Optimal Risk Bounds of Stochastic Gradient Descent for Shallow ReLU Networks*

Yuqing Liu[1,2] **Yunwen Lei**[2†]
[1] RPTU Kaiserslautern-Landau, Kaiserslautern, Germany
[2]Department of Mathematics, The University of Hong Kong, Hong Kong, China
yuliu6@hku.hk leiyw@hku.hk

Abstract

Recent work [21, 41] shows that gradient descent (GD) can achieve almost optimal risk bounds for shallow ReLU networks with a logarithmic width. However, these discussions require $O(n^2)$ gradient computations to achieve this optimality, which is computationally expensive for modern learning problems with a large sample size n. In this paper, we significantly improve the gradient computation complexity by showing that stochastic gradient descent (SGD) can achieve comparable risk bounds with only $O(n)$ stochastic gradient computations for shallow ReLU networks under logarithmic width conditions. Compared to GD, SGD introduces additional stochastic fluctuations along the optimization process, making the localization analysis more challenging. By leveraging concentration inequalities to control these fluctuations, we show that the entire SGD trajectory stays around its initialization point with high probability. Under an NTK separability condition with margin γ, we prove that SGD achieves near-optimal risk bounds $\widetilde{O}(1/(n\gamma^2))$ with a logarithmic width.

1 Introduction

Neural networks have become a powerful tool over the past few decades for addressing many challenging problems in modern machine learning and related applications. Despite their remarkable empirical success, the theoretical foundations underlying their optimization and generalization behavior remain incomplete. In particular, understanding shallow ReLU networks remains difficult since the associated objective function is nonsmooth and nonconvex. Although substantial progress has been made, many existing analyses still rely on either strong assumptions on overparameterization, smooth activation functions [41], or only imply guarantees in expectation [41]. For example, initial analyses often require the width to grow as a polynomial function of the sample size n and the iteration number T to obtain meaningful theoretical guarantees [1, 2, 49].

The seminal work [16] substantially advances the state of the art by showing that gradient descent (GD) can achieve global convergence when training shallow ReLU networks, provided that the width grows as a *logarithmic* function of n and T. This motivates several subsequent works on the convergence and generalization of GD in the training of neural networks with logarithmic width [21, 22, 41]. For example, the work [21] shows that optimization-dependent localization provides an effective framework for analyzing GD in shallow ReLU networks. The core idea is to control the optimization trajectory, use this control to define a localized hypothesis class, and then exploit the activation geometry of ReLU units to derive refined Rademacher complexity bounds. In particular, under an NTK separability assumption, this approach shows that a ReLU network

*The work of Yunwen Lei was supported by the Hong Kong Research Grants Council under the ECS project 22303723 and the GRF project 17304526.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

with polylogarithmic width can achieve a near-optimal rate of order $1/(n\gamma^2)$, improving upon the classical rate $1/(\sqrt{n}\gamma^2)$ established in [2, 6, 16, 32].

While the works [21, 22, 41] develop optimal risk bounds for GD, they require $O(n)$ iterations and therefore a total $O(n^2)$ gradient computations since GD requires going through all training examples per iteration. This computational complexity makes the result not applicable to large-scale problems often encountered in practice. A natural choice to relax the computation burden is to consider stochastic gradient descent (SGD) [29], where the true gradient is approximated by a stochastic gradient. In this way, the gradient complexity per iteration is reduced from $O(n)$ to $O(1)$. While SGD is more computationally efficient, it introduces random noise and fluctuations in the trajectory. These fluctuations create additional challenges in the theoretical analysis. For example, it is unclear whether the SGD trajectory can be controlled in a way similar to GD since the stochastic gradient noise adds significant fluctuation in the optimization process. This challenge asks for the development of new techniques in the localization analysis of SGD. Furthermore, it is also not clear whether the handling of fluctuation requires smaller step sizes to stabilize the optimization process, which in turn may lead to more iterations.

In this paper, we address the above-mentioned challenges by showing that SGD with $O(n)$ gradient computations can achieve almost optimal risk bounds. To our knowledge, this is the first result in a fixed-sample setting. Our contributions are summarized below.

1. We develop a high-probability optimization framework for SGD on shallow ReLU networks under a general self-bounding and smooth loss. In particular, we show that, conditional on a suitable initialization event, the SGD trajectory admits simultaneous localization in both the Euclidean norm and the $(2,\infty)$-norm. Moreover, the averaged empirical risk satisfies an optimization bound of order $\widetilde{O}\big(\frac{\mathfrak{C}_S(\mathbf{W}^*)}{\eta T}\big)$[3], where η is the step size, T is the iteration number and $\mathfrak{C}_S(\mathbf{W}^*)$ is an optimization-dependent radius determined by a reference model $\mathbf{W}^*$, the initialization $\mathbf{W}_1$, and the stochastic fluctuations of SGD. As compared to gradient descent, the analysis with SGD is more challenging since the SGD trajectory admits more fluctuations due to the noise in building stochastic gradients.
2. We combine this optimization-dependent localization with an initialization-dependent Rademacher complexity analysis to derive population-risk bounds for SGD. Specifically, we prove that the localized hypothesis class containing the SGD iterates has complexity of order $\widetilde{O}\big(\sqrt{\frac{\mathfrak{C}(\mathbf{W}^*)}{n}}\big)$, which leads to an averaged population-risk bound of order $\widetilde{O}\big(\frac{\mathfrak{C}(\mathbf{W}^*)}{n}\big)$. Here, n is the sample size and $\mathfrak{C}(\mathbf{W}^*)$ is a data-independent counterpart of $\mathfrak{C}_S(\mathbf{W}^*)$.
3. We apply the general theory to realizable and NTK-separable settings. Under NTK separability with margin γ, we show that shallow ReLU networks trained by SGD with $O(n)$ iterations achieve high-probability risk bounds of order $\widetilde{O}\big(\frac{1}{n\gamma^2}\big)$, which is minimax optimal up to logarithmic factors [37]. Significantly, our SGD analysis considers a fixed training set and repeatedly samples from it, in contrast to the one-pass online SGD analysis of Ji and Telgarsky [16], where each update uses a distinct sample drawn from the data distribution. Moreover, this substantially improves the $O(n^2)$ gradient computations required for GD to achieve similar risk bounds [21, 22, 41]. Our analyses only require a logarithmic width, which is consistent with the existing work [21, 22, 41].

The paper is structured as follows. Section 2 reviews related work on risk analysis of neural networks. The framework is introduced in Section 3. Section 4 presents the main optimization and generalization results. We conclude the paper in Section 5, and put all proofs in appendix.

2 Related Work

Optimization analysis. The optimization of overparameterized neural networks has been extensively studied through the neural tangent kernel (NTK) framework introduced by Jacot et al. [15]. This framework provides a linearization of neural networks around initialization and enables global convergence analyses for gradient-based methods in both shallow and deep architectures. Early results established convergence guarantees for GD applied to shallow ReLU networks under strong

[3] We use $\widetilde{O}$ to hide logarithmic factors.

overparameterization assumptions, typically requiring the network width to scale polynomially with the sample size [1, 2, 9]. These requirements have subsequently been relaxed by exploiting finer structural properties of the loss landscape, such as semi-smoothness and refined curvature bounds [33, 39, 49]. The seminal work [16] substantially relaxes the width condition by showing that shallow ReLU networks with a logarithmic width condition can achieve a global convergence. This extension were extended to GD with more general self-bounding loss functions [21] and GD with large step sizes [5, 44].

Generalization analysis. Generalization in neural networks has been widely studied through capacity-based approaches, especially via Rademacher complexity and covering numbers. These methods provide uniform convergence bounds for rich hypothesis classes and have led to a large body of results for networks [4, 7, 12, 13, 18, 19, 25, 48]. In particular, path-norm were often used to control the complexity of neural networks [10, 24, 30, 45]. For gradient-based training, such analyses typically yield bounds of order $\widetilde{O}(n^{-1/2})$ after combining optimization control with uniform convergence arguments [2, 16, 32].

A central difficulty in this direction is that classical capacity bounds are often too coarse to reflect the behavior of trained networks, especially in highly overparameterized regimes. Recent work has therefore sought finer, optimization-dependent complexity measures. In particular, the work [21, 22] showed that for GD on ReLU networks, the optimization trajectory can be localized and the resulting localized class admits a refined Rademacher analysis. More recently, the work [20] developed initialization-dependent and non-vacuous generalization bounds for overparameterized shallow networks by exploiting path-norm-type control relative to initialization.

Other than the uniform convergence analyses, there are also some recent work on developing generalization bounds of gradient methods on neural networks based on algorithmic stability [8, 26, 35, 42, 43, 46]. However, the stability approach often implies bounds in expectation, which cannot illustrate the high-order information of the SGD trajectory.

3 Problem Setup

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 the input space and $\mathcal{Y} = \{-1, 1\}$ is the output space. Let $S = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ be a dataset of n i.i.d. samples drawn independently from ρ, where each $\mathbf{z} = (\mathbf{x}, y) \in \mathcal{Z}$. We study one-hidden-layer ReLU networks of the form

$$\Phi(\mathbf{W}; \mathbf{x}) = \sum_{j=1}^{m} a_j \sigma(\mathbf{w}_j^\top \mathbf{x}), \qquad \sigma(a) = \max\{a, 0\}, \tag{3.1}$$

where m is the width of the hidden layer. Here, $\mathbf{w}_j \in \mathbb{R}^d$ denotes the connection weight for the j-th hidden unit, and $\mathbf{W} = (\mathbf{w}_1^\top, \ldots, \mathbf{w}_m^\top)^\top \in \mathbb{R}^{dm}$ collects all hidden layer weights $\mathbf{w}_j$ into a column vector. The coefficients $a_j \in \{-1/\sqrt{m}, 1/\sqrt{m}\}$ are fixed connection weights from the j-th hidden unit to the output node. Throughout the paper, we treat $\{a_j\}_{j=1}^m$ as given and only optimize the hidden weights $\mathbf{W}$, which is widely adopted in the existing analyses [2, 6, 11, 16, 32, 41].

Given a loss function $\ell : \mathbb{R} \mapsto \mathbb{R}_+$, we define the sample-wise loss at $\mathbf{z} = (\mathbf{x}, y)$ by

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

The empirical risk and population risk are defined as

$$F_S(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{W}; \mathbf{z}_i) \quad \text{and} \quad F(\mathbf{W}) = \mathbb{E}_{\mathbf{z}}[f(\mathbf{W}; \mathbf{z})] \tag{3.2}$$

where $\mathbb{E}_{\mathbf{z}}[\cdot]$ denotes the expectation with respect to (w.r.t.) $\mathbf{z} \sim \rho$. Let $N(0, I_d)$ denote a standard Gaussian distribution in $\mathbb{R}^d$.

Definition 1 (Initialization)**.** We adopt a symmetric initialization scheme that plays a central role in the analysis sections. Assume m is even. For $j = 1, \ldots, m/2$, we sample

$$\mathbf{w}_{1,j} \sim N(0, I_d) \quad \text{and set} \quad \mathbf{w}_{1,j+m/2} = \mathbf{w}_{1,j},$$

and assign output weights in opposite pairs:

$$a_{j+m/2} = -a_j, \qquad |a_j| = \frac{1}{\sqrt{m}}.$$

Under this construction, the network is initialized at zero: $\Phi(\mathbf{W}_1;\mathbf{x}) = 0, \forall \mathbf{x} \in \mathcal{X}$. We train the hidden weights $\mathbf{W}$ by stochastic gradient descent (SGD) with a constant step size $\eta > 0$.

Definition 2 (Stochastic gradient descent)**.** Let $\eta > 0$. Let $\mathbf{W}_1$ be initialized by the above symmetric random scheme. At iteration t, sample i_t uniformly from $[n] := \{1, \dots, n\}$ and update

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \nabla f(\mathbf{W}_t; \mathbf{z}_{i_t}), \quad t = 1, \dots, T. \tag{3.3}$$

Since ReLU activation is not differentiable everywhere, we interpret the gradient in Eq. (3.3) as a subgradient. In particular, the update can be written component-wise as

$$\mathbf{w}_{t+1,j} = \mathbf{w}_{t,j} - \eta a_j y_{i_t} \ell'(y_{i_t} \Phi(\mathbf{W}_t; \mathbf{x}_{i_t})) \mathbb{I}_{[\mathbf{w}_{t,j}^\top \mathbf{x}_{i_t} \geq 0]} \mathbf{x}_{i_t}, \quad \forall j \in [m], \tag{3.4}$$

where $\ell'(a)$ denotes the derivative of ℓ at a and $\mathbb{I}_{[\cdot]}$ is the indicator function, i.e., returning 1 if the argument holds and 0 otherwise. We impose standard assumptions on the loss function and the input data. We call ℓ is convex if $\ell(a) \geq \ell(a') + (a - a')\ell'(a')$ for any $a, a' \in \mathbb{R}$.

Assumption 1. Suppose 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 $\widetilde{L} > 0$ and $|\ell'(a)| \leq G$ for some $G > 0$.

Assumption 1 is satisfied, for example, by the logistic loss $\ell(a) = \log(1 + \exp(-a))$, as well as by polynomially-tailed and sub-exponential-tailed losses [36]. For simplicity, we set $\widetilde{L} = 1$ throughout the paper. We let $\|\cdot\|_2$ denote the Euclidean norm, which becomes the Frobenius norm if applied to a matrix. The following assumption is standard in risk analysis of SNNs [2, 14, 16].

