Towards optimistic convergence of gradients for nonconvex learning

Yunwen Lei[1]
[1]Department of Mathematics, The University of Hong Kong
leiyw@hku.hk

Abstract

Uniform convergence of gradients is important to understand the generalization behavior of machine learning models in a nonconvex setting. Existing uniform convergence suffers from some drawbacks such as the dependence on the dimensionality or the inefficiency to imply fast rates. In this paper, we develop an optimistic bound for the uniform convergence of gradients for nonconvex problems. Our bound is dimension-free and interpolates between the standard rate of order $1/\sqrt{n}$ and the fast rate of order $1/n$, where n is the sample size. Our result is very general since it relates the uniform convergence to Rademacher complexities of two real-valued function classes. We justify the applicability by applying our result to shallow neural networks with Lipschitz activation functions.

1 Introduction

Recent years have witnessed a remarkable success of machine learning (ML) in solving various practical problems, which motivates an increasing interest to build theoretical guarantees to safeguard the performance of ML models. An important topic in this direction is the generalization analysis of ML models, i.e., how the behavior of ML models on training can be generalized to testing [3, 19, 39, 47]. Understanding the generalization gap between testing and training is useful to explain how different parameters of learning algorithms affect the true behavior of the resulting ML models.

A popular approach to studying the generalization gap is the uniform convergence approach [3, 12, 20, 36, 39, 44], where the generalization gap is controlled by the uniform deviation between empirical and population risks over the hypothesis space. This approach yields bounds depending on the complexity of the hypothesis space, for which various concepts have been introduced such as VC dimension, covering numbers and Rademacher complexities [3, 34, 47]. This generalization bound together with convergence rates of an optimization algorithm [6] further yields risk bounds to certify the prediction ability of the constructed models.

In modern ML we often encounter nonconvex problems for which the performance of an optimization algorithm is often measured by the magnitude of gradients [8, 11, 14, 21]. For example, a classical result shows that SGD can find a model at which the gradient of the empirical risk decays with the order of $T^{-\frac{1}{4}}$ [14], where T is the iteration number. In this case, it is natural to measure the generalization behavior via the gradient of the population risk, leading to the generalization gap measured by the deviation between the population and empirical gradients [9]. This observation motivates the recent growing interests in studying the uniform convergence of gradients.

To our knowledge, the first uniform convergence of gradients was studied in a seminal paper [27], where the uniform convergence of order $\sqrt{d}/\sqrt{n}$ was established based on covering numbers. Here d is the dimensionality of the model and n is the sample size. This uniform convergence was improved by localization analysis (i.e., considering only the complexity of a subset of the original hypothesis space to which the output model belongs) for shifted objectives [42, 43, 45]. The work [12] built a

chain rule for gradient Rademacher complexity, which implies bounds independent of d for generalized linear models. While these uniform convergence results imply interesting generalization bounds for nonconvex problems, there is still room for further improvements. For example, the discussions in [12, 27] do not imply optimistic bounds which interpolate between fast and slow rates depending on the difficulty of the learning problem. The discussions in [43, 45] require the model to be close to the best model for an optimistic bound, which is difficult to realize in nonconvex problems. Furthermore, the discussions in [27, 43, 45] only imply uniform convergence rates with a square-root dependency on d, which is not appealing for high-dimensional problems.

In this paper, we aim to develop dimension-free optimistic uniform convergence of gradients for nonconvex learning problems. Our main contributions are summarized as follows.

- We develop the first dimension-free optimistic bounds for the uniform convergence of gradients which interpolate between the slow rate $\widetilde{O}_n(1/\sqrt{n})$ and the fast rate $\widetilde{O}_n(1/n)$[1], depending on the difficulty of the problem.
- We introduce a new localization strategy to get the optimistic rates, where the localization function is built in terms of the derivative of the training errors. As compared to the existing localization analysis based on either the training errors [36] or the distance from the best model [43], our localization function is smaller and yields better rates.
- We apply our analysis to shallow neural networks (SNNs) with Lipschitz activation functions, for which we develop uniform convergence interpolating the rate $\widetilde{O}_n(\sqrt{1/n})$ and the fast rate $\widetilde{O}_n(1/n)$.

2 Related work

In this section, we present related work on two categories: the uniform convergence of gradients and optimistic bounds.

2.1 Uniform Convergence of Gradients

The work [27] pioneered the uniform convergence analysis under some assumptions on Hessian statistical noise and Hessian regularity, which was improved based on a Bernstein inequality for random variables in Hilbert spaces [45]. A localization analysis was recently conducted based on a generic chaining for a process with mixed tail increments [43]. These works [43, 45] require the model to be close to the best model for optimistic bounds. Furthermore, the above mentioned works are based on covering numbers of balls in $\mathbb{R}^d$ and therefore imply bounds with a square-root dependency on d. To remove the dependency on the dimensionality, the work [12] developed a chain rule for Rademacher complexity of vector-valued functions and used it to study excess risks under a Polyak-Łojasiewicz (PL) condition [12]. The uniform convergence was used to study the generalization behavior of stochastic optimization for smooth and nonconvex problems [23], differentially private optimization [1], iterative recovery [32] and minimax problems [46].

The uniform convergence of gradients often requires a smoothness-type condition on loss functions. For nonsmooth problems, the gradients may not exist and therefore another concept of uniform convergence is needed. The work [9] considered weakly convex and nonsmooth problems, and established uniform convergence for the gradient of Moreau envelope. The uniform convergence of Hessian matrices was also studied [13, 27] under a Lipschitz continuity assumption on the Hessian matrix.

Other than the uniform convergence approach, generalization measured by gradients was also studied based on algorithmic stability [10, 50], which introduced uniform gradient stability and developed generalization bounds with high probability.

[1]We use $O_n(\cdot)$ and $\widetilde{O}_n(\cdot)$ to indicate that we only focus on the dependency on n, where we ignore constant factors and logarithmic factors, respectively.

2.2 Optimistic Bounds

Optimistic bounds refer to bounds that can interpolate between slow and fast rates, depending on the difficulty of the problem [36]. Optimistic bounds were developed for smooth learning problems based on *worst-case* Rademacher complexities [36]. The key observation is that the Rademacher complexity of a local function space with empirical risks less than r scales as a square-root function of r. To this aim, the paper [36] used extensively the self-bounding property of smooth functions. The discussions [28] extended this result to multi-output learning, which include several tasks such as multi-class classification and multi-label learning. Optimistic rates were recently developed for multi-task representation learning [41], where a chain rule was developed for local Rademacher complexities. Optimistic bounds were also developed for online learning [36, 49].

These optimistic bounds were developed based on complexity measures [28, 36, 41]. Another approach to develop optimistic bounds is the stability approach. Optimistic bounds for empirical risk minimization were developed for smooth and strongly convex problems [33]. Stability-based optimistic bounds were also developed for SGD [25, 30], and nonconvex problems such as SNNs [29, 37] and deep neural networks [38].

3 Problem Setting

3.1 Background

Let $\mathbb{P}$ be a probability measure defined over a sample space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^{\tilde{d}}$ is an input space and $\mathcal{Y}$ is an output space. We consider classification problems with $\mathcal{Y} = \{+1, -1\}$. Let $S = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ be independently drawn from $\mathbb{P}$, based on which we aim to build a prediction function h from $\mathcal{X}$ to $\mathcal{Y}$. We consider parametric models $\Psi(\mathbf{w}; \mathbf{x})$ indexed by a parameter $\mathbf{w}$ in a parameter space $\mathcal{W} \subseteq \mathbb{R}^d$, where d is the dimensionality. The performance of $\Psi(\mathbf{w}; \cdot)$ on a single example $\mathbf{z} = (\mathbf{x}, y)$ is quantified by a loss function f of the form $f(\mathbf{w}; \mathbf{z}) = \ell\big(y, \Psi(\mathbf{w}; \mathbf{x})\big)$, where $\ell : \mathcal{Y} \times \mathbb{R} \mapsto \mathbb{R}_+$. The empirical and population risks of a model $\mathbf{w}$ are then quantified as

$$F_S(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{w}; \mathbf{z}_i) \quad \text{and} \quad F(\mathbf{w}) = \mathbb{E}_{\mathbf{z}}[f(\mathbf{w}; \mathbf{z})]. \tag{3.1}$$

We often apply an optimization algorithm $A : \cup_n \mathcal{Z}^n \mapsto \mathcal{W}$ to approximately minimize the empirical risk. Then we are interested in its prediction behavior as quantified by $F(A(S))$. The gap between testing and training, i.e., $F(A(S)) - F_S(A(S))$ is referred to as the generalization gap, which is a central concept in statistical learning theory.

For nonconvex problems, we often measure the convergence of an optimization algorithm by the magnitude of gradients. For example, convergence rates have been developed for $\|\nabla F_S(A(S))\|_2$ [14], where $A(S)$ is the output by SGD, $\|\cdot\|_2$ denotes the Euclidean norm and ∇ means the gradient operator. Then, it is natural to quantify the prediction performance via $\|\nabla F(A(S))\|_2$, according to which it is natural to consider the generalization gap $\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2$.

A popular approach to studying the generalization gap is the uniform convergence approach, i.e., controlling $\sup_{\mathbf{w} \in \mathcal{W}} \|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\|_2$. The underlying motivation is to remove the dependency of $A(S)$ on S. The uniform convergence of gradients has already been studied in the literature [12, 23, 27]. In this paper, we aim to build optimistic uniform convergence which can interpolate between the slow rate $\widetilde{O}_n(1/\sqrt{n})$ and the fast rate $\widetilde{O}_n(1/n)$, depending on the difficulty of the problem.

3.2 Complexity Measures

The uniform convergence of gradients depends on the complexity of the involved function spaces. Various complexity measures have been introduced in the literature, including covering number [5, 9, 27], Rademacher and Gaussian complexity [3, 4, 12, 36]. In this paper, we will use these complexity

measures to conduct the uniform convergence analysis. Let $\boldsymbol{\sigma} = (\sigma_i)_{i\in[n]} \sim \{\pm 1\}^n$ be independent Rademacher variables, i.e., $\Pr\{\sigma_i = 1\} = \Pr\{\sigma_i = -1\} = 1/2$, where $[n] := \{1, \ldots, n\}$. Let $\mathcal{F}'$ be a real-valued function class over a space $\mathcal{Z}'$ and $S' = \{\mathbf{z}'_i\}_{i=1}^n \subseteq \mathcal{Z}'$. Let $|S'|$ be the cardinality of S'.

Definition 1 (Empirical Rademacher Complexity)**.** The *empirical* Rademacher complexity of $\mathcal{F}'$ with respect to (w.r.t.) S' is defined as $\mathfrak{R}_{S'}(\mathcal{F}') = \mathbb{E}_{\boldsymbol{\sigma}}\big[\sup_{f\in\mathcal{F}'} \frac{1}{n}\sum_{i\in[n]} \sigma_i f(\mathbf{z}'_i)\big]$. We define the *worst-case* Rademacher complexity as $\mathfrak{R}_{\mathcal{Z}',n}(\mathcal{F}') = \sup_{S'\subseteq\mathcal{Z}':|S'|=n} \mathfrak{R}_{S'}(\mathcal{F}')$.

Rademacher complexity expresses the capability of the function space to correlate Rademacher noises. Another way to define complexity is through the number of balls required to approximate the function space up to an accuracy.

