_2^2 \geq \max\{\eta T F_S(\mathbf{W}^*), \eta F_S(\mathbf{W}_1)\}$ and $m \geq 18^2 L_\sigma^2 \|\mathbf{W}^* - \mathbf{W}_1\|_2^4$, then it was shown there

$$\|\mathbf{W}_t - \mathbf{W}^*\|_2 \leq 4\|\mathbf{W}^* - \mathbf{W}_1\|_2 \quad \text{and} \quad \eta \sum_{k=1}^{t} F_S(\mathbf{W}_k) \lesssim \mathfrak{C}_S(\mathbf{W}^*). \tag{4.6}$$

Note that the assumption $\|\mathbf{W}^* - \mathbf{W}_1\|_2^2 \geq \eta T F_S(\mathbf{W}^*)$ implies that

$$\mathfrak{C}_S(\mathbf{W}^*) = 3\eta T F_S(\mathbf{W}^*) + \|\mathbf{W}_1 - \mathbf{W}^*\|_2^2 \leq 4\|\mathbf{W}^* - \mathbf{W}_1\|_2^2 \leq 4\mathfrak{C}_S(\mathbf{W}^*).$$

Therefore, our bound in Eq. (4.4) for ReLU networks matches the optimization error bound in Taheri and Thrampoulidis (2024) for networks with a smooth activation. While their discussions did not consider the $\|\cdot\|_{2,\infty}$ norm, their bound immediately implies such a result. Indeed, similar to Eq. (3.2), we know the following identity for all $j \in [m]$

$$\mathbf{w}_{t+1,j} = \mathbf{w}_{1,j} - \sum_{k=1}^{t} \frac{\eta a_j}{n} \sum_{i=1}^{n} y_i \ell'(y_i \Phi(\mathbf{W}_k; \mathbf{x}_i)) \sigma'(\mathbf{w}_{k,j}^\top \mathbf{x}_i) \mathbf{x}_i.$$

Work	$\|\mathbf{W}_t - \mathbf{W}^*\|_2$	$\|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty}$	$\eta\sum_{k=1}^{t-1} F_S(\mathbf{W}_k)$
Taheri and Thrampoulidis (2024) with smooth activation	$O\big(\mathfrak{C}_S^{\frac{1}{2}}(\mathbf{W}^*)\big)$	$O\big(\mathfrak{C}_S(\mathbf{W}^*)/\sqrt{m}\big)$	$O\big(\mathfrak{C}_S(\mathbf{W}^*)\big)$
Ours with ReLU activation	$O\big(\mathfrak{C}_S^{\frac{1}{2}}(\mathbf{W}^*)\big)$	$O\big(\mathfrak{C}_S(\mathbf{W}^*)/\sqrt{m}\big)$	$O\big(\mathfrak{C}_S(\mathbf{W}^*)\big)$

Table 1: Comparison on convergence analysis. Both Taheri and Thrampoulidis (2024) and our work develop similar bounds on $\|\mathbf{W}_t - \mathbf{W}^*\|_2, \|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty}$ and $\eta\sum_{k=1}^{t-1} F_S(\mathbf{W}_k)$. Taheri and Thrampoulidis (2024) consider smooth and Lipschtiz activation function, and their analysis depends on a crucial self-bounded weak convexity. Our work focuses on the ReLU activation function, for which the self-bounded weak convexity does not hold.

It then follows from the inequality $|\ell'(a)| \leq \ell(a)$ that

$$\begin{aligned}\|\mathbf{w}_{t+1,j} - \mathbf{w}_{1,j}\|_2 &\leq \sum_{k=1}^{t}\frac{\eta}{n\sqrt{m}}\sum_{i=1}^{n}\Big\|y_i\ell'(y_i\Phi(\mathbf{W}_k;\mathbf{x}_i))\sigma'(\mathbf{w}_{k,j}^\top\mathbf{x}_i)\mathbf{x}_i\Big\|_2 \\ &\leq \sum_{k=1}^{t}\frac{\eta G_\sigma}{n\sqrt{m}}\sum_{i=1}^{n}\big|\ell'(y_i\Phi(\mathbf{W}_k;\mathbf{x}_i))\big| \leq \sum_{k=1}^{t}\frac{\eta G_\sigma}{n\sqrt{m}}\sum_{i=1}^{n}\ell(y_i\Phi(\mathbf{W}_k;\mathbf{x}_i)) \\ &= \frac{\eta G_\sigma}{\sqrt{m}}\sum_{k=1}^{t}F_S(\mathbf{W}_k) \lesssim \frac{G_\sigma\mathfrak{C}_S(\mathbf{W}^*)}{\sqrt{m}},\end{aligned}$$

where we have used Eq. (4.6) in the last step. Therefore, the analysis in Taheri and Thrampoulidis (2024) also implicitly establishes the $(2,\infty)$-norm constraint in Eq. (4.5). However, their approach relies critically on a self-bounded weak convexity condition: $\lambda_{\min}(\nabla^2 F_S(\mathbf{W})) \geq -\frac{L_\sigma}{\sqrt{m}}F_S(\mathbf{W})$, which only holds for smooth activation functions. As a comparison, we consider the nonsmooth ReLU activation function, for which the self-bounded weak convexity does not hold. To address this challenge, we leverage concentration inequalities and Gaussian initialization to approximate the ReLU function by linear functions around the initialization. We summarize the comparison in Table 1.

For the optimization error bounds discussed above, our overparameterization requirement depends on the quantity $\mathfrak{C}_S(\mathbf{W}^*)$. To clarify this constraint, we impose the following assumption of realizability of data by SNNs, meaning that there exists a model in the neighborhood of the initialization point with a small training error. This assumption has been considered in the literature for linear models (Schliserman and Koren, 2022) and SNNs with smooth activation functions (Taheri and Thrampoulidis, 2024). We will verify this assumption for SNNs with the ReLU activation function by constructing a function g under a margin condition. We consider $\mathbf{W}^* = \mathbf{W}^{\frac{1}{T}}$ in the proof of Theorem 6. For simplicity, we always assume $T \geq 4/L$. The proof of Theorem 6 will be given in Section 5.1.

Assumption 3 (Realizability) *Assume there exists a decreasing function $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that for any $\epsilon > 0$ there exists $\mathbf{W}^\epsilon \in \mathbb{R}^{dm}$ with*

$$F_S(\mathbf{W}^\epsilon) \leq \epsilon \quad \text{and} \quad \|\mathbf{W}^\epsilon - \mathbf{W}_1\|_2 \leq g(\epsilon).$$

Theorem 6 (Optimization under Realizability) *Let Assumptions 1, 2, 3 hold. If $\eta \leq 4/(5L)$, the event E_1 holds and $m \geq 64\log(n/\delta)\big(3/L + g^2(1/T)\big)^2$, then we can find $\mathbf{W}^{\frac{1}{T}}$ such that*

$$\|\mathbf{W}_t - \mathbf{W}^{\frac{1}{T}}\|_2^2 \leq 3\eta + g^2(1/T) \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T} F_S(\mathbf{W}_t) \leq \frac{3\eta + g^2(1/T)}{\eta T}. \tag{4.7}$$

Furthermore, there holds

$$\|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty} \leq \frac{3\eta + g^2(1/T)}{\sqrt{m}}. \tag{4.8}$$

Theorem 6 requires the happening of E_1. In the following corollary, we remove this assumption by showing that E_1 happens with high probability. As we will see for the NTK separable data, $\|\mathbf{W}^{\frac{1}{T}} - \mathbf{W}_1\|_{2,\infty}$ is of order $1/\sqrt{m}$, and thus Eq. (4.9) holds under an appropriate overparameterization assumption. For any $a, b \in \mathbb{R}$, we denote $a \vee b = \max\{a, b\}$. The proof will be given in Section 5.1.

Corollary 7 *Let $\delta \in (0,1)$ and Assumptions 1, 2, 3 hold. If $\eta \leq 4/(5L)$, $m \geq 64\log(n/\delta) \times \big(3/L + g^2(1/T)\big)^2 \vee \frac{256}{\pi}\big(3/L + g^2(1/T)\big)^4$ and*

$$\|\mathbf{W}^{\frac{1}{T}} - \mathbf{W}_1\|_{2,\infty} \leq \frac{\sqrt{\pi}}{16\big(3/L + g^2(1/T)\big)}, \tag{4.9}$$

then with probability at least $1-\delta$, the inequalities (4.7), (4.8) hold.

If $g(\epsilon) \lesssim \log(1/\epsilon)$, then Corollary 7 implies that

$$\|\mathbf{W}_t - \mathbf{W}^{\frac{1}{T}}\|_2^2 \lesssim \log^2 T, \quad \frac{1}{T}\sum_{t=1}^{T} F_S(\mathbf{W}_t) \lesssim \frac{\log^2 T}{\eta T} \quad \text{and} \quad \|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty} \lesssim \frac{\log^2 T}{\sqrt{m}}.$$

The constraint on m becomes $m \gtrsim \log^8 T$. This shows that a polylogarithmic width is sufficient to guarantee the convergence of gradient descent.

4.2 Generalization Analysis

In this subsection, we present generalization analysis based on the uniform convergence approach. To this aim, we use Rademacher complexity to measure the complexity of a function space.

Definition 8 (Rademacher complexity) *Let $\widetilde{\mathcal{F}}$ be a class of real-valued functions over a space $\mathcal{Z}$ and $\widetilde{S} = \{\mathbf{z}_i\}_{i=1}^n \subseteq \mathcal{Z}$. We define the* empirical *Rademacher complexity as*

$$\mathfrak{R}_{\widetilde{S}}(\widetilde{\mathcal{F}}) = \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{f\in\widetilde{\mathcal{F}}} \frac{1}{n}\sum_{i\in[n]} \epsilon_i f(\mathbf{z}_i)\Big],$$

where $\boldsymbol{\epsilon} = (\epsilon_i)_{i\in[n]} \sim \{\pm 1\}^n$ are independent Rademacher variables, i.e., taking values in $\{\pm 1\}$ with the same probability.

Let $b_* \geq \sup_{\mathbf{z}} f(\mathbf{W}^*; \mathbf{z})$ and

$$\mathfrak{C}(\mathbf{W}^*) = 3\eta T\Big(2F(\mathbf{W}^*) + \frac{7b_* \log(2/\delta)}{6n}\Big) + \|\mathbf{W}_1 - \mathbf{W}^*\|_2^2. \tag{4.10}$$

In this subsection, we assume the reference model $\mathbf{W}^*$ is independent of S, which, as we will see, is the case for learning with linear-separable data and NTK-separable data. We consider the following function space

$$\mathcal{F} := \{\mathbf{x} \mapsto \Psi(\mathbf{W}; \mathbf{x}) : \mathbf{W} \in \mathcal{W}\}, \tag{4.11}$$

where the set $\mathcal{W}$ is defined as follows

$$\mathcal{W} = \Big\{\mathbf{W} \in \mathbb{R}^{dm} : \|\mathbf{W} - \mathbf{W}^*\|_2^2 \leq \mathfrak{C}(\mathbf{W}^*), \|\mathbf{W} - \mathbf{W}_1\|_{2,\infty} \leq \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}\Big\}. \tag{4.12}$$

Here we use $\mathfrak{C}(\mathbf{W}^*)$ instead of $\mathfrak{C}_S(\mathbf{W}^*)$ to get a data-independent hypothesis space, which is essential for the Rademacher complexity analysis. As we will see in the proof, $\mathfrak{C}(\mathbf{W}^*)$ is a high-probability bound of $\mathfrak{C}_S(\mathbf{W}^*)$. Therefore, according to the discussions in the previous subsection, the gradient descent iterates will fall onto $\mathcal{W}$ with high probability. The following lemma, to be proved in Section 5.2, estimates the Rademacher complexity of $\mathcal{F}$ in terms of several parameters such as $L, \mathfrak{C}(\mathbf{W}^*)$, m and n. Let R_2 be a number independent of $\mathbf{W}_1$. Let E_2 be the event (w.r.t. $\mathbf{W}_1$) that Eq. (4.1) holds with $R = R_2$.

Lemma 9 *Let Assumptions 1, 2 hold. Let $\mathcal{F}$ and $\mathcal{W}$ be defined in Eq.* (4.11) *and Eq.* (4.12)*, respectively. If the event E_2 holds and $R_2 \geq \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}$, then*

$$\mathfrak{R}_{S,n}(\mathcal{F}) \leq \frac{\mathfrak{C}(\mathbf{W}^*)\big(m^{\frac{1}{4}}\sqrt{2R_2} + \log^{\frac{1}{4}}(n/\delta)\big)}{\sqrt{n}m^{\frac{1}{4}}} + \frac{\sqrt{\mathfrak{C}(\mathbf{W}^*)}}{\sqrt{n}}, \tag{4.13}$$

where

$$\mathfrak{R}_{S,n}(\mathcal{F}) = \sup_{\widetilde{S}\subseteq S:|\widetilde{S}|=n} \mathfrak{R}_{\widetilde{S}}(\mathcal{F}).$$

Remark 10 (Comparison) A typical choice of R_2 satisfies $R_2 \asymp \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}$ (we can construct R_2 independent of $\mathbf{W}_1$ such that $R_2 \asymp \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}$ with high probability). In this case, Eq. (4.13) becomes

$$\mathfrak{R}_{S,n}(\mathcal{F}) \lesssim \frac{\mathfrak{C}(\mathbf{W}^*)\big(\mathfrak{C}^{\frac{1}{2}}(\mathbf{W}^*) + \log^{\frac{1}{4}}(n/\delta)\big)}{\sqrt{n}m^{\frac{1}{4}}} + \frac{\sqrt{\mathfrak{C}(\mathbf{W}^*)}}{\sqrt{n}}. \tag{4.14}$$

We now compare this result with existing complexity analysis of SNNs. It was shown that $\mathfrak{R}_{\widetilde{S}}(\widetilde{\mathcal{F}}) \leq \mathfrak{C}(\mathbf{W}^*)/\sqrt{n}$ in Ji and Telgarsky (2019), where

$$\widetilde{\mathcal{F}} := \Big\{\mathbf{x} \mapsto \Psi(\mathbf{W}; \mathbf{x}) : \mathbf{W} \in \mathbb{R}^{md}, \|\mathbf{W} - \mathbf{W}_1\|_{2,\infty} \leq \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}\Big\}.$$

This space does not incorporate a constraint on $\|\mathbf{W} - \mathbf{W}^*\|_2^2$. We use a special property of the gradient descent to build $\mathcal{F}$, and use the constraint on $\|\mathbf{W} - \mathbf{W}^*\|_2^2$ to give an improved Rademacher complexity estimate in Lemma 9. Note that the first term on the right-hand

side of Eq. (4.14) outperforms $\mathfrak{C}(\mathbf{W}^*)/\sqrt{n}$ by a factor of $\mathfrak{C}^{\frac{1}{2}}(\mathbf{W}^*)/m^{\frac{1}{4}}$, while the last term is better by a factor of $\mathfrak{C}^{-\frac{1}{2}}(\mathbf{W}^*)$. Therefore, our Rademacher complexity bound improves the existing one by a factor of $\mathfrak{C}^{\frac{1}{2}}(\mathbf{W}^*)/m^{\frac{1}{4}} \vee \mathfrak{C}^{-\frac{1}{2}}(\mathbf{W}^*)$. As we will see, this improved Rademacher complexity estimate plays a key role in deriving almost optimal risk bounds under a NTK separability assumption.