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 [21]

$$\|\nabla f(\mathbf{W}; \mathbf{z})\|_2 = \|\nabla \Phi(\mathbf{W}; \mathbf{x})\|_2 |\ell'(y\Phi(\mathbf{W}; \mathbf{x}))| \leq \ell(y\Phi(\mathbf{W}; \mathbf{x})) = f(\mathbf{W}; \mathbf{z}), \tag{3.5}$$

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

4 Main Results

Our goal is to control the average population risk along the SGD trajectory. The overall decomposition picture can be expressed as

$$\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{W}_t) = \frac{1}{T}\sum_{t=1}^{T} \big(F(\mathbf{W}_t) - 2F_S(\mathbf{W}_t)\big) + \frac{2}{T}\sum_{t=1}^{T} F_S(\mathbf{W}_t).$$

The second term is controlled by the optimization analysis from Section 4.1. For the first term, we will derive a uniform bound over a localized hypothesis class containing the SGD iterates, by combining a Rademacher complexity estimate with an optimistic generalization bound.

4.1 Optimization Analysis

We first present the optimization analysis, which provides high-probability control of the SGD trajectory. This control will later be used to define a localized hypothesis class for the generalization analysis. The following lemma shows the behavior of the network initialization.

Lemma 1 ([16])**.** *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*

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

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

In our analysis, we often encounter an event on the initialization w.r.t. $\mathbf{W}_1$, which we will show with high probability. For brevity of reference, we give a formal definition. For any $\mathbf{W}, \mathbf{W}'$, we denote $\|\mathbf{W} - \mathbf{W}'\|_{2,\infty} = \max_{j \in [m]} \|\mathbf{w}_j - \mathbf{w}'_j\|_2$. Let

$$\mathfrak{C}_S(\mathbf{W}^*) := 2\|\mathbf{W}_1 - \mathbf{W}^*\|_2^2 + 4\eta T F_S(\mathbf{W}^*) + 36\eta^2 G^2 \log^2 \frac{2T}{\delta} \tag{4.2}$$

and

$$\widehat{R}_1(\mathfrak{C}_S(\mathbf{W}^*)) := \Big(1+\frac{2L}{G}\Big)\frac{\mathfrak{C}_S(\mathbf{W}^*)}{\sqrt{m}} + \frac{G\eta\log(mT/\delta)}{\sqrt{m}}. \tag{4.3}$$

Definition 3 (Initialization event E_1)**.** Fix a reference model $\mathbf{W}^*$ and let $R_1 > 0$ be a number independent of $\mathbf{W}_1$. Define E_1 (w.r.t. $\mathbf{W}_1$) as the event that:

(i) For all $i \in [n]$,

$$\big|\{j\in[m] : |\mathbf{w}_{1,j}^\top \mathbf{x}_i| \le R_1\}\big| \le \frac{2R_1 m}{\sqrt{\pi}} + \Big(\frac{m\log(n/\delta)}{2}\Big)^{\frac{1}{2}} =: C_{R_1}. \tag{4.4}$$

(ii) We have

$$\max\Big\{\widehat{R}_1(\mathfrak{C}_S(\mathbf{W}^*)),\ \|\mathbf{W}^*-\mathbf{W}_1\|_{2,\infty}\Big\} \le R_1 \le \frac{\sqrt{\pi}}{16\,\mathfrak{C}_S(\mathbf{W}^*)}. \tag{4.5}$$

By Lemma 1, part (i) holds with probability at least $1-\delta$. In the NTK section, we verify that part (ii) also holds with high probability under an NTK separability assumption.

Theorem 2, to be proved in Section A.1, illustrates the convergence behavior of SGD. Eq. (4.6) gives convergence rates for an average of empirical risks under the constraint $m \gtrsim \mathfrak{C}_S^2(\mathbf{W}^*)$.

Theorem 2 (Optimization rate for SGD)**.** *Let Assumptions 1, 2 hold. Let R_1 and the event E_1 be as in Definition 3 and $\{\mathbf{W}_k\}_{k=1}^{T+1}$ be generated by SGD with step size $\eta \le \frac{1}{256L\log(2T/\delta)}$. Assume the width satisfies $m \ge 32\mathfrak{C}_S^2(\mathbf{W}^*)\log\frac{n}{\delta}$. Then, conditional on E_1, with probability at least $1-2\delta$ w.r.t. $\{i_t\}$, for all $t \in [T+1]$, we have*

$$\|\mathbf{W}_t-\mathbf{W}^*\|_2^2 \le \mathfrak{C}_S(\mathbf{W}^*), \qquad \eta\sum_{k=1}^{t-1} F_S(\mathbf{W}_k) \le \mathfrak{C}_S(\mathbf{W}^*). \tag{4.6}$$

Furthermore, for all $t \in [T+1]$, there holds

$$\|\mathbf{W}_t-\mathbf{W}_1\|_{2,\infty} \le \widehat{R}_1\big(\mathfrak{C}_S(\mathbf{W}^*)\big). \tag{4.7}$$

Remark 1. Theorem 2 shows not only empirical-risk control but also pathwise localization of SGD. The Euclidean bound in (4.6) and the $(2,\infty)$ bound in (4.7) together ensure that the SGD iterates stay in a controlled local class, which is the basis for the generalization analysis. Compared with deterministic GD analyses such as [16, 21, 28, 41], the SGD proof must additionally control martingale fluctuations from sample-wise updates.

For the above optimization error bounds, our overparameterization requirement assumption involves $\mathfrak{C}_S(\mathbf{W}^*)$. To interpret this constraint, we introduce the following realizability assumption for SNNs. This assumption has been considered in the literature for linear models [36] and SNNs with smooth activation functions [41]. We will verify this assumption for SNNs with the ReLU activation function. The proof of Theorem 3 is given in Section A.1.

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

$$F_S(\mathbf{W}^\epsilon) \le \epsilon, \qquad \|\mathbf{W}^\epsilon-\mathbf{W}_1\|_2 \le g(\epsilon)$$

and

$$\|\mathbf{W}^\epsilon-\mathbf{W}_1\|_{2,\infty} \le \frac{\sqrt{\pi}}{16\mathfrak{C}_S(\mathbf{W}^\epsilon)}.$$

Theorem 3. *Let Assumptions 1, 2, 3 hold. Suppose $m \ge 32\log\frac{n}{\delta}\big(2g^2(\frac{1}{T})+\frac{1}{64L\log\frac{2T}{\delta}}+\frac{9G^2}{16384L^2}\big)^2$ and event E_1 holds. If $\eta \le \frac{1}{256L\log(2T/\delta)}$, then conditional on E_1, with probability at least $1-2\delta$ w.r.t. $\{i_t\}$, we can find $\mathbf{W}^{\frac{1}{T}}$ such that for all $t \in [T+1]$*

$$\|\mathbf{W}_t-\mathbf{W}^{\frac{1}{T}}\|_2^2 \le 2g^2(\frac{1}{T})+4\eta+36\eta^2G^2\log^2\frac{2T}{\delta} \tag{4.8}$$

and

$$\frac{1}{T}\sum_{k=1}^{T} F_S(\mathbf{W}_k) \le \frac{2g^2(1/T)}{\eta T} + \frac{4}{T} + \frac{36\eta G^2}{T}\log^2\frac{2T}{\delta}. \tag{4.9}$$

Furthermore,

$$\|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty} \le \Big(1+\frac{2L}{G}\Big)\frac{2g^2(\frac{1}{T}) + 4\eta + 36\eta^2G^2\log^2\frac{2T}{\delta}}{\sqrt{m}} + \frac{G\eta\log(mT/\delta)}{\sqrt{m}}, \qquad \forall t \in [T+1]. \tag{4.10}$$

Corollary 4 is a direct application of Theorem 3, where we remove the assumption on the event E_1.

Corollary 4. *Let*

$$\bar{B}_T := 2g^2(1/T) + \frac{1}{64L\log\frac{2T}{\delta}} + \frac{9G^2}{16384L^2}.$$

Suppose that Assumptions 1, 2, and 3 hold. Assume $m \ge 32\log\frac{n}{\delta}\,\bar{B}_T^2$, $\eta \le \frac{1}{256L\log(2T/\delta)}$,

$$\|\mathbf{W}^{\frac{1}{T}} - \mathbf{W}_1\|_{2,\infty} \le \frac{\sqrt{\pi}}{16\bar{B}_T} \tag{4.11}$$

and

$$\Big(1+\frac{2L}{G}\Big)\frac{\bar{B}_T}{\sqrt{m}} + \frac{G\eta\log(mT/\delta)}{\sqrt{m}} \le \frac{\sqrt{\pi}}{16\bar{B}_T}. \tag{4.12}$$

Then, with probability at least $1-3\delta$*, the inequalities (4.8),(4.9) and (4.10) hold for all* $t \in [T+1]$.

Remark 2. If $g(\epsilon) \lesssim \log(1/\epsilon)$[4], then by Corollary 4, we have that with probability at least $1-3\delta$, for all $t \in [T+1]$,

$$\|\mathbf{W}_t - \mathbf{W}^{\frac{1}{T}}\|_2^2 = \widetilde{O}(\log^2 T), \quad \frac{1}{T}\sum_{k=1}^{T} F_S(\mathbf{W}_k) = \frac{\widetilde{O}(\log^2 T)}{\eta T}$$

and

$$\|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty} = \frac{\widetilde{O}(\log^2 T)}{\sqrt{m}}.$$

In this case, the width constraint on the SNN is sufficient up to a polylogarithmic factor. Specifically, by combining the constraints in Eq. (4.12) and $m \ge 32\log\frac{n}{\delta}$, we obtain $m \gtrsim \log^8 T$.

4.2 Generalization Analysis

We now turn to the generalization analysis. The key idea is to combine the optimization-dependent localization with a uniform convergence argument over the induced hypothesis class, which depends on the complexity of the space.

Definition 4 (Rademacher complexity [3])**.** Let $\mathcal{F}$ be a class of real-valued functions over a space $\mathcal{Z}$ and let $\mathcal{S} = \{z_i\}_{i=1}^n \subset \mathcal{Z}$. The empirical Rademacher complexity of $\mathcal{F}$ w.r.t. $\mathcal{S}$ is defined as

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

where $\boldsymbol{\epsilon} = (\epsilon_i)_{i\in[n]} \sim \{\pm 1\}^n$ are independent Rademacher variables.

We assume that the reference model $\mathbf{W}^*$ is independent of the sample S throughout this section. Let $b_* \ge \sup_z f(\mathbf{W}^*; z)$ be independent of S and $\mathbf{W}_1$. Define the data-independent complexity radius

$$\mathfrak{C}(\mathbf{W}^*) := 2\|\mathbf{W}_1 - \mathbf{W}^*\|_2^2 + 4\eta T\Big(2F(\mathbf{W}^*) + \frac{7\,b_*\log(2/\delta)}{6n}\Big) + 36\eta^2G^2\log^2\frac{2T}{\delta}. \tag{4.13}$$

We consider the localized parameter class

[4]For nonnegative quantities A and B, we write $A \lesssim B$ or $A \gtrsim B$ if there is a universal constant $\kappa > 0$ such that $A \le \kappa B$ or $A \ge \kappa B$. We denote $A \asymp B$ if there exist universal constants κ_1, κ_2 such that $\kappa_1 A \le B \le \kappa_2 A$.

$$\mathcal{W} := \Big\{ \mathbf{W} \in \mathbb{R}^{md} : \|\mathbf{W} - \mathbf{W}^*\|_2^2 \le \mathfrak{C}(\mathbf{W}^*) \Big\} \tag{4.14}$$

and the hypothesis class

$$\mathcal{F} := \{ \mathbf{x} \mapsto \Phi(\mathbf{W}; \mathbf{x}) : \mathbf{W} \in \mathcal{W} \}. \tag{4.15}$$

For the sample S, define

$$\mathfrak{R}_{S,n}(\mathcal{F}) := \sup_{\widetilde{S} \subseteq S : |\widetilde{S}| = n} \mathfrak{R}_{\widetilde{S}}(\mathcal{F}). \tag{4.16}$$

The following lemma gives an estimate of the Rademacher complexity of the localized class.

Lemma 5. *Let Assumption 2 hold. Let $\mathcal{W}$ and $\mathcal{F}$ be defined as in* (4.14) *and* (4.15)*. Then*

$$\mathfrak{R}_{S,n}(\mathcal{F}) = \widetilde{O}\Big(\sqrt{\frac{\mathfrak{C}(\mathbf{W}^*)}{n}} \Big). \tag{4.17}$$

Remark 3. The work [16] considered the function space

$$\mathcal{F}_1 := \{ \mathbf{x} \mapsto \Phi(\mathbf{W}; \mathbf{x}) : \|\mathbf{W} - \mathbf{W}_1\|_{2,\infty} \le \mathfrak{C}(\mathbf{W}^*)/\sqrt{m} \}$$

and showed that

$$\mathfrak{R}_{S,n}(\mathcal{F}_1) \le \mathfrak{C}(\mathbf{W}^*)/\sqrt{n}. \tag{4.18}$$

The recent work [21] considered the function space

$$\mathcal{F}_2 := \{ \mathbf{x} \mapsto \Phi(\mathbf{W}; \mathbf{x}) : \|\mathbf{W} - \mathbf{W}^*\|_2^2 \le \mathfrak{C}(\mathbf{W}^*), \|\mathbf{W} - \mathbf{W}_1\|_{2,\infty} \le \mathfrak{C}(\mathbf{W}^*)/\sqrt{m} \}.$$

Under a condition that Eq. (4.1) holds with $R \asymp \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}$, it was shown there [21] that

$$\mathfrak{R}_{S,n}(\mathcal{F}_2) \lesssim \frac{\mathfrak{C}(\mathbf{W}^*)\big(\mathfrak{C}^{\frac{1}{2}}(\mathbf{W}^*) + \log^{\frac{1}{4}}(n/\delta)\big)}{\sqrt{n} m^{\frac{1}{4}}} + \frac{\sqrt{\mathfrak{C}(\mathbf{W}^*)}}{\sqrt{n}}. \tag{4.19}$$

As compared to Eq. (4.18), we consider a local space with a constraint by $\|\cdot\|_2$ instead of $\|\cdot\|_{2,\infty}$. Furthermore, our bound improves Eq. (4.18) by replacing $\mathfrak{C}(\mathbf{W}^*)$ with $\mathfrak{C}^{\frac{1}{2}}(\mathbf{W}^*)$. As compared to Eq. (4.19), our Rademacher complexity bound removes the constraint $\|\mathbf{W} - \mathbf{W}_1\|_{2,\infty} \le \mathfrak{C}(\mathbf{W}^*)/\sqrt{m}$ in the function space, the term $\frac{\mathfrak{C}(\mathbf{W}^*)\left(\mathfrak{C}^{\frac{1}{2}}(\mathbf{W}^*) + \log^{\frac{1}{4}}(n/\delta)\right)}{\sqrt{n} m^{\frac{1}{4}}}$ in the upper bound and also the condition on Eq. (4.1).