Definition 2 (Covering number)**.** For any $\epsilon > 0$ and $p \geq 1$, the empirical ℓ_p-norm covering number $\mathcal{N}_p(\epsilon, \mathcal{F}', S')$ w.r.t. S' is defined as the smallest number m of a collection of vectors $\mathbf{v}^1, \ldots, \mathbf{v}^m \in \{(f(\mathbf{z}'_1), \ldots, f(\mathbf{z}'_n)) : f \in \mathcal{F}'\}$ such that $\sup_{f\in\mathcal{F}'} \min_{j\in[m]} \big(\frac{1}{n}\sum_{i\in[n]} |f(\mathbf{z}'_i) - \mathbf{v}_i^j|^p\big)^{1/p} \leq \epsilon$, where $\mathbf{v}_i^j$ is the i-th component of the vector $\mathbf{v}^j$. In this case, we call $\{\mathbf{v}^1, \ldots, \mathbf{v}^m\}$ an (ϵ, ℓ_p)-cover of $\mathcal{F}'$ w.r.t. S'.

The above definition involves a parameter p, which is used to define a metric $d_{p,S'}$ on the function space as $d_{p,S'}(f, g) := \big(\frac{1}{n}\sum_{i=1}^n \big|f(\mathbf{z}'_i) - g(\mathbf{z}'_i)\big|^p\big)^{1/p}, p \geq 1$. By the monotonicity of $d_{p,S'}$ w.r.t. p, we know that $\mathcal{N}_p(\epsilon, \mathcal{F}', S') \leq \mathcal{N}_q(\epsilon, \mathcal{F}', S')$ if $p \leq q$. In this paper, we will use the localization trick to study optimistic bounds, where the localization is achieved by L_S and L defined below

$$L_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n |\ell'(y_i, \Psi(\mathbf{w}; \mathbf{x}_i))|^2, \quad L(\mathbf{w}) := \mathbb{E}_{\mathbf{z}}[|\ell'(y, \Psi(\mathbf{w}; \mathbf{x}))|^2], \tag{3.2}$$

where $\ell'(y, a)$ means the derivative of ℓ w.r.t. a. These notions lead to the concept of local Rademacher complexity [4], which is a powerful technique to develop a fast decay of risk bounds.

4 Main Results

We require some assumptions on the loss function for our analysis. Eq. (4.1) is a self-bounding property and holds for the logistic loss, exponential loss, polynomially tailed losses and sub-exponential tailed losses. In Section B, we will provide details to show how these loss functions satisfy Assumption 1 and the Lipschitz continuity. While the exponential loss is not Lipschitz continuous, it is locally Lipschitz continuous and our analysis still applies if one projects the output of model to a bounded interval. Let $\langle\cdot,\cdot\rangle$ denote the dot product.

Definition 3. Let $g : \mathcal{W} \mapsto \mathbb{R}$ and $G, L \geq 0$. We say g is convex if $g(\mathbf{w}) \geq g(\mathbf{w}') + \langle \mathbf{w} - \mathbf{w}', \nabla g(\mathbf{w}')\rangle$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$. We say g is G-Lipschitz continuous if $|g(\mathbf{w}) - g(\mathbf{w}')| \leq G\|\mathbf{w} - \mathbf{w}'\|_2$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$. We say g is L-smooth if $\|\nabla g(\mathbf{w}) - \nabla g(\mathbf{w}')\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$.

Assumption 1. Assume there exists $\beta > 0$ such that

$$|\ell''(y, a)| \leq \beta|\ell'(y, a)|, \quad \forall a \in \mathbb{R}, y \in \mathcal{Y}. \tag{4.1}$$

Assume for all y, we have $|\ell'(y, a)| \leq \max\{|\ell'(y, a_1)|, |\ell'(y, a_2)|\}$ if $a \in [a_1, a_2]$.

The following assumption imposes boundedness conditions on the gradient of Ψ and ℓ. We will show that $\|\nabla\Psi(\mathbf{w}, \mathbf{x})\|_2 \leq G_\Psi$ holds for SNNs with appropriate activation functions, where $\nabla\Psi(\mathbf{w}, \mathbf{x})$ denotes the gradient w.r.t. $\mathbf{w}$. We always assume $G_\ell \geq 1, G_\Psi \geq 1$, which is a mild assumption as we can always take the maximum of them and 1.

Assumption 2. Let $G_\Psi, G_\ell \geq 0$. Assume $\|\nabla\Psi(\mathbf{w}, \mathbf{x})\|_2 \leq G_\Psi$ for all $\mathbf{w} \in \mathcal{W}$ and $\mathbf{x} \in \mathcal{X}$. Assume $|\ell'(y, a)| \leq G_\ell$ for all $a \in \mathbb{R}$ and $y \in \mathcal{Y}$.

We will use the following (empirical) gradient Rademacher complexity to study the uniform convergence of gradients, which is an extension of Rademacher complexity to handle gradients [12].

Definition 4 (Empirical Gradient Rademacher Complexity)**.** The *empirical* gradient Rademacher complexity of $\mathcal{F}'$ w.r.t. S' is defined as $\widetilde{\mathfrak{R}}_{S'}(\mathcal{F}') = \mathbb{E}_{\boldsymbol{\sigma}}\big[\sup_{f\in\mathcal{F}'} \big\|\frac{1}{n}\sum_{i=1}^n \sigma_i \nabla f(\mathbf{z}_i')\big\|_2\big]$.

4.1 Optimistic Bounds

In this section, we conduct a localization analysis to get optimistic bounds. Theorem 1, to be proved in Section 7.1, establishes the uniform convergence of gradients over a localized space

$$\mathcal{W}_r = \big\{\mathbf{w} : \mathbf{w}\in\mathcal{W}, L(\mathbf{w}) \le r\big\}, \quad r \ge 0, \tag{4.2}$$

where we use $L(\cdot)$ in Eq. (3.2) to achieve the localization. Both $\mathbb{E}_{\boldsymbol{\sigma}}\big[\sup_{\mathbf{w}\in\mathcal{W}_r} \big\|\frac{1}{n}\sum_{i=1}^n \sigma_i\nabla f(\mathbf{w};\mathbf{z}_i)\big\|_2\big]$ and $\big(\frac{2r\log(2/\delta)}{n}\big)^{1/2}$ are increasing functions of r. This clearly shows how the radius affects the uniform convergence rates. While the term $\frac{32G_\ell G_\Psi \log(2/\delta)}{3n}$ does not depend on r, it is not the dominant term and can be ignored.

Theorem 1 (Uniform convergence of gradients)**.** *Let Assumption 2 hold. Then, for any $\delta\in(0,1)$, with probability (w.r.t. S) at least $1-\delta$ we have*

$$\sup_{\mathbf{w}\in\mathcal{W}_r} \big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 \le 6\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_r}\Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i\nabla f(\mathbf{w};\mathbf{z}_i)\Big\|_2\Big] + G_\Psi\Big(\frac{2r\log(2/\delta)}{n}\Big)^{\frac{1}{2}} + \frac{32G_\ell G_\Psi\log(2/\delta)}{3n}.$$

According to Theorem 1, it suffices to estimate the gradient Rademacher complexity under the localization scheme via $L(\cdot)$. However, as we will see, it is more convenient to consider the localization via its empirical counterpart $L_S(\cdot)$. That is, we now consider the gradient Rademacher complexity of $\{\mathbf{z}\mapsto f(\mathbf{w};\mathbf{z}) : \mathbf{w}\in\widehat{\mathcal{W}}_r\}$, where

$$\widehat{\mathcal{W}}_r = \big\{\mathbf{w} : \mathbf{w}\in\mathcal{W}, L_S(\mathbf{w}) \le r\big\}, \quad r \ge 0. \tag{4.3}$$

The underlying reason is that we need to consider empirical covering numbers, which, according to Definition 2, are more related to L_S instead of L.

We now estimate $\widetilde{\mathfrak{R}}_S(\{\mathbf{z}\mapsto f(\mathbf{w};\mathbf{z}) : \mathbf{w}\in\widehat{\mathcal{W}}_r\})$, and our basic idea is to relate it to a standard Rademacher complexity. To this aim, denote

$$\mathcal{V} := \big\{\mathbf{v}\in\mathbb{R}^d : \|\mathbf{v}\|_2 \le 1\big\} \tag{4.4}$$

and introduce an expanded dataset

$$\widetilde{X} := \big\{(\mathbf{x}_1,1),(\mathbf{x}_1,2),(\mathbf{x}_2,1),(\mathbf{x}_2,2),\ldots,(\mathbf{x}_n,1),(\mathbf{x}_n,2)\big\}. \tag{4.5}$$

For any $\mathbf{w}\in\mathcal{W}, \mathbf{v}\in\mathcal{V}$, we introduce the function $\tilde{f}_{\mathbf{w},\mathbf{v}} : \widetilde{X}\mapsto\mathbb{R}$

$$\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x},k) = \begin{cases} \langle\nabla\Psi(\mathbf{w},\mathbf{x}),\mathbf{v}\rangle, & \text{if } k=1 \\ 2G_\Psi\beta\Psi(\mathbf{w};\mathbf{x}), & \text{if } k=2. \end{cases} \tag{4.6}$$

Note that $\tilde{f}$ is defined on $\widetilde{X}$, which is constructed by concatenating elements of $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ and $\{1,2\}$ together. Theorem 2 relates the gradient Rademacher complexity to the Rademacher complexity of $\widetilde{\mathcal{F}}$:

$$\widetilde{\mathcal{F}} = \big\{(\mathbf{x},k)\mapsto\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x},k) : (\mathbf{w},\mathbf{v})\in\mathcal{W}\times\mathcal{V}\big\}. \tag{4.7}$$

Note the nonlinear function ℓ is not involved in $\tilde{f}_{\mathbf{w},\mathbf{v}}$. Therefore, $\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})$ is more convenient to handle as compared to the gradient Rademacher complexity of $\{\mathbf{z}\mapsto f(\mathbf{w};\mathbf{z}) : \mathbf{w}\in\widehat{\mathcal{W}}_r\}$. We will provide explicit examples on how to control $\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})$. The proof of Theorem 2 is given in Section 7.2.

Theorem 2 (Gradient Rademacher complexity bounds)**.** *Let Assumption 1 hold and* $\|\nabla\Psi(\mathbf{w},\mathbf{x})\|_2 \le G_\Psi$ *for all* $\mathbf{w}\in\mathcal{W}$ *and* $\mathbf{x}\in\mathcal{X}$*. Then*

$$\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\widehat{\mathcal{W}}_r}\Big\|\frac{1}{n}\sum_{i=1}^n\sigma_i\nabla f(\mathbf{w};\mathbf{z}_i)\Big\|_2\Big] \le \max\big\{G_\Psi\sqrt{r}, 3\sqrt{3r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\big\}\Big(\frac{6}{\sqrt{n}}+\frac{2\max\big\{3\sqrt{3}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}}),1/\sqrt{n}\big\}}{G_\Psi}\Big) + 48\sqrt{6r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\Big(\widetilde{N}\log(48eB^2n/G_\Psi^2)+2\widetilde{N}^2\Big),$$

where $\widetilde{N}:=\lceil\log_2\big(G_\Psi\sqrt{n}\big)\rceil, b:=\sup_{\mathbf{w}\in\mathcal{W},\mathbf{x}\in\mathcal{X}}|\Psi(\mathbf{w};\mathbf{x})|$ *and* $B:=G_\Psi\max\{1,2\beta b\}$.