Remark 11 (Idea) A key challenge in estimating the Rademacher complexity is to handle the constraint involving the reference model. Indeed, due to the implicit regularization of gradient descent, the distance from the reference model may be much smaller than the norm itself (Dziugaite and Roy, 2017). We cannot use the homogeneity of the ReLU activation to fully use these constraints, and have to introduce new techniques. Let $\widetilde{S} = \{\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n\} \subseteq S$. Our key idea in estimating $\mathfrak{R}_{\widetilde{S}}(\mathcal{F})$ is to introduce the following decomposition

$$\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{j\in[m]}\sum_{i\in[n]}\epsilon_i a_j\big(\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_j)-\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_{1,j})\big)\Big] \le \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{j\in[m]}\sum_{i\in\widetilde{S}_j}\epsilon_i a_j\big(\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_j)-\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_{1,j})\big)\Big]$$
$$+ \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{j\in[m]}\sum_{i\notin\widetilde{S}_j}\epsilon_i a_j\big(\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_j)-\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_{1,j})\big)\Big] := I_1 + I_2, \quad (4.15)$$

where we introduce $\widetilde{S}_j := \big\{i \in [n] : |\tilde{\mathbf{x}}_i^\top\mathbf{w}_{1,j}| \le R_2\big\}$ for any $j \in [m]$. For I_1, the term $|\widetilde{S}_j|$ is relatively small and a crude estimate is enough. We use the constraint $\|\mathbf{W} - \mathbf{W}_1\|_{2,\infty} \le \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}$ to show that $I_1 = \widetilde{O}\big(\mathfrak{C}(\mathbf{W}^*)\sqrt{R_2}/\sqrt{n}\big)$, where the notation $\widetilde{O}(\cdot)$ ignores logarithmic factors. For I_2, a key observation is that $\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_j) - \sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_{1,j}) = \xi_{ij}\tilde{\mathbf{x}}_i^\top(\mathbf{w}_j - \mathbf{w}_{1,j})$, where $\xi_{ij} \in \{0, 1\}$. This shows that the summands in I_2 enjoy a linear property (one can exchange σ and the substraction), and we can use the Rademacher complexity for linear function classes to show that $I_2 \le \sqrt{\mathfrak{C}(\mathbf{W}^*)}/\sqrt{n}$. This linear property allows us to use the constraint $\|\mathbf{W} - \mathbf{W}^*\|_2^2 \le \mathfrak{C}(\mathbf{W}^*)$ in the estimation of I_2, where the correlation among $\mathbf{w}_j$ is preserved under this constraint.

We combine Lemma 9 with the optimistic bounds for smooth loss functions established in Srebro et al. (2010) to obtain generalization bounds. By further combining these generalization bounds with the optimization error bounds in Theorem 3, we derive the following risk guarantees. Let E_3 be the event that the following inequality holds simultaneously for all $\mathbf{W} \in \mathcal{W}$

$$F(\mathbf{W}) - F_S(\mathbf{W}) \lesssim F_S^{\frac{1}{2}}(\mathbf{W})\Big(\sqrt{L\log^3 n}\,\mathfrak{R}_{S,n}(\mathcal{F}) + \Big(\frac{b\log(2/\delta)}{n}\Big)^{\frac{1}{2}}\Big) + L\log^3 n\,\mathfrak{R}_{S,n}^2(\mathcal{F}) + \frac{b\log(2/\delta)}{n}$$

and

$$\big|F_S(\mathbf{W}^*) - F(\mathbf{W}^*)\big| \le \frac{2b_*\log(2/\delta)}{3n} + \Big(\frac{2b_*F(\mathbf{W}^*)\log(2/\delta)}{n}\Big)^{\frac{1}{2}},$$

where $b := 2b_* + L\mathfrak{C}(\mathbf{W}^*)$. The proof will be given in Section 5.2.

Theorem 12 (Risk bounds) *Let assumptions in Theorem 3 hold and $R_2 \ge \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}$. Under the event $E_1 \cap E_2 \cap E_3$, we have*

$$\frac{1}{T}\sum_{t=1}^T F(\mathbf{W}_t) \lesssim \frac{L\log^3 n}{n}\Big(\frac{\mathfrak{C}^2(\mathbf{W}^*)(\sqrt{m}R_2 + \log^{\frac{1}{2}}(n/\delta))}{\sqrt{m}} + \mathfrak{C}(\mathbf{W}^*)\Big) + \frac{b_*\log(2/\delta)}{n} + \frac{\mathfrak{C}(\mathbf{W}^*)}{\eta T}.$$

Remark 13 By the definition of $\mathfrak{C}(\mathbf{W}^*)$ and Bernstein inequality (i.e., Eq. (5.27) below), we know that

$$\mathfrak{C}(\mathbf{W}^*) \leq 3\eta T\Big(4F_S(\mathbf{W}^*) + \frac{8b_* \log(2/\delta)}{n}\Big) + \|\mathbf{W}_1 - \mathbf{W}^*\|_2^2 \leq 4\mathfrak{C}_S(\mathbf{W}^*) + \frac{24\eta T b_* \log(2/\delta)}{n}.$$

Therefore, if $\eta T \asymp n$ and $m \gtrsim (\eta T b_*/n)^4 \asymp b_*^4$, the constraint $m \gtrsim \mathfrak{C}_S^4(\mathbf{W}^*)$ implies

$$\frac{\mathfrak{C}^2(\mathbf{W}^*)}{\sqrt{m}} \lesssim 1 + \frac{\eta^2 T^2 b_*^2 \log^2(2/\delta)}{n^2\sqrt{m}} \lesssim 1.$$

Then, if we choose $R_2 = \widetilde{O}(\mathfrak{C}(\mathbf{W}^*)/\sqrt{m})$, Theorem 12 implies that

$$\frac{1}{T}\sum_{t=1}^T F(\mathbf{W}_t) = \widetilde{O}\Big(\frac{\mathfrak{C}(\mathbf{W}^*)}{n} + \frac{b_* \log(2/\delta)}{n}\Big) = \widetilde{O}\Big(F(\mathbf{W}^*) + \frac{\|\mathbf{W}_1 - \mathbf{W}^*\|_2^2 + b_* \log(2/\delta)}{\eta T}\Big). \tag{4.16}$$

We can choose $\mathbf{W}^* = \arg\min_{\mathbf{W}} \big\{F(\mathbf{W}) + \frac{\|\mathbf{W}_1 - \mathbf{W}\|_2^2}{\eta T}\big\}$. This shows that our analysis implies effective excess risk bounds as long as $\inf_{\mathbf{W}} \big\{F(\mathbf{W}) + \frac{\|\mathbf{W}_1 - \mathbf{W}\|_2^2}{\eta T}\big\}$ is small, which shows an implicit regularization of gradient descent in favoring good models around the initialization (Oymak and Soltanolkotabi, 2019).

Finally, we provide risk bounds in an interpolation setting. The proof will be given in Section 5.2.

Theorem 14 (Risk bounds under Realizability) *Let Assumptions 1, 2, 3 hold with $\mathbf{W}^\epsilon$ independent of S. Let $R_2 \geq \mathfrak{C}(\mathbf{W}^{\frac{1}{T}})/\sqrt{m}$ and ℓ be Lipscthiz continuous (i.e. $|\ell(a) - \ell(b)| \lesssim |a-b|$). If $\eta \leq 4L/5, \eta T \asymp n$ and $m \geq 64\log(n/\delta)\big(3/L + g^2(1/T)\big)^2 \vee \frac{256}{\pi}\big(3/L + g^2(1/T)\big)^4$, then under the event $E_1 \cap E_2 \cap E_3$ we have*

$$\frac{1}{T}\sum_{t=1}^T F(\mathbf{W}_t) = \widetilde{O}\Big(\frac{L\sqrt{m}R_2}{n} + \frac{g^2(1/T)}{n}\Big). \tag{4.17}$$

To get the risk bounds in Theorem 14, we require the width to satisfy $m \gtrsim g^8(1/n)$. If $g(\epsilon) \lesssim \log\frac{1}{\epsilon}$ (we will show this under either a linear separability or a NTK separability assumption), then a polylogarithmic width $m \gtrsim \log^8 n$ is able to achieve the risk bounds of order $1/n$.

We show that the event $E_1 \cap E_2 \cap E_3$ happens with high probability and derive the following corollary. If $g(\epsilon) \lesssim \log(1/\epsilon)$, then it gives risk bounds of order $\widetilde{O}(1/n)$, which are called fast rates in the literature (Srebro et al., 2010). The proof will be given in Section 5.2.

Corollary 15 *Let $\delta \in (0,1)$ and Assumptions 1, 2, 3 hold with $\mathbf{W}^\epsilon$ independent of S. If $\eta \leq 4/(5L), \eta T \asymp n$, $m \geq 64\log(n/\delta)\big(3/L + g^2(1/T)\big)^2 \vee \frac{256}{\pi}\big(3/L + g^2(1/T)\big)^4$ and Eq. (4.9) holds, then with probability at least $1-\delta$, we have*

$$\frac{1}{T}\sum_{t=1}^T F(\mathbf{W}_t) = \widetilde{O}\Big(\frac{Lg^2(1/T)}{n}\Big).$$

Example 1 (Linearly separable data) Consider the logistic loss $\ell(a) = \log(1+\exp(-a))$ and a linearly separable data with margin $\gamma > 0$, i.e., there exists a unit-norm vector $\mathbf{v}^* \in \mathbb{R}^d$ such that $\min_{i\in[n]} y_i \mathbf{x}_i^\top \mathbf{v}^* \geq \gamma$. We now show Assumption 3 holds under this linear separability assumption. We assume $a_j = 1/\sqrt{m}$ if $j \leq m/2$ and $a_j = -1/\sqrt{m}$ otherwise. For any $\epsilon > 0$, define $\alpha_\epsilon = 4\log(1/\epsilon)/\gamma$, and $\mathbf{W}^\epsilon = ((\mathbf{w}_1^\epsilon)^\top, \ldots, (\mathbf{w}_m^\epsilon)^\top)^\top$ with

$$\mathbf{w}_j^\epsilon = \begin{cases} \mathbf{w}_{1,j} + \alpha_\epsilon \mathbf{v}^*/\sqrt{m}, & \text{if } j \leq m/2 \\ \mathbf{w}_{1,j} - \alpha_\epsilon \mathbf{v}^*/\sqrt{m}, & \text{otherwise.} \end{cases}$$

Then, we have $\|\mathbf{W}^\epsilon - \mathbf{W}_1\|_2 = \frac{\alpha_\epsilon}{\sqrt{m}}\sqrt{m}\|\mathbf{v}^*\|_2 = 4\log(1/\epsilon)/\gamma$ and $\|\mathbf{W}^\epsilon - \mathbf{W}_1\|_{2,\infty} = \frac{4\log(1/\epsilon)}{\sqrt{m}\gamma}$. Furthermore, we know

$$\begin{aligned} y_i\Phi(\mathbf{W}^\epsilon; \mathbf{x}_i) &= y_i \sum_{j=1}^{m/2} a_j \sigma\big(\mathbf{x}_i^\top \mathbf{w}_{1,j} + \alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m}\big) + y_i \sum_{j=m/2+1}^{m} a_j \sigma\big(\mathbf{x}_i^\top \mathbf{w}_{1,j} - \alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m}\big) \\ &= \frac{1}{m}\sum_{j=1}^{m/2} \underbrace{y_i\sqrt{m}\Big(\sigma\big(\mathbf{x}_i^\top \mathbf{w}_{1,j} + \alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m}\big) - \sigma\big(\mathbf{x}_i^\top \mathbf{w}_{1,j} - \alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m}\big)\Big)}_{:=\xi_j}, \end{aligned}$$

where we have used the identity $\mathbf{w}_{1,j} = \mathbf{w}_{1,j-m/2}$ if $j > m/2$. It is clear that

$$\begin{aligned} |\xi_j| &\leq \sqrt{m}\Big|\big(\mathbf{x}_i^\top \mathbf{w}_{1,j} + \alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m}\big) - \big(\mathbf{x}_i^\top \mathbf{w}_{1,j} - \alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m}\big)\Big| \\ &\leq 2\sqrt{m}\big|\alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m}\big| \leq 2\alpha_\epsilon. \end{aligned} \tag{4.18}$$

We now consider two cases to control $\mathbb{E}[\xi_j]$.
• If $y_i = 1$, then we know $\mathbf{x}_i^\top \mathbf{v}^* \geq \gamma$ and define a random variable

$$\xi_j' = \begin{cases} y_i\sqrt{m}\big(\mathbf{x}_i^\top \mathbf{w}_{1,j} + \alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m}\big), & \text{if } |\mathbf{x}_i^\top \mathbf{w}_{1,j}| \leq \alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m} \\ 0, & \text{otherwise.} \end{cases}$$

It is clear that $\xi_j = \xi_j'$ if $|\mathbf{x}_i^\top \mathbf{w}_{1,j}| \leq \alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m}$ and $\xi_j \geq 0$ otherwise. Therefore, we have $\xi_j \geq \xi_j'$ and

$$\mathbb{E}_{\mathbf{w}_{1,j}}[\xi_j] \geq \mathbb{E}_{\mathbf{w}_{1,j}}[\xi_j'] = y_i\sqrt{m}\alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m} \geq \alpha_\epsilon\gamma, \tag{4.19}$$

where we have used the symmetry of Gaussian distribution in the identity.
• If $y_i = -1$, then we know $\mathbf{x}_i^\top \mathbf{v}^* \leq -\gamma$ and define a random variable

$$\xi_j' = \begin{cases} -y_i\sqrt{m}\big(\mathbf{x}_i^\top \mathbf{w}_{1,j} - \alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m}\big), & \text{if } |\mathbf{x}_i^\top \mathbf{w}_{1,j}| \leq -\alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m} \\ 0, & \text{otherwise.} \end{cases}$$

It is clear that $\xi_j = \xi_j'$ if $|\mathbf{x}_i^\top \mathbf{w}_{1,j}| \leq -\alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m}$ and $\xi_j \geq 0$ otherwise. Therefore, we have $\xi_j \geq \xi_j'$ and

$$\mathbb{E}_{\mathbf{w}_{1,j}}[\xi_j] \geq \mathbb{E}_{\mathbf{w}_{1,j}}[\xi_j'] = y_i\sqrt{m}\alpha_\epsilon \mathbf{x}_i^\top \mathbf{v}^*/\sqrt{m} \geq \alpha_\epsilon\gamma.$$

In both two cases, we show that $\mathbb{E}_{\mathbf{w}_{1,j}}[\xi_j] \geq \alpha_\epsilon \gamma$. Therefore, we can apply Eq. (4.18) and Hoeffding's inequality to derive the following inequality with probability at least $1 - \delta/n$

$$y_i \Phi(\mathbf{W}^\epsilon; \mathbf{x}_i) = \frac{1}{m} \sum_{j=1}^{m/2} \xi_j \geq \frac{\alpha_\epsilon \gamma}{2} - \frac{2\alpha_\epsilon \log^{\frac{1}{2}}(n/\delta)}{\sqrt{m}}.$$

Therefore, if $m \geq 64 \log(n/\delta)/\gamma^2$, with probability at least $1 - \delta$ we have $y_i \Phi(\mathbf{W}^\epsilon; \mathbf{x}_i) \geq \alpha_\epsilon \gamma/4$ simultaneously for all $i \in [n]$. It then follows that

$$F_S(\mathbf{W}^\epsilon) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \Phi(\mathbf{W}^\epsilon; \mathbf{x}_i))) \leq \exp(-\alpha_\epsilon \gamma/4) = \epsilon,$$

where we have used the inequality $\log(1+x) \leq x$ for all $x \geq 0$. Therefore, with probability at least $1-\delta$, Assumption 3 holds with $g(\epsilon) = 4\log(1/\epsilon)/\gamma$ provided that $m \geq 64 \log(n/\delta)/\gamma^2$. Then, we can apply Corollary 15 to show $\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{W}_t) = \widetilde{O}(1/(n\gamma^2))$ for training SNNs by gradient descent on linearly separable dataset with margin γ.