For brevity, we define another event that holds with high probability. With high probability, Eq. (4.20) follows by local Rademacher complexity analysis [40], while Eq. (4.21) follows by a standard Bernstein inequality.

Definition 5 (Generalization event E_2)**.** Let $b := 2b_* + L\mathfrak{C}(\mathbf{W}^*)$. We define E_2 as the event that the following inequalities hold simultaneously:

(i) for all $\mathbf{W} \in \mathcal{W}$,

$$F(\mathbf{W}) - F_S(\mathbf{W}) \lesssim F_S(\mathbf{W})^{1/2} \Big(\sqrt{L \log^3 n} \mathfrak{R}_{S,n}(\mathcal{F}) + \Big(\frac{b \log(2/\delta)}{n} \Big)^{1/2} \Big) + L \log^3 n \, \mathfrak{R}_{S,n}^2(\mathcal{F}) + \frac{b \log(2/\delta)}{n}; \tag{4.20}$$

(ii) for the reference model $\mathbf{W}^*$,

$$|F_S(\mathbf{W}^*) - F(\mathbf{W}^*)| \le \frac{2b_* \log(2/\delta)}{3n} + \Big(\frac{2b_* F(\mathbf{W}^*) \log(2/\delta)}{n} \Big)^{1/2}. \tag{4.21}$$

We are now ready to present population risk bounds under the event $E_1 \cap E_2$.

Theorem 6 (Risk bounds for SGD)**.** *Let the assumptions of Theorem 2 hold. Then, on the event $E_1 \cap E_2$, we have*

$$\frac{1}{T} \sum_{t=1}^{T} F(\mathbf{W}_t) = \widetilde{O}\Big(\frac{\mathfrak{C}(\mathbf{W}^*)}{n} + \frac{b \log(2/\delta)}{n} + \frac{\mathfrak{C}(\mathbf{W}^*)}{\eta T} \Big). \tag{4.22}$$

Theorem 7 (Risk bounds under realizability)**.** *Let Assumptions 1, 2, and 3 hold. Assume* $\eta \le \frac{1}{256L\log(2T/\delta)}$ *and* $\eta T \asymp n$*. Then on* $E_1 \cap E_2$*, we have*

$$\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{W}_t) = \widetilde{O}\Big(\frac{g^2(1/T)}{n}\Big). \tag{4.23}$$

Theorem 6 and Theorem 7 require $E_1 \cap E_2$ to hold. Corollary 8 removes this requirement. To our knowledge, this gives the first excess risk bounds of order $\widetilde{O}(1/n)$ for SGD on ReLU networks.

Corollary 8. *Let* $\delta \in (0,1)$*, and Assumptions 1, 2, 3 hold with* $\mathbf{W}^\epsilon$ *independent of* S*. Assume* $\eta \le \frac{1}{256L\log(2T/\delta)}$*,* $\eta T \asymp n$*,* $m \ge \frac{128}{\pi}\bar{B}_T^4 \ \vee \ 32\log\frac{n}{\delta}\,\bar{B}_T^2$ *and Eq.* (4.11)*, Eq.* (4.12) *hold. Then, with probability at least* $1-4\delta$*, we have*

$$\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{W}_t) = \widetilde{O}\Big(\frac{g^2(1/T)}{n}\Big). \tag{4.24}$$

In particular, if $g(\epsilon) \lesssim \log(1/\epsilon)$*, then* $\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{W}_t) = \widetilde{O}(1/n)$*.*

Remark 4. Our results provide a high-probability framework for analyzing SGD in shallow ReLU networks through optimization-dependent localization. Unlike analyses based on global landscape conditions or purely linearized NTK dynamics, our approach derives a local hypothesis class directly from the stochastic training trajectory and combines it with refined Rademacher complexity bounds. Our results complement several modern directions in the theory of stochastic optimization for neural networks. In contrast to analyses based on global geometric conditions such as Polyak–Łojasiewicz or Łojasiewicz-type inequalities [1, 17, 23], our approach does not assume a favorable landscape a priori; instead, it derives localization directly from the SGD dynamics.

4.3 NTK separability

In this subsection, we connect our optimization and generalization analysis for SGD to the classical NTK separability condition, which assumes that the data distribution is separable by the random NTK feature map at initialization with margin γ. Our goal is to verify that the event E_1 in Definition 3 holds with high probability under NTK separability and symmetric initialization. As a consequence, our general results recover the population risk bound $\widetilde{O}(1/(n\gamma^2))$.

Assumption 4 (Separability by NTK [32])**.** Let $\gamma, \beta > 0$. Assume there exists $\mathbf{v}_* \in \mathbb{R}^{md}$ with

$$\|\mathbf{v}_*\|_2 = 1, \qquad \|\mathbf{v}_*\|_{2,\infty} \le \beta,$$

such that

$$y\big\langle \nabla\Phi(\mathbf{W}_1;\mathbf{x}),\, \mathbf{v}_*\big\rangle \ge \gamma \quad \text{almost surely for } (\mathbf{x},y)\sim\rho.$$

Note that the reference model $\mathbf{W}^*$ is assumed to be independent of the training sample S throughout Section 4.2. Under Assumption 4, $\mathbf{v}_*$ may depend on the initialization $\mathbf{W}_1$, but is independent of S. Hence, the reference model $\mathbf{W}^\epsilon = \mathbf{W}_1 + \alpha_\epsilon \mathbf{v}_*$ is also independent of S.

The next lemma shows that Assumption 3 is implied by Assumption 4 in the case of the logistic loss. In particular, it yields the choice $g(\epsilon) \lesssim \log(1/\epsilon)/\gamma$.

Lemma 9 (Realizability from NTK separability [21])**.** *Suppose Assumption 4 holds and let* ℓ *be the logistic loss. For any* $\epsilon \in (0,1)$*, define*

$$\mathbf{W}^\epsilon = \mathbf{W}_1 + \alpha_\epsilon \mathbf{v}_*, \qquad \alpha_\epsilon = \frac{2\log(1/\epsilon)}{\gamma}. \tag{4.25}$$

If

$$\alpha_\epsilon \beta^2 \le \frac{\gamma\sqrt{\pi}}{8\sqrt{m}}, \qquad \beta \le \frac{\gamma}{4\log^{1/2}(n/\delta)}, \tag{4.26}$$

then with probability at least $1-\delta$ *over the randomness of* $\mathbf{W}_1$*, we have*

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

Moreover,

$$\|\mathbf{W}^\epsilon - \mathbf{W}_1\|_2 = \alpha_\epsilon, \qquad \|\mathbf{W}^\epsilon - \mathbf{W}_1\|_{2,\infty} \le \alpha_\epsilon\beta.$$

Taking $\epsilon = 1/T$ in Lemma 9, set $\alpha = 2\log T/\gamma$ and $\mathbf{W}^{1/T} = \mathbf{W}_1 + \alpha \mathbf{v}_*$. Then, with high probability, we know

$$F_S(\mathbf{W}^{1/T}) \le \frac{1}{T}, \qquad \|\mathbf{W}^{1/T} - \mathbf{W}_1\|_2 = \alpha, \qquad \|\mathbf{W}^{1/T} - \mathbf{W}_1\|_{2,\infty} \le \alpha\beta.$$

Consequently,

$$\mathfrak{C}_S(\mathbf{W}^{1/T}) \le 2\alpha^2 + 4\eta + 36\eta^2 G^2 \log^2 \frac{2T}{\delta} =: \bar{C}_T. \tag{4.27}$$

The corresponding population-radius bound for $\mathfrak{C}(\mathbf{W}^{1/T})$ is derived in Appendix A.3. Recall $\widehat{R}_1(\cdot)$ from Definition 3(ii). We choose

$$R_1 := \max\Big\{\widehat{R}_1(\bar{C}_T), \alpha\beta\Big\}, \tag{4.28}$$

where $\widehat{R}_1(\bar{C}_T) = \big(1 + \frac{2L}{G}\big)\frac{\bar{C}_T}{\sqrt{m}} + \frac{G\eta\log(mT/\delta)}{\sqrt{m}}$. The next lemma verifies that the localization event E_1 holds under explicit NTK-type conditions.

Lemma 10. *Under Assumptions 2 and 4, and* (4.26) *with $\epsilon = 1/T$, assume*

$$\alpha\beta\,\bar{C}_T \le \frac{\sqrt{\pi}}{16}, \qquad \widehat{R}_1(\bar{C}_T)\,\bar{C}_T \le \frac{\sqrt{\pi}}{16},$$

and

$$m \ge \max\Big\{\frac{128}{\pi}\bar{C}_T^4,\ 32\bar{C}_T^2 \log\frac{n}{\delta}\Big\}.$$

Then with probability at least $1 - 2\delta$, the event E_1 holds with reference model $\mathbf{W}^* = \mathbf{W}^{1/T}$.

We now state the resulting population risk bound under NTK separability.

Theorem 11 (Risk bound under NTK separability)**.** *Let $\delta \in (0,1)$, and suppose Assumptions 1, 2, and 4 hold. Assume $\eta \le \frac{1}{256L\log(2T/\delta)}$, $\eta T \asymp n$. Set $\alpha = \frac{2\log T}{\gamma}$. If Eq.* (4.26) *and the assumptions of Lemma 10 hold with $\epsilon = 1/T$, then with probability at least $1 - \delta$,*

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

Remark 5. Under the NTK separability, Lemma 9 shows that Assumption 3 holds with $g(\epsilon) \lesssim \log(1/\epsilon)/\gamma$. Hence Theorem 7 immediately implies the rate $\widetilde{O}(1/(n\gamma^2))$, which, according to the discussions in [37, 38] is optimal up to a logarithmic factor. This rate significantly improves the existing rate of order $\widetilde{O}(1/\sqrt{n}\gamma^2)$ [16]. The recent work [21, 41] also gave rates of order $\widetilde{O}(1/(n\gamma^2))$ for GD with $T \asymp n$. Since GD requires passing over all the training examples to compute the gradient, this requires $O(n^2)$ gradient computations to achieve the rate of order $\widetilde{O}(1/(n\gamma^2))$. As a comparison, Theorem 11 shows that SGD with $T \asymp n$ can achieve the same rate under the same NTK separability condition assumption, which requires $O(n)$ gradient computations since we only compute the gradient once per iteration. Therefore, SGD significantly improves the gradient complexity in [21, 22, 41] to achieve almost optimal risk bounds. We summarize the comparison in Table 1.

The work [16] also considered SGD in an online/one-pass setting, where each training example arrives sequentially from ρ and is used only once. In their setting, there is no need to consider generalization issues since $\mathbf{z}_t$ has no dependency with $\mathbf{w}_t$. As a comparison, we consider a fixed training set and allow each training example to be used several times, which is often the choice in practice to fully utilize the training datasets. Indeed, multi-pass SGD often empirically lead to better generalization than the one-pass counterpart in practice [34]. The multi-pass setting introduces an additional generalization issue between training and testing, which is controlled by our Rademacher complexity analysis. Furthermore, the analysis in [34] only gives a high-probability bound $\|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty} = \widetilde{O}(1/(\sqrt{m}\lambda^2))$. As a comparison, we also give a high-probability bound on $\|\mathbf{W}_t - \mathbf{W}_1\|_2$, which is crucial to build a localized parameter space $\{\mathbf{W} : \|\mathbf{W} - \mathbf{W}^*\|_2^2 \le C(\mathbf{W}^*)\}$ in Eq. (4.14) and to conduct our Rademacher complexity analysis. Indeed, the constraint on $\|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty}$ ignores the correlation among different components of $\mathbf{W}_t$, and Rademacher complexity analysis with this constraint incurs a factor of $\sqrt{m}$, as shown in Lemma B.3 of [34].

Work	Method	Bound	Activation	Width	Type	GC
[32]	GD	$\widetilde{O}\big(1/(n^{\frac{1}{4}}\gamma^{\frac{1}{2}})\big)$	Smooth	polynomial	high-probability	$n^{\frac{3}{2}}/\gamma$
[16]	GD	$\widetilde{O}\big(1/(\sqrt{n}\gamma^2)\big)$	ReLU	logarithmic	high-probability	n^2
[16]	One-pass SGD	$\widetilde{O}(1/(n\gamma^2))$	ReLU	logarithmic	high-probability	n
[41]	GD	$\widetilde{O}(1/(n\gamma^2))$	Smooth	logarithmic	expectation	n^2
[21]	GD	$\widetilde{O}(1/(n\gamma^2))$	ReLU	logarithmic	high-probability	n^2
Thm. 11	Multi-pass SGD	$\widetilde{O}(1/(n\gamma^2))$	ReLU	logarithmic	high-probability	n

Table 1: Comparison of the risk bounds under NTK separability condition with margin γ. "GC" means the gradient complexity. The work [41] used stability analysis to derive risk bounds in expectation, while other works used Rademacher complexity analysis to derive high-probability risk bounds. The one-pass SGD analysis in [16] assumes that each training example arrives sequentially from ρ and is used only once. In their setting, there is no need to consider generalization issues since $\mathbf{z}_t$ has no dependency with $\mathbf{w}_t$.