Remark 1. The above complexity bound grows as a square-root function of r. This clearly shows how the localization affects the complexity. If $\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\lesssim 1/\sqrt{n}$ (which, as we will show, is the case for SNNs), then Theorem 2 shows that

$$\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\widehat{\mathcal{W}}_r}\Big\|\frac{1}{n}\sum_{i=1}^n\sigma_i\nabla f(\mathbf{w};\mathbf{z}_i)\Big\|_2\Big] = \widetilde{O}\Big(\frac{G_\Psi\sqrt{r}}{\sqrt{n}}+\sqrt{r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\Big)=\widetilde{O}\Big(\frac{G_\Psi\sqrt{r}}{\sqrt{n}}\Big).$$

Here we denote $A\lesssim B$ if $A\le c\cdot B$, and $A\gtrsim B$ if $A\ge c\cdot B$ for some universal constant $c>0$, where c is independent of problem parameters such as $n,d,\tilde{d},\beta,G_\ell,G_\Psi,G_\gamma$. We use $\widetilde{O}(\cdot)$ to ignore logarithmic factors of problem parameters. If $L_S(A(S))=O_n(1/n)$, then intuitively we can apply Theorem 2 with $r\lesssim 1/n$, which leads to bounds of order $\widetilde{O}_n(1/n)$. This implies fast rates.

Below we present our main result on optimistic bounds for the uniform convergence of gradients, the proof of which is given in Section 7.4. For simplicity of presentation, we introduce

$$\mathfrak{R}_{2n}(\widetilde{\mathcal{F}}):=\max\big\{\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}}),G_\Psi/\sqrt{n}\big\}\quad\text{and}\quad\mathfrak{R}_n(\mathcal{G})=\max\big\{\mathfrak{R}_{S_{\mathbf{x}},n}(\mathcal{G}),1/(\beta\sqrt{n})\big\},\tag{4.8}$$

where we denote $S_{\mathbf{x}}:=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ and

$$\mathcal{G}:=\Big\{\mathbf{x}\mapsto\Psi(\mathbf{w};\mathbf{x}):\mathbf{w}\in\mathcal{W}\Big\}.\tag{4.9}$$

Theorem 3 (Optimistic bounds)**.** *Let Assumption 1 and Assumption 2 hold. Assume that* $\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\lesssim G_\Psi$*. Then, for any* $\delta\in(0,1)$*, with probability (w.r.t. S) at least* $1-\delta$ *we have the following inequality uniformly for all* $\mathbf{w}\in\mathcal{W}$

$$\big\|\nabla F_S(\mathbf{w})-\nabla F(\mathbf{w})\big\|_2=\widetilde{O}\Big(\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\big(L_S^{\frac{1}{2}}(\mathbf{w})+\beta G_\ell\mathfrak{R}_n(\mathcal{G})\log(1/\delta)\big)\Big).$$

Remark 2. A notable property is that the upper bound involves $L_S(\mathbf{w})$ when controlling $\big\|\nabla F_S(\mathbf{w})-\nabla F(\mathbf{w})\big\|_2$. Then, to control the generalization gap for the output model $A(S)$, we only need to estimate $L_S(A(S))=\frac{1}{n}\sum_{i=1}^n|\ell'(y_i,\Psi(A(S);\mathbf{x}_i))|^2$. Under Assumptions 1 and 2, we know $|\ell''(y,a)|\le\beta G_\ell$ for all $y\in\mathcal{Y},a\in\mathbb{R}$. By the self-bounding property of ℓ [36] (i.e., $|\ell'(y,a)|^2\le 2\beta G_\ell\ell(y,a)$ for all $y\in\mathcal{Y},a\in\mathbb{R}$) we know

$$L_S(A(S))\le\frac{2\beta G_\ell}{n}\sum_{i=1}^n\ell(y_i,\Psi(A(S);\mathbf{x}_i))=2\beta G_\ell F_S(A(S)).$$

We often apply A to approximately minimize F_S. Therefore, it is expected that $F_S(A(S))$ is often small in practice. In an interpolation setting, we have $F_S(A(S))=0$, then Theorem 3 implies that $\big\|\nabla F_S(A(S))-\nabla F(A(S))\big\|_2=\widetilde{O}\big(\beta G_\ell\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\mathfrak{R}_n(\mathcal{G})\big)$. This leads to a fast rate of order $\widetilde{O}_n(1/n)$ since Rademacher complexity is typically of order $O_n(1/\sqrt{n})$ [3]. It should be mentioned that the bounds in Theorem 3 do not show an explicit dependency on d. To our knowledge, this is the first optimistic and dimension-free bound of order $\widetilde{O}_n(1/n)$ for the uniform convergence of gradients.

Our analysis involves $\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})$. The following lemma shows that this Rademacher complexity can be reduced to other two Rademacher complexities which are more convenient to handle. The proof is given in Section 7.4.

Lemma 4. *Let $\widetilde{\mathcal{F}}$ be defined in Eq. (4.7). We have*

$$\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}}) \leq \max_{I\subseteq[2n]} \Big(\frac{|I|\mathfrak{R}_{S_{\mathbf{x}},|I|}(\widetilde{\mathcal{G}})}{2n} + \frac{2G_\Psi \beta |\bar{I}| \mathfrak{R}_{S_{\mathbf{x}},|\bar{I}|}(\mathcal{G})}{2n} \Big),$$

where $\bar{I}$ denotes the complement of I and

$$\widetilde{\mathcal{G}} = \Big\{ \mathbf{x} \mapsto \langle \nabla \Psi(\mathbf{w};\mathbf{x}), \mathbf{v} \rangle : (\mathbf{w},\mathbf{v}) \in \mathcal{W} \times \mathcal{V} \Big\}. \tag{4.10}$$

4.2 Comparisons on Uniform Convergence of Gradients

In this section, we compare our results with the literature. The discussion on the uniform convergence of gradients was initialized in the seminal paper [27]. Under some regularity assumption on gradient and Hessian, it was shown there

$$\sup_{\mathbf{w}\in\mathbb{R}^d:\|\mathbf{w}\|_2\leq 1} \big\| \nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w}) \big\|_2 \lesssim \widetilde{G} \Big(\frac{d \log n \log \widetilde{G}}{n} \Big)^{\frac{1}{2}}, \tag{4.11}$$

where $\widetilde{G}$ is a sub-Gaussian parameter. For convex f, the following useful uniform localized convergence of gradient was developed simultaneously for all $\mathbf{w} \in \{\mathbf{w}\in\mathbb{R}^d : \|\mathbf{w}\|_2 \leq 1\}$ [45]

$$\big\| \big(\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big) - \big(\nabla F_S(\mathbf{w}^*) - \nabla F(\mathbf{w}^*)\big) \big\|_2 \lesssim \frac{L_f d \|\mathbf{w}-\mathbf{w}^*\|_2 \log n}{n} + \Big(\frac{L_f d (F(\mathbf{w}) - F(\mathbf{w}^*)) \log n}{n} \Big)^{\frac{1}{2}}, \tag{4.12}$$

where L_f is the smoothness parameter of f and $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$. If $\frac{\nabla f(\mathbf{w};\mathbf{z}) - \nabla f(\mathbf{w}';\mathbf{z})}{\|\mathbf{w}-\mathbf{w}'\|_2}$ is $\widetilde{L}_f$-subexponential for any $\mathbf{w},\mathbf{w}'$, a similar result was also recently derived [43]

$$\big\| \big(\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big) - \big(\nabla F_S(\mathbf{w}^*) - \nabla F(\mathbf{w}^*)\big) \big\|_2 \lesssim \widetilde{L}_f \max\{\|\mathbf{w}-\mathbf{w}^*\|_2, 1/n\} \Big(\frac{d+\log n}{n} + \Big(\frac{d+\log n}{n} \Big)^{\frac{1}{2}} \Big). \tag{4.13}$$

We hide universal constant factors in the above $\lesssim$ notations. We now compare Theorem 3 with Eq. (4.11), (4.12) and Eq. (4.13) from two aspects: the dependency on d and the optimistic property.

1. Eq. (4.11), (4.12) and Eq. (4.13) enjoy a square-root dependency on the dimensionality d, and therefore are not appealing for high-dimensional problems. As a comparison, our optimistic bounds involve the Rademacher complexity, which imply dimension-free bounds if there is a norm control on the parameters (as we will see, this is the case for SNNs with smooth activations). Therefore, Theorem 3 is preferable for high-dimensional problems.

2. We now consider the optimistic property, e.g., whether these bounds can imply fast rates under a low-noise condition. The upper bound in Eq. (4.11) does not involve either $F_S(\cdot)$ or $L_S(\cdot)$, and therefore is not optimistic. Both Eq. (4.12) and Eq. (4.13) imply fast rates only if $\mathbb{E}_{\mathbf{z}}[\|\nabla f(\mathbf{w}^*;\mathbf{z})\|_2^2] = o_n(1)$ (they still need to control $\|\nabla F_S(\mathbf{w}^*) - \nabla F(\mathbf{w}^*)\|_2$ by the Bernstein inequality) [43, 43, 45, 45] or $\|\mathbf{w}-\mathbf{w}^*\|_2 = o_n(1)$, where $o_n(\cdot)$ is a small-o notation focusing only the dependency on n. That is, the parameter is close enough to the best parameter, which is a bit restrictive for nonconvex problems. Generally, an algorithm cannot guarantee the convergence in parameters for nonconvex problems [8, 11, 14, 21]. Indeed, these discussions [43, 45] used $T(\mathbf{w}) = \|\mathbf{w}-\mathbf{w}^*\|_2$ as the localization function to develop a localized upper bound on the concentration of gradients. As a comparison, we use $L(\mathbf{w})$ as the localization function, and our analysis implies fast rates if $L(\mathbf{w}) = o_n(1)$, which is milder than the condition $\|\mathbf{w}-\mathbf{w}^*\|_2 = o_n(1)$ and $\mathbb{E}_{\mathbf{z}}[\|\nabla f(\mathbf{w}^*;\mathbf{z})\|_2^2] = o_n(1)$ [43, 45].

Dimension-free uniform convergence of gradients was also derived in [12]. They developed a chain rule for Rademacher complexity, which, for loss functions of the form $f(\mathbf{w};\mathbf{z}) = \ell(y, \Psi(\mathbf{w};\mathbf{x}))$, reduces to

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{w}\in\mathcal{W}} \Big\| \sum_{i=1}^n \sigma_i \nabla f(\mathbf{w};\mathbf{z}_i) \Big\|_2 \leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}} \Big[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^n \sigma_i \ell'(y_i, \Psi(\mathbf{w};\mathbf{x}_i)) \Big] + \frac{G_\ell}{n} \mathbb{E}_{\boldsymbol{\sigma}} \Big\| \sum_{i=1}^n \sigma_i \mathbf{x}_i \Big\|_2.$$

If we assume ℓ is L_ℓ-smooth, the Talagrand's contraction principle further implies

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{w}\in\mathcal{W}} \Big\| \sum_{i=1}^n \sigma_i \nabla f(\mathbf{w};\mathbf{z}_i)\Big\|_2 \leq \frac{L_\ell}{n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^n \sigma_i \Psi(\mathbf{w};\mathbf{x}_i)\Big] + \frac{G_\ell}{n}\Big(\sum_{i=1}^n \|\mathbf{x}_i\|_2^2\Big)^{\frac{1}{2}}. \tag{4.14}$$

Eq. (4.14) also does not directly involve the dimensionality in the upper bound, and can be dimension-free if there is a norm control. However, Eq. (4.14) only implies a slow rate of order $O_n(1/\sqrt{n})$ for the uniform convergence. As a comparison, we use the localization technique and derive optimistic bounds that can interpolate between a slow rate $\widetilde{O}_n(1/\sqrt{n})$ and a fast rate $\widetilde{O}_n(1/n)$, depending on the difficulty of the problem at hand.