Example 2 We now show that our analysis also implies excess risk bounds under a separability assumption on a feature not induced by gradients of SNNs. Specifically, we consider the following feature map

$$\Psi_{\mathbf{u}}(\mathbf{x}) = \big(u_1 \mathbf{x}^\top, u_2 \mathbf{x}^\top, \ldots, u_m \mathbf{x}^\top\big)^\top, \tag{4.20}$$

where $\mathbf{u} = (u_1, \ldots, u_m)^\top \in \mathbb{R}^m$ satisfies $\|\mathbf{u}\|_2 \leq 1$. This feature is not induced by the gradient of SNNs since it does not involve the indicator function, and the parameter $\mathbf{u}$ can be arbitrarily chosen as long as $\|\mathbf{u}\|_2 \leq 1$. We can derive a similar risk bound under a separability assumption on the feature map $\Psi_{\mathbf{u}}(\cdot)$, i.e., there exists $\mathbf{V}_* = ((\mathbf{v}_1^*)^\top, \ldots, (\mathbf{v}_m^*)^\top)^\top \in \mathbb{R}^{md}$ with $\|\mathbf{V}_*\|_2 = 1$ and

$$y_i \langle \Psi_{\mathbf{u}}(\mathbf{x}_i), \mathbf{V}_* \rangle \geq \gamma, \quad \forall i \in [n]. \tag{4.21}$$

To this aim, we first show that Eq. (4.21) and $\|\mathbf{V}_*\|_2 = 1$ imply a linear separability assumption in Example 1. Indeed, we can construct $\mathbf{v}^* = \sum_{j\in[m]} u_j \mathbf{v}_{j,*}$, for which we have

$$y_i \mathbf{x}_i^\top \mathbf{v}^* = y_i \sum_{j\in[m]} u_j \mathbf{x}_i^\top \mathbf{v}_{j,*} = y_i \langle \Psi_{\mathbf{u}}(\mathbf{x}_i), \mathbf{V}_* \rangle \geq \gamma$$

and (by Schwarz's inequality)

$$\|\mathbf{v}^*\|_2^2 = \Big\| \sum_{j\in[m]} u_j \mathbf{v}_{j,*} \Big\|_2^2 \leq \Big(\sum_{j\in[m]} u_j^2 \Big) \Big(\sum_{j\in[m]} \|\mathbf{v}_{j,*}\|_2^2 \Big) \leq \|\mathbf{V}_*\|_2^2 = 1.$$

Therefore, the dataset is linearly separable with margin γ and our analysis in Example 1 works, which verifies Assumption 3 and implies risk bounds of order $\widetilde{O}(1/(n\gamma^2))$.

4.3 Connection to NTK Separability

In this subsection, we show the connection of our generalization and optimization analysis to that based on NTK separability, which means that the dataset is separable by NTK feature with a margin γ (Chen et al., 2021; Nitanda et al., 2019; Ji and Telgarsky, 2019; Nacson et al., 2019). As shown in Nitanda et al. (2019), this separability assumption is weaker than the positivity assumption on the Gram-matrix of NTK considered in the literature (Arora et al., 2019; Du et al., 2019; Zou et al., 2020), and is reasonable by the universal approximation ability of neural tangent models. Our aim is to show that our general analysis can improve the existing analysis when applied to the specific problem setup in Ji and Telgarsky (2019). Let $\langle \cdot, \cdot \rangle$ denote the dot product.

Assumption 4 (Separability by NTK) *Let $\gamma, \beta > 0$. Assume there exists $\mathbf{W}_* \in \mathbb{R}^{md}$ with $\|\mathbf{W}_*\|_2 = 1$ and $\|\mathbf{W}_*\|_{2,\infty} \leq \beta$ such that the following inequality holds*

$$y_i \big\langle \nabla \Phi(\mathbf{W}_1; \mathbf{x}_i), \mathbf{W}_* \big\rangle \geq \gamma, \quad \forall i \in [n].$$

In the following lemma to be proved in Section 5.3, we show that the realizability assumption holds under the NTK separability assumption. In this case, we choose $g(\epsilon) \lesssim \log(1/\epsilon)/\gamma$. A similar result has been derived for SNNs with smooth activation functions (Taheri and Thrampoulidis, 2024). Note $\mathbf{W}^\epsilon$ defined below is independent of S.

Lemma 16 *Let Assumption 4 hold. Let ℓ be the logistic loss. For any $\epsilon > 0$, we choose $\mathbf{W}^\epsilon = \mathbf{W}_1 + \alpha_\epsilon \mathbf{W}_*$ with $\alpha_\epsilon = (2 \log 1/\epsilon)/\gamma$. If $\alpha_\epsilon \beta^2 \leq \frac{\gamma\sqrt{\pi}}{8\sqrt{m}}$ and $\beta \leq \gamma/(4 \log^{\frac{1}{2}}(n/\delta))$, then with probability at least $1 - \delta$ over the randomness of the initialization we have*

$$f(\mathbf{W}^\epsilon; \mathbf{z}_i) \leq \epsilon, \quad \forall i \in [n].$$

Furthermore, we have $\|\mathbf{W}^\epsilon - \mathbf{W}_1\|_2 = \alpha_\epsilon$ and $\|\mathbf{W}^\epsilon - \mathbf{W}_1\|_{2,\infty} \leq \alpha_\epsilon \beta$.

Remark 17 Assumption 4 was essentially considered in the literature with $\beta = 1/\sqrt{m}$, see, e.g., Eq. (2.2) in Ji and Telgarsky (2019). In this case, the condition $\alpha_\epsilon \beta^2 \leq \frac{\gamma\sqrt{\pi}}{8\sqrt{m}}$ becomes $\alpha_\epsilon \leq \frac{\gamma\sqrt{\pi m}}{8} \iff m \geq \frac{64}{\pi} \alpha_\epsilon^2/\gamma^2$, and the condition $\beta \leq \gamma/(4 \log^{\frac{1}{2}}(n/\delta))$ becomes $m \geq 16 \log(n/\delta)/\gamma^2$.

We consider $\epsilon = 1/T$, choose $\mathbf{W}^* = \mathbf{W}^{\frac{1}{T}}$ and set $\eta T = n$ for simplicity. Then, according to Lemma 16, the definition of $\mathfrak{C}_S$ in Eq. (4.2) and Eq. (5.28), with probability at least $1 - 3\delta$ we have

$$\mathfrak{C}_S(\mathbf{W}^{\frac{1}{T}}) \leq 3\eta + \alpha_\epsilon^2 \quad \text{and} \quad \mathfrak{C}(\mathbf{W}^{\frac{1}{T}}) \leq 12\eta + 24 b_* \log(2/\delta) + \alpha_\epsilon^2, \tag{4.22}$$

where we have used $\eta T F_S(\mathbf{W}^{\frac{1}{T}}) \leq \eta$ by Lemma 16. Suppose the event in Lemma 16 holds and $\eta \leq 1/L$. Then, we can choose R_1 as

$$R_1 = \max\Big\{ 3/(\sqrt{m} L) + \alpha_\epsilon^2/\sqrt{m}, \alpha_\epsilon \beta \Big\}$$

and (since we require $R_2 \geq \mathfrak{C}(\mathbf{W}^{\frac{1}{T}})/\sqrt{m}$)

$$R_2 = \frac{1}{\sqrt{m}} \Big(12\eta + 24 b_* \log(2/\delta) + \alpha_\epsilon^2 \Big). \tag{4.23}$$

Recall $\alpha_\epsilon = (2\log T)/\gamma$. Therefore, if

$$\frac{2\log T}{\gamma}\big(3/L + 4\log^2 T/\gamma^2\big)\beta \le \frac{\sqrt{\pi}}{16} \tag{4.24}$$

and

$$m \ge \max\Big\{\frac{256}{\pi}\Big(3/L + 4\log^2 T/\gamma^2\Big)^4, 64\log(n/\delta)\Big(3/L + 4\log^2 T/\gamma^2\Big)^2\Big\}, \tag{4.25}$$

we have

$$R_1\mathfrak{C}_S(\mathbf{W}^{\frac{1}{T}}) \le \max\Big\{\big(3/L + \alpha_\epsilon^2\big)/\sqrt{m}, \alpha_\epsilon\beta\Big\}\big(3/L + \alpha_\epsilon^2\big) \le \frac{\sqrt{\pi}}{16}, \tag{4.26}$$

where $\alpha_\epsilon\beta\big(3/L + \alpha_\epsilon^2\big) \le \frac{\sqrt{\pi}}{16}$ is due to Eq. (4.24), and $\frac{1}{\sqrt{m}}(3/L + \alpha_\epsilon^2)\big(3/L + \alpha_\epsilon^2\big) \le \frac{\sqrt{\pi}}{16}$ is due to Eq. (4.25). Therefore, the event E_1 holds with probability at least $1 - 2\delta$. Similarly, $R_2 \ge \mathfrak{C}(\mathbf{W}^{\frac{1}{T}})/\sqrt{m}$ with probability at least $1 - 2\delta$. We now apply Theorem 14 to derive the risk bounds under the NTK separability condition. The proof will be given in Section 5.3.

Theorem 18 (Risk bounds under NTK Separability) *Let Assumptions 2 and 4 hold. Let ℓ be the logistic loss and $\delta \in (0, 1)$. Let Eq. (4.24) and Eq. (4.25) hold. If $\eta \le 16/5$, $(2\log T)\beta^2 \le \frac{\gamma^2\sqrt{\pi}}{8\sqrt{m}}$ and $\beta \le \gamma/(4\log^{\frac{1}{2}}(n/\delta))$, then for $\eta T = n$ with probability at least $1 - \delta$ there holds $\frac{1}{T}\sum_{t=1}^T F(\mathbf{W}_t) = \widetilde{O}\big(1/(n\gamma^2)\big)$.*

Remark 19 (Comparison) Under a similar assumption, the risk bounds of order $\widetilde{O}\big(\frac{1}{\gamma^2\sqrt{n}}\big)$ were developed in Ji and Telgarsky (2019). Theorem 18 improves the rate to $\widetilde{O}(\frac{1}{n\gamma^2})$, which, according to the discussions in Shamir (2021); Schliserman and Koren (2024) is optimal up to a logarithmic factor. We use two techniques to achieve this improvement. First, we use the localization technique in Srebro et al. (2010) to derive optimistic bounds by using the smoothness of the loss function. However, this technique alone with the estimate $\mathfrak{R}_{\widetilde{S}}(\mathcal{F}) \le \mathfrak{C}(\mathbf{W}^*)/\sqrt{n}$ (Ji and Telgarsky, 2019) only implies bounds of the order $\widetilde{O}(\frac{1}{n\gamma^4})$. The second technique is an improved Rademacher complexity estimate of $\mathfrak{R}_{S,n}(\mathcal{F})$ in Lemma 9, which improves the dependency on the margin from $1/\gamma^4$ to $1/\gamma^2$. For the NTK-separable data, we have $\beta = 1/\sqrt{m}$ (Ji and Telgarsky, 2019) and therefore the constraint on m becomes $m = \widetilde{\Omega}(1/\gamma^8)$. This matches the overparameterization requirement in Ji and Telgarsky (2019). The recent work (Taheri and Thrampoulidis, 2024) also derived risk bounds of order $\widetilde{O}(1/(n\gamma^2))$. However, their discussion considers Lipschitz and smooth activation functions. Furthermore, they derived generalization bounds by using the stability analysis, which are stated in expectation. As a comparison, we use Rademacher complexity and our analysis implies bounds with high-probability, which are stronger than bounds in expectation. Indeed, high-probability bounds offers strong assurances that the bounds will hold with high confidence, while bounds in expectation only guarantee that the bounds hold in the average case, which do not rule out the possibility of rare, large deviations. We summarize the comparisons in Table 2.

Work	Bound	Activation	Width	Type
Ji and Telgarsky (2019)	$\widetilde{O}\big(1/(\sqrt{n}\gamma^2)\big)$	ReLU	logarithmic	high-probability
Taheri and Thrampoulidis (2024)	$\widetilde{O}(1/(n\gamma^2))$	Smooth	logarithmic	expectation
Ours	$\widetilde{O}(1/(n\gamma^2))$	ReLU	logarithmic	high-probability

Table 2: Comparison of the risk bounds under NTK separability condition with margin γ. Taheri and Thrampoulidis (2024) used algorithmic stability to derive bounds in expectation, and their discussions considered smooth and Lipschitz activation functions. As a comparison, both Ji and Telgarsky (2019) and our work used Rademacher complexity to derive high-probability bounds, which considered the ReLU activation function. High-probability bounds are stronger than bounds in expectation: high-probability bounds offer strong assurances with high confidence, while bounds in expectation only offer guarantees in the average case.