5 Conclusion

We developed a high-probability framework for analyzing SGD in shallow ReLU networks through optimization-dependent localization. Under realizability or NTK separability, the resulting population-risk bounds are nearly optimal while requiring only $O(n)$ stochastic-gradient computations when $T \asymp n$. The scope of the framework, together with several directions for extending it to other architectures and optimization methods, are discussed in Appendix B.

Acknowledgments

We are grateful to the anonymous reviewers and area chairs for their insightful and constructive comments. The work of Yunwen Lei was supported by the Hong Kong Research Grants Council under the ECS project 22303723 and the GRF project 17304526.

References

[1] Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[2] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332, 2019.

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

[4] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

[5] Y. Cai, J. Wu, S. Mei, M. Lindsey, and P. Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. In *Advances in Neural Information Processing Systems*, volume 37, pages 71306–71351, 2024.

[6] Z. Chen, Y. Cao, D. Zou, and Q. Gu. How much over-parameterization is sufficient to learn deep relu networks? In *International Conference on Learning Representation*, 2021.

[7] K. Cheng, J. Fan, L. Song, and D.-X. Zhou. Learning frechet differentiable operators via prespecified neural operators. *Applied and Computational Harmonic Analysis*, page 101878, 2026.

[8] P. Deora, R. Ghaderi, H. Taheri, and C. Thrampoulidis. On the optimization and generalization of multi-head attention. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[9] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

[10] W. E, C. Ma, L. Wu, and S. Wojtowytsch. Towards a mathematical understanding of neural network-based machine learning: What we know and what we don't. *CSIAM Transactions on Applied Mathematics*, 1(4):561–615, 2020.

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

[12] N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

[13] N. M. Hieu and A. Ledent. Generalization analysis for supervised contrastive representation learning under non-iid settings. In *International Conference on Machine Learning*, pages 23179–23218, 2025.

[14] R. Huang, Y. Liang, and J. Yang. Non-asymptotic convergence of training transformers for next-token prediction. In *Advances in Neural Information Processing Systems*, volume 37, pages 80634–80673, 2024.

[15] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[16] Z. Ji and M. Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2019.

[17] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-Łojasiewicz condition. *Machine Learning and Knowledge Discovery in Databases*, page 795–811, 2016.

[18] A. Ledent and R. Alves. Generalization analysis of deep non-linear matrix completion. In *International Conference on Machine Learning*, pages 26290–26360, 2024.

[19] A. Ledent, R. Alves, and Y. Lei. Generalization bounds for rank-sparse neural networks. In *Advances in Neural Information Processing Systems*, pages 147927–147996, 2025.

[20] Y. Lei and Y. Xie. Towards initialization-dependent and non-vacuous generalization bounds for overparameterized shallow neural networks. *arXiv preprint arXiv:2604.00505*, 2026.

[21] Y. Lei, P. Wang, Y. Ying, and D.-X. Zhou. Optimization and generalization of gradient descent for shallow relu networks with minimal width. *Journal of Machine Learning Research*, 27(34): 1–35, 2026.

[22] Y. Li, Y. Lei, Z.-C. Guo, and Y. Ying. Optimal rates for generalization of gradient descent for deep relu classification. In *Advances in Neural Information Processing Systems*, page 28623–28663, 2025.

[23] C. Liu, L. Zhu, and M. Belkin. Loss landscapes and optimization in over-parameterized nonlinear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

[24] F. Liu, L. Dadi, and V. Cevher. Learning with norm constrained, over-parameterized, two-layer neural networks. *Journal of Machine Learning Research*, 25(138):1–42, 2024.

[25] Y. Liu, T. Mao, and D.-X. Zhou. Approximation of functions from korobov spaces by shallow neural networks. *Information Sciences*, 670:120573, 2024.

[26] W. Ma, Y. Sui, J. Teng, B. Wang, J. Xu, and J. Yang. Generalization bounds of stochastic gradient descent in homogeneous neural networks. *arXiv preprint arXiv:2602.22936*, 2026.

[27] R. Magen and O. Shamir. Initialization-dependent sample complexity of linear predictors and neural networks. *Advances in Neural Information Processing Systems*, 36:7632–7658, 2023.

[28] M. S. Nacson, J. Lee, S. Gunasekar, P. H. P. Savarese, N. Srebro, and D. Soudry. Convergence of gradient descent on separable data. In *International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

[29] G. Neu, G. K. Dziugaite, M. Haghifam, and D. M. Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pages 3526–3545. PMLR, 2021.

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

[31] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019.

[32] A. Nitanda, G. Chinot, and T. Suzuki. Gradient descent can learn less over-parameterized two-layer neural networks on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

[33] S. Oymak and M. Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

[34] L. Pillaud-Vivien, A. Rudi, and F. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.

[35] D. Richards and I. Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[36] M. Schliserman and T. Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In *Conference on Learning Theory*, pages 3380–3394, 2022.

[37] M. Schliserman and T. Koren. Tight risk bounds for gradient descent on separable data. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

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

[39] M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

[40] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010.

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

[42] P. Wang, Y. Lei, D. Wang, Y. Ying, and D.-X. Zhou. Generalization guarantees of gradient descent for shallow neural networks. *Neural Computation*, 37(2):344–402, 2025.

[43] P. Wang, J. Zhou, P. Liznerski, and M. Kloft. Optimization, generalization and differential privacy bounds for gradient descent on kolmogorov–arnold networks. In *International Conference on Machine Learning*, pages 132982–133019, 2026.

[44] J. Wu, P. L. Bartlett, M. Telgarsky, and B. 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. PMLR, 2024.

[45] Y. Yang and D.-X. Zhou. Nonparametric regression using over-parameterized shallow relu neural networks. *Journal of Machine Learning Research*, 25(165):1–35, 2024.

[46] K. Zhang, Y. Wang, and R. Arora. Stability and generalization of adversarial training for shallow neural networks with smooth activation. In *Advances in Neural Information Processing Systems*, volume 37, pages 16160–16193, 2024.

[47] T. Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *Conference on Learning Theory*, pages 173–187, 2005.

[48] T.-Y. Zhou and X. Huo. Learning ability of interpolating deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 68:101582, 2024.

[49] D. Zou, Y. Cao, D. Zhou, and Q. Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109:467–492, 2020.

A Proofs

We give the proof details for Section 4. The analysis proceeds in three steps.

First, we prove a pathwise optimization result: on the initialization event E_1, the SGD iterates remain localized around a reference model $\mathbf{W}^*$ in Euclidean norm and around the initialization $\mathbf{W}_1$ in $(2,\infty)$-norm. The latter control is essential for the nonlinear activation function ReLU to handle activation changes of ReLU units.

Second, we use this localization to define the hypothesis class $\mathcal{F}$. Its Rademacher complexity is bounded by embedding the localized class into a Euclidean ball centered at initialization and applying the complexity estimate of [20]. The event E_2 turns this complexity bound into a uniform population-risk estimate. Bernstein's inequality is used to replace the sample-dependent radius $\mathfrak{C}_S(\mathbf{W}^*)$ by the data-independent radius $\mathfrak{C}(\mathbf{W}^*)$.

Finally, under NTK separability, we construct the reference model $\mathbf{W}^{1/T} = \mathbf{W}_1 + \alpha \mathbf{v}_*$ with $\alpha = 2\log T/\gamma$. The margin γ controls the loss of this reference model, while β controls its neuron-wise displacement. This verifies E_1 and yields the NTK-separable risk bound.

Table 2 collects the main quantities appearing in the local optimization control and the subsequent uniform generalization analysis.

Table 2: Main quantities used in the proofs.

Symbol	Definition
$\mathfrak{C}_S(\mathbf{W}^*)$	$2\|\mathbf{W}_1 - \mathbf{W}^*\|_2^2 + 4\eta T F_S(\mathbf{W}^*) + 36\eta^2 G^2 \log^2 \frac{2T}{\delta}$
$\widehat{R}_1(C)$	$\Big(1 + \frac{2L}{G}\Big)\frac{C}{\sqrt{m}} + \frac{G\eta \log(mT/\delta)}{\sqrt{m}}$
R_1	Initialization radius in E_1; in the NTK section, $R_1 = \max\{\widehat{R}_1(\bar{C}_T), \alpha\beta\}$.
C_{R_1}	$\frac{2R_1 m}{\sqrt{\pi}} + \Big(\frac{m\log(n/\delta)}{2}\Big)^{1/2}$; upper bound on the number of near-flipping neurons.
$\mathfrak{C}(\mathbf{W}^*)$	$2\|\mathbf{W}_1 - \mathbf{W}^*\|_2^2 + 4\eta T\Big(2F(\mathbf{W}^*) + \frac{7b_* \log(2/\delta)}{6n}\Big) + 36\eta^2 G^2 \log^2 \frac{2T}{\delta}$
$\mathcal{W}$	$\Big\{\mathbf{W} :\ \|\mathbf{W} - \mathbf{W}^*\|_2^2 \le \mathfrak{C}(\mathbf{W}^*)\Big\}$
$\mathcal{F}$	$\{x \mapsto \Phi(\mathbf{W}; x) :\ \mathbf{W} \in \mathcal{W}\}$
E_1	Initialization and localization event in Definition 3
E_2	Generalization event in Definition 5
b_*	$b_* \ge \sup_z f(\mathbf{W}^*; z)$
b	$b = 2b_* + L\mathfrak{C}(\mathbf{W}^*)$
$\bar{C}_T$	$2\alpha^2 + 4\eta + 36\eta^2 G^2 \log^2 \frac{2T}{\delta}$
ξ_k, ξ_k'	Martingale terms in the SGD recursion

A.1 Proofs on Optimization

In this subsection, we present the proofs for the optimization analysis. Lemma 14 controls the distance between the SGD iterates and the initialization, as measured by the $(2,\infty)$-norm. The proof requires concentration inequalities on martingales.

Lemma 12 ([47]). *Let $z_1, \dots, z_n$ be a sequence of independent random variables. Consider a sequence of functionals $\xi_k(z_1, \dots, z_k), k = 1, \dots, n$.*

(a) Assume that $|\xi_k - \mathbb{E}_{z_k}[\xi_k]| \le b_k$ for each k. Let $\delta \in (0,1)$. With probability at least $1-\delta$, we have

$$\sum_{k=1}^n \xi_k - \sum_{k=1}^n \mathbb{E}_{z_k}[\xi_k] \le \Big(2\sum_{k=1}^n b_k^2 \log\frac{1}{\delta}\Big)^{\frac{1}{2}}. \tag{A.1}$$

(b) Let $\sigma_n^2 = \sum_{k=1}^n \mathbb{E}_{z_k}\big[\big(\xi_k - \mathbb{E}_{z_k}[\xi_k]\big)^2\big]$ be the conditional variance. Assume that $\xi_k - \mathbb{E}_{z_k}[\xi_k] \le b$ for each k. Let $\rho \in (0,1]$ and $\delta \in (0,1)$. With probability at least $1-\delta$, we have

$$\sum_{k=1}^n \xi_k - \sum_{k=1}^n \mathbb{E}_{z_k}[\xi_k] \le \frac{\rho\sigma_n^2}{b} + \frac{b\log\frac{1}{\delta}}{\rho}. \tag{A.2}$$

Lemma 13 (*Self-bounding* property[40]). *If ℓ is nonnegative and L-smooth, then $|\ell'(a)|^2 \le 2L\ell(a)$.*

The following lemma provides a pathwise, neuron-wise bound of the displacement, ensuring that the SGD iterates stay within the localized ReLU regime. This constitutes the central stochastic component in the optimization analysis.

Lemma 14. *If Assumption 1 and Assumption 2 hold, then with probability at least $1-\delta$, we have the following hold for all $j \in [m]$ and $t \in [T]$*

$$\|\mathbf{w}_{t+1,j} - \mathbf{w}_{1,j}\|_2 \leq \frac{2L\eta}{G\sqrt{m}} \sum_{k=1}^{t} F_S(\mathbf{W}_k) + \frac{G\eta \log(mT/\delta)}{\sqrt{m}} + \sum_{k=1}^{t} \frac{\eta}{\sqrt{m}} F_S(\mathbf{W}_k)$$

Proof. According to Eq. (3.4) and $|\ell'(a)| \leq \ell(a)$, we know

$$\begin{aligned}
&\mathbb{E}_{i_t}[\|\mathbf{w}_{t+1,j} - \mathbf{w}_{t,j}\|_2] \leq \frac{\eta}{\sqrt{m}} \mathbb{E}_{i_t}\Big[\Big\| y_{i_t} \ell'(y_{i_t} \Phi(\mathbf{W}_t; \mathbf{x}_{i_t})) \mathbb{I}_{[\mathbf{w}_{t,j}^\top \mathbf{x}_{i_t} \geq 0]} \mathbf{x}_{i_t} \Big\|_2\Big] \\
&= \frac{\eta}{n\sqrt{m}} \sum_{i=1}^{n} \Big\| y_i \ell'(y_i \Phi(\mathbf{W}_t; \mathbf{x}_i)) \mathbb{I}_{[\mathbf{w}_{t,j}^\top \mathbf{x}_i \geq 0]} \mathbf{x}_i \Big\|_2 \leq \frac{\eta}{n\sqrt{m}} \sum_{i=1}^{n} \big| \ell'(y_i \Phi(\mathbf{W}_t; \mathbf{x}_i)) \big| \\
&\leq \frac{\eta}{n\sqrt{m}} \sum_{i=1}^{n} \ell(y_i \Phi(\mathbf{W}_t; \mathbf{x}_i)) = \frac{\eta}{\sqrt{m}} F_S(\mathbf{W}_t).
\end{aligned} \tag{A.3}$$