In summary, existing studies either suffer from a square-root dependency on the dimensionality or only imply a slow rate of order $O_n(1/\sqrt{n})$. We improve the existing analysis by developing a uniform convergence of gradients that is both dimension-free and optimistic. A novelty of our analysis is to use $L_S(\cdot)$ as the localization function, which removes the requirement of $\|\mathbf{w}-\mathbf{w}^*\|_2 = o_n(1)$ for optimistic rates based on the localization function $T(\mathbf{w}) = \|\mathbf{w}-\mathbf{w}^*\|_2$ [43]. It also outperforms the localization analysis based on $F_S(\cdot)$ [36] since $L_S(\cdot) \lesssim \beta F_S(\cdot)$ by the self-bounding property. We summarize the comparisons in Table 1.

Table 1: Comparison of uniform convergence of gradients. Here, DG means deviation of gradients $\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\|_2$, while DEG means deviation of excess gradients $\|(\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})) - (\nabla F_S(\mathbf{w}^*) - \nabla F(\mathbf{w}^*))\|_2$. Notations: d means dimensionality and n is the sample size. The works [12, 27] consider global complexity analyses, while the other works consider localized complexity analyses. We ignore logarithmic factors in the table.

Reference	Type	Assumption	Bound
[27]	DG	$\widetilde{G}$-Subgaussian gradient, Lipschitz Hessian	$\widetilde{G}\big(\frac{d}{n}\big)^{\frac{1}{2}}$
[45]	DEG	Convex and L_f-smooth f	$\frac{L_f d\|\mathbf{w}-\mathbf{w}^*\|_2}{n} + \big(\frac{L_f d(F(\mathbf{w})-F(\mathbf{w}^*))}{n}\big)^{\frac{1}{2}}$
[43]	DEG	$\widetilde{L}_f$-subexponential $\frac{\nabla f(\mathbf{w};\mathbf{z})-\nabla f(\mathbf{w}';\mathbf{z})}{\|\mathbf{w}-\mathbf{w}'\|}$	$\widetilde{L}_f \max\{\|\mathbf{w}-\mathbf{w}^*\|_2, 1/n\}\big(\frac{d}{n} + \big(\frac{d}{n}\big)^{\frac{1}{2}}\big)$
[12]	DG	L_ℓ-Smooth and G_ℓ-Lipschitz ℓ	$L_\ell \mathfrak{R}_n(\mathcal{G}) + \frac{G_\ell}{\sqrt{n}}$
Ours	DG	β-self-bounding and G_ℓ-Lipschitz ℓ	$\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\big(L_S^{\frac{1}{2}}(\mathbf{w}) + \beta G_\ell \mathfrak{R}_n(\mathcal{G})\big)$

5 Idea and Novelty of the Proof

Our optimistic analysis is motivated by the discussions in [36], where optimistic bounds were developed for $F(\mathbf{w}) - F_S(\mathbf{w})$ uniformly for all $\mathbf{w}\in\mathcal{W}$. We extend their discussions to the convergence of gradients. This extension requires several new techniques. First, we need to control the $d_{2,S}$ distance between $g_{\mathbf{w},\mathbf{v}}$ and $g_{\mathbf{w}',\mathbf{v}'}$

$$\frac{1}{n}\sum_{i=1}^n \big(g_{\mathbf{w},\mathbf{v}}(\mathbf{z}_i) - g_{\mathbf{w}',\mathbf{v}'}(\mathbf{z}_i)\big)^2 = \frac{1}{n}\sum_{i=1}^n \Big(\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w};\mathbf{x}_i),\mathbf{v}\rangle - \ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\Big)^2,$$

where, for any $\mathbf{w},\mathbf{v}$, we define $g_{\mathbf{w},\mathbf{v}} : \mathcal{Z}\mapsto\mathbb{R}$ by

$$g_{\mathbf{w},\mathbf{v}}(\mathbf{z}) := \langle\nabla f(\mathbf{w};\mathbf{z}),\mathbf{v}\rangle = \ell'(y,\Psi(\mathbf{w};\mathbf{x}))\langle\nabla\Psi(\mathbf{w};\mathbf{x}),\mathbf{v}\rangle. \tag{5.1}$$

Unlike the case with loss in [36], the case with gradients yields a product of two terms: $\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))$ and $\langle\nabla\Psi(\mathbf{w};\mathbf{x}_i),\mathbf{v}\rangle$. We introduce a decomposition to handle this by two terms, for each of which we use the localization constraint to get bounds depending on the square root of the localization parameter. Second, we introduce the function $\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x},k)$ to conduct the covering number analysis w.r.t. the

ℓ_∞-norm. As we will see in the proof, the d_∞ metric for $\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x},k)$ dominates the d_∞ metric for two functions introduced in our decomposition. Below, we sketch our basic idea in the analysis.

Our basic idea to study the gradient Rademacher complexity is to relate it to the Rademacher complexity of a real-valued function class

$$\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\widehat{\mathcal{W}}_r}\Big\|\frac{1}{n}\sum_{i=1}^n\sigma_i\nabla f(\mathbf{w};\mathbf{z}_i)\Big\|_2\Big]=\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{(\mathbf{w},\mathbf{v})\in\widehat{\mathcal{W}}_r\times\mathcal{V}}\Big\langle\frac{1}{n}\sum_{i=1}^n\sigma_i\nabla f(\mathbf{w};\mathbf{z}_i),\mathbf{v}\Big\rangle\Big]=\mathfrak{R}_S(\widehat{\mathcal{G}}_r),\tag{5.2}$$

where we introduce

$$\widehat{\mathcal{G}}_r=\Big\{\mathbf{z}\mapsto\langle\nabla f(\mathbf{w};\mathbf{z}),\mathbf{v}\rangle:(\mathbf{w},\mathbf{v})\in\widehat{\mathcal{W}}_r\times\mathcal{V}\Big\}.\tag{5.3}$$

It suffices to estimate the Rademacher complexity of $\widehat{\mathcal{G}}_r$, which, according to Dudley's entropy integral, further reduces to the covering number estimation of $\widehat{\mathcal{G}}_r$, towards which a key step is as follows (see Eq. (7.7))

$$\frac{1}{n}\sum_{i=1}^n\big(g_{\mathbf{w},\mathbf{v}}(\mathbf{z}_i)-g_{\mathbf{w}',\mathbf{v}'}(\mathbf{z}_i)\big)^2\leq 2r\max_{i\in[n]}\Big(\langle\nabla\Psi(\mathbf{w};\mathbf{x}_i),\mathbf{v}\rangle-\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\Big)^2+4G_\Psi^2\beta^2r\max_{i\in[n]}\big(\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i)\big)^2.$$

A notable property is that the right-hand side is a linear function of r, which allows us to conduct localization analysis to get improved bounds. The above inequality shows that

$$d_{2,S}(g_{\mathbf{w},\mathbf{v}},g_{\mathbf{w}',\mathbf{v}'})\leq\sqrt{2r}d_{\infty,S_{\mathbf{x}}}\big(\langle\nabla\Psi(\mathbf{w};\cdot),\mathbf{v}\rangle,\langle\nabla\Psi(\mathbf{w}';\cdot),\mathbf{v}'\rangle\big)+2G_\Psi\beta\sqrt{r}d_{\infty,S_{\mathbf{x}}}\big(\Psi(\mathbf{w};\cdot),\Psi(\mathbf{w}';\cdot)\big).$$

Our motivation in introducing $\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x},k)$ is based on the observation

$$\sqrt{2}d_{\infty,S_{\mathbf{x}}}\big(\langle\nabla\Psi(\mathbf{w};\cdot),\mathbf{v}\rangle,\langle\nabla\Psi(\mathbf{w}';\cdot),\mathbf{v}'\rangle\big)+2G_\Psi\beta d_{\infty,S_{\mathbf{x}}}\big(\Psi(\mathbf{w};\cdot),\Psi(\mathbf{w}';\cdot)\big)\leq(\sqrt{2}+1)d_{\infty,\widetilde{X}}\big(\tilde{f}_{\mathbf{w},\mathbf{v}},\tilde{f}_{\mathbf{w}',\mathbf{v}'}\big).$$

Indeed, it is clear that

$$\max_{i\in[n]}\big|\langle\nabla\Psi(\mathbf{w};\mathbf{x}_i),\mathbf{v}\rangle-\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\big|=\max_{i\in[n]}\big|\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x}_i,1)-\tilde{f}_{\mathbf{w}',\mathbf{v}'}(\mathbf{x}_i,1)\big|\leq\max_{i\in[n],k\in[2]}\big|\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x}_i,k)-\tilde{f}_{\mathbf{w}',\mathbf{v}'}(\mathbf{x}_i,k)\big|$$

and

$$2G_\Psi\beta\max_{i\in[n]}\big|\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i)\big|=\max_{i\in[n]}\big|\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x}_i,2)-\tilde{f}_{\mathbf{w}',\mathbf{v}'}(\mathbf{x}_i,2)\big|\leq\max_{i\in[n],k\in[2]}\big|\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x}_i,k)-\tilde{f}_{\mathbf{w}',\mathbf{v}'}(\mathbf{x}_i,k)\big|.$$

The above analysis leads to the following structural lemma on covering numbers, which relates the covering number of $\widehat{\mathcal{G}}_r$ w.r.t. the metric $d_{2,S}$ to that of $\widetilde{\mathcal{F}}$ w.r.t. the metric $d_{\infty,\widetilde{X}}$. The proof is given in Section 7.2.

Lemma 5. *Let $\widehat{\mathcal{G}}_r$ be defined in Eq. (5.3) and $\widetilde{\mathcal{F}}$ be defined in Eq. (4.7). Let Assumption 1 hold and $\|\nabla\Psi(\mathbf{w},\mathbf{x})\|_2\leq G_\Psi$ for all $\mathbf{w}\in\mathcal{W}$ and $\mathbf{x}\in\mathcal{X}$. Then*

$$\mathcal{N}_2\big(\epsilon,\widehat{\mathcal{G}}_r,S\big)\leq\mathcal{N}_\infty(\epsilon/\sqrt{3r},\widetilde{\mathcal{F}},\widetilde{X}).\tag{5.4}$$

By the standard argument [36], one can control the covering number of $\widetilde{\mathcal{F}}$ by the worst-case Rademacher complexity of $\widetilde{\mathcal{F}}$ (see Lemma 14)

$$\log\mathcal{N}_2\big(\sqrt{3r}\epsilon,\widehat{\mathcal{G}}_r,S\big)\leq\log\mathcal{N}_\infty(\epsilon,\widetilde{\mathcal{F}},\widetilde{X})\leq 1+\frac{64n\mathfrak{R}^2_{\widetilde{X},n}(\widetilde{\mathcal{F}})}{\epsilon^2}\Big(\log^2\frac{8eB^2n}{\epsilon^2}\Big).$$

Then, we apply an improved Dudley's entropy integral to relate the above covering number bounds to Rademacher complexity bounds of $\widehat{\mathcal{G}}_r$. This leads to Theorem 2 by Eq. (5.2).

Now we plug this Rademacher complexity bound into Theorem 1 and get the following bound (see Lemma 17)

$$\sup_{\mathbf{w}\in\mathcal{W}_r}\big\|\nabla F_S(\mathbf{w})-\nabla F(\mathbf{w})\big\|_2\lesssim\Big(\sqrt{r}+\beta G_\ell\mathfrak{R}_n(\mathcal{G})\log^2(G_\ell b\beta n)\Big)\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn)\log^{\frac{1}{2}}(1/\delta)+\frac{G_\ell G_\Psi\log(1/\delta)}{n}.$$

The right-hand side involves the radius parameter r instead of an individual $\mathbf{w}$. Finally, we apply the standard peeling argument to transform this bound to Theorem 3 [4, 43], where we replace r by $L_S(\mathbf{w})$ that can be computed from the data.