Example: Noisy 2-XOR Data. We now consider a specific problem where the NTK separability condition holds. Specifically, we consider the noisy 2-XOR distribution introduced in Wei et al. (2019), which is the uniform distribution over the following 2^d points

$$(x_1, x_2, y, x_3, \ldots, x_d) \in \Big\{\Big(\frac{1}{\sqrt{d-1}}, 0, 1\Big), \Big(0, \frac{1}{\sqrt{d-1}}, -1\Big), \Big(\frac{-1}{\sqrt{d-1}}, 0, 1\Big), \Big(0, \frac{-1}{\sqrt{d-1}}, -1\Big)\Big\} \times \Big\{\frac{-1}{\sqrt{d-1}}, \frac{1}{\sqrt{d-1}}\Big\}^{d-2},$$

where the factor $1/\sqrt{d-1}$ is introduced to ensure $\|\mathbf{x}\|_2 = 1$ and $\times$ denotes the Cartesin product. For the 2-XOR data, the label y depends only on the first two coordinates of the input $\mathbf{x}$. For any $(\mathbf{x}, y)$ sampled from the noisy 2-XOR distribution and $d \geq 3$, it was shown that there exists $\bar{\omega}(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^d$ with $\|\bar{\omega}(\mathbf{w})\|_2 \leq 1$ for all $\mathbf{w} \in \mathbb{R}^d$ and all $(\mathbf{x}, y) \sim \rho$ (Ji and Telgarsky, 2019)

$$y \int_{\mathbb{R}^d} \mathbb{I}_{[\mathbf{w}^\top \mathbf{x} \geq 0]} \cdot \mathbf{x}^\top \bar{\omega}(\mathbf{w}) d\mu_N(\mathbf{w}) \geq \frac{1}{60d}, \tag{4.27}$$

where $\mu_N(\cdot)$ denotes the standard Gaussian measure. Then, we choose $\mathbf{W}_* = (\mathbf{w}_{1,*}^\top, \ldots, \mathbf{w}_{m,*}^\top)^\top$ with $\mathbf{w}_{j,*} = a_j \bar{\omega}(\mathbf{w}_{1,j})$ and get

$$\begin{aligned} y_i \big\langle \nabla \Phi(\mathbf{W}_1; \mathbf{x}_i), \mathbf{W}_* \big\rangle &= y_i \sum_{j=1}^m a_j \mathbf{x}_i^\top \mathbf{w}_{j,*} \mathbb{I}_{[\mathbf{w}_{1,j}^\top \mathbf{x}_i \geq 0]} = y_i \sum_{j=1}^m a_j \mathbf{x}_i^\top \big(a_j \bar{\omega}(\mathbf{w}_{1,j})\big) \mathbb{I}_{[\mathbf{w}_{1,j}^\top \mathbf{x}_i \geq 0]} \\ &= \frac{2y_i}{m} \sum_{j=1}^{m/2} \mathbf{x}_i^\top \bar{\omega}(\mathbf{w}_{1,j}) \mathbb{I}_{[\mathbf{w}_{1,j}^\top \mathbf{x}_i \geq 0]}, \end{aligned}$$

where we have used Eq. (3.3) in the computation of $\nabla \Phi(\mathbf{W}_1; \mathbf{x}_i)$ and the symmetric initialization. Note that $|y_i \mathbf{x}_i^\top \bar{\omega}(\mathbf{w}_{1,j}) \mathbb{I}_{[\mathbf{w}_{1,j}^\top \mathbf{x}_i \geq 0]}| \leq 1$. According to Hoeffding's inequality, with

probability at least $1-\delta$ we have

$$y_i\big\langle \nabla\Phi(\mathbf{W}_1;\mathbf{x}_i), \mathbf{W}_*\big\rangle \geq \frac{1}{60d} - \frac{2\log^{\frac{1}{2}}(n/\delta)}{\sqrt{m}}, \quad \forall i \in [n].$$

Therefore, if $m \geq (240d)^2\log(n/\delta)$, we know Assumption 4 holds with probability at least $1-\delta$, $\gamma = \frac{1}{120d}$ and $\beta = 1/\sqrt{m}$. Then, we immediately derive the following risk bounds for the XOR data.

Corollary 20 *Consider the 2-XOR distribution. Let ℓ be the logistic loss and $\delta \in (0,1)$. Let Eq. (4.24) and Eq. (4.25) hold with $\gamma = \frac{1}{120d}$. If $m \geq (120d)^2\max\{\frac{256\cdot(120d)^2\log^2 T}{\pi}, 16\log(n/\delta)\}$ and $\eta \leq 16/5$, then for $\eta T = n$ with probability $1-\delta$ we get $\frac{1}{T}\sum_{t=1}^T F(\mathbf{W}_t) = \widetilde{O}\big(d^2/n\big)$.*

5 Proofs

5.1 Proofs on Optimization

In this subsection, we present the proofs related to the optimization error analysis. The following lemma controls the distance between the gradient descent iterates and the initialization, as measured by the $(2,\infty)$-norm.

Lemma 21 (Ji and Telgarsky 2019) *If Assumption 1 and Assumption 2 hold, then for any $j \in [m]$ we have*

$$\|\mathbf{w}_{t+1,j} - \mathbf{w}_{1,j}\|_2 \leq \frac{\eta}{\sqrt{m}}\sum_{k=1}^t F_S(\mathbf{W}_k).$$

Lemma 21 shows that any convergence analysis on $\sum_{k=1}^t F_S(\mathbf{W}_k)$ automatically implies a bound on the $(2,\infty)$-norm. This motivates us to use $(2,\infty)$-norm as a metric to study both optimization and generalization. Indeed, a key property of the ReLU activation in our analysis is expressed in terms of the $(2,\infty)$-norm: if $\|\mathbf{W}-\mathbf{W}_1\|_{2,\infty} \leq R$ and $|\mathbf{w}_{1,j}^\top\mathbf{x}| > R$, then $\mathbb{I}_{[\mathbf{w}_j^\top\mathbf{x}\geq 0]} = \mathbb{I}_{[\mathbf{w}_{1,j}^\top\mathbf{x}\geq 0]}$ and $\sigma(\mathbf{w}_j^\top\mathbf{x}) - \sigma(\mathbf{w}_{1,j}^\top\mathbf{x}) = \mathbb{I}_{[\mathbf{w}_{1,j}^\top\mathbf{x}\geq 0]}(\mathbf{w}_j - \mathbf{w}_{1,j})^\top\mathbf{x}$. This shows that the ReLU function behaves similarly to a linear function under a $(2,\infty)$-norm constraint and some conditions on the initialization. As we will see, we use this key property to control $\langle \mathbf{W}' - \mathbf{W}, \nabla F_S(\mathbf{W}')\rangle$ (Lemma 22) and the Rademacher complexity (Lemma 9), which are key lemmas in convergence and generalization analysis, respectively.

The following lemma shows a weak-convexity property of F_S on a ball around the initialization point. It shows that as the width becomes larger, the objective function becomes more convex. For any $R > 0$, define

$$B_R = \big\{\mathbf{W} \in \mathbb{R}^{md} : \max_{j\in[m]} \|\mathbf{w}_j - \mathbf{w}_{1,j}\|_2 \leq R\big\}. \tag{5.1}$$

For any $i \in [n]$, we define

$$S_i = \big\{j \in [m] : |\mathbf{w}_{1,j}^\top\mathbf{x}_i| \leq R_1\big\}. \tag{5.2}$$

Recall Lemma 2 gives probabilistic bounds on $|S_i|$, which denotes the cardinality of S_i.

Lemma 22 *Let $\mathbf{W}, \mathbf{W}' \in B_{R_1}$. Under the event E_1, we have*

$$\langle \mathbf{W}' - \mathbf{W}, \nabla F_S(\mathbf{W}') \rangle \geq F_S(\mathbf{W}') - F_S(\mathbf{W}) + \frac{1}{2Ln} \sum_{i=1}^{n} \big(\ell'(y_i \Phi(\mathbf{W}; \mathbf{x}_i)) - \ell'(y_i \Phi(\mathbf{W}'; \mathbf{x}_i))\big)^2 - \frac{C_{R_1}^{\frac{1}{2}} \|\mathbf{W} - \mathbf{W}'\|_2 F_S(\mathbf{W}')}{\sqrt{m}}.$$

Proof For any $\mathbf{W}$ and $\mathbf{x}$, we know (Ji and Telgarsky, 2019)

$$\langle \mathbf{W}, \nabla \Phi(\mathbf{W}; \mathbf{x}) \rangle = \Phi(\mathbf{W}; \mathbf{x}). \tag{5.3}$$

It then follows that for any $\mathbf{W}, \mathbf{W}' \in B_{R_1}$

$$y_i \Big\langle \mathbf{W}' - \mathbf{W}, \ell'(y_i \Phi(\mathbf{W}'; \mathbf{x}_i)) \nabla \Phi(\mathbf{W}'; \mathbf{x}_i) \Big\rangle = y_i \ell'(y_i \Phi(\mathbf{W}'; \mathbf{x}_i)) \big(\Phi(\mathbf{W}'; \mathbf{x}_i) - \langle \nabla \Phi(\mathbf{W}'; \mathbf{x}_i), \mathbf{W} \rangle\big)$$
$$= y_i \ell'(y_i \Phi(\mathbf{W}'; \mathbf{x}_i)) \Big(\Phi(\mathbf{W}'; \mathbf{x}_i) - \Phi(\mathbf{W}; \mathbf{x}_i) + \langle \nabla \Phi(\mathbf{W}; \mathbf{x}_i) - \nabla \Phi(\mathbf{W}'; \mathbf{x}_i), \mathbf{W} \rangle\Big). \tag{5.4}$$

By the co-coercivity of ℓ, i.e., $\ell(a) \geq \ell(b) + (a-b)\ell'(b) + \frac{1}{2L}(\ell'(a) - \ell'(b))^2$, due to its convexity and smoothness, we further know

$$y_i \Big\langle \mathbf{W}' - \mathbf{W}, \ell'(y_i \Phi(\mathbf{W}'; \mathbf{x}_i)) \nabla \Phi(\mathbf{W}'; \mathbf{x}_i) \Big\rangle \geq \ell(y_i \Phi(\mathbf{W}'; \mathbf{x}_i)) - \ell(y_i \Phi(\mathbf{W}; \mathbf{x}_i)) + \frac{1}{2L} \big(\ell'(y_i \Phi(\mathbf{W}'; \mathbf{x}_i)) - \ell'(y_i \Phi(\mathbf{W}; \mathbf{x}_i))\big)^2 + y_i \ell'(y_i \Phi(\mathbf{W}'; \mathbf{x}_i)) \langle \nabla \Phi(\mathbf{W}; \mathbf{x}_i) - \nabla \Phi(\mathbf{W}'; \mathbf{x}_i), \mathbf{W} \rangle. \tag{5.5}$$

Furthermore, we have

$$y_i \ell'(y_i \Phi(\mathbf{W}'; \mathbf{x}_i)) \langle \nabla \Phi(\mathbf{W}; \mathbf{x}_i) - \nabla \Phi(\mathbf{W}'; \mathbf{x}_i), \mathbf{W} \rangle = y_i \sum_{j=1}^{m} a_j \big(\mathbb{I}_{[\mathbf{x}_i^\top \mathbf{w}_j \geq 0]} - \mathbb{I}_{[\mathbf{x}_i^\top \mathbf{w}_j' \geq 0]}\big) \ell'(y_i \Phi(\mathbf{W}'; \mathbf{x}_i)) \mathbf{x}_i^\top \mathbf{w}_j. \tag{5.6}$$

If $|\mathbf{x}_i^\top \mathbf{w}_j| > \|\mathbf{w}_j - \mathbf{w}_j'\|_2$, then

$$|\mathbf{x}_i^\top \mathbf{w}_j' - \mathbf{x}_i^\top \mathbf{w}_j| \leq \|\mathbf{x}_i\|_2 \|\mathbf{w}_j' - \mathbf{w}_j\|_2 \leq \|\mathbf{w}_j' - \mathbf{w}_j\|_2 < |\mathbf{x}_i^\top \mathbf{w}_j|.$$

This shows that $\mathbb{I}_{[\mathbf{x}_i^\top \mathbf{w}_j' \geq 0]} = \mathbb{I}_{[\mathbf{x}_i^\top \mathbf{w}_j \geq 0]}$ and therefore

$$\big(\mathbb{I}_{[\mathbf{x}_i^\top \mathbf{w}_j' \geq 0]} - \mathbb{I}_{[\mathbf{x}_i^\top \mathbf{w}_j \geq 0]}\big) \Big\langle \mathbf{x}_i, \ell'(y_i \Phi(\mathbf{W}'; \mathbf{x}_i)) \mathbf{w}_j \Big\rangle = 0. \tag{5.7}$$

If $|\mathbf{x}_i^\top \mathbf{w}_j| \leq \|\mathbf{w}_j - \mathbf{w}_j'\|_2$, then

$$\Big|\big(\mathbb{I}_{[\mathbf{x}_i^\top \mathbf{w}_j \geq 0]} - \mathbb{I}_{[\mathbf{x}_i^\top \mathbf{w}_j' \geq 0]}\big) \ell'(y_i \Phi(\mathbf{W}'; \mathbf{x}_i)) \mathbf{x}_i^\top \mathbf{w}_j\Big| \leq \big|\mathbb{I}_{[\mathbf{x}_i^\top \mathbf{w}_j \geq 0]} - \mathbb{I}_{[\mathbf{x}_i^\top \mathbf{w}_j' \geq 0]}\big| \cdot |\ell'(y_i \Phi(\mathbf{W}'; \mathbf{x}_i))| \|\mathbf{w}_j - \mathbf{w}_j'\|_2.$$

We plug the above two inequalities back into Eq. (5.6), and derive

$$\begin{aligned}&\big|y_i\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))\langle\nabla\Phi(\mathbf{W};\mathbf{x}_i)-\nabla\Phi(\mathbf{W}';\mathbf{x}_i),\mathbf{W}\rangle\big|\\&\leq\frac{|\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))|}{\sqrt{m}}\sum_{j=1}^m\|\mathbf{w}_j-\mathbf{w}_j'\|_2|\mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_j\geq0]}-\mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_j'\geq0]}|.\end{aligned}$$

Note if $j\notin S_i$, the condition $\mathbf{W}\in B_{R_1}$ implies

$$|\mathbf{x}_i^\top\mathbf{w}_j-\mathbf{x}_i^\top\mathbf{w}_{1,j}|\leq\|\mathbf{x}_i\|_2\|\mathbf{w}_j-\mathbf{w}_{1,j}\|_2\leq R_1<|\mathbf{w}_{1,j}^\top\mathbf{x}_i|. \tag{5.8}$$

This shows that the sign of $\mathbf{x}_i^\top\mathbf{w}_j$ is the same as that of $\mathbf{x}_i^\top\mathbf{w}_{1,j}$. Similarly, the condition $\mathbf{W}'\in B_{R_1}$ implies that the sign of $\mathbf{x}_i^\top\mathbf{w}_j'$ is the same as that of $\mathbf{x}_i^\top\mathbf{w}_{1,j}$ for $j\notin S_i$. Therefore, we have $\mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_j\geq0]}-\mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_j'\geq0]}=0$ and therefore

$$\begin{aligned}&\big|y_i\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))\langle\nabla\Phi(\mathbf{W};\mathbf{x}_i)-\nabla\Phi(\mathbf{W}';\mathbf{x}_i),\mathbf{W}\rangle\big|\\&\leq\frac{|\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))|}{\sqrt{m}}\sum_{j\in S_i}\|\mathbf{w}_j-\mathbf{w}_j'\|_2|\mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_j\geq0]}-\mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_j'\geq0]}|.\end{aligned}$$