Define

$$\xi_t = \|\mathbf{w}_{t+1,j} - \mathbf{w}_{t,j}\|_2 - \mathbb{E}_{i_t}[\|\mathbf{w}_{t+1,j} - \mathbf{w}_{t,j}\|_2].$$

Then, we have

$$\xi_t \leq \frac{\eta}{\sqrt{m}} \Big\| y_{i_t} \ell'(y_{i_t} \Phi(\mathbf{W}_t; \mathbf{x}_{i_t})) \mathbb{I}_{[\mathbf{w}_{t,j}^\top \mathbf{x}_{i_t} \geq 0]} \mathbf{x}_{i_t} \Big\|_2 \leq \frac{G\eta}{\sqrt{m}}.$$

Moreover, by the self-bounding property $|\ell'(a)|^2 \leq 2L\ell(a)$, we know

$$\begin{aligned}
\mathrm{Var}_{i_t}[\xi_t] &\leq \mathbb{E}_{i_t}[\|\mathbf{w}_{t+1,j} - \mathbf{w}_{t,j}\|_2^2] \leq \frac{\eta^2}{m} \mathbb{E}_{i_t}\Big[\Big\| y_{i_t} \ell'(y_{i_t} \Phi(\mathbf{W}_t; \mathbf{x}_{i_t})) \mathbb{I}_{[\mathbf{w}_{t,j}^\top \mathbf{x}_{i_t} \geq 0]} \mathbf{x}_{i_t} \Big\|_2^2\Big] \\
&\leq \frac{\eta^2}{m} \mathbb{E}_{i_t}\Big[|\ell'(y_{i_t} \Phi(\mathbf{W}_t; \mathbf{x}_{i_t}))|^2\Big] \leq \frac{2L\eta^2}{m} \mathbb{E}_{i_t}\big[\ell(y_{i_t} \Phi(\mathbf{W}_t; \mathbf{x}_{i_t}))\big] = \frac{2L\eta^2 F_S(\mathbf{W}_t)}{m}.
\end{aligned}$$

According to Part (b) of Lemma 12, with $\rho = 1$ and confidence $\delta/(mT)$, we have that with probability at least $1 - \delta/(mT)$

$$\begin{aligned}
\sum_{k=1}^{t} \xi_k &\leq \frac{\sqrt{m}}{G\eta} \sum_{k=1}^{t} \frac{2L\eta^2 F_S(\mathbf{W}_k)}{m} + \frac{G\eta \log(mT/\delta)}{\sqrt{m}} \\
&= \frac{2L\eta}{G\sqrt{m}} \sum_{k=1}^{t} F_S(\mathbf{W}_k) + \frac{G\eta \log(mT/\delta)}{\sqrt{m}}.
\end{aligned}$$

Combining it with (A.3), we have

$$\begin{aligned}
\sum_{k=1}^{t} \|\mathbf{w}_{k+1,j} - \mathbf{w}_{k,j}\|_2 &= \sum_{k=1}^{t} \big(\xi_k + \mathbb{E}_{i_k}[\|\mathbf{w}_{k+1,j} - \mathbf{w}_{k,j}\|_2]\big) \\
&\leq \frac{2L\eta}{G\sqrt{m}} \sum_{k=1}^{t} F_S(\mathbf{W}_k) + \frac{G\eta \log(mT/\delta)}{\sqrt{m}} + \sum_{k=1}^{t} \frac{\eta}{\sqrt{m}} F_S(\mathbf{W}_k).
\end{aligned}$$

Finally, by the triangle inequality, we know

$$\|\mathbf{w}_{t+1,j} - \mathbf{w}_{1,j}\|_2 \leq \sum_{k=1}^{t} \|\mathbf{w}_{k+1,j} - \mathbf{w}_{k,j}\|_2.$$

Taking a union bound over all (t, j) yields the stated claim with probability at least $1 - \delta$. □

The following lemma shows a weak-convexity property of F_S on a ball around the initialization point. It shows that as the network width increases, the objective function becomes progressively closer to convex. Define the ball

$$B_R = \big\{\mathbf{W} \in \mathbb{R}^{md} : \|\mathbf{W} - \mathbf{W}_1\|_{2,\infty} \le R\big\}. \tag{A.4}$$

Lemma 15 ([21]). *Let $\mathbf{W}, \mathbf{W}' \in B_{R_1}$. Under the event E_1, we have*

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

The following lemma follows directly from the self-bounding property of nonnegative and smooth loss functions. That is, if g is nonnegative and L-smooth, we have $\|\nabla g(\mathbf{w})\|^2 \le 2Lg(\mathbf{w})$.

Lemma 16. *Let $g_k = \ell'(y_{i_k}\Phi(\mathbf{W}_k; x_{i_k}))$. Under Assumptions 1, 2, we know*

$$\big|g_k - \mathbb{E}_k[g_k]\big| \le 2G \quad \textit{and} \quad \mathbb{E}_k\big[g_k^2\big] \le 2LF_S(\mathbf{W}_k). \tag{A.5}$$

Proof of Theorem 2. We divide the proof into three steps: a one-step progress inequality, martingale decompositions, and an induction argument on a high-probability event.

Step 1: One-step progress inequality.

By Eq. (3.3) and $\|\nabla\Phi(\mathbf{W};\mathbf{x})\|_2 \le 1$ [21], we know

$$\begin{aligned}&\big\|\mathbf{W}_{k+1} - \mathbf{W}^*\big\|_2^2 = \big\|\mathbf{W}_k - \eta\nabla f(\mathbf{W}_k;\mathbf{z}_{i_k}) - \mathbf{W}^*\big\|_2^2 \\ &= \big\|\mathbf{W}_k - \mathbf{W}^*\big\|_2^2 + \eta^2\|\nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\|_2^2 + 2\eta\big\langle \mathbf{W}^* - \mathbf{W}_k, \nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\big\rangle \\ &= \big\|\mathbf{W}_k - \mathbf{W}^*\big\|_2^2 + \eta^2\|y_{i_k}\ell'(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))\nabla\Phi(\mathbf{W}_k;\mathbf{x}_{i_k})\|_2^2 + 2\eta\big\langle \mathbf{W}^* - \mathbf{W}_k, \nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\big\rangle \\ &\le \big\|\mathbf{W}_k - \mathbf{W}^*\big\|_2^2 + \eta^2|\ell'(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))|^2 + 2\eta\big\langle \mathbf{W}^* - \mathbf{W}_k, \nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\big\rangle.\end{aligned}$$

Summing the above inequality, we obtain

$$\big\|\mathbf{W}_{t+1} - \mathbf{W}^*\big\|_2^2 \le \|\mathbf{W}_1 - \mathbf{W}^*\|_2^2 + \eta^2\sum_{k=1}^t |\ell'(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))|^2 + 2\eta\sum_{k=1}^t \big\langle \mathbf{W}^* - \mathbf{W}_k, \nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\big\rangle. \tag{A.6}$$

Step 2: Martingale decompositions.

For $k = 1, \ldots, T$, define ξ_k and ξ_k' as

$$\begin{aligned}\xi_k &= \Big(\big\langle \mathbf{W}^* - \mathbf{W}_k, \nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\big\rangle - \mathbb{E}_{i_k}\big[\big\langle \mathbf{W}^* - \mathbf{W}_k, \nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\big\rangle\big]\Big)\mathbb{I}_{[\|\mathbf{W}^* - \mathbf{W}_k\|_2 \le \sqrt{\mathfrak{C}_S(\mathbf{W}^*)}]} \\ \xi_k' &= |\ell'(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))|^2 - \mathbb{E}_{i_k}\big[|\ell'(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))|^2\big].\end{aligned}$$

Then, since $\|\nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\|_2 \le G$ we know

$$\begin{aligned}|\xi_k| &\le 2\|\mathbf{W}^* - \mathbf{W}_k\|_2\|\nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\|_2\mathbb{I}_{[\|\mathbf{W}^* - \mathbf{W}_k\|_2 \le \sqrt{\mathfrak{C}_S(\mathbf{W}^*)}]} \\ &\le 2G\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}\mathbb{I}_{[\|\mathbf{W}^* - \mathbf{W}_k\|_2 \le \sqrt{\mathfrak{C}_S(\mathbf{W}^*)}]}.\end{aligned}$$

Moreover, by the self-bounding property $|\ell'(a)|^2 \le 2L\ell(a)$ and the fact $\|\nabla\Phi(\mathbf{W};\mathbf{x}_{i_k})\|_2 \le 1$, we know

$$\begin{aligned}\mathrm{Var}_{i_k}[\xi_k] &\le \mathbb{E}_{i_k}\big[\xi_k^2\big] \\ &\le \mathbb{E}_{i_k}\Big[\big\langle \mathbf{W}^* - \mathbf{W}_k, \nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\big\rangle^2\mathbb{I}_{[\|\mathbf{W}^* - \mathbf{W}_k\|_2 \le \sqrt{\mathfrak{C}_S(\mathbf{W}^*)}]}\Big] \\ &\le \|\mathbf{W}^* - \mathbf{W}_k\|_2^2\,\mathbb{E}_{i_k}\big[\|\nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\|_2^2\big]\mathbb{I}_{[\|\mathbf{W}^* - \mathbf{W}_k\|_2 \le \sqrt{\mathfrak{C}_S(\mathbf{W}^*)}]} \\ &\le \|\mathbf{W}^* - \mathbf{W}_k\|_2^2\mathbb{E}_{i_k}\big[\|\nabla\Phi(\mathbf{W}_k;\mathbf{x}_{i_k})\|_2^2|\ell'(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))|^2\big]\mathbb{I}_{[\|\mathbf{W}^* - \mathbf{W}_k\|_2 \le \sqrt{\mathfrak{C}_S(\mathbf{W}^*)}]} \\ &\le 2L\mathfrak{C}_S(\mathbf{W}^*)\mathbb{E}_{i_k}[\ell(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))]\mathbb{I}_{[\|\mathbf{W}^* - \mathbf{W}_k\|_2 \le \sqrt{\mathfrak{C}_S(\mathbf{W}^*)}]} \\ &\le 2L\mathfrak{C}_S(\mathbf{W}^*)F_S(\mathbf{W}_k)\mathbb{I}_{[\|\mathbf{W}^* - \mathbf{W}_k\|_2 \le \sqrt{\mathfrak{C}_S(\mathbf{W}^*)}]}.\end{aligned}$$

Applying Lemma 12(b) with $\rho_1 \in (0,1]$,

$$b = 2G\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}, \quad \sigma^2 = 2L\mathfrak{C}_S(\mathbf{W}^*)\sum_{k=1}^{t} F_S(\mathbf{W}_k)\mathbb{I}_{[\|\mathbf{W}^*-\mathbf{W}_k\|_2 \le \sqrt{\mathfrak{C}_S(\mathbf{W}^*)}]},$$

with confidence level $\delta/(2T)$ and taking a union bound over $t = 1, \dots, T$, we obtain that with probability at least $1 - \frac{\delta}{2}$,

$$\sum_{k=1}^{t} \xi_k \le \frac{\rho_1 L}{G}\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}\sum_{k=1}^{t} F_S(\mathbf{W}_k)\mathbb{I}_{[\|\mathbf{W}^*-\mathbf{W}_k\|_2 \le \sqrt{\mathfrak{C}_S(\mathbf{W}^*)}]} + \frac{2G}{\rho_1}\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}\log\frac{2T}{\delta}. \quad \text{(A.7)}$$

By Lemma 16, $|\xi_k'| \le 2G^2$, and the self-bounding property $\mathbb{E}_{i_k}[|\ell'(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))|^2] \le 2L\mathbb{E}_{i_k}[\ell(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))]$, we know

$$\begin{aligned}\mathrm{Var}_{i_k}[\xi_k'] &\le \mathbb{E}_{i_k}\left[(\xi_k')^2\right] \le \mathbb{E}_{i_k}\left[|\ell'(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))|^4\right] \\ &\le G^2\mathbb{E}_{i_k}\left[|\ell'(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))|^2\right] \le 2G^2LF_S(\mathbf{W}_k).\end{aligned}$$

Applying Lemma 12 (b) again with $b = 2G^2$, $\sigma^2 = 2LG^2\sum_{k=1}^{t} F_S(\mathbf{W}_k)$, $\rho_2 \in (0,1]$, with confidence level $\delta/(2T)$ and taking a union bound over $t = 1, \dots, T$, we have with probability at least $1 - \frac{\delta}{2}$,

$$\sum_{k=1}^{t} \xi_k' \le \rho_2 L\sum_{k=1}^{t} F_S(\mathbf{W}_k) + \frac{2G^2}{\rho_2}\log\frac{2T}{\delta}. \quad \text{(A.8)}$$

Step 3: Induction on a high-probability event.

Let E_ξ denote the event that Eq. (A.7) holds for all $t \le T$, and let $E_{\xi'}$ denote the event that Eq. (A.8) holds for all $t \le T$. Let E_w denote the event that Lemma 14 holds for all $t \le T$ and all $j \in [m]$. Define $\mathcal{G} := E_1 \cap E_\xi \cap E_{\xi'} \cap E_w$. We prove Eq. (4.6) on $\mathcal{G}$ by induction on t.