Our analysis uses two localization schemes: one based on L and one based on L_S. Each of these two localization schemes has its own advantage. The localization based on L leads to a data-independent hypothesis space, which is convenient in applying the symmetry argument to relate the uniform convergence to the Rademacher complexity [4]. The localization based on L_S is convenient to establish a structural result on covering numbers since the definition of covering number involves the metric defined on the training dataset (i.e., $d_{p,S}$). The following lemma shows that these two localization schemes are essentially the same. One can transform one localization scheme to the other, at the cost of introducing some terms of order $\widetilde{O}_n(1/n)$. The proof of Lemma 6 is given in Section 7.3.

Lemma 6. *Let Assumption 1 hold. Assume $|\ell'(y,a)| \le G_\ell$ for all $a \in \mathbb{R}, y \in \mathcal{Y}$ and some $G_\ell \ge 0$. Let $\delta \in (0,1)$ and $\mathfrak{R}_n(\mathcal{G})\beta \lesssim 1$. Then, with probability (w.r.t. S) at least $1 - 2\exp(-x)$ the following inequality holds simultaneously for all $\mathbf{w} \in \mathcal{W}$*

$$\max\big\{L_S(\mathbf{w}) - 2L(\mathbf{w}), L(\mathbf{w}) - 2L_S(\mathbf{w})\big\} \lesssim \beta^2 G_\ell^2 \mathfrak{R}_n^2(\mathcal{G}) \log^4(G_\ell b \beta n) + \frac{G_\ell^2(x + \log\log n)}{n}.$$

6 Applications to Shallow Neural Networks

In this section, we apply our uniform convergence to study the generalization behavior of SNNs. Assume $\mathcal{X} \subseteq \mathbb{R}^{\tilde{d}}$ for some $\tilde{d} \in \mathbb{N}$. In this case, the prediction function Ψ takes the form: $\Psi(\mathbf{w};\mathbf{x}) = \sum_{j=1}^m \mu_j \gamma(\mathbf{x}^\top \mathbf{w}^{(j)})$, where m is the number of neurons in the hidden layer, μ_j is the connection weight between the output and the j-th hidden neuron, $\mathbf{w}^{(j)} \in \mathbb{R}^{\tilde{d}}$ denotes the weight between the input and the j-th hidden neuron and $\gamma : \mathbb{R} \mapsto \mathbb{R}$ is an activation function. We collect all the weights into a matrix $\mathbf{w} = (\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}) \in \mathbb{R}^{\tilde{d} \times m}$. If $m = 1$, then the SNN becomes a generalized linear model [27, 43]. Therefore, our analysis here also directly implies optimistic bounds for generalized linear models. We fix $\mu_j \in \{\frac{1}{\sqrt{m}}, -\frac{1}{\sqrt{m}}\}$, and only optimize $\mathbf{w}$, which is standard in the theoretical analysis of SNNs [2, 15, 17, 35, 37]. It is clear that

$$\nabla\Psi(\mathbf{w};\mathbf{x}) = \begin{pmatrix} \mu_1 \mathbf{x} \gamma'(\mathbf{x}^\top \mathbf{w}^{(1)}) \\ \vdots \\ \mu_m \mathbf{x} \gamma'(\mathbf{x}^\top \mathbf{w}^{(m)}) \end{pmatrix}. \tag{6.1}$$

Assume that $|\gamma'(a)| \le G_\gamma$ for all a, which is satisfied for popular activation functions such as the sigmoid function and the hyperbolic tangent activation function. It is then clear that $\|\nabla\Psi(\mathbf{w};\mathbf{x})\|_2^2 = \sum_{j=1}^m \mu_j^2 \|\mathbf{x}\|_2^2 \big(\gamma'(\mathbf{x}^\top \mathbf{w}^{(j)})\big)^2 \le \|\mathbf{x}\|_2^2 G_\gamma^2$. This shows that if we assume $|\ell'(y,a)| \le G_\ell$ and $|\gamma'(a)| \le G_\gamma$, then Assumption 2 holds with $G_\Psi = \sup_{\mathbf{x} \in S_\mathbf{x}} \|\mathbf{x}\|_2 G_\gamma$. Let $\mathbf{w}_0 \in \mathbb{R}^{\tilde{d} \times m}$ be any point independent of S, e.g., the initialization of gradient methods, and consider the initialization-dependent space $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^{\tilde{d} \times m} : \|\mathbf{w} - \mathbf{w}_0\|_F \le R\}$ [24, 26], where $\|\cdot\|_F$ denotes the Frobenius norm. We divide our discussions into two cases: SNNs with smooth activations and SNNs with the ReLU activation function.

6.1 Smooth Activations

We first present the uniform convergence of gradients for SNNs with smooth activation functions. The following lemma provides Rademacher complexity bounds with a logarithmic dependency on the width for $\mathcal{G}$ and $\widetilde{\mathcal{G}}$ with initialization-dependent constraints. The proofs are given in Section C.

Lemma 7. *Let $\mathcal{G}$ be defined in Eq. (4.9) and $\widetilde{\mathcal{G}}$ be defined in Eq. (4.10). Let $\gamma(\cdot)$ be G_γ-Lipschitz and L_γ-smooth. Suppose $\sup_{\mathbf{x} \in S_\mathbf{x}} \|\mathbf{x}\|_2 \le R_\mathbf{x}$ for some $R_\mathbf{x} > 0$ and $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^{\tilde{d} \times m} : \|\mathbf{w} - \mathbf{w}_0\|_F \le R\}$*

for some $R > 0$. Then, we have

$$\mathfrak{R}_S(\mathcal{G}) = \widetilde{O}\Big(\frac{G_\gamma R R_{\mathbf{x}}}{\sqrt{n}}\Big), \quad \mathfrak{R}_S(\widetilde{\mathcal{G}}) \leq \frac{(\sqrt{\pi}+1)G_\gamma + \sqrt{\pi}L_\gamma R_{\mathbf{x}} R}{n}\Big(\sum_{i=1}^n \|\mathbf{x}_i\|_2^2\Big)^{\frac{1}{2}}.$$

We plug Rademacher complexity bounds in Lemma 7 back into Theorem 3, and derive the following theorem on uniform convergence for SNNs with smooth activations. The proof is given in Section C.

Theorem 8. *Let Assumptions 1 hold. Let $\gamma(\cdot)$ be G_γ-Lipschitz and L_γ-smooth. Suppose $\sup_{\mathbf{x}\in S_{\mathbf{x}}}\|\mathbf{x}\|_2 \leq R_{\mathbf{x}}$ for some $R_{\mathbf{x}} > 0$ and $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^{\tilde{d}\times m} : \|\mathbf{w}-\mathbf{w}_0\|_F \leq R\}$ for some $R > 0$. Then for any $\delta \in (0,1)$, with probability (w.r.t. S) at least $1-\delta$ the following inequality holds simultaneously for all $\mathbf{w} \in \mathcal{W}$*

$$\big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 = \widetilde{O}\Big(\frac{G_\gamma + L_\gamma R_{\mathbf{x}} R + G_\Psi \beta G_\gamma R}{\sqrt{n}}\Big(L_S^{\frac{1}{2}}(\mathbf{w}) + \frac{\beta G_\ell G_\gamma R R_{\mathbf{x}} \log(1/\delta)}{\sqrt{n}}\Big)R_{\mathbf{x}}\Big).$$

If we only focus on the dependency on m and n, then Theorem 8 shows $\sup_{\mathbf{w}\in\mathcal{W}}\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 = \widetilde{O}_{m,n}(\frac{1}{\sqrt{n}}(L_S^{\frac{1}{2}}(\mathbf{w}) + \frac{1}{\sqrt{n}}))$. If $L_S(\mathbf{w}) = \widetilde{O}_n(1/n)$, this further implies a fast uniform convergence $\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 = \widetilde{O}_{m,n}(1/n)$.

6.2 ReLU Activation

We now turn to the ReLU activation function $\gamma(t) = \max\{0, t\}$. Since $\gamma(\cdot)$ is not differentiable, we consider the convergence of subgradients and choose $\gamma'(t) = \mathbb{I}_{[t\geq 0]}$.

Lemma 9. *Let $\mathcal{G}$ be defined in Eq. (4.9) and $\widetilde{\mathcal{G}}$ be defined in Eq. (4.10) for SNNs with the ReLU activation. Suppose $\sup_{\mathbf{x}\in S_{\mathbf{x}}}\|\mathbf{x}\|_2 \leq R_{\mathbf{x}}, \mathcal{W} = \{\mathbf{w} \in \mathbb{R}^{\tilde{d}\times m} : \|\mathbf{w}-\mathbf{w}_0\|_F \leq R\}$ for some $R > 0$. Then $\mathfrak{R}_S(\mathcal{G}) = \widetilde{O}\big(\frac{RR_{\mathbf{x}}}{\sqrt{n}}\big)$ and $\mathfrak{R}_S(\widetilde{\mathcal{G}}) \lesssim \frac{R_{\mathbf{x}}\sqrt{\tilde{d}}}{\sqrt{n}}$.*

We plug the above complexity bounds back to Lemma 4, and get $\mathfrak{R}_{2n}(\widetilde{\mathcal{F}}) = \widetilde{O}\big(\frac{R_{\mathbf{x}}\sqrt{\tilde{d}}}{\sqrt{n}} + \frac{G_\Psi R R_{\mathbf{x}}\beta}{\sqrt{n}}\big)$. Based on these Rademacher complexity bounds, we immediately derive the following convergence of gradients. We omit the detailed analysis due to its similarity to the proof of Theorem 8.

Theorem 10. *Let Assumption 1 hold. Let $\gamma(\cdot)$ be the ReLU activation. Suppose $\sup_{\mathbf{x}\in S_{\mathbf{x}}}\|\mathbf{x}\|_2 \leq R_{\mathbf{x}}$ for some $R_{\mathbf{x}} > 0$. Then for any $\delta \in (0,1)$, with probability (w.r.t. S) at least $1-\delta$ the following inequality holds simultaneously for all $\mathbf{w} \in \mathcal{W} = \{\mathbf{w} \in \mathbb{R}^{\tilde{d}\times m} : \|\mathbf{w}-\mathbf{w}_0\|_F \leq R\}$*

$$\big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 = \widetilde{O}\Big(\Big(\frac{R_{\mathbf{x}}\sqrt{\tilde{d}}}{\sqrt{n}} + \frac{G_\Psi R R_{\mathbf{x}}\beta}{\sqrt{n}}\Big)\Big(L_S^{\frac{1}{2}}(\mathbf{w}) + \frac{\beta G_\ell R R_{\mathbf{x}}}{\sqrt{n}}\Big)\Big).$$

Our localization analysis still implies rates of order $\widetilde{O}_{m,n}(1/n)$ if $L_S(\mathbf{w}) = \widetilde{O}_n(1/n)$. Note Corollary 10 also enjoys a square-root dependency on $\tilde{d}$ for *nonsmooth* activation functions. Generally, this square-root dependency on $\tilde{d}$ is not avoidable. Indeed, the work [12] considered $f(\tilde{\mathbf{w}}; \mathbf{z}) = \max\{-y\tilde{\mathbf{w}}^\top\mathbf{x}, 0\}$ for $\mathbf{x} \in \mathbb{R}^{\tilde{d}}$ with $\|\mathbf{x}\|_2 \leq 1$ and showed

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\tilde{\mathbf{w}}\in\mathbb{R}^{\tilde{d}}:\|\tilde{\mathbf{w}}\|_2\leq 1} \big\|\sum_{i=1}^n \sigma_i \nabla f(\tilde{\mathbf{w}}; \mathbf{z}_i)\big\|_2 \gtrsim \min\big\{\big(\tilde{d}/n\big)^{\frac{1}{2}}, 1\big\}.$$

This shows that the uniform convergence of gradients is more challenging than the uniform convergence of functions as the latter can be dimension-independent under a norm constraint even for nonsmooth activation functions [16]. Our analysis in Theorem 10 captures this square-root dependency on $\tilde{d}$ for nonsmooth activation functions. Note that this bound has a logarithmic-dependency on m, while the existing uniform-deviation analyses [27, 43] imply at least a square-root dependency on m when applied to ReLU networks.