We combine Eq. (5.5) and the above inequality together, and derive the following bound

$$\begin{aligned}y_i\Big\langle\mathbf{W}'-\mathbf{W},\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))\nabla\Phi(\mathbf{W}';\mathbf{x}_i)\Big\rangle&\geq\ell(y_i\Phi(\mathbf{W}';\mathbf{x}_i))-\ell(y_i\Phi(\mathbf{W};\mathbf{x}_i))\\&+\frac{1}{2L}\big(\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))-\ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i))\big)^2-\frac{|\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))|}{\sqrt{m}}\sum_{j\in S_i}\|\mathbf{w}_j-\mathbf{w}_j'\|_2.\end{aligned}$$

By the Schwarz's inequality, we know

$$\sum_{j\in S_i}\|\mathbf{w}_j-\mathbf{w}_j'\|_2\leq\sqrt{|S_i|}\Big(\sum_{j=1}^m\|\mathbf{w}_j-\mathbf{w}_j'\|_2^2\Big)^{\frac{1}{2}}=\sqrt{|S_i|}\|\mathbf{W}'-\mathbf{W}\|_2. \tag{5.9}$$

It then follows that ($|\ell'(a)|\leq\ell(a)$)

$$\begin{aligned}&\langle\mathbf{W}'-\mathbf{W},\nabla F_S(\mathbf{W}')\rangle\geq\frac{1}{n}\sum_{i=1}^n\Big(\ell(y_i\Phi(\mathbf{W}';\mathbf{x}_i))-\ell(y_i\Phi(\mathbf{W};\mathbf{x}_i))\Big)+\\&\frac{1}{2Ln}\sum_{i=1}^n\big(\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))-\ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i))\big)^2-\frac{\|\mathbf{W}'-\mathbf{W}\|_2}{n\sqrt{m}}\sum_{i=1}^n\sqrt{|S_i|}\ell(y_i\Phi(\mathbf{W}';\mathbf{x}_i)).\end{aligned}$$

The proof is completed by Lemma 2 on bounds of $|S_i|$. ■

Remark 23 If we use the inequality $\sum_{j\in S_i}\|\mathbf{w}_j-\mathbf{w}_j'\|_2\leq|S_i|\|\mathbf{W}-\mathbf{W}'\|_{2,\infty}$ instead of Eq. (5.9), then the proof of Lemma 22 implies that

$$\begin{aligned}\langle\mathbf{W}'-\mathbf{W},\nabla F_S(\mathbf{W}')\rangle&\geq F_S(\mathbf{W}')-F_S(\mathbf{W})\\&+\frac{1}{2Ln}\sum_{i=1}^n\big(\ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i))-\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))\big)^2-\frac{C_{R_1}\|\mathbf{W}-\mathbf{W}'\|_{2,\infty}F_S(\mathbf{W}')}{\sqrt{m}}.\end{aligned}$$

The following lemma shows a one-step progress inequality for gradient descent applied to SNNs with the ReLU activation function. It shows how the distance between gradient descent iterators and the reference model $\mathbf{W}^*$ would change after a single gradient descent update.

Lemma 24 *Let Assumptions 1, 2 hold. Assume the event E_1 happens. Then, for $\eta \leq 4/(5L)$ and $\mathbf{W}, \mathbf{W}' \in B_{R_1}$ we have*

$$\big\|\mathbf{W} - \eta\nabla F_S(\mathbf{W}) - \mathbf{W}'\big\|_2^2 \leq \|\mathbf{W} - \mathbf{W}'\|_2^2 - 2\eta\big(F_S(\mathbf{W}) - F_S(\mathbf{W}')\big) + \frac{2\eta C_{R_1}^{\frac{1}{2}}\|\mathbf{W} - \mathbf{W}'\|_2 F_S(\mathbf{W})}{\sqrt{m}} + 5\eta^2\widetilde{\mathfrak{C}}_S^2(\mathbf{W}'). \quad (5.10)$$

Proof For any $\mathbf{W}$, we know $\|\nabla\Phi(\mathbf{W};\mathbf{x})\|_2 \leq 1$ and therefore

$$\begin{aligned} n^2\|\nabla F_S(\mathbf{W})\|_2^2 &= \sum_{i=1}^n\sum_{j=1}^n y_i y_j \ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i))\ell'(y_j\Phi(\mathbf{W};\mathbf{x}_j))\Big\langle\nabla\Phi(\mathbf{W};\mathbf{x}_i), \nabla\Phi(\mathbf{W};\mathbf{x}_j)\Big\rangle \\ &\leq \sum_{i=1}^n\sum_{j=1}^n |\ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i))\ell'(y_j\Phi(\mathbf{W};\mathbf{x}_j))| = \Big(\sum_{i=1}^n |\ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i))|\Big)^2. \end{aligned}$$

It then follows that

$$\begin{aligned} \|\nabla F_S(\mathbf{W})\|_2^2 &\leq \Big(\frac{1}{n}\sum_{i=1}^n |\ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i))|\Big)^2 \\ &= \Big(\frac{1}{n}\sum_{i=1}^n \Big(|\ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i))| - |\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))| + |\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))|\Big)\Big)^2 \\ &\leq \frac{5}{4}\Big(\frac{1}{n}\sum_{i=1}^n \big(|\ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i))| - |\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))|\big)\Big)^2 + 5\Big(\frac{1}{n}\sum_{i=1}^n |\ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))|\Big)^2, \end{aligned}$$

where we have used the standard inequality $(a+b)^2 \leq 5a^2 + 5b^2/4$. We then use this inequality and Lemma 22 to derive the following inequality

$$\begin{aligned} &\big\|\mathbf{W} - \eta\nabla F_S(\mathbf{W}) - \mathbf{W}'\big\|_2^2 = \|\mathbf{W} - \mathbf{W}'\|_2^2 + \eta^2\|\nabla F_S(\mathbf{W})\|_2^2 - 2\eta\langle\mathbf{W} - \mathbf{W}', \nabla F_S(\mathbf{W})\rangle \\ &\leq \|\mathbf{W} - \mathbf{W}'\|_2^2 - 2\eta\big(F_S(\mathbf{W}) - F_S(\mathbf{W}')\big) + \frac{2\eta C_{R_1}^{\frac{1}{2}}\|\mathbf{W} - \mathbf{W}'\|_2 F_S(\mathbf{W})}{\sqrt{m}} + 5\eta^2\widetilde{\mathfrak{C}}_S^2(\mathbf{W}') \\ &+ \frac{5\eta^2}{4}\Big(\frac{1}{n}\sum_{i=1}^n |\ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i)) - \ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))|\Big)^2 - \frac{\eta}{Ln}\sum_{i=1}^n \big(\ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i)) - \ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))\big)^2. \end{aligned}$$

Since $\eta \leq 4/(5L)$ and

$$\Big(\frac{1}{n}\sum_{i=1}^n |\ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i)) - \ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))|\Big)^2 \leq \frac{1}{n}\sum_{i=1}^n \big(\ell'(y_i\Phi(\mathbf{W};\mathbf{x}_i)) - \ell'(y_i\Phi(\mathbf{W}';\mathbf{x}_i))\big)^2,$$

we further get the stated bound. The proof is completed. ■

Proof of Theorem 3 We use the induction strategy to prove the result. It is clear that $\|\mathbf{W}_1 - \mathbf{W}^*\|_2^2 \leq \mathfrak{C}_S(\mathbf{W}^*)$. This shows that Eq. (4.4) holds with $t = 1$. We now assume it holds for $t \leq k$ and will prove that it also holds for $t = k + 1$. Note that the induction and Lemma 21 imply that

$$\|\mathbf{w}_{t',j} - \mathbf{w}_{1,j}\|_2 \leq \frac{\eta}{\sqrt{m}} \sum_{k=1}^{t'-1} F_S(\mathbf{W}_k) \leq \frac{\mathfrak{C}_S(\mathbf{W}^*)}{\sqrt{m}}, \quad \forall j \in [m], t' \in [k]. \tag{5.11}$$

Furthermore, it is clear that $\|\mathbf{w}_j^* - \mathbf{w}_{1,j}\|_2 \leq R_1$ according to Eq. (4.3). Therefore, $\mathbf{W}_t \in B_{R_1}$ and $\mathbf{W}^* \in B_{R_1}$ for $t = 1, \ldots, k$. We choose $\mathbf{W} = \mathbf{W}_k$ and $\mathbf{W}' = \mathbf{W}^*$ in Lemma 24 to derive

$$\big\|\mathbf{W}_{k+1} - \mathbf{W}^*\big\|_2^2 + 2\eta\big(F_S(\mathbf{W}_k) - F_S(\mathbf{W}^*)\big) \leq \|\mathbf{W}_k - \mathbf{W}^*\|_2^2 + \frac{2\eta C_{R_1}^{\frac{1}{2}} \|\mathbf{W}_k - \mathbf{W}^*\|_2 F_S(\mathbf{W}_k)}{\sqrt{m}} + 5\eta^2 \widetilde{\mathfrak{C}}_S^2(\mathbf{W}^*).$$

Taking a summation implies further (note $\widetilde{\mathfrak{C}}_S(\mathbf{W}^*) \leq F_S(\mathbf{W}^*)$)

$$\|\mathbf{W}_{k+1} - \mathbf{W}^*\|_2^2 + 2\eta \sum_{t=1}^{k} \big(F_S(\mathbf{W}_t) - F_S(\mathbf{W}^*)\big) \leq \|\mathbf{W}_1 - \mathbf{W}^*\|_2^2 + \frac{2\eta C_{R_1}^{\frac{1}{2}}}{\sqrt{m}} \sum_{t=1}^{k} F_S(\mathbf{W}_t)\|\mathbf{W}_t - \mathbf{W}^*\|_2 + 5\eta^2 k \widetilde{\mathfrak{C}}_S(\mathbf{W}^*) F_S(\mathbf{W}^*).$$

The above inequality implies that

$$\|\mathbf{W}_{k+1} - \mathbf{W}^*\|_2^2 + 2\eta \sum_{t=1}^{k} F_S(\mathbf{W}_t)\Big(1 - \frac{C_{R_1}^{\frac{1}{2}} \max_{t\in[k]} \|\mathbf{W}_t - \mathbf{W}^*\|_2}{\sqrt{m}}\Big) \leq \Big(2 + 5\eta\widetilde{\mathfrak{C}}_S(\mathbf{W}^*)\Big)\eta k F_S(\mathbf{W}^*) + \|\mathbf{W}_1 - \mathbf{W}^*\|_2^2. \tag{5.12}$$

We now show $m \geq 4C_{R_1}\mathfrak{C}_S(\mathbf{W}^*)$, which, according to the definition of C_{R_1} in Eq. (4.1), becomes

$$m \geq \frac{8\mathfrak{C}_S(\mathbf{W}^*)R_1 m}{\sqrt{\pi}} + 4\sqrt{m\log(n/\delta)}\mathfrak{C}_S(\mathbf{W}^*). \tag{5.13}$$

By Eq. (4.3), we know that Eq. (5.13) holds if

$$m \geq 8\sqrt{m\log(n/\delta)}\mathfrak{C}_S(\mathbf{W}^*) \Longleftrightarrow m \geq 64\log(n/\delta)\mathfrak{C}_S^2(\mathbf{W}^*),$$

which holds by our assumption. Therefore, it holds that $m \geq 4C_{R_1}\mathfrak{C}_S(\mathbf{W}^*)$, which together with the induction hypothesis $\max_{t\in[k]} \|\mathbf{W}_t - \mathbf{W}^*\|_2^2 \leq \mathfrak{C}_S(\mathbf{W}^*)$, implies $C_{R_1}^{\frac{1}{2}} \max_{t\in[k]} \|\mathbf{W}_t - \mathbf{W}^*\|_2 \leq \sqrt{m}/2$. We put this inequality in Eq. (5.12) and get

$$\|\mathbf{W}_{k+1} - \mathbf{W}^*\|_2^2 + \eta \sum_{t=1}^{k} F_S(\mathbf{W}_t) \leq 3\eta k F_S(\mathbf{W}^*) + \|\mathbf{W}_1 - \mathbf{W}^*\|_2^2,$$

where we have used the assumption $\eta \leq 1/(5\widetilde{\mathfrak{C}}_S(\mathbf{W}^*))$. This shows

$$\|\mathbf{W}_{k+1} - \mathbf{W}^*\|_2^2 \leq \mathfrak{C}_S(\mathbf{W}^*) \quad \text{and} \quad \eta \sum_{t=1}^{k} F_S(\mathbf{W}_t) \leq \mathfrak{C}_S(\mathbf{W}^*).$$

Therefore, the induction hypothesis holds with $t = k+1$ and finishes the proof of Eq. (4.4).

The stated inequality in Eq. (4.5) holds according to Eq. (5.11) and Eq. (4.4). The proof is completed. ■

Finally, we prove Theorem 6 as a simple application of Theorem 3 in an interpolation setting.

Proof of Theorem 6 By Assumption 3 with $\epsilon = 1/T$, we can find $\mathbf{W}^{\frac{1}{T}}$ with

$$F_S(\mathbf{W}^{\frac{1}{T}}) \leq 1/T \quad \text{and} \quad \|\mathbf{W}^{\frac{1}{T}} - \mathbf{W}_1\|_2 \leq g(1/T).$$

It then follows that

$$\mathfrak{C}_S(\mathbf{W}^{\frac{1}{T}}) = 3\eta T F_S(\mathbf{W}^{\frac{1}{T}}) + \|\mathbf{W}_1 - \mathbf{W}^{\frac{1}{T}}\|_2^2 \leq 3\eta + g^2(1/T) \leq 3/L + g^2(1/T). \tag{5.14}$$

Since $|\ell'(a)| \leq \ell(a)$, we know $\widetilde{\mathfrak{C}}_S(\mathbf{W}^{\frac{1}{T}}) \leq F_S(\mathbf{W}^{\frac{1}{T}}) \leq 1/T$. Then, the constraint $\eta \leq \min\{4/(5L), 1/(5\widetilde{\mathfrak{C}}_S(\mathbf{W}^{\frac{1}{T}}))\}$ holds if $\eta \leq \min\{4/(5L), T/5\} = 4/(5L)$. The stated bound then follows directly from Theorem 3 by using the above bound on $\widetilde{\mathfrak{C}}_S(\mathbf{W}^{\frac{1}{T}})$. The proof is completed. ■

Proof of Corollary 7 We define

$$R_1 = \frac{\sqrt{\pi}}{16\big(3/L + g^2(1/T)\big)}.$$

Since $m \geq \frac{256}{\pi}\big(3/L + g^2(1/T)\big)^4$ and Eq. (5.14), we know that

$$\frac{\mathfrak{C}_S(\mathbf{W}^{\frac{1}{T}})}{\sqrt{m}} \leq R_1 = \frac{\sqrt{\pi}}{16\big(3/L + g^2(1/T)\big)} \leq \frac{\sqrt{\pi}}{16\mathfrak{C}_S(\mathbf{W}^{\frac{1}{T}})}.$$

Furthermore, the assumption $\|\mathbf{W}^{\frac{1}{T}} - \mathbf{W}_1\|_{2,\infty} \leq \frac{\sqrt{\pi}}{16(3/L+g^2(1/T))}$ implies that Eq. (4.3) holds. According to Lemma 2, we know the event E_1 happens with probability at least $1-\delta$. The stated bounds directly follow from Theorem 6. ■

5.2 Proofs on Generalization

In this subsection, we present proofs on generalization analysis. We require the following lemma on the contraction property of the Rademacher complexity, which is useful to handle the nonlinear Lipschitz function ϕ.