The base case holds since $\|\mathbf{W}_1 - \mathbf{W}^*\|_2^2 \le \mathfrak{C}_S(\mathbf{W}^*)$ by Eq. (4.2) and $\sum_{k=1}^{0} F_S(\mathbf{W}_k) = 0$. Assume for all $k \le t$,

$$\|\mathbf{W}_k - \mathbf{W}^*\|_2 \le \sqrt{\mathfrak{C}_S(\mathbf{W}^*)} \quad \text{and} \quad \sum_{k=1}^{t-1} \eta F_S(\mathbf{W}_k) \le \mathfrak{C}_S(\mathbf{W}^*). \quad \text{(A.9)}$$

Then $\mathbb{I}_{[\|\mathbf{W}^*-\mathbf{W}_k\|_2 \le \sqrt{\mathfrak{C}_S(\mathbf{W}^*)}]} = 1$ for all $k \le t$, and

$$\begin{aligned}|\ell'(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))|^2 &= \mathbb{E}_{i_k}[|\ell'(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))|^2] + \xi_k' \\ &\le 2LF_S(\mathbf{W}_k) + \xi_k', \end{aligned} \quad \text{(A.10)}$$

together with Eq. (A.8), we have

$$\eta^2\sum_{k=1}^{t}|\ell'(y_{i_k}\Phi(\mathbf{W}_k;\mathbf{x}_{i_k}))|^2 \le 2\eta^2L\sum_{k=1}^{t}F_S(\mathbf{W}_k) + L\eta^2\rho_2\sum_{k=1}^{t}F_S(\mathbf{W}_k) + \frac{2\eta^2G^2}{\rho_2}\log\frac{2T}{\delta}. \quad \text{(A.11)}$$

Meanwhile, since $\mathbb{E}_{i_k}[\nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})] = \nabla F_S(\mathbf{W}_k)$, we have

$$\mathbb{E}_{i_k}\left[\left\langle \mathbf{W}^* - \mathbf{W}_k, \nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\right\rangle\right] = \left\langle \mathbf{W}^* - \mathbf{W}_k, \nabla F_S(\mathbf{W}_k)\right\rangle.$$

On the event $\mathcal{G}$, for any $k \le t$, applying Lemma 14 with t replaced by $k-1$ and using $\eta\sum_{r=1}^{k-1}F_S(\mathbf{W}_r) \le \mathfrak{C}_S(\mathbf{W}^*)$ from the inductive hypothesis, we obtain

$$\|\mathbf{W}_k - \mathbf{W}_1\|_{2,\infty} \le \left(1 + \frac{2L}{G}\right)\frac{\mathfrak{C}_S(\mathbf{W}^*)}{\sqrt{m}} + \frac{G\eta\log(mT/\delta)}{\sqrt{m}} =: \widehat{R}_1(\mathfrak{C}_S(\mathbf{W}^*)). \quad \text{(A.12)}$$

On the event E_1, we have $\widehat{R}_1(\mathfrak{C}_S(\mathbf{W}^*)) \le R_1$, which implies $\|\mathbf{W}_k - \mathbf{W}_1\|_{2,\infty} \le R_1$. Thus, we have $\mathbf{W}_k \in B_{R_1}$ for all $k \le t$. Also, on E_1 we have $\|\mathbf{W}^* - \mathbf{W}_1\|_{2,\infty} \le R_1$, hence $\mathbf{W}^* \in B_{R_1}$. We can now apply Lemma 15 with $\mathbf{W} = \mathbf{W}^*$ and $\mathbf{W}' = \mathbf{W}_k$ to obtain

$$\langle \mathbf{W}^* - \mathbf{W}_k, \nabla F_S(\mathbf{W}_k)\rangle \le F_S(\mathbf{W}^*) - F_S(\mathbf{W}_k) + \frac{C_{R_1}^{1/2}\|\mathbf{W}_k - \mathbf{W}^*\|_2}{\sqrt{m}}F_S(\mathbf{W}_k). \quad \text{(A.13)}$$

Summing Eq. (A.13) and combining Eq. (A.7), noting $\mathbb{I}_{[\|\mathbf{W}^*-\mathbf{W}_k\|_2\le\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}]}=1$, we have

$$
\begin{aligned}
&2\eta\sum_{k=1}^{t}\big\langle\mathbf{W}^*-\mathbf{W}_k,\nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\big\rangle\le 2\eta\sum_{k=1}^{t}\mathbb{E}_{i_k}\big[\big\langle\mathbf{W}^*-\mathbf{W}_k,\nabla f(\mathbf{W}_k;\mathbf{z}_{i_k})\big\rangle\big]+2\eta\sum_{k=1}^{t}\xi_k\\
&\le\frac{2\eta\rho_1 L}{G}\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}\sum_{k=1}^{t}F_S(\mathbf{W}_k)+\frac{4\eta G}{\rho_1}\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}\log\frac{2T}{\delta}+2\eta tF_S(\mathbf{W}^*)\\
&-2\eta\sum_{k=1}^{t}F_S(\mathbf{W}_k)\Big(1-\frac{C_{R_1}^{1/2}\|\mathbf{W}_k-\mathbf{W}^*\|_2}{\sqrt{m}}\Big).
\end{aligned}
\tag{A.14}
$$

Combining Eqs. (A.14), (A.11) and (A.6), we obtain

$$
\begin{aligned}
&\big\|\mathbf{W}_{t+1}-\mathbf{W}^*\big\|_2^2+2\eta\sum_{k=1}^{t}F_S(\mathbf{W}_k)\Big(1-\frac{C_{R_1}^{1/2}\|\mathbf{W}_k-\mathbf{W}^*\|_2}{\sqrt{m}}\Big)\\
&\le\|\mathbf{W}_1-\mathbf{W}^*\|_2^2+(2+\rho_2)\eta^2L\sum_{k=1}^{t}F_S(\mathbf{W}_k)+\frac{2\eta^2G^2}{\rho_2}\log\frac{2T}{\delta}+\frac{2\eta\rho_1L}{G}\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}\sum_{k=1}^{t}F_S(\mathbf{W}_k)\\
&+\frac{4\eta G\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}}{\rho_1}\log\frac{2T}{\delta}+2\eta tF_S(\mathbf{W}^*).
\end{aligned}
\tag{A.15}
$$

On the event E_1, we have the upper bound in Definition 3(ii) $R_1\le\frac{\sqrt{\pi}}{16\mathfrak{C}_S(\mathbf{W}^*)}$. Also, we have

$$
C_{R_1}=\frac{2R_1m}{\sqrt{\pi}}+\Big(\frac{m\log(n/\delta)}{2}\Big)^{1/2}.
$$

Since $m\ge 32\mathfrak{C}_S^2(\mathbf{W}^*)\log\frac{n}{\delta}$, the second term is at most $m/(8\mathfrak{C}_S(\mathbf{W}^*))$. Using the upper bound above on R_1, the first term is also at most $m/(8\mathfrak{C}_S(\mathbf{W}^*))$. Hence $C_{R_1}\le m/(4\mathfrak{C}_S(\mathbf{W}^*))$, which implies

$$
\frac{\sqrt{C_{R_1}}\,\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}}{\sqrt{m}}\le\frac{1}{2}.
\tag{A.16}
$$

By Eqs. (A.9) and (A.16), we have

$$
1-\frac{\max_{k\in[t]}C_{R_1}^{1/2}\|\mathbf{W}_k-\mathbf{W}^*\|_2}{\sqrt{m}}\ge\frac{1}{2}.
$$

Now, take $\rho_1=\min\Big\{1,\frac{G}{16L\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}}\Big\}$ and $\rho_2=1$. Together with the assumption $\eta\le\frac{1}{256L\log(2T/\delta)}$, this yields

$$
(2+\rho_2)\eta^2L\le 3\eta^2L\le\frac{\eta}{8}\quad\text{and}\quad\frac{2\eta\rho_1L}{G}\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}\le\frac{\eta}{8}.
$$

Hence, Eq. (A.15) implies

$$
\begin{aligned}
\big\|\mathbf{W}_{t+1}-\mathbf{W}^*\big\|_2^2+\frac{3}{4}\eta\sum_{k=1}^{t}F_S(\mathbf{W}_k)&\le\|\mathbf{W}_1-\mathbf{W}^*\|_2^2+2\eta tF_S(\mathbf{W}^*)\\
&+2\eta^2G^2\log\frac{2T}{\delta}+\frac{4\eta G\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}}{\rho_1}\log\frac{2T}{\delta}
\end{aligned}
\tag{A.17}
$$

For the term $\frac{4\eta G\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}}{\rho_1}\log\frac{2T}{\delta}$, if $\rho_1=1$, by Young's inequality we have

$$
4\eta G\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}\log\frac{2T}{\delta}\le\frac{1}{4}\mathfrak{C}_S(\mathbf{W}^*)+16\eta^2G^2\log^2\frac{2T}{\delta}.
$$

If $\rho_1=\frac{G}{16L\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}}$, by $\eta\le\frac{1}{256L\log(2T/\delta)}$, we know

$$\frac{4\eta G\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}}{\rho_1}\log\frac{2T}{\delta} \le \frac{64\mathfrak{C}_S(\mathbf{W}^*)L\log(2T/\delta)}{256L\log(2T/\delta)} \le \frac{1}{4}\mathfrak{C}_S(\mathbf{W}^*).$$

Hence, in both cases, we have $\frac{4\eta G\sqrt{\mathfrak{C}_S(\mathbf{W}^*)}}{\rho_1}\log\frac{2T}{\delta} \le \frac{1}{4}\mathfrak{C}_S(\mathbf{W}^*) + 16\eta^2G^2\log^2\frac{2T}{\delta}$. Combining the above bounds with Eq. (A.17), we have

$$\begin{aligned}\left\|\mathbf{W}_{t+1}-\mathbf{W}^*\right\|_2^2 + \frac{3}{4}\eta\sum_{k=1}^{t}F_S(\mathbf{W}_k) \le \\ \|\mathbf{W}_1-\mathbf{W}^*\|_2^2 + 2\eta t F_S(\mathbf{W}^*) + \frac{1}{4}\mathfrak{C}_S(\mathbf{W}^*) + 18\eta^2G^2\log^2\frac{2T}{\delta}. \quad \text{(A.18)}\end{aligned}$$

By definition of $\mathfrak{C}_S(\mathbf{W}^*)$ in Eq. (4.2) and $t \le T$,

$$\|\mathbf{W}_1-\mathbf{W}^*\|_2^2 + 2\eta t F_S(\mathbf{W}^*) + 18\eta^2G^2\log^2\frac{2T}{\delta} \le \frac{1}{2}\mathfrak{C}_S(\mathbf{W}^*).$$

Thus,

$$\left\|\mathbf{W}_{t+1}-\mathbf{W}^*\right\|_2^2 + \frac{3}{4}\eta\sum_{k=1}^{t}F_S(\mathbf{W}_k) \le \frac{3}{4}\mathfrak{C}_S(\mathbf{W}^*). \quad \text{(A.19)}$$

Hence, we conclude

$$\left\|\mathbf{W}_{t+1}-\mathbf{W}^*\right\|_2^2 \le \mathfrak{C}_S(\mathbf{W}^*), \qquad \eta\sum_{k=1}^{t}F_S(\mathbf{W}_k) \le \mathfrak{C}_S(\mathbf{W}^*). \quad \text{(A.20)}$$

Therefore, conditional on E_1, the induction closes with probability at least $1-2\delta$, and (4.6) holds for all $t \in [T+1]$. On the same event, the case $t=1$ in (4.7) is trivial. For $t \in \{2,\dots,T+1\}$, applying Lemma 14 with $t-1$ in place of t, together with the bound $\eta\sum_{k=1}^{t-1}F_S(\mathbf{W}_k) \le \mathfrak{C}_S(\mathbf{W}^*)$, we obtain

$$\|\mathbf{W}_t-\mathbf{W}_1\|_{2,\infty} \le \Big(1+\frac{2L}{G}\Big)\frac{\mathfrak{C}_S(\mathbf{W}^*)}{\sqrt{m}} + \frac{G\eta\log(mT/\delta)}{\sqrt{m}} = \widehat{R}_1\big(\mathfrak{C}_S(\mathbf{W}^*)\big).$$

This proves Eq. (4.7). □

With Theorem 2 and the realizability assumption, we now prove Theorem 3 as an application.

Proof of Theorem 3. By Assumption 3 with $\epsilon = 1/T$, there exists $\mathbf{W}^{\frac{1}{T}}$ such that

$$F_S\big(\mathbf{W}^{\frac{1}{T}}\big) \le \frac{1}{T}, \qquad \left\|\mathbf{W}_1-\mathbf{W}^{\frac{1}{T}}\right\|_2 \le g\Big(\frac{1}{T}\Big). \quad \text{(A.21)}$$

Hence,

$$\begin{aligned}\mathfrak{C}_S\big(\mathbf{W}^{\frac{1}{T}}\big) &= 2\|\mathbf{W}^{\frac{1}{T}}-\mathbf{W}_1\|_2^2 + 4\eta T F_S(\mathbf{W}^{\frac{1}{T}}) + 36\eta^2G^2\log^2\frac{2T}{\delta} \\ &\le 2g^2\Big(\frac{1}{T}\Big) + 4\eta + 36\eta^2G^2\log^2\frac{2T}{\delta}.\end{aligned} \quad \text{(A.22)}$$

We take $\mathbf{W}^* = \mathbf{W}^{\frac{1}{T}}$ in Theorem 2, which gives

$$\|\mathbf{W}_t-\mathbf{W}^{\frac{1}{T}}\|_2^2 \le \mathfrak{C}_S\big(\mathbf{W}^{\frac{1}{T}}\big), \qquad \eta\sum_{k=1}^{t-1}F_S(\mathbf{W}_k) \le \mathfrak{C}_S\big(\mathbf{W}^{\frac{1}{T}}\big), \quad \text{(A.23)}$$

and

$$\|\mathbf{W}_t-\mathbf{W}_1\|_{2,\infty} \le \widehat{R}_1\big(\mathfrak{C}_S\big(\mathbf{W}^{\frac{1}{T}}\big)\big). \quad \text{(A.24)}$$

Combining Eqs. (A.22) with (A.23), we obtain

$$\|\mathbf{W}_t-\mathbf{W}^{\frac{1}{T}}\|_2^2 \le 2g^2\Big(\frac{1}{T}\Big) + 4\eta + 36\eta^2G^2\log^2\frac{2T}{\delta}, \qquad \forall t \in [T+1],$$

which proves Eq. (4.8). Next, taking $t = T + 1$ in Eq. (A.23) gives $\eta \sum_{k=1}^{T} F_S(\mathbf{W}_k) \le \mathfrak{C}_S\big(\mathbf{W}^{\frac{1}{T}}\big)$. Dividing by ηT and using Eq. (A.22), we obtain

$$\frac{1}{T}\sum_{k=1}^{T} F_S(\mathbf{W}_k) \le \frac{2g^2(1/T)}{\eta T} + \frac{4}{T} + \frac{36\eta G^2}{T}\log^2\frac{2T}{\delta},$$

which proves Eq. (4.9).
Finally, since $\widehat{R}_1(C) = \big(1 + \frac{2L}{G}\big)\frac{C}{\sqrt{m}} + \frac{G\eta\log(mT/\delta)}{\sqrt{m}}$ is increasing in C, (A.22) and (A.24) imply

$$\begin{aligned}\|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty} &\le \widehat{R}_1\big(\mathfrak{C}_S\big(\mathbf{W}^{\frac{1}{T}}\big)\big) \\ &\le \Big(1 + \frac{2L}{G}\Big)\frac{2g^2(\frac{1}{T}) + 4\eta + 36\eta^2G^2\log^2\frac{2T}{\delta}}{\sqrt{m}} + \frac{G\eta\log(mT/\delta)}{\sqrt{m}},\end{aligned}$$

for all $t \in [T+1]$. This proves Eq. (4.10).