Remark 3 (Comparison)**.** Most existing uniform convergence of neural networks focus on the convergence of function values [2, 5, 16, 22]. An existing uniform convergence of gradients was studied in [48], which was used to study the convergence of gradient descent in learning a teacher SNN. Specifically, the work [48] showed that with high probability

$$\sup_{\mathbf{w}\in\mathbb{R}^{\tilde{d}\times m}:\|\mathbf{w}-\mathbf{w}^*\|_2\leq c^*/m^2}\|\nabla\hat{F}(\mathbf{w})-\nabla\hat{F}_S(\mathbf{w})\|_2\lesssim m^{\frac{3}{2}}\|\mathbf{w}^*\|_2\Big(\frac{\tilde{d}\log n}{n}\Big)^{\frac{1}{2}},\tag{6.2}$$

where $\mathbf{w}^*\in\mathbb{R}^{\tilde{d}\times m}$ is a teacher model, c^* is a constant depending on $\mathbf{w}^*$ and

$$\hat{F}(\mathbf{w})=\mathbb{E}_{\mathbf{x}}\Big[\Big(\sum_{j=1}^m\mu_j\big(\gamma(\mathbf{x}^\top\mathbf{w}^{(j)})-\gamma(\mathbf{x}^\top(\mathbf{w}^*)^{(j)})\big)\Big)^2\Big],\ \hat{F}_S(\mathbf{w})=\frac{1}{n}\sum_{i=1}^n\Big(\sum_{j=1}^m\mu_j\big(\gamma(\mathbf{x}_i^\top\mathbf{w}^{(j)})-\gamma(\mathbf{x}_i^\top(\mathbf{w}^*)^{(j)})\big)\Big)^2.$$

Note this corresponds to training SNNs with the least-square loss and $y_i=\sum_{j=1}^m\mu_j\gamma(\mathbf{x}_i^\top(\mathbf{w}^*)^{(j)})$. The bound in Eq. (6.2) has a polynomial dependency on the width m and cannot imply optimistic bounds. As a comparison, our bound in Theorem 10 enjoys a logarithmic dependency on m and can imply fast rates in an interpolation setting. It should be mentioned that the bound in Eq. (6.2) considers the least square loss, while Theorem 10 considers self-bounding and Lipschtiz loss functions.

7 Proofs on Optimistic Bounds

7.1 Proof of Theorem 1

In this section, we prove Theorem 1 by local Rademacher complexity analysis. To this aim, we introduce the following uniform convergence of functions established in [4]. The notation $\mathrm{Var}(\cdot)$ denotes the variance.

Lemma 11 ([4])**.** *Let $\mathcal{F}'$ be a class of functions that map $\mathcal{Z}'$ into $[a,b]$. Let $S':=\{\mathbf{z}_1',\ldots,\mathbf{z}_n'\}\subseteq\mathcal{Z}'$ be independently drawn from the same distribution. Assume there exists an $r'>0$ such that for every $f\in\mathcal{F}'$, $\mathrm{Var}[f(\mathbf{z}_i)]\leq r'$. Then, for every $x>0$, with probability at least $1-2\exp(-x)$ we have*

$$\sup_{f\in\mathcal{F}'}\big(\mathbb{E}_{\mathbf{z}'}[f(\mathbf{z}')]-\frac{1}{n}\sum_{i=1}^n f(\mathbf{z}_i')\big)\leq\inf_{\alpha\in(0,1)}\Big(\frac{2(1+\alpha)}{1-\alpha}\mathfrak{R}_{S'}(\mathcal{F}')+\Big(\frac{2r'x}{n}\Big)^{\frac{1}{2}}+\frac{(b-a)x}{n}\Big(\frac{1}{3}+\frac{1}{\alpha}+\frac{1+\alpha}{2\alpha(1-\alpha)}\Big)\Big).$$

Moreover, the same result holds for the quantity $\sup_{f\in\mathcal{F}'}\big(\frac{1}{n}\sum_{i=1}^n f(\mathbf{z}_i')-\mathbb{E}_{\mathbf{z}'}[f(\mathbf{z}')]\big)$.

For any $r>0$, define

$$\mathcal{G}_r=\Big\{\mathbf{z}\mapsto\langle\nabla f(\mathbf{w};\mathbf{z}),\mathbf{v}\rangle:(\mathbf{w},\mathbf{v})\in\mathcal{W}_r\times\mathcal{V}\Big\}.\tag{7.1}$$

Proof of Theorem 1. It is clear that

$$\nabla f(\mathbf{w};\mathbf{z})=\ell'(y,\Psi(\mathbf{w};\mathbf{x}))\nabla\Psi(\mathbf{w};\mathbf{x}).\tag{7.2}$$

Then, for any $(\mathbf{w},\mathbf{v})\in\mathcal{W}_r\times\mathcal{V}$, we have

$$|\langle\nabla f(\mathbf{w};\mathbf{z}),\mathbf{v}\rangle|\leq\|\nabla f(\mathbf{w};\mathbf{z})\|_2=\|\ell'(y,\Psi(\mathbf{w};\mathbf{x}))\nabla\Psi(\mathbf{w};\mathbf{x})\|_2\leq G_\ell G_\Psi.$$

Furthermore, for any $(\mathbf{w},\mathbf{v})\in\mathcal{W}_r\times\mathcal{V}$, we have

$$\mathrm{Var}_{\mathbf{z}}[\langle\nabla f(\mathbf{w};\mathbf{z}),\mathbf{v}\rangle]\leq\mathbb{E}_{\mathbf{z}}[\langle\nabla f(\mathbf{w};\mathbf{z}),\mathbf{v}\rangle^2]\leq\mathbb{E}_{\mathbf{z}}\Big[\|\ell'(y,\Psi(\mathbf{w};\mathbf{x}))\nabla\Psi(\mathbf{w};\mathbf{x})\|_2^2\Big]\leq G_\Psi^2\mathbb{E}_{\mathbf{z}}\big[|\ell'(y,\Psi(\mathbf{w};\mathbf{x}))|^2\big]\leq G_\Psi^2 r.$$

We apply Lemma 11 with $\mathcal{F}'=\mathcal{G}_r,\alpha=1/2,r'=G_\Psi^2 r,S'=S$ and derive the following inequality with probability at least $1-2\exp(-x)$

$$\sup_{(\mathbf{w},\mathbf{v})\in(\mathcal{W}_r\times\mathcal{V})}\Big[\mathbb{E}_{\mathbf{z}}[\langle\nabla f(\mathbf{w};\mathbf{z}),\mathbf{v}\rangle]-\frac{1}{n}\sum_{i=1}^n\langle\nabla f(\mathbf{w};\mathbf{z}_i),\mathbf{v}\rangle\Big]\leq 6\mathfrak{R}_S\big(\mathcal{G}_r\big)+\Big(\frac{2G_\Psi^2 rx}{n}\Big)^{\frac{1}{2}}+\frac{32G_\ell G_\Psi x}{3n}.$$

It is also clear that

$$\sup_{(\mathbf{w},\mathbf{v})\in(\mathcal{W}_r\times\mathcal{V})}\Big[\mathbb{E}_{\mathbf{z}}[\langle\nabla f(\mathbf{w};\mathbf{z}),\mathbf{v}\rangle]-\frac{1}{n}\sum_{i=1}^n\langle\nabla f(\mathbf{w};\mathbf{z}_i),\mathbf{v}\rangle\Big]$$
$$=\sup_{(\mathbf{w},\mathbf{v})\in\mathcal{W}_r\times\mathcal{V}}\Big[\Big\langle\mathbb{E}_{\mathbf{z}}[\nabla f(\mathbf{w};\mathbf{z})]-\frac{1}{n}\sum_{i=1}^n\nabla f(\mathbf{w};\mathbf{z}_i),\mathbf{v}\Big\rangle\Big]=\sup_{\mathbf{w}\in\mathcal{W}_r}\Big[\Big\|\mathbb{E}_{\mathbf{z}}[\nabla f(\mathbf{w};\mathbf{z})]-\frac{1}{n}\sum_{i=1}^n\nabla f(\mathbf{w};\mathbf{z}_i)\Big\|_2\Big],$$

where we can interchange $\mathbb{E}_{\mathbf{z}}[\cdot]$ and the dot product since $\|\nabla f(\mathbf{w};\mathbf{z})\|_2\leq G_\ell G_\Psi$ and $\|\mathbf{v}\|_2\leq 1$ for any $\mathbf{v}\in\mathcal{V}$. That is, with probability at least $1-2\exp(-x)$ we have

$$\sup_{\mathbf{w}\in\mathcal{W}_r}\big\|\nabla F_S(\mathbf{w})-\nabla F(\mathbf{w})\big\|_2\leq 6\mathfrak{R}_S\big(\mathcal{G}_r\big)+G_\Psi\Big(\frac{2rx}{n}\Big)^{\frac{1}{2}}+\frac{32G_\ell G_\Psi x}{3n}.$$

Furthermore, we know

$$\mathfrak{R}_S\big(\mathcal{G}_r\big)=\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{(\mathbf{w},\mathbf{v})\in\mathcal{W}_r\times\mathcal{V}}\Big\langle\frac{1}{n}\sum_{i=1}^n\sigma_i\nabla f(\mathbf{w};\mathbf{z}_i),\mathbf{v}\Big\rangle\Big]=\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_r}\Big\|\frac{1}{n}\sum_{i=1}^n\sigma_i\nabla f(\mathbf{w};\mathbf{z}_i)\Big\|_2\Big].$$

The proof is complete by combining the above two inequalities together (taking $x=\log 2/\delta$). □

7.2 Gradient Rademacher Complexity Bounds

In this section, we present the proof of Theorem 2 on gradient Rademacher complexity bounds. As illustrated in Section 5, a key step is to establish a structural result on covering numbers in Lemma 5. Recall that $g_{\mathbf{w},\mathbf{v}}$ is defined in Eq. (5.1).