Lemma 25 (Contraction Lemma (Bartlett and Mendelson, 2002)) *Suppose $\phi : \mathbb{R} \mapsto \mathbb{R}$ is contractive in the sense that $|\phi(t) - \phi(\tilde{t})| \leq |t - \tilde{t}|$. Then the following inequality holds for any $\widetilde{\mathcal{F}}$*

$$\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f \in \widetilde{\mathcal{F}}} \sum_{i=1}^{n} \epsilon_i \phi\big(f(x_i)\big) \leq \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f \in \widetilde{\mathcal{F}}} \sum_{i=1}^{n} \epsilon_i f(x_i).$$

Proof of Lemma 9 Let $\widetilde{S} = \{\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n\}$ with $\tilde{\mathbf{z}}_i = (\tilde{\mathbf{x}}_i, \tilde{y}_i) \in S$ for $i \in [n]$. According to Lemma 25, we know

$$\begin{aligned}
&\mathfrak{R}_{\widetilde{S}}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W} \in \mathcal{W}} \frac{1}{n} \sum_{i \in [n]} \epsilon_i \Phi(\mathbf{W}; \tilde{\mathbf{x}}_i)\Big] \\
&\leq \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W} \in \mathcal{W}} \frac{1}{n} \sum_{j \in [m]} \sum_{i \in [n]} \epsilon_i a_j \big(\sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_j) - \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_{1,j})\big)\Big] + \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\frac{1}{n} \sum_{j \in [m]} \sum_{i \in [n]} \epsilon_i a_j \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_{1,j})\Big] \\
&= \frac{1}{n} \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W} \in \mathcal{W}} \sum_{j \in [m]} \sum_{i \in [n]} \epsilon_i a_j \big(\sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_j) - \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_{1,j})\big)\Big]. \qquad (5.15)
\end{aligned}$$

For any $i \in [n]$, we define

$$S_i = \big\{j \in [m] : |\tilde{\mathbf{x}}_i^\top \mathbf{w}_{1,j}| \leq \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}\big\}.$$

For any $j \in [m]$, we define

$$\widetilde{S}_j := \big\{i \in [n] : |\tilde{\mathbf{x}}_i^\top \mathbf{w}_{1,j}| \leq \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}\big\}.$$

If $i \notin \widetilde{S}_j$, then analogous to Eq. (5.8), we know that the sign of $\tilde{\mathbf{x}}_i^\top \mathbf{w}_j$ is the same as the sign of $\tilde{\mathbf{x}}_i^\top \mathbf{w}_{1,j}$ (note $\|\mathbf{w}_j - \mathbf{w}_{1,j}\|_2 \leq \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}$). Therefore

$$\sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_j) - \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_{1,j}) = \xi_{ij} \tilde{\mathbf{x}}_i^\top (\mathbf{w}_j - \mathbf{w}_{1,j}),$$

where $\xi_{ij} := \mathbb{I}_{[\tilde{\mathbf{x}}_i^\top \mathbf{w}_{1,j} \geq 0]}$. Therefore, we know that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W} \in \mathcal{W}} \sum_{j \in [m]} \sum_{i \in [n]} \epsilon_i a_j \big(\sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_j) - \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_{1,j})\big)\Big] \\
&\leq \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W} \in \mathcal{W}} \sum_{j \in [m]} \sum_{i \in \widetilde{S}_j} \epsilon_i a_j \big(\sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_j) - \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_{1,j})\big)\Big] + \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W} \in \mathcal{W}} \sum_{j \in [m]} \sum_{i \notin \widetilde{S}_j} \epsilon_i a_j \big(\sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_j) - \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_{1,j})\big)\Big] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W} \in \mathcal{W}} \sum_{j \in [m]} \sum_{i \in \widetilde{S}_j} \epsilon_i a_j \big(\sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_j) - \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}_{1,j})\big)\Big] + \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W} \in \mathcal{W}} \sum_{j \in [m]} \sum_{i \notin \widetilde{S}_j} \epsilon_i a_j \xi_{ij} \tilde{\mathbf{x}}_i^\top \big(\mathbf{w}_j - \mathbf{w}_{1,j}\big)\Big]. \\
&\qquad (5.16)
\end{aligned}$$

Note that

$$\sum_{j \in [m]} |\widetilde{S}_j|^{\frac{1}{2}} \leq \Big(m \sum_{j \in [m]} |\widetilde{S}_j|\Big)^{\frac{1}{2}} = \Big(m \sum_{i \in [n]} |S_i|\Big)^{\frac{1}{2}}. \qquad (5.17)$$

For the first term in Eq. (5.16), Lemma 25 implies that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{j\in[m]}\sum_{i\in\widetilde{S}_j}\epsilon_i a_j\big(\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_j)-\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_{1,j})\big)\Big]\\
&\le\frac{1}{\sqrt{m}}\sum_{j\in[m]}\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{i\in\widetilde{S}_j}\epsilon_i\big(\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_j)-\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_{1,j})\big)\Big]\\
&=\frac{1}{\sqrt{m}}\sum_{j\in[m]}\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{i\in\widetilde{S}_j}\epsilon_i\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_j)\Big]\le\frac{1}{\sqrt{m}}\sum_{j\in[m]}\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{i\in\widetilde{S}_j}\epsilon_i\tilde{\mathbf{x}}_i^\top\mathbf{w}_j\Big],
\end{aligned}$$

where we have used the fact that $\epsilon_i a_j$ has the same distribution of ϵ_i. It follows that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{j\in[m]}\sum_{i\in\widetilde{S}_j}\epsilon_i a_j\big(\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_j)-\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_{1,j})\big)\Big]&\le\frac{1}{\sqrt{m}}\sum_{j\in[m]}\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{i\in\widetilde{S}_j}\epsilon_i\tilde{\mathbf{x}}_i^\top(\mathbf{w}_j-\mathbf{w}_{1,j})\Big]\\
&\le\frac{1}{\sqrt{m}}\sum_{j\in[m]}\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{w}_j:\|\mathbf{w}_j-\mathbf{w}_{1,j}\|_2\le\frac{\mathfrak{C}(\mathbf{W}^*)}{\sqrt{m}}}\sum_{i\in\widetilde{S}_j}\epsilon_i\tilde{\mathbf{x}}_i^\top(\mathbf{w}_j-\mathbf{w}_{1,j})\Big]\\
&\le\frac{1}{\sqrt{m}}\sum_{j\in[m]}\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{w}_j:\|\mathbf{w}_j-\mathbf{w}_{1,j}\|_2\le\frac{\mathfrak{C}(\mathbf{W}^*)}{\sqrt{m}}}\Big\|\sum_{i\in\widetilde{S}_j}\epsilon_i\tilde{\mathbf{x}}_i\Big\|_2\|\mathbf{w}_j-\mathbf{w}_{1,j}\|_2\Big]\\
&\le\frac{\mathfrak{C}(\mathbf{W}^*)}{m}\sum_{j\in[m]}\mathbb{E}_{\boldsymbol{\epsilon}}\Big\|\sum_{i\in\widetilde{S}_j}\epsilon_i\tilde{\mathbf{x}}_i\Big\|_2\le\frac{\mathfrak{C}(\mathbf{W}^*)}{m}\sum_{j\in[m]}|\widetilde{S}_j|^{\frac{1}{2}}\\
&\le\frac{\mathfrak{C}(\mathbf{W}^*)}{\sqrt{m}}\Big(\sum_{j\in[m]}|\widetilde{S}_j|\Big)^{\frac{1}{2}}=\frac{\mathfrak{C}(\mathbf{W}^*)}{\sqrt{m}}\Big(\sum_{i\in[n]}|S_i|\Big)^{\frac{1}{2}},
\end{aligned}$$

where we have used Eq. (5.17) and the standard result

$$\mathbb{E}_{\boldsymbol{\epsilon}}\Big\|\sum_{i\in\widetilde{S}_j}\epsilon_i\tilde{\mathbf{x}}_i\Big\|_2\le\Big(\mathbb{E}_{\boldsymbol{\epsilon}}\Big\|\sum_{i\in\widetilde{S}_j}\epsilon_i\tilde{\mathbf{x}}_i\Big\|_2^2\Big)^{\frac{1}{2}}=\Big(\sum_{i\in\widetilde{S}_j}\|\tilde{\mathbf{x}}_i\|_2^2\Big)^{\frac{1}{2}}\le|\widetilde{S}_j|^{\frac{1}{2}}.\tag{5.18}$$

For the second term in Eq. (5.16), we know (note ξ_{ij} is independent of ϵ_i)

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{j\in[m]}\sum_{i\notin\widetilde{S}_j}\epsilon_i a_j\xi_{ij}\tilde{\mathbf{x}}_i^\top\big(\mathbf{w}_j-\mathbf{w}_{1,j}\big)\Big] &= \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{j\in[m]}a_j\big(\mathbf{w}_j-\mathbf{w}_{1,j}\big)^\top\Big(\sum_{i\notin\widetilde{S}_j}\epsilon_i\xi_{ij}\tilde{\mathbf{x}}_i\Big)\Big]\\
&\le \frac{1}{\sqrt{m}}\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{j\in[m]}\big\|\mathbf{w}_j-\mathbf{w}_{1,j}\big\|_2\Big\|\sum_{i\notin\widetilde{S}_j}\epsilon_i\xi_{ij}\tilde{\mathbf{x}}_i\Big\|_2\Big]\\
&\le \frac{1}{\sqrt{m}}\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\Big(\sum_{j\in[m]}\big\|\mathbf{w}_j-\mathbf{w}_{1,j}\big\|_2^2\Big)^{\frac{1}{2}}\Big(\sum_{j\in[m]}\Big\|\sum_{i\notin\widetilde{S}_j}\epsilon_i\xi_{ij}\tilde{\mathbf{x}}_i\Big\|_2^2\Big)^{\frac{1}{2}}\Big]\\
&= \frac{1}{\sqrt{m}}\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\|\mathbf{W}-\mathbf{W}_1\|_2\Big(\sum_{j\in[m]}\Big\|\sum_{i\notin\widetilde{S}_j}\epsilon_i\xi_{ij}\tilde{\mathbf{x}}_i\Big\|_2^2\Big)^{\frac{1}{2}}\Big]\\
&\le \frac{\sqrt{\mathfrak{C}(\mathbf{W}^*)}}{\sqrt{m}}\Big(\sum_{j\in[m]}\mathbb{E}_{\boldsymbol{\epsilon}}\Big\|\sum_{i\notin\widetilde{S}_j}\epsilon_i\xi_{ij}\tilde{\mathbf{x}}_i\Big\|_2^2\Big)^{\frac{1}{2}}\\
&\le \frac{\sqrt{\mathfrak{C}(\mathbf{W}^*)}\sqrt{mn}}{\sqrt{m}} = \sqrt{\mathfrak{C}(\mathbf{W}^*)}\sqrt{n},
\end{aligned}$$

where we have used the Schwarz's inequality and Eq. (5.18). We plug the above two inequalities back into Eq. (5.16), and get

$$\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{j\in[m]}\sum_{i\in[n]}\epsilon_i a_j\big(\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_j)-\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_{1,j})\big)\Big] \le \frac{\mathfrak{C}(\mathbf{W}^*)}{\sqrt{m}}\Big(\sum_{i\in[n]}|S_i|\Big)^{\frac{1}{2}}+\sqrt{\mathfrak{C}(\mathbf{W}^*)}\sqrt{n}.$$

Since $\tilde{\mathbf{z}}_i\in S$ and $R_2\ge\mathfrak{C}(\mathbf{W}^*)/\sqrt{m}$, under the event E_2, we have

$$|S_i|\le 2R_2m/\sqrt{\pi}+\big(m\log(n/\delta)\big)^{\frac{1}{2}},\quad \forall i\in[n].$$

Therefore, under the event E_2, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{j\in[m]}\sum_{i\in[n]}\epsilon_i a_j\big(\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_j)-\sigma(\tilde{\mathbf{x}}_i^\top\mathbf{w}_{1,j})\big)\Big] \le\\
&\qquad\qquad \frac{\sqrt{n}\mathfrak{C}(\mathbf{W}^*)}{\sqrt{m}}\Big(\frac{2R_2m}{\sqrt{\pi}}+\big(m\log(n/\delta)\big)^{\frac{1}{2}}\Big)^{\frac{1}{2}}+\sqrt{\mathfrak{C}(\mathbf{W}^*)}\sqrt{n}.
\end{aligned}$$

We combine the above inequality and Eq. (5.15) and derive that

$$\mathfrak{R}_{\widetilde{S}}(\mathcal{F})\le\frac{\mathfrak{C}(\mathbf{W}^*)}{\sqrt{n}m^{\frac{1}{4}}}\Big(\frac{2\sqrt{m}R_2}{\sqrt{\pi}}+\big(\log(n/\delta)\big)^{\frac{1}{2}}\Big)^{\frac{1}{2}}+\frac{\sqrt{\mathfrak{C}(\mathbf{W}^*)}}{\sqrt{n}}.$$

The proof is completed. ■

To prove Theorem 12, we require the following two probabilistic inequalities for the deviation between empirical risks and population risks. Lemma 26 is the classical Bernstein inequality, while Lemma 27 shows optimistic bounds by using the smoothness of the loss function.

Lemma 26 (Bernstein inequality) *Let $\{\xi_i\}_{i=1}^n$ be a sequence of i.i.d. random variables. Let b be a constant such that $|\xi_i| \leq b$ and the variance $Var(\xi_i) < \infty$. Then, for any $0 < \delta < 1$ with probability at least $1-\delta$ there holds*

$$\Big|\frac{1}{n}\sum_{i=1}^n \big(\xi_i - \mathbb{E}[\xi_i]\big)\Big| \leq \frac{2b\log\frac{2}{\delta}}{3n} + \Big(\frac{2\,Var(\xi_i)\log\frac{2}{\delta}}{n}\Big)^{\frac{1}{2}}.$$

The following lemma gives optimistic generalization bounds for learning with smooth loss functions. Note that $\mathfrak{R}_{S,n}(\mathcal{F})$ is a *worst-case* Rademacher complexity restricted to S. In the supremum of $\widetilde{S} \subseteq S$ under the constraint $|\widetilde{S}| = n$, we allow $\widetilde{S}$ contains repeated elements of S. This construction is used to relate the fat-shattering dimension to Rademacher complexity (Srebro et al., 2010).