□

Proof of Corollary 4. Let

$$R_1 := \frac{\sqrt{\pi}}{16\bar{B}_T}, \qquad \bar{B}_T := 2g^2\Big(\frac{1}{T}\Big) + \frac{1}{64L\log\frac{2T}{\delta}} + \frac{9G^2}{16384L^2}.$$

By Eq. (A.22) and $\eta \le \frac{1}{256L\log(2T/\delta)}$, we know

$$\begin{aligned}\mathfrak{C}_S\big(\mathbf{W}^{\frac{1}{T}}\big) &= 2\|\mathbf{W}^{\frac{1}{T}} - \mathbf{W}_1\|_2^2 + 4\eta T F_S(\mathbf{W}^{\frac{1}{T}}) + 36\eta^2G^2\log^2\frac{2T}{\delta} \\ &\le 2g^2\Big(\frac{1}{T}\Big) + \frac{1}{64L\log\frac{2T}{\delta}} + \frac{9G^2}{16384L^2} = \bar{B}_T.\end{aligned}$$

Therefore, $m \ge 32\bar{B}_T^2\log\frac{n}{\delta} \ge 32\mathfrak{C}_S^2\big(\mathbf{W}^{\frac{1}{T}}\big)\log\frac{n}{\delta}$. Since $\widehat{R}_1(C)$ is increasing in C, the assumption

$$\Big(1 + \frac{2L}{G}\Big)\frac{\bar{B}_T}{\sqrt{m}} + \frac{G\eta\log(mT/\delta)}{\sqrt{m}} \le \frac{\sqrt{\pi}}{16\bar{B}_T}$$

implies

$$\widehat{R}_1\big(\mathfrak{C}_S\big(\mathbf{W}^{\frac{1}{T}}\big)\big) \le \widehat{R}_1(\bar{B}_T) \le R_1 \le \frac{\sqrt{\pi}}{16\,\mathfrak{C}_S\big(\mathbf{W}^{\frac{1}{T}}\big)}.$$

Moreover, by assumption,

$$\|\mathbf{W}^{\frac{1}{T}} - \mathbf{W}_1\|_{2,\infty} \le R_1 \le \frac{\sqrt{\pi}}{16\,\mathfrak{C}_S\big(\mathbf{W}^{\frac{1}{T}}\big)}.$$

Thus, part (ii) of Definition 3 holds for the reference model $\mathbf{W}^* = \mathbf{W}^{\frac{1}{T}}$ and the deterministic radius R_1. By Lemma 1, part (i) of E_1 holds with probability at least $1 - \delta$. Hence E_1 holds with probability at least $1 - \delta$. Applying Theorem 3 and taking a union bound, we conclude that (4.8), (4.9), and (4.10) hold for all $t \in [T+1]$ with probability at least $1 - 3\delta$.

□

A.2 Proofs of Generalization

In this subsection, we establish the proof of the generalization analysis. To complete our proof, we introduce a sequence of lemmas in the following paragraph. Lemma 17 provides Rademacher complexity bounds for neural networks under initialization-dependent constraints, which can capture the implicit regularization of gradient methods bias towards models close to initialization [21, 27, 31, 32].

Lemma 17 (Complexity Bounds, [20, Theorem 1]). *Let $R > 0$ and $\mathcal{G} := \{\mathbf{x} \mapsto \Phi(\mathbf{W}; \mathbf{x}) : \|\mathbf{W} - \mathbf{W}_1\|_2 \leq R\}$ be the hypothesis space, where $\Phi(\mathbf{W}; \mathbf{x})$ is defined in Eq.* (3.1)*. Then,*

$$\mathfrak{R}_S(\mathcal{G}) \leq R\Big(\frac{3+\sqrt{5}}{n}\Big(\sum_{i=1}^{n}\|\mathbf{x}_i\|_2^2\Big)^{\frac{1}{2}} + \frac{c_m\left\|\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right\|_\sigma^{\frac{1}{2}}}{n}\Big), \tag{A.25}$$

where $\|\cdot\|_\sigma$ denotes the spectral norm and

$$c_m = 2\sqrt{2}\Big(1+\frac{1}{2\log(2m)}\Big)\log^{\frac{1}{2}}\Big(2m\lceil\log_2(2m^{\frac{1}{2}})\rceil\Big).$$

Remark 6. In the original statement in [20], there is a term $\frac{1}{n}\big(\sum_{j=1}^{m}\sum_{i=1}^{n}\sigma^2(\mathbf{x}_i^\top\mathbf{w}_{1,j})\big)^{\frac{1}{2}}$ in the upper bound. However, this term can be removed since the weight a_j is fixed in our setting. The above theorem provides a flexible complexity bound that allows us to avoid additional structural assumptions and significantly simplifies the generalization analysis compared to the framework of [21].

Proof of Lemma 5. For any $\mathbf{W} \in \mathcal{W}$, we have

$$\|\mathbf{W} - \mathbf{W}_1\|_2 \leq \|\mathbf{W} - \mathbf{W}^*\|_2 + \|\mathbf{W}^* - \mathbf{W}_1\|_2 \leq 2\sqrt{\mathfrak{C}(\mathbf{W}^*)},$$

where the last inequality follows from (4.13) and (4.14). Applying Lemma 17 with $R = 2\sqrt{\mathfrak{C}(\mathbf{W}^*)}$, for any sample of size n drawn from S, we have

$$\mathfrak{R}_S(\mathcal{F}) \leq 2\sqrt{\mathfrak{C}(\mathbf{W}^*)}\Big(\frac{3+\sqrt{5}}{n}\Big(\sum_{i=1}^{n}\|\mathbf{x}_i\|_2^2\Big)^{1/2} + \frac{c_m}{n}\Big\|\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\Big\|_\sigma^{1/2}\Big).$$

Under Assumption 2, we know $\sum_{i=1}^{n}\|\mathbf{x}_i\|_2^2 = n$ and $\big\|\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\big\|_\sigma \leq n$. Therefore,

$$\mathfrak{R}_S(\mathcal{F}) \leq \frac{2(3+\sqrt{5}+c_m)\sqrt{\mathfrak{C}(\mathbf{W}^*)}}{\sqrt{n}}.$$

The bound is uniform over all samples of size n contained in S. Taking the supremum in the definition of $\mathfrak{R}_{S,n}(\mathcal{F})$ and absorbing c_m into the $\widetilde{O}(\cdot)$ notation yields

$$\mathfrak{R}_{S,n}(\mathcal{F}) = \widetilde{O}\Big(\sqrt{\frac{\mathfrak{C}(\mathbf{W}^*)}{n}}\Big). \tag{A.26}$$

The proof is complete. □

Lemma 18 (Bernstein inequality). *Let $\{\zeta_i\}_{i=1}^n$ be a sequence of i.i.d. random variables. Let b be a constant such that $|\zeta_i| \leq b$ and the variance $\mathrm{Var}(\zeta_i) < \infty$. Then, for any $0 < \delta < 1$ with probability at least $1-\delta$, there holds*

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

Lemma 19 ([40]). *Let $\mathcal{W} \subset \mathbb{R}^{md}$ be a parameter set, and define the function class*

$$\mathcal{F} := \{\mathbf{x} \mapsto \Phi(\mathbf{W}; \mathbf{x}) : \mathbf{W} \in \mathcal{W}\}.$$

Let $b := \sup_{\mathbf{z},\mathbf{W}\in\mathcal{W}} f(\mathbf{W}; \mathbf{z})$ denote a uniform bound. Then for any $\delta \in (0,1)$, with probability at least $1-\delta/2$ over a sample S, the following holds simultaneously for all $\mathbf{W} \in \mathcal{W}$:

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

We first control the loss function value. Since ℓ is nonnegative and L-smooth, it satisfies the self-bounding inequality $|\ell'(a)| \le \sqrt{2L\ell(a)}, \forall a \in \mathbb{R}$. Moreover,

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

Hence, for any $(\mathbf{x}, y)$ by the L-smoothness of ℓ we know

$$\begin{aligned}
&\ell(y\Phi(\mathbf{W};x))\\
&\le \ell(y\Phi(\mathbf{W}^*;x)) + \ell'(y\Phi(\mathbf{W}^*;x))\,y(\Phi(\mathbf{W};x)-\Phi(\mathbf{W}^*;x)) + \frac{L}{2}(\Phi(\mathbf{W};x)-\Phi(\mathbf{W}^*;x))^2\\
&\le \ell(y\Phi(\mathbf{W}^*;x)) + \frac{1}{2L}|\ell'(y\Phi(\mathbf{W}^*;x))|^2 + L(\Phi(\mathbf{W};x)-\Phi(\mathbf{W}^*;x))^2\\
&\le 2\ell(y\Phi(\mathbf{W}^*;x)) + L\|\mathbf{W}-\mathbf{W}^*\|_2^2. \qquad \text{(A.27)}
\end{aligned}$$

Therefore, for any $\mathbf{W} \in \mathcal{W}$,

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

Let $\xi_i := f(\mathbf{W}^*; z_i)$. Then we have

$$\mathrm{Var}(\xi_i) \le \mathbb{E}[\xi_i^2] \le b_*\mathbb{E}[\xi_i] = b_*F(\mathbf{W}^*). \qquad \text{(A.28)}$$

By Lemma 18, Lemma 19, and a union bound, the event E_2 holds with probability at least $1-\delta$. Applying Bernstein's inequality together with (A.28), we obtain

$$F_S(\mathbf{W}^*) - F(\mathbf{W}^*) \le \frac{2b_*\log(2/\delta)}{3n} + \left(\frac{2b_*F(\mathbf{W}^*)\log(2/\delta)}{n}\right)^{1/2}.$$

Using Young's inequality in the form $uv \le u^2 + \frac{1}{4}v^2$, with $u = \sqrt{F(\mathbf{W}^*)}$ and $v = \big(\frac{2b_*\log(2/\delta)}{n}\big)^{1/2}$, we get

$$\left(\frac{2b_*F(\mathbf{W}^*)\log(2/\delta)}{n}\right)^{1/2} \le F(\mathbf{W}^*) + \frac{b_*\log(2/\delta)}{2n}. \qquad \text{(A.29)}$$

Therefore,

$$F_S(\mathbf{W}^*) \le 2F(\mathbf{W}^*) + \frac{7b_*\log(2/\delta)}{6n}. \qquad \text{(A.30)}$$

Together with Eq.(4.2) and Eq. (4.13), the above inequality immediately implies the high-probability bound

$$\mathfrak{C}_S(\mathbf{W}^*) \le \mathfrak{C}(\mathbf{W}^*) \qquad \text{(A.31)}$$

Similarly, using Young's inequality $uv \le \frac{1}{2}u^2 + \frac{1}{2}v^2$ on the mixed term and absorbing constants, we obtain

$$F(\mathbf{W}) - 2F_S(\mathbf{W}) \lesssim L\log^3 n\,\mathfrak{R}_{S,n}^2(\mathcal{F}) + \frac{b\log(2/\delta)}{n}, \qquad \forall \mathbf{W} \in \mathcal{W}. \qquad \text{(A.32)}$$

We are in a position to prove Theorem 6.

Proof of Theorem 6. By Theorem 2 and Eq. (A.31), on the event E_1, for all $t \in [T]$,

$$\|\mathbf{W}_t - \mathbf{W}^*\|_2^2 \le \mathfrak{C}(\mathbf{W}^*), \qquad \|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty} \le \widehat{R}_1\big(\mathfrak{C}(\mathbf{W}^*)\big). \qquad \text{(A.33)}$$

Hence $\mathbf{W}_t \in \mathcal{W}$ for all $t \in [T]$. By (A.27), the uniform bound in Lemma 19 holds with $b = 2b_* + L\mathfrak{C}(\mathbf{W}^*)$. On the event E_2, Eq. (A.32) implies

$$F(\mathbf{W}_t) - 2F_S(\mathbf{W}_t) = \widetilde{O}\Big(\mathfrak{R}_{S,n}^2(\mathcal{F}) + \frac{b\log(2/\delta)}{n}\Big), \qquad \forall t \in [T].$$

Using Lemma 5, we further obtain

$$F(\mathbf{W}_t) - 2F_S(\mathbf{W}_t) = \widetilde{O}\Big(\frac{\mathfrak{C}(\mathbf{W}^*)}{n} + \frac{b\log(2/\delta)}{n}\Big), \qquad \forall t \in [T].$$

Summing over $t = 1, \dots, T$ and multiplying by η, we get

$$\eta \sum_{t=1}^{T} F(\mathbf{W}_t) = \widetilde{O}\Big(\eta T \frac{\mathfrak{C}(\mathbf{W}^*)}{n} + \frac{\eta T\, b \log(2/\delta)}{n} + \eta \sum_{t=1}^{T} F_S(\mathbf{W}_t)\Big).$$

By Theorem 2 and Eq. (A.31), we know $\eta \sum_{t=1}^{T} F_S(\mathbf{W}_t) \le \mathfrak{C}(\mathbf{W}^*)$. Therefore,

$$\eta \sum_{t=1}^{T} F(\mathbf{W}_t) = \widetilde{O}\Big(\eta T \frac{\mathfrak{C}(\mathbf{W}^*)}{n} + \frac{\eta T\, b \log(2/\delta)}{n} + \mathfrak{C}(\mathbf{W}^*)\Big).$$

Dividing both sides by ηT proves (4.22). □

We now prove the generalization result of SGD with the realizability assumption.