Proof of Lemma 5. By the standard inequality $(a+b)^2\leq 2a^2+2b^2$ and Eq. (7.2), we get the following inequality for any $(\mathbf{w},\mathbf{v}),(\mathbf{w}',\mathbf{v}')\in\widehat{\mathcal{W}}_r\times\mathcal{V}$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\big(g_{\mathbf{w},\mathbf{v}}(\mathbf{z}_i)-g_{\mathbf{w}',\mathbf{v}'}(\mathbf{z}_i)\big)^2=\frac{1}{n}\sum_{i=1}^n\Big(\langle\nabla f(\mathbf{w};\mathbf{z}_i),\mathbf{v}\rangle-\langle\nabla f(\mathbf{w}';\mathbf{z}_i),\mathbf{v}'\rangle\Big)^2\\
&=\frac{1}{n}\sum_{i=1}^n\Big(\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w};\mathbf{x}_i),\mathbf{v}\rangle-\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\Big)^2\\
&\leq\frac{2}{n}\sum_{i=1}^n\Big(\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w};\mathbf{x}_i),\mathbf{v}\rangle-\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\Big)^2\\
&+\frac{2}{n}\sum_{i=1}^n\Big(\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle-\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\Big)^2.
\end{aligned}\tag{7.3}$$

For the first term on the right-hand side of the above inequality, we know

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^n\Big(\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w};\mathbf{x}_i),\mathbf{v}\rangle-\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\Big)^2\\
&\leq\frac{2}{n}\sum_{i=1}^n|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|^2\max_{i\in[n]}\Big(\langle\nabla\Psi(\mathbf{w};\mathbf{x}_i),\mathbf{v}\rangle-\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\Big)^2\\
&=2L_S(\mathbf{w})\max_{i\in[n]}\Big(\langle\nabla\Psi(\mathbf{w};\mathbf{x}_i),\mathbf{v}\rangle-\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\Big)^2,
\end{aligned}\tag{7.4}$$

where $L_S(\mathbf{w})$ is defined in Eq. (3.2). For the second term, we know

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\Big(\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle-\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\Big)^2\\
&=\frac{2}{n}\sum_{i=1}^{n}\Big(\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))-\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))\Big)^2\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle^2\\
&\leq\frac{2G_\Psi^2}{n}\sum_{i=1}^{n}\Big(\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))-\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))\Big)^2,
\end{aligned}\tag{7.5}$$

where we have used the inequality $|\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle|\leq\|\nabla\Psi(\mathbf{w}';\mathbf{x}_i)\|_2\|\mathbf{v}'\|_2\leq G_\Psi$. Furthermore, by the mean-value theorem, we know the existence of $a\in\mathbb{R}$ between $\Psi(\mathbf{w};\mathbf{x}_i)$ and $\Psi(\mathbf{w}';\mathbf{x}_i)$ such that

$$\big|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))-\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))\big|=\big|\ell''(y_i,a)(\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i))\big|\leq\beta\big|\ell'(y_i,a)(\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i))\big|,$$

where we have used Assumption 1. Applying Assumption 1 again, we derive

$$\big|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))-\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))\big|\leq\beta\max\big\{|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|,|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|\big\}\big|\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i)\big|.\tag{7.6}$$

We plug the above inequality back into Eq. (7.5), and derive

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\Big(\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle-\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\Big)^2\\
&\leq\frac{2G_\Psi^2\beta^2}{n}\sum_{i=1}^{n}\max\big\{|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|^2,|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|^2\big\}\big(\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i)\big)^2\\
&\leq\frac{2G_\Psi^2\beta^2}{n}\sum_{i=1}^{n}\max\big\{|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|^2,|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|^2\big\}\max_{i\in[n]}\big(\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i)\big)^2\\
&\leq\frac{2G_\Psi^2\beta^2}{n}\sum_{i=1}^{n}\big(|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|^2+|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|^2\big)\max_{i\in[n]}\big(\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i)\big)^2\\
&=2G_\Psi^2\beta^2\big(L_S(\mathbf{w})+L_S(\mathbf{w}')\big)\max_{i\in[n]}\big(\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i)\big)^2.
\end{aligned}$$

We plug Eq. (7.4) and the above inequality back into Eq. (7.3), and derive

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\big(g_{\mathbf{w},\mathbf{v}}(\mathbf{z}_i)-g_{\mathbf{w}',\mathbf{v}'}(\mathbf{z}_i)\big)^2\leq{}&2L_S(\mathbf{w})\max_{i\in[n]}\Big(\langle\nabla\Psi(\mathbf{w};\mathbf{x}_i),\mathbf{v}\rangle-\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\Big)^2\\
&+2G_\Psi^2\beta^2\big(L_S(\mathbf{w})+L_S(\mathbf{w}')\big)\max_{i\in[n]}\big(\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i)\big)^2.
\end{aligned}$$

Since $L_S(\mathbf{w})\leq r$ and $L_S(\mathbf{w}')\leq r$ for $\mathbf{w},\mathbf{w}'\in\widehat{\mathcal{W}}_r$, we further get

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\big(g_{\mathbf{w},\mathbf{v}}(\mathbf{z}_i)-g_{\mathbf{w}',\mathbf{v}'}(\mathbf{z}_i)\big)^2\leq{}&2r\max_{i\in[n]}\Big(\langle\nabla\Psi(\mathbf{w};\mathbf{x}_i),\mathbf{v}\rangle-\langle\nabla\Psi(\mathbf{w}';\mathbf{x}_i),\mathbf{v}'\rangle\Big)^2\\
&+4G_\Psi^2\beta^2r\max_{i\in[n]}\big(\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i)\big)^2.
\end{aligned}\tag{7.7}$$

For any $r\geq0$, introduce

$$\widetilde{\mathcal{F}}_r=\big\{(\mathbf{x},k)\mapsto\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x},k):(\mathbf{w},\mathbf{v})\in\widehat{\mathcal{W}}_r\times\mathcal{V}\big\},\tag{7.8}$$

where $\tilde{f}$ is defined in Eq. (4.6). Let $m=\mathcal{N}_\infty(\epsilon,\widetilde{\mathcal{F}}_r,\widetilde{X})$. By the definition of covering numbers, there exist $(\mathbf{w}_1,\mathbf{v}_1),\ldots,(\mathbf{w}_m,\mathbf{v}_m)$ such that

$$\Big\{\big(\tilde{f}_{\mathbf{w}_j,\mathbf{v}_j}(\mathbf{x}_1,1),\tilde{f}_{\mathbf{w}_j,\mathbf{v}_j}(\mathbf{x}_1,2),\ldots,\tilde{f}_{\mathbf{w}_j,\mathbf{v}_j}(\mathbf{x}_n,1),\tilde{f}_{\mathbf{w}_j,\mathbf{v}_j}(\mathbf{x}_n,2)\big):j\in[m]\Big\}$$

is an (ϵ, ℓ_∞)-cover of the set

$$\Big\{\big(\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x}_1,1), \tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x}_1,2), \ldots, \tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x}_n,1), \tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x}_n,2)\big) : \mathbf{w} \in \widehat{\mathcal{W}}_r, \mathbf{v} \in \mathcal{V}\Big\}.$$

Then, for any $(\mathbf{w}, \mathbf{v}) \in \widehat{\mathcal{W}}_r \times \mathcal{V}$, there exists $j \in [m]$ such that

$$\max_{i\in[n]} \max_{k\in[2]} \big|\tilde{f}_{\mathbf{w}_j,\mathbf{v}_j}(\mathbf{x}_i,k) - \tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x}_i,k)\big| \le \epsilon,$$

which amounts to saying that

$$\max_{i\in[n]} \big|\langle \nabla\Psi(\mathbf{w},\mathbf{x}_i), \mathbf{v}\rangle - \langle \nabla\Psi(\mathbf{w}_j,\mathbf{x}_i), \mathbf{v}_j\rangle\big| \le \epsilon \quad \text{and} \quad 2G_\Psi\beta \max_{i\in[n]} \big|\Psi(\mathbf{w},\mathbf{x}_i) - \Psi(\mathbf{w}_j,\mathbf{x}_i)\big| \le \epsilon.$$

It then follows from Eq. (7.7) that $\frac{1}{n}\sum_{i=1}^n \big(g_{\mathbf{w},\mathbf{v}}(\mathbf{z}_i) - g_{\mathbf{w}_j,\mathbf{v}_j}(\mathbf{z}_i)\big)^2 \le 2r\epsilon^2 + r\epsilon^2$. It then follows that $\big\{\big(g_{\mathbf{w}_j,\mathbf{v}_j}(\mathbf{z}_1), \ldots, g_{\mathbf{w}_j,\mathbf{v}_j}(\mathbf{z}_n)\big) : j \in [m]\big\}$ is an $(\sqrt{3r}\epsilon, \ell_2)$-cover of $\widehat{\mathcal{G}}_r$ w.r.t. S, where $\widehat{\mathcal{G}}_r$ is defined in Eq. (5.3). That is, $\mathcal{N}_2\big(\sqrt{3r}\epsilon, \widehat{\mathcal{G}}_r, S\big) \le m$ and therefore Eq. (5.4) holds by noting that $\mathcal{N}_\infty(\epsilon, \widetilde{\mathcal{F}}_r, \widetilde{X}) \le \mathcal{N}_\infty(\epsilon, \widetilde{\mathcal{F}}, \widetilde{X})$. The proof is complete. □

The following lemma controls the radius of the function class $\{\mathbf{z} \mapsto g_{\mathbf{w},\mathbf{v}}(\mathbf{z}) : (\mathbf{w},\mathbf{v}) \in \widehat{\mathcal{W}}_r \times \mathcal{V}\}$ on the dataset S, which is required in applying the entropy integral.

Lemma 12. *Let Assumption 1 hold and $\|\nabla\Psi(\mathbf{w},\mathbf{x})\|_2 \le G_\Psi$ for all $\mathbf{w} \in \mathcal{W}$ and $\mathbf{x} \in \mathcal{X}$. Then, for $g_{\mathbf{w},\mathbf{v}}$ defined in Eq. (5.1), we have $\frac{1}{n}\sum_{i=1}^n g^2_{\mathbf{w},\mathbf{v}}(\mathbf{z}_i) \le G^2_\Psi r$ for all $(\mathbf{w},\mathbf{v}) \in \widehat{\mathcal{W}}_r \times \mathcal{V}$.*

Proof. For any $(\mathbf{w},\mathbf{v}) \in \widehat{\mathcal{W}}_r \times \mathcal{V}$, we know

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n g^2_{\mathbf{w},\mathbf{v}}(\mathbf{z}_i) = \frac{1}{n}\sum_{i=1}^n \Big(\langle \nabla f(\mathbf{w};\mathbf{z}_i), \mathbf{v}\rangle\Big)^2 = \frac{1}{n}\sum_{i=1}^n \Big(\ell'(y_i, \Psi(\mathbf{w};\mathbf{x}_i))\langle \nabla\Psi(\mathbf{w};\mathbf{x}_i), \mathbf{v}\rangle\Big)^2 \\
&\le \frac{1}{n}\sum_{i=1}^n \big(\ell'(y_i, \Psi(\mathbf{w};\mathbf{x}_i))\big)^2 \|\nabla\Psi(\mathbf{w};\mathbf{x}_i)\|_2^2 \|\mathbf{v}\|_2^2 \le \frac{G^2_\Psi}{n}\sum_{i=1}^n \big(\ell'(y_i, \Psi(\mathbf{w};\mathbf{x}_i))\big)^2 = G^2_\Psi L_S(\mathbf{w}) \le G^2_\Psi r.
\end{aligned}$$

The proof is complete. □

To prove Theorem 2, we further introduce two lemmas. The following lemma is an improved entropy integral, which controls the Rademacher complexity via covering numbers. The lemma was essentially developed in [36]. The version we provide here is Theorem 9 in [18]. It holds for any $(a_1, \ldots, a_n) \in \mathbb{R}^n$, which is due to the shift-invariance of Rademacher complexity, i.e., $\mathbb{E}_\sigma \sup_f \sum_{i=1}^n \sigma_i f(\mathbf{z}'_i) = \mathbb{E}_\sigma \sup_f \sum_{i=1}^n \sigma_i\big(f(\mathbf{z}'_i) - a_i\big)$ for any $a_i \in \mathbb{R}$.