Lemma 27 (Srebro et al. 2010) *Let $\mathcal{W}$ be a set in $\mathbb{R}^{md}$. Let $\mathcal{F} := \{\mathbf{x} \mapsto \Psi(\mathbf{W};\mathbf{x}), \mathbf{W} \in \mathcal{W}\}$ and $\tilde{b} = \sup_{\mathbf{z},\mathbf{W}\in\mathcal{W}} f(\mathbf{W};\mathbf{z})$. For any $\delta \in (0,1)$, we have with probability at least $1-\delta/2$ over S, for any $\mathbf{W} \in \mathcal{W}$*

$$F(\mathbf{W}) - F_S(\mathbf{W}) \lesssim F_S^{\frac{1}{2}}(\mathbf{W})\Big(\sqrt{L\log^3 n}\mathfrak{R}_{S,n}(\mathcal{F}) + \Big(\frac{\tilde{b}\log(2/\delta)}{n}\Big)^{\frac{1}{2}}\Big) + L\log^3 n\mathfrak{R}_{S,n}^2(\mathcal{F}) + \frac{\tilde{b}\log(2/\delta)}{n}.$$

Let $\xi_i = f(\mathbf{W}^*;\mathbf{z}_i)$. We know that

$$\mathrm{Var}(\xi_i) \leq \mathbb{E}[\xi_i^2] \leq b_*\mathbb{E}[\xi_i] = b_*F(\mathbf{W}^*). \tag{5.19}$$

According to Lemma 27 and Lemma 26, we know that the event E_3 happens with probability at least $1-\delta$. By Schwarz's inequality, under the event E_3 we know

$$F(\mathbf{W}) - 2F_S(\mathbf{W}) \lesssim L\log^3 n\mathfrak{R}_{S,n}^2(\mathcal{F}) + \frac{b\log(2/\delta)}{n}, \quad \forall \mathbf{W} \in \mathcal{W} \tag{5.20}$$

and

$$F_S(\mathbf{W}^*) \leq 2F(\mathbf{W}^*) + \frac{7b_*\log(2/\delta)}{6n}. \tag{5.21}$$

We are now ready to prove Theorem 12. Since ℓ is nonnegative and smooth, we also have the following self-bounding property (Srebro et al., 2010)

$$|\ell'(a)| \leq \sqrt{2L\ell(a)}, \quad \forall a \in \mathbb{R}. \tag{5.22}$$

Furthermore, for any $(\mathbf{x}, y)$, the L-smoothness of ℓ implies

$$\begin{aligned}
&\ell(y\Phi(\mathbf{W};\mathbf{x})) \\
&\leq \ell(y\Phi(\mathbf{W}^*;\mathbf{x})) + y(\Phi(\mathbf{W};\mathbf{x}) - \Phi(\mathbf{W}^*;\mathbf{x}))\ell'(y\Phi(\mathbf{W}^*;\mathbf{x})) + \frac{L}{2}\big(\Phi(\mathbf{W};\mathbf{x}) - \Phi(\mathbf{W}^*;\mathbf{x})\big)^2 \\
&\leq \ell(y\Phi(\mathbf{W}^*;\mathbf{x})) + \frac{1}{2L}|\ell'(y\Phi(\mathbf{W}^*;\mathbf{x}))|^2 + L\big(\Phi(\mathbf{W};\mathbf{x}) - \Phi(\mathbf{W}^*;\mathbf{x})\big)^2 \\
&\leq 2\ell(y\Phi(\mathbf{W}^*;\mathbf{x})) + L\|\mathbf{W} - \mathbf{W}^*\|_2^2,
\end{aligned} \tag{5.23}$$

where we have used the self-bounding property in Eq. (5.22) and

$$(\Phi(\mathbf{W};\mathbf{x}) - \Phi(\mathbf{W}';\mathbf{x}))^2 \leq \frac{m}{m}\sum_{j=1}^m \big(\sigma(\mathbf{x}^\top\mathbf{w}_j) - \sigma(\mathbf{x}^\top\mathbf{w}_j')\big)^2 \leq \sum_{j=1}^m \|\mathbf{w}_j - \mathbf{w}_j'\|_2^2 = \|\mathbf{W} - \mathbf{W}'\|_2^2.$$

We apply the above inequality for each $(\mathbf{x}, y)$ in the dataset and use $\eta \le 1/L$ to derive that

$$F_S(\mathbf{W}) \le 2F_S(\mathbf{W}^*) + \frac{\|\mathbf{W} - \mathbf{W}^*\|_2^2}{\eta}.$$

Proof of Theorem 12 Let $\mathcal{W}$ be defined in Eq. (4.12) and $\mathcal{F} := \{\mathbf{x} \mapsto \Psi(\mathbf{W}; \mathbf{x}), \mathbf{W} \in \mathcal{W}\}$. According to Theorem 3, we know that if the events E_1 and E_3 hold, then the following inequalities hold for all $t \in [T]$

$$\|\mathbf{W}_t - \mathbf{W}^*\|_2^2 \le \mathfrak{C}(\mathbf{W}^*) \quad \text{and} \quad \|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty} \le \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}, \tag{5.24}$$

where we control $F_S(\mathbf{W}^*)$ in $\mathfrak{C}_S(\mathbf{W}^*)$ by $F(\mathbf{W}^*)$ via Eq. (5.21). Then, Eq. (5.24) implies that, under the event $E_1 \cap E_3$, we have $\mathbf{W}_t \in \mathcal{W}$. For any $\mathbf{W} \in \mathcal{W}$, Eq. (5.23) implies that

$$f(\mathbf{W}; \mathbf{z}) \le 2f(\mathbf{W}^*; \mathbf{z}) + L\|\mathbf{W} - \mathbf{W}^*\|_2^2 \le 2b_* + L\mathfrak{C}(\mathbf{W}^*) := b.$$

According to Lemma 9, we know

$$\mathfrak{R}_{S,n}(\mathcal{F}) \le \frac{\mathfrak{C}(\mathbf{W}^*)\big(m^{\frac{1}{4}}\sqrt{2R_2} + \log^{\frac{1}{4}}(n/\delta)\big)}{\sqrt{n}m^{\frac{1}{4}}} + \frac{\sqrt{\mathfrak{C}(\mathbf{W}^*)}}{\sqrt{n}}. \tag{5.25}$$

Under the event $E_1 \cap E_2 \cap E_3$, Eq. (5.20) and Eq. (5.25) imply

$$F(\mathbf{W}_t) - 2F_S(\mathbf{W}_t) \lesssim L\log^3 n\Big(\frac{\mathfrak{C}(\mathbf{W}^*)\big(m^{\frac{1}{4}}\sqrt{R_2} + \log^{\frac{1}{4}}(n/\delta)\big)}{\sqrt{n}m^{\frac{1}{4}}} + \frac{\sqrt{\mathfrak{C}(\mathbf{W}^*)}}{\sqrt{n}}\Big)^2 + \frac{b\log(2/\delta)}{n}.$$

Eq. (4.4) further implies that

$$\begin{aligned}
&\eta\sum_{t=1}^T F(\mathbf{W}_t) = \eta\sum_{t=1}^T \big(F(\mathbf{W}_t) - 2F_S(\mathbf{W}_t)\big) + 2\eta\sum_{t=1}^T F_S(\mathbf{W}_t) \\
&\lesssim \frac{L\eta T\log^3 n}{n}\Big(\frac{\mathfrak{C}^2(\mathbf{W}^*)(\sqrt{m}R_2 + \log^{\frac{1}{2}}(n/\delta))}{\sqrt{m}} + \mathfrak{C}(\mathbf{W}^*)\Big) + \frac{\eta T b\log(2/\delta)}{n} + \mathfrak{C}_S(\mathbf{W}^*).
\end{aligned}$$

The proof is completed by noting $b = 2b_* + L\mathfrak{C}(\mathbf{W}^*)$. ■

To prove Theorem 14, we introduce the following elementary lemma. We omit the proof for simplicity.

Lemma 28 *Let $a, b \ge 0$. If $x^2 \le ax + b$, then $x^2 \le a^2 + 2b$.*

Proof of Theorem 14 We choose $\mathbf{W}^* = \mathbf{W}^\epsilon$ with $\epsilon = 1/T$. We know

$$\begin{aligned}
&\sup_{\mathbf{z}} \ell(y\Phi(\mathbf{W}^\epsilon; \mathbf{x})) \le \sup_{\mathbf{z}} \big(\ell(y\Phi(\mathbf{w}^\epsilon; \mathbf{x})) - \ell(y\Phi(\mathbf{W}_1; \mathbf{x}))\big) + \sup_{\mathbf{z}} \ell(y\Phi(\mathbf{W}_1; \mathbf{x})) \\
&\le \sup_{\mathbf{x}} \big|\Phi(\mathbf{W}^\epsilon; \mathbf{x}) - \Phi(\mathbf{W}_1; \mathbf{x})\big| + \log 2 \le \sup_{\mathbf{x}} \sum_{j=1}^m |a_j||\sigma(\mathbf{x}^\top \mathbf{w}_j^\epsilon) - \sigma(\mathbf{w}_{1,j}^\top \mathbf{x})| + \log 2 \\
&\le \frac{1}{\sqrt{m}} \sup_{\mathbf{x}} \sum_{j=1}^m \big|(\mathbf{w}_j^\epsilon - \mathbf{w}_{1,j})^\top \mathbf{x}\big| + \log 2 \le \frac{1}{\sqrt{m}} \sum_{j=1}^m \big\|\mathbf{w}_j^\epsilon - \mathbf{w}_{1,j}\big\|_2 + \log 2 \\
&\le \Big(\sum_{j=1}^m \|\mathbf{w}_j^\epsilon - \mathbf{w}_{1,j}\|_2^2\Big)^{\frac{1}{2}} + \log 2 \le g(\epsilon) + \log 2,
\end{aligned} \tag{5.26}$$

where we have used the $\Phi(\mathbf{W}_1;\mathbf{x}) = 0$ and Schwarz's inequality. Therefore, we choose $b^* = g(\epsilon) + \log 2$. By Eq. (5.14), we know

$$\mathfrak{C}_S(\mathbf{W}^{\frac{1}{T}}) \leq 3\eta + g^2(1/T) \leq 3/L + g^2(1/T).$$

Under the event E_3 with $\mathbf{W}^* = \mathbf{W}^{\frac{1}{T}}$, we know

$$F(\mathbf{W}^{\frac{1}{T}}) \leq F_S(\mathbf{W}^{\frac{1}{T}}) + \frac{2b_*\log(2/\delta)}{3n} + \Big(\frac{2b_*F(\mathbf{W}^{\frac{1}{T}})\log(2/\delta)}{n}\Big)^{\frac{1}{2}}.$$

Solving the above quadratic inequality of $F^{\frac{1}{2}}(\mathbf{W}^{\frac{1}{T}})$ yields that (Lemma 28)

$$F(\mathbf{W}^{\frac{1}{T}}) \leq 2F_S(\mathbf{W}^{\frac{1}{T}}) + \frac{4b_*\log(2/\delta)}{3n} + \frac{2b_*\log(2/\delta)}{n} \leq \frac{2}{T} + \frac{10b_*\log(2/\delta)}{3n}. \tag{5.27}$$

Recall that $\|\mathbf{W}_1 - \mathbf{W}^{\frac{1}{T}}\|_2 \leq g(1/T)$ and the definition in Eq. (4.10), which implies

$$\begin{aligned}\mathfrak{C}(\mathbf{W}^{\frac{1}{T}}) &\leq 3\eta T\Big(\frac{4}{T} + \frac{20b_*\log(2/\delta)}{3n} + \frac{7b_*\log(2/\delta)}{6n}\Big) + \|\mathbf{W}_1 - \mathbf{W}^{\frac{1}{T}}\|_2^2 \\ &\leq 12\eta + \frac{24b_*\eta T\log(2/\delta)}{n} + \|\mathbf{W}_1 - \mathbf{W}^{\frac{1}{T}}\|_2^2 \lesssim b_*\log(1/\delta) + g^2(1/T). \end{aligned} \tag{5.28}$$

We then apply Theorem 12 to get

$$\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{W}_t) \lesssim \frac{L\log^3 n}{n}\Big(\frac{\mathfrak{C}^2(\mathbf{W}^{\frac{1}{T}})(\sqrt{m}R_2 + \log^{\frac{1}{2}}(n/\delta))}{\sqrt{m}} + \mathfrak{C}(\mathbf{W}^{\frac{1}{T}})\Big) + \frac{b_*\log(1/\delta)}{n} + \frac{g^2(1/T)}{\eta T}.$$

By the constraint on $m \geq \frac{256}{\pi}\big(3/L + g^2(1/T)\big)^4$, Eq. (5.26) and Eq. (5.28), we know

$$\frac{\mathfrak{C}^2(\mathbf{W}^{\frac{1}{T}})}{\sqrt{m}} \lesssim \frac{g^4(1/T) + b_*^2\log^2(1/\delta)}{\sqrt{m}} \lesssim 1 + \frac{b_*^2\log^2(1/\delta)}{\sqrt{m}} \lesssim 1$$

and therefore

$$\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{W}_t) = \widetilde{O}\Big(\frac{L\sqrt{m}R_2}{n} + \frac{g^2(1/T)}{n}\Big).$$

The proof is completed. ■

Proof of Corollary 15 According to the proof of Corollary 7, we can choose $R_1 = \frac{\sqrt{\pi}}{16(3/L+g^2(1/T))}$ such that the event E_1 happens with probability at least $1-\delta$. By Eq. (5.26) we can choose $b^* = g(1/T) + \log 2$. According to Eq. (5.28), we can choose

$$R_2 = \frac{12\eta + 24b_*\eta T n^{-1}\log(2/\delta) + g^2(1/T)}{\sqrt{m}}, \tag{5.29}$$

which is independent of $\mathbf{W}_1$. Therefore, the event E_2 happens with probability at least $1-\delta$. By Lemma 26 and Lemma 27, the event E_3 happens with probability at least $1-2\delta$.

Therefore, all the events E_1, E_2, E_3 happen, and we can apply Theorem 14 with R_2 given in Eq. (5.29) to derive the following inequality with probability at least $1 - 4\delta$

$$\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{W}_t) = \widetilde{O}\Big(\frac{L\big(b_* \log(1/\delta) + g^2(1/T)\big)}{n} + \frac{g^2(1/T)}{n}\Big).$$

The proof is completed by Eq. (5.26). ■

5.3 Proofs on NTK Separability

In this subsection, we present the proof on the connection with NTK separability.