Proof of Theorem 7. We work on the event $E_1 \cap E_2$. Choose $\mathbf{W}^* := \mathbf{W}^{1/T}$. By Assumption 3,

$$F_S(\mathbf{W}^{1/T}) \le \frac{1}{T}, \qquad \|\mathbf{W}^{1/T} - \mathbf{W}_1\|_2 \le g\Big(\frac{1}{T}\Big). \tag{A.34}$$

Since ℓ is Lipschitz continuous, $\Phi(\mathbf{W}_1; x) = 0$, σ is 1-Lipschitz, $|a_j| = 1/\sqrt{m}$, and $\|x\|_2 = 1$, we have

$$\begin{aligned}
\sup_{\mathbf{z}} f(\mathbf{W}^{1/T}; \mathbf{z}) &\le \sup_{\mathbf{z}} \Big(\ell(y\Phi(\mathbf{W}^{1/T}; \mathbf{x})) - \ell(y\Phi(\mathbf{W}_1; \mathbf{x}))\Big) + \ell(0) \\
&\le G \sup_{\mathbf{x}} \big|\Phi(\mathbf{W}^{1/T}; \mathbf{x}) - \Phi(\mathbf{W}_1; \mathbf{x})\big| + \ell(0) \\
&\le \frac{G}{\sqrt{m}} \sum_{j=1}^{m} \|\mathbf{w}_j^{1/T} - \mathbf{w}_{1,j}\|_2 + \ell(0) \\
&\le G\|\mathbf{W}^{1/T} - \mathbf{W}_1\|_2 + \ell(0) \le Gg\Big(\frac{1}{T}\Big) + \ell(0).
\end{aligned} \tag{A.35}$$

Hence, we may take $b_* = Gg(1/T) + \ell(0)$. By Eq. (A.30) with $\mathbf{W}^* = \mathbf{W}^{1/T}$, together with $F_S(\mathbf{W}^{1/T}) \le 1/T$, we obtain

$$F(\mathbf{W}^{1/T}) \le \frac{2}{T} + \frac{7b_* \log(2/\delta)}{6n}. \tag{A.36}$$

By Eq. (4.13), Eq. (A.34), and Eq. (A.36),

$$\begin{aligned}
\mathfrak{C}(\mathbf{W}^{1/T}) &= 2\|\mathbf{W}_1 - \mathbf{W}^{1/T}\|_2^2 + 4\eta T\Big(2F(\mathbf{W}^{1/T}) + \frac{7b_* \log(2/\delta)}{6n}\Big) + 36\eta^2 G^2 \log^2 \frac{2T}{\delta} \\
&\lesssim g^2\Big(\frac{1}{T}\Big) + \eta T\Big(\frac{1}{T} + \frac{b_* \log(2/\delta)}{n}\Big) + \eta^2 G^2 \log^2 \frac{2T}{\delta}.
\end{aligned} \tag{A.37}$$

Using the stepsize assumption and $\eta T \asymp n$, we get

$$\mathfrak{C}(\mathbf{W}^{1/T}) \lesssim g^2\Big(\frac{1}{T}\Big) + b_* \log(1/\delta). \tag{A.38}$$

Applying Theorem 6 with $\mathbf{W}^* = \mathbf{W}^{1/T}$, we obtain

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

Combining this with (A.38), (A.35), and $\eta T \asymp n$, we conclude

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

The proof is complete. □

Proof of Corollary 8. By Corollary 4, under the stated assumptions, the event E_1 holds and the bounds (4.8), (4.9), and (4.10) hold simultaneously for all $t \in [T+1]$ with probability at least $1-3\delta$. By Eq. (A.35), we may choose $b_* = Gg(1/T) + \ell(0)$. By the argument above, the event E_2 holds with probability at least $1-\delta$. Therefore, by a union bound, $E_1 \cap E_2$ holds with probability at least $1-4\delta$. Applying Theorem 7, we obtain

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

In particular, if $g(\epsilon) \lesssim \log(1/\epsilon)$, then $g^2(1/T) \lesssim \log^2 T$, which is absorbed into the $\widetilde{O}(\cdot)$ notation, and thus $\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{W}_t) = \widetilde{O}(1/n)$. The proof is complete. □

A.3 Proofs on NTK Separability

In this subsection, we present the proofs of the connection with NTK separability. The proof of Lemma 9 is identical to the corresponding NTK realizability argument in the gradient descent setting[21], since it depends only on the NTK separability assumption, the construction of $\mathbf{W}^\epsilon$, and symmetric initialization, but not on the optimization algorithm. We therefore omit the details.

A.3.1 Additional derivation for NTK separability

We provide the derivation of the population-radius bound used in Section 4. Recall that

$$\alpha = \frac{2\log T}{\gamma}, \qquad \mathbf{W}^{1/T} = \mathbf{W}_1 + \alpha \mathbf{v}_*.$$

By Lemma 9,

$$F_S(\mathbf{W}^{1/T}) \le \frac{1}{T}, \qquad \|\mathbf{W}^{1/T} - \mathbf{W}_1\|_2 = \alpha, \qquad \|\mathbf{W}^{1/T} - \mathbf{W}_1\|_{2,\infty} \le \alpha\beta.$$

Thus, by the definition of $\mathfrak{C}_S(\cdot)$,

$$\mathfrak{C}_S(\mathbf{W}^{1/T}) \le 2\alpha^2 + 4\eta + 36\eta^2 G^2 \log^2 \frac{2T}{\delta} =: \bar{C}_T.$$

Since ℓ is Lipschitz continuous, $\Phi(\mathbf{W}_1; \mathbf{x}) = 0$ by symmetric initialization, σ is 1-Lipschitz, $|a_j| = 1/\sqrt{m}$, and $\|\mathbf{x}\|_2 = 1$, we have

$$\sup_{\mathbf{z}} f(\mathbf{W}^{1/T}; \mathbf{z}) \le \alpha + \log 2. \tag{A.39}$$

Hence we may take $b_* = \alpha + \log 2$.

By Eq. (A.30) with $\mathbf{W}^* = \mathbf{W}^{1/T}$, together with $F_S(\mathbf{W}^{1/T}) \le 1/T$, we obtain

$$F(\mathbf{W}^{1/T}) \le \frac{2}{T} + \frac{7b_*\log(2/\delta)}{6n}. \tag{A.40}$$

Therefore, by the definition of $\mathfrak{C}(\cdot)$,

$$\begin{aligned}\mathfrak{C}(\mathbf{W}^{1/T}) &= 2\|\mathbf{W}_1 - \mathbf{W}^{1/T}\|_2^2 + 4\eta T\Big(2F(\mathbf{W}^{1/T}) + \frac{7b_*\log(2/\delta)}{6n}\Big) + 36\eta^2 G^2 \log^2\frac{2T}{\delta} \\ &\lesssim \alpha^2 + \eta T\Big(\frac{1}{T} + \frac{b_*\log(2/\delta)}{n}\Big) + \eta^2 G^2 \log^2\frac{2T}{\delta}. \end{aligned} \tag{A.41}$$

This is the population-radius bound used in the proof of Theorem 11.

Proof of Lemma 10. By Lemma 9, with probability at least $1-\delta$,

$$F_S(\mathbf{W}^{1/T}) \le \frac{1}{T}, \qquad \|\mathbf{W}^{1/T} - \mathbf{W}_1\|_{2,\infty} \le \alpha\beta.$$

Hence, (4.27) implies $\mathfrak{C}_S(\mathbf{W}^{1/T}) \le \bar{C}_T$. Since $\widehat{R}_1(\cdot)$ is increasing, we know

$$\widehat{R}_1(\mathfrak{C}_S(\mathbf{W}^{1/T})) \le \widehat{R}_1(\bar{C}_T) \le R_1, \qquad \|\mathbf{W}^{1/T} - \mathbf{W}_1\|_{2,\infty} \le \alpha\beta \le R_1.$$

Thus, the lower bound in Definition 3(ii) holds. Moreover, by the assumptions of the lemma,

$$R_1 \bar{C}_T = \max\{\widehat{R}_1(\bar{C}_T), \alpha\beta\}\, \bar{C}_T \le \frac{\sqrt{\pi}}{16}.$$

Since $\mathfrak{C}_S(\mathbf{W}^{1/T}) \le \bar{C}_T$, it follows that $R_1\, \mathfrak{C}_S(\mathbf{W}^{1/T}) \le \frac{\sqrt{\pi}}{16}$, which is the upper bound in Definition 3(ii).

Finally, by Lemma 1 and the width condition, Definition 3(i) holds with probability at least $1-\delta$. A union bound completes the proof. □

Proof of Theorem 11. By Lemma 9, Assumption 3 holds with

$$g(\epsilon) \lesssim \frac{\log(1/\epsilon)}{\gamma}, \qquad \text{which implies} \qquad g(1/T) \lesssim \frac{\log T}{\gamma}.$$

Lemma 10 shows that E_1 holds with probability at least $1-2\delta$. By the construction of E_2, that event holds with probability at least $1-\delta$. Applying Theorem 7 and then using $g(1/T) \lesssim \log T/\gamma$, we get

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

The probability follows by combining Lemma 10, the event E_2, and the high-probability optimization event from Theorem 2, followed by a union bound. □

B Future directions

Several directions remain open. First, extending the present analysis beyond shallow ReLU networks would require new tools for controlling activation changes and localization across different architectures. For smooth activations, the localization framework is expected to extend more directly, since changes in a smooth activation function can be controlled through standard smoothness arguments. For example, Lemma 3.1 of [41] shows that SNNs with a smooth activation function satisfy a self-bound weak convexity condition, which is then used to derive risk bounds of order $\widetilde{O}(1/(n\gamma^2))$ for GD. This suggests that the self-bounded weak convexity condition could also be used to extend our analysis to smooth activation functions. For convolutional neural networks, the Euclidean localization condition $\|\mathbf{W} - \mathbf{W}_1\|_{2,\infty}$ alone may not fully exploit the convolutional structure. A suitable extension would likely require filter-wise localization conditions and complexity estimates adapted to convolutional architectures. For ResNets and other deep architectures, the extension is more challenging, as perturbations may propagate across layers. New techniques or additional structural assumptions may therefore be needed to control this accumulation. Second, the optimization-dependent localization framework may apply beyond standard SGD, including variance-reduced methods, momentum-based methods, and adaptive-gradient algorithms. Understanding whether similar high-probability population-risk guarantees can be obtained for these algorithms is an interesting direction. Third, while our NTK-separable setting provides a clean structural condition for fast rates, recent work on feature learning and representation dynamics suggests that sharper guarantees may be possible beyond the near-initialization regime. Developing high-probability analyses that capture such feature-learning effects remains an important open problem.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's theoretical contributions: a high-probability analysis of SGD for shallow ReLU networks, optimization-dependent localization of the SGD trajectory, and nearly optimal population-risk bounds under realizability or NTK separability. The stated claims are matched by the formal theorems in Section 4 and the proofs in the appendix.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the scope of the results through the stated assumptions, including shallow ReLU architectures, realizability or NTK separability, and the high-probability setting considered in the analysis. Additional limitations and future directions, including extensions to deep networks, other stochastic optimization methods, and feature-learning regimes beyond NTK separability, are discussed in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theorem states or references the required assumptions, including the loss, data, realizability, and NTK separability assumptions. Complete proofs of the optimization, generalization, and NTK-separability results are provided in the appendix, with the main quantities and events cross-referenced throughout.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper is theoretical and does not include experimental results. The main claims are mathematical statements supported by formal assumptions and proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does not include experiments requiring data or code. The contribution is a theoretical analysis. All assumptions, algorithms, and proof details needed to verify the results are stated in the paper and appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include empirical experiments. The optimizer, step-size assumptions, sample complexity, and model assumptions relevant to the theoretical results are specified in the theorem statements.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in the appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not report empirical experiments. Instead, the results are high-probability theoretical guarantees with explicit probability statements.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include computational experiments. The computational comparison concerns theoretical gradient-computation complexity, such as the $O(n)$ stochastic-gradient computations required by SGD when $T \asymp n$.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is a mathematical analysis and does not involve human subjects, private or sensitive data, deployed systems, or released models. To the best of the authors' knowledge, the research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is foundational theoretical work on the optimization and generalization properties of SGD for shallow ReLU networks. It does not introduce a deployed system, dataset, model, or application with direct societal impacts.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data, pretrained models, generative systems, or scraped datasets, and therefore does not pose the release risks described in the question.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use external datasets, codebases, pretrained models, or other experimental assets. Prior mathematical and algorithmic works used for comparison are cited in the related discussion and references.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new datasets, code, models, benchmarks, or other research assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subject research, crowdsourcing, or data collection from participants; thus, IRB approval or equivalent review is not applicable.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology and theoretical results do not use LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.