Proof of Lemma 16 Since $y_i\Phi(\mathbf{W}^\epsilon; \mathbf{x}_i) = y_i\langle\nabla\Phi(\mathbf{W}^\epsilon; \mathbf{x}_i), \mathbf{W}^\epsilon\rangle$ by Eq. (5.3), we know

$$\begin{aligned}
&y_i\Phi(\mathbf{W}^\epsilon; \mathbf{x}_i)\\
&= y_i\langle\nabla\Phi(\mathbf{W}^\epsilon; \mathbf{x}_i) - \nabla\Phi(\mathbf{W}_1; \mathbf{x}_i), \mathbf{W}^\epsilon\rangle + y_i\langle\nabla\Phi(\mathbf{W}_1; \mathbf{x}_i); \mathbf{W}^\epsilon - \mathbf{W}_1\rangle + y_i\langle\nabla\Phi(\mathbf{W}_1; \mathbf{x}_i), \mathbf{W}_1\rangle\\
&= y_i\langle\nabla\Phi(\mathbf{W}^\epsilon; \mathbf{x}_i) - \nabla\Phi(\mathbf{W}_1; \mathbf{x}_i), \mathbf{W}^\epsilon\rangle + y_i\langle\nabla\Phi(\mathbf{W}_1; \mathbf{x}_i); \mathbf{W}^\epsilon - \mathbf{W}_1\rangle, \qquad (5.30)
\end{aligned}$$

where we have used the fact that $\langle\nabla\Phi(\mathbf{W}_1; \mathbf{x}_i), \mathbf{W}_1\rangle = \Phi(\mathbf{W}_1; \mathbf{x}_i) = 0$ due to our symmetric initialization. For the first term, we know

$$y_i\langle\nabla\Phi(\mathbf{W}^\epsilon; \mathbf{x}_i) - \nabla\Phi(\mathbf{W}_1; \mathbf{x}_i), \mathbf{W}^\epsilon\rangle = y_i\sum_{j=1}^{m} a_j\Big(\mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_j^\epsilon\geq 0]} - \mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_{1,j}\geq 0]}\Big)\mathbf{x}_i^\top\mathbf{w}_j^\epsilon.$$

If $|\mathbf{x}_i^\top\mathbf{w}_j^\epsilon| > \|\mathbf{w}_j^\epsilon - \mathbf{w}_{1,j}\|_2$, then analogous to Eq. (5.7), we have $\mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_j^\epsilon\geq 0]} - \mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_{1,j}\geq 0]} = 0$. Otherwise, we have

$$\Big|\Big(\mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_j^\epsilon\geq 0]} - \mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_{1,j}\geq 0]}\Big)\mathbf{x}_i^\top\mathbf{w}_j^\epsilon\Big| \leq \Big|\mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_j^\epsilon\geq 0]} - \mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_{1,j}\geq 0]}\Big|\|\mathbf{w}_j^\epsilon - \mathbf{w}_{1,j}\|_2.$$

Note that $\|\mathbf{W}^\epsilon - \mathbf{W}_1\|_{2,\infty} \leq \alpha_\epsilon\beta$. Let $\widetilde{S}_i = \big\{j \in [m] : |\mathbf{w}_{1,j}^\top\mathbf{x}_i| \leq \alpha_\epsilon\beta\big\}$. By Lemma 2, with probability at least $1 - \delta$ we have simultaneously the following inequality for all $i \in [n]$

$$|\widetilde{S}_i| \leq 2\alpha_\epsilon\beta m/\sqrt{\pi} + \big(m\log(n/\delta)\big)^{\frac{1}{2}}. \qquad (5.31)$$

Then, if $j \notin \widetilde{S}_i$, we have

$$|\mathbf{x}_i^\top\mathbf{w}_j^\epsilon - \mathbf{x}_i^\top\mathbf{w}_{1,j}| \leq \|\mathbf{w}_j^\epsilon - \mathbf{w}_{1,j}\|_2 \leq \alpha_\epsilon\beta < |\mathbf{w}_{1,j}^\top\mathbf{x}_i|,$$

which means that $\mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_j^\epsilon\geq 0]} - \mathbb{I}_{[\mathbf{x}_i^\top\mathbf{w}_{1,j}]} = 0$. Therefore, we have

$$\begin{aligned}
\big|y_i\langle\nabla\Phi(\mathbf{W}^\epsilon; \mathbf{x}_i) - \nabla\Phi(\mathbf{W}_1; \mathbf{x}_i), \mathbf{W}^\epsilon\rangle\big| &\leq \frac{1}{\sqrt{m}}\sum_{j\in\widetilde{S}_i}\|\mathbf{w}_j^\epsilon - \mathbf{w}_{1,j}\|_2\\
&\leq \frac{|\widetilde{S}_i|\|\mathbf{W}^\epsilon - \mathbf{W}_1\|_{2,\infty}}{\sqrt{m}} \leq \frac{|\widetilde{S}_i|\beta\alpha_\epsilon}{\sqrt{m}}. \qquad (5.32)
\end{aligned}$$

For the second term, by the construction of $\mathbf{W}^\epsilon$, we know

$$y_i\langle\nabla\Phi(\mathbf{W}_1;\mathbf{x}_i),\mathbf{W}^\epsilon-\mathbf{W}_1\rangle=\alpha_\epsilon y_i\langle\nabla\Phi(\mathbf{W}_1;\mathbf{x}_i),\mathbf{W}_*\rangle\geq\alpha_\epsilon\gamma, \tag{5.33}$$

where we have used Assumption 4. We plug Eq. (5.32) and Eq. (5.33) back into Eq. (5.30), and get

$$\begin{aligned}y_i\Phi(\mathbf{W}^\epsilon;\mathbf{x}_i)&\geq y_i\langle\nabla\Phi(\mathbf{W}_1;\mathbf{x}_i);\mathbf{W}^\epsilon-\mathbf{W}_1\rangle-\big|y_i\langle\nabla\Phi(\mathbf{W}^\epsilon;\mathbf{x}_i)-\nabla\Phi(\mathbf{W}_1;\mathbf{x}_i),\mathbf{W}^\epsilon\rangle\big|\\&\geq\alpha_\epsilon\gamma-\frac{|\widetilde{S}_i|\alpha_\epsilon\beta}{\sqrt{m}}\geq\alpha_\epsilon\gamma-\frac{2\alpha_\epsilon^2\beta^2\sqrt{m}}{\sqrt{\pi}}-\alpha_\epsilon\beta\log^{\frac{1}{2}}(n/\delta)\\&\geq\alpha_\epsilon\gamma-\alpha_\epsilon\gamma/4-\alpha_\epsilon\gamma/4=\alpha_\epsilon\gamma/2=\log(1/\epsilon),\end{aligned}$$

where we have used the assumption $\alpha_\epsilon\beta^2\leq\frac{\gamma\sqrt{\pi}}{8\sqrt{m}}$ and $\beta\leq\gamma/(4\log^{\frac{1}{2}}(n/\delta))$. This shows that

$$f(\mathbf{W}^\epsilon;\mathbf{z}_i)=\log(1+\exp(-y_i\Phi(\mathbf{W}^\epsilon;\mathbf{x}_i)))\leq\exp(-y_i\Phi(\mathbf{W}^\epsilon;\mathbf{x}_i))\leq\exp(-\log(1/\epsilon))=\epsilon.$$

The proof is completed. ■

Proof of Theorem 18 The logistic loss satisfies Assumption 1 with $L=1/4$. By Lemma 2 and Eq. (4.26), with probability at least $1-3\delta$ the event E_1 happens and $R_2\geq\mathfrak{C}(\mathbf{W}^{\frac{1}{T}})/\sqrt{m}$. According to Lemma 2, the events E_2 happen with probability at least $1-\delta$. According to Lemma 27 and Lemma 26, the event E_3 happens with probability at least $1-2\delta$. We now assume all these events happen simultaneously. We choose $\epsilon=1/T$. The constraint $\alpha_\epsilon\beta^2\leq\frac{\gamma\sqrt{\pi}}{8\sqrt{m}}$ becomes $(2\log T)\beta^2\leq\frac{\gamma^2\sqrt{\pi}}{8\sqrt{m}}$. By Lemma 16, Assumption 3 holds with $g(\epsilon)=(2\log 1/\epsilon)/\gamma$. Then $g(1/T)=\alpha_\epsilon=(2\log T)/\gamma$. By Eq. (4.22), we know $\mathfrak{C}_S(\mathbf{W}^{\frac{1}{T}})\leq 3/L+g^2(1/T)$. By Theorem 14, we know

$$\frac{1}{T}\sum_{t=1}^T F(\mathbf{W}_t)=\widetilde{O}\Big(\frac{\sqrt{m}R_2}{n}+\frac{g^2(1/T)}{n}\Big).$$

The proof is finished by the definition of R_2 in Eq. (4.23) and the bound of b_* in Eq. (5.26). ■

6 Conclusions

In this paper, we present some optimization and generalization analysis for gradient methods to train shallow ReLU networks with polylogarithmic width. We establish optimization error bounds of order $\mathfrak{C}_S(\mathbf{W}^*)/(\eta T)$, which shows an implicit regularization effect of gradient descent. We develop improved Rademacher complexity estimates for a hypothesis space motivated by our optimization analysis, which yields generalization bounds of order $\widetilde{O}(\mathfrak{C}(\mathbf{W}^*)/n)$. We apply our general result under a NTK separability condition, and derive risk bounds of order $\widetilde{O}(1/(n\gamma^2))$.

There remain several interesting problems for further study. First, we only consider SNNs. It would be interesting to investigate whether our analysis can be extended to deep ReLU networks. Second, we only consider fully-connected neural networks. It would be interesting to study the generalization and optimization of neural networks with other structures, e.g., recurrent neural networks and convolutional neural networks (Zhou, 2020).

Acknowledgment

The work of Yunwen Lei is supported by the Research Grants Council of Hong Kong [Project No. 22303723, 17302624]. The work of Puyu Wang is supported by the Alexander von Humboldt Foundation. The work of Yiming Ying is supported by Discovery Project (DP250101359) of Australian Research Council. The work of Ding-Xuan Zhou is supported by Discovery Project (DP240101919) of the Australian Research Council. The corresponding author is Yunwen Lei.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems*, volume 32, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019b.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332, 2019.

Yajie Bao, Amarda Shehu, and Mingrui Liu. Global convergence analysis of local sgd for two-layer neural network without overparameterization. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Francesca Bartolucci, Ernesto De Vito, Lorenzo Rosasco, and Stefano Vigogna. Understanding neural networks with reproducing kernel banach spaces. *Applied and Computational Harmonic Analysis*, 62:194–236, 2023.

Simone Bombari, Mohammad Hossein Amani, and Marco Mondelli. Memorization and optimization in deep neural networks with minimum over-parameterization. *Advances in Neural Information Processing Systems*, 35:7628–7640, 2022.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. *Advances in Neural Information Processing Systems*, 33:13363–13373, 2020.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep ReLU networks? In *International Conference on Learning Representation*, 2021.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

Puneesh Deora, Rouzbeh Ghaderi, Hossein Taheri, and Christos Thrampoulidis. On the optimization and generalization of multi-head attention. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *Journal of Machine Learning Research*, 24(303):1–49, 2023.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

David Holzmüller and Ingo Steinwart. Training two-layer relu networks with gradient descent is inconsistent. *Journal of Machine Learning Research*, 23(181):1–82, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. In *International Conference on Learning Representations*, 2019.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM*, 68(2):1–29, 2021.

Ilja Kuzborskij and Csaba Szepesvári. Learning lipschitz functions by gd-trained shallow overparameterized relu neural networks. *arXiv preprint arXiv:2212.13848*, 2022.

Antoine Ledent, Waleed Mustafa, Yunwen Lei, and Marius Kloft. Norm-based generalisation bounds for deep multi-class convolutional neural networks. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 8279–8287, 2021.

Yunwen Lei. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *Annual Conference on Learning Theory*, pages 191–227, 2023.

Yunwen Lei, Rong Jin, and Yiming Ying. Stability and generalization analysis of gradient methods for shallow neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 38557–38570, 2022.

Chaoyue Liu, Libin Zhu, and Misha Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33:15954–15964, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Fanghui Liu, Leello Dadi, and Volkan Cevher. Learning with norm constrained, over-parameterized, two-layer neural networks. *Journal of Machine Learning Research*, 25 (138):1–42, 2024.

Tong Mao and Ding-Xuan Zhou. Rates of approximation by ReLU shallow neural networks. *Journal of Complexity*, 79:101784, 2023.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

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

Atsushi Nitanda, Geoffrey Chinot, and Taiji Suzuki. Gradient descent can learn less over-parameterized two-layer neural networks on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Rahul Parhi and Robert D Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Transactions on Information Theory*, 69(2):1125–1140, 2022.

Dominic Richards and Ilja Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. *Advances in Neural Information Processing Systems*, 34, 2021.

Itay M Safran, Gilad Yehudai, and Ohad Shamir. The effects of mild over-parameterization on the optimization landscape of shallow relu neural networks. In *Conference on Learning Theory*, pages 3889–3934. PMLR, 2021.

Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In *Conference on Learning Theory*, pages 3380–3394, 2022.

Matan Schliserman and Tomer Koren. Tight risk bounds for gradient descent on separable data. *Advances in Neural Information Processing Systems*, 36, 2024.

Ohad Shamir. Gradient methods never overfit on separable data. *Journal of Machine Learning Research*, 22(85):1–20, 2021.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010.

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

Hossein Taheri and Christos Thrampoulidis. Generalization and stability of interpolating neural networks with minimal width. *Journal of Machine Learning Research*, 25(156): 1–41, 2024.

Puyu Wang, Yunwen Lei, Di Wang, Yiming Ying, and Ding-Xuan Zhou. Generalization guarantees of gradient descent for shallow neural networks. *Neural Computation*, 37(2): 344–402, 2025.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems*, 32, 2019.

Jingfeng Wu, Peter L Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *Annual Conference on Learning Theory*, pages 5019–5073, 2024.

Yunfei Yang and Ding-Xuan Zhou. Nonparametric regression using over-parameterized shallow ReLU neural networks. *Journal of Machine Learning Research*, 25(165):1–35, 2024.

Yunfei Yang and Ding-Xuan Zhou. Optimal rates of approximation by shallow relu k neural networks and applications to nonparametric regression. *Constructive Approximation*, 62 (2):329–360, 2025.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Tong Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020.

Mo Zhou, Rong Ge, and Chi Jin. A local convergence theory for mildly over-parameterized two-layer neural network. In *Conference on Learning Theory*, pages 4577–4632. PMLR, 2021.

Tian-Yi Zhou and Xiaoming Huo. Learning ability of interpolating deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 68:101582, 2024.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109:467–492, 2020.