Lemma 13. *Let $\mathcal{F}'$ be a real-valued function class defined over a space $\mathcal{Z}'$. Let $S' := \{\mathbf{z}'_1, \ldots, \mathbf{z}'_n\} \subseteq \mathcal{Z}'$. Let $(\epsilon_j)_{j=0}^\infty$ be a monotone sequence decreasing to 0 and any $(a_1, \ldots, a_n) \in \mathbb{R}^n$. If $\epsilon_0 \ge \big(\frac{1}{n}\sup_{f\in\mathcal{F}'}\sum_{i=1}^n \big(f(\mathbf{z}'_i) - a_i\big)^2\big)^{\frac{1}{2}}$, then for any non-negative integer N we have*

$$\mathfrak{R}_{S'}(\mathcal{F}') \le 2\sum_{j=1}^N (\epsilon_j + \epsilon_{j-1})\Big(\frac{\log \mathcal{N}_2(\epsilon_j, \mathcal{F}', S')}{n}\Big)^{\frac{1}{2}} + \epsilon_N. \tag{7.9}$$

The following lemma controls covering numbers in terms of worst-case Rademacher complexities. Note that B only appears in a logarithmic function.

Lemma 14 ([36])**.** *Let $\mathcal{F}'$ be a real-valued function class defined over a space $\mathcal{Z}'$ taking values in $[-B, B]$. Let $S' := \{\mathbf{z}'_1, \ldots, \mathbf{z}'_n\} \subseteq \tilde{\mathcal{Z}}'$. For any $\epsilon \ge 3\mathfrak{R}_{\mathcal{Z}',n}(\mathcal{F}')$, we have*

$$\log \mathcal{N}_\infty(\epsilon, \mathcal{F}', S') \le 1 + \frac{64n\mathfrak{R}^2_{\mathcal{Z}',n}(\mathcal{F}')}{\epsilon^2}\Big(\log^2 \frac{8eB^2n}{\epsilon^2}\Big).$$

Recall that $B = G_\Psi \max\{1, 2\beta b\}$. Then, it is clear from the definition of $\tilde{f}$ in Eq. (4.6) that $\big|\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x}, k)\big| \le B$, for all $\mathbf{x} \in \mathcal{X}, k \in \{1, 2\}$. Now, we are ready to prove Theorem 2.

Proof of Theorem 2. Motivated by Lemma 12, we choose $\epsilon_0 = \max\big\{G_\Psi\sqrt{r}, 3\sqrt{3r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\big\}$. Let

$$N = \Big\lfloor \log_2 \frac{G_\Psi\sqrt{r}}{\max\big\{3\sqrt{3r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}}), \sqrt{r}/\sqrt{n}\big\}} \Big\rfloor$$

and define $\epsilon_j = 2^{-j}\epsilon_0$, for all $j \in [N]$. Then it is clear that $\epsilon_j \ge 3\sqrt{3r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})$ for $j = 0, \ldots, N$. Lemma 14 then shows that

$$\log \mathcal{N}_\infty(\epsilon_j/\sqrt{3r}, \widetilde{\mathcal{F}}, \widetilde{X}) \le 1 + \frac{64 \cdot 3r(2n)\mathfrak{R}^2_{\widetilde{X},2n}(\widetilde{\mathcal{F}})}{\epsilon_j^2}\Big(\log^2 \frac{48erB^2n}{\epsilon_j^2}\Big), \quad \forall j = 0, \ldots, N.$$

By Lemma 12, ϵ_0 defined above satisfies the condition in Lemma 13. By Eq. (7.9) and Lemma 5, we know (recall $\widehat{\mathcal{G}}_r$ is defined in Eq. (5.3))

$$\begin{aligned}
\mathfrak{R}_S(\widehat{\mathcal{G}}_r) &\le 2\sum_{j=1}^N (\epsilon_j + \epsilon_{j-1})\Big(\frac{\log \mathcal{N}_2(\epsilon_j, \widehat{\mathcal{G}}_r, S)}{n}\Big)^{\frac{1}{2}} + \epsilon_N \\
&\le 2\sum_{j=1}^N (\epsilon_j + \epsilon_{j-1})\Big(\frac{\log \mathcal{N}_\infty(\epsilon_j/\sqrt{3r}, \widetilde{\mathcal{F}}, \widetilde{X})}{n}\Big)^{\frac{1}{2}} + \epsilon_N \\
&\le 2\sum_{j=1}^N (\epsilon_j + \epsilon_{j-1})\Big(\frac{1}{\sqrt{n}} + \frac{8\sqrt{6r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\log(48erB^2n/\epsilon_j^2)}{\epsilon_j}\Big) + \epsilon_N,
\end{aligned}$$

where we have used $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. Since $\epsilon_{j-1} = 2\epsilon_j$, we further know

$$\begin{aligned}
\mathfrak{R}_S(\widehat{\mathcal{G}}_r) &\le \frac{6}{\sqrt{n}}\sum_{j=1}^N \epsilon_j + \epsilon_N + 48\sqrt{6r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\sum_{j=1}^N \log(48erB^2n/\epsilon_j^2) \\
&\le \frac{6\epsilon_0}{\sqrt{n}} + \epsilon_N + 48\sqrt{6r}N\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\log(48erB^2n) + 96\sqrt{6r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\sum_{j=1}^N \log 1/\epsilon_j,
\end{aligned}$$

where we have used $\sum_{j=1}^N \epsilon_j = \epsilon_0 \sum_{j=1}^N 2^{-j} \le \epsilon_0$. We know

$$\sum_{j=1}^N \log 1/\epsilon_j = \sum_{j=1}^N \log 2^j/\epsilon_0 = N\log 1/\epsilon_0 + \sum_{j=1}^N j\log 2 = N\log 1/\epsilon_0 + 2^{-1}N(N+1)\log 2. \tag{7.10}$$

It then follows that

$$\mathfrak{R}_S(\widehat{\mathcal{G}}_r) \le \frac{6\epsilon_0}{\sqrt{n}} + \epsilon_N + 48\sqrt{6r}N\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\log(48erB^2n) + 96\sqrt{6r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\Big(N\log 1/\epsilon_0 + N^2\Big).$$

According to the definition of N, we know

$$\epsilon_N = 2^{-N}\epsilon_0 \le 2\epsilon_0 2^{-\log_2 \frac{G_\Psi\sqrt{r}}{\max\left\{3\sqrt{3r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}}), \sqrt{r}/\sqrt{n}\right\}}} = \frac{2\epsilon_0 \max\big\{3\sqrt{3r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}}), \sqrt{r}/\sqrt{n}\big\}}{G_\Psi\sqrt{r}}$$

and $N \le \big\lceil \log_2 \frac{G_\Psi\sqrt{r}}{\sqrt{r}/\sqrt{n}} \big\rceil = \big\lceil \log_2(G_\Psi\sqrt{n}) \big\rceil$. Combining the above three inequalities together shows

$$\begin{aligned}
\mathfrak{R}_S(\widehat{\mathcal{G}}_r) \le \max\big\{G_\Psi\sqrt{r}, 3\sqrt{3r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\big\}\Big(\frac{6}{\sqrt{n}} + \frac{2\max\big\{3\sqrt{3}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}}), 1/\sqrt{n}\big\}}{G_\Psi}\Big) \\
+ 48\sqrt{6r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\Big(\widetilde{N}\log(48erB^2n) + 2\widetilde{N}\log(1/(G_\Psi\sqrt{r})) + 2\widetilde{N}^2\Big).
\end{aligned}$$

The stated bound then follows from Eq. (5.2) and

$$\widetilde{N}\log(48erB^2n)+2\widetilde{N}\log(1/(G_\Psi\sqrt{r}))=\widetilde{N}\big(\log(48erB^2n)+\log(1/(G_\Psi^2 r))\big)=\widetilde{N}\log(48eB^2n/G_\Psi^2).$$

The proof is complete. □

7.3 Bounds on Squared Derivatives

To prove Lemma 6, we introduce Lemma 15 established in [7], which provides a uniform connection between the expectation and empirical average over a function space.

Definition 5 (Sub-root function)**.** We say $\phi:[0,\infty)\mapsto[0,\infty]$ is a sub-root function if it is non-decreasing, not identically zero and $r\mapsto\phi(r)/\sqrt{r}$ is non-increasing. We define the fixed point of ϕ as the unique point r^* such that $\phi(r^*)=r^*$.

Lemma 15 ([7])**.** *Let $\mathcal{F}'$ be a class of functions that map $\mathcal{Z}'$ to $[0,b']$. Let ϕ_n be a sub-root function satisfying $\mathbb{E}_{\boldsymbol{\sigma}}\big[\sup_{f\in\mathcal{F}':\frac{1}{n}\sum_{i=1}^n f(\mathbf{z}_i)\le r}\frac{1}{n}\sum_{i=1}^n\sigma_i f(\mathbf{z}_i)\big]\le\phi_n(r)$. Then for any $x>0$, with probability at least $1-2\exp(-x)$, the following inequalities hold simultaneously for all $f\in\mathcal{F}'$*

$$\frac{1}{n}\sum_{i=1}^n f(\mathbf{z}_i)\le 2\mathbb{E}_{\mathbf{z}}[f(\mathbf{z})]+108r_n^*+20r_0+38\sqrt{r_0r_n^*},\quad \mathbb{E}_{\mathbf{z}}[f(\mathbf{z})]\le\frac{2}{n}\sum_{i=1}^n f(\mathbf{z}_i)+106r_n^*+48r_0,$$

where r_n^* is the fixed point of ϕ_n, and we introduce $r_0=b'(x+6\log\log n)/n$.

To apply Lemma 15, we present a lemma in estimating local Rademacher complexities.

Lemma 16. *Let Assumption 1 hold. Assume $|\ell'(y,a)|\le G_\ell$ for all $a\in\mathbb{R},y\in\mathcal{Y}$ and some $G_\ell\ge 0$. Define $\widehat{\mathcal{L}}_r=\big\{\mathbf{z}\mapsto|\ell'(y,\Psi(\mathbf{w};\mathbf{x}))|^2:\mathbf{w}\in\widehat{\mathcal{W}}_r\big\}$. Then, for $\bar{N}=\lfloor\log_2(G_\ell\sqrt{n})\rfloor$ we have*

$$\begin{aligned}\mathfrak{R}_S(\widehat{\mathcal{L}}_r)\le\max\big\{G_\ell\sqrt{r},6G_\ell\beta\sqrt{r}\mathfrak{R}_n(\mathcal{G})\big\}\Big(\frac{6}{\sqrt{n}}+\frac{2\max\big\{6G_\ell\beta\mathfrak{R}_n(\mathcal{G}),1/\sqrt{n}\big\}}{G_\ell}\Big)\\+96\beta G_\ell\sqrt{r}\mathfrak{R}_n(\mathcal{G})\Big(\bar{N}\log(32eb^2\beta^2n)+2\bar{N}^2\Big)=:\bar{\phi}_n(r).\end{aligned}$$

Proof. For any $\mathbf{w},\mathbf{w}'\in\widehat{\mathcal{W}}_r$, we know

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\big(|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|^2-|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|^2\big)^2\\&=\frac{1}{n}\sum_{i=1}^n\big(|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|+|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|\big)^2\big(|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|-|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|\big)^2\\&\le\frac{2}{n}\sum_{i=1}^n\big(|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|^2+|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|^2\big)\big(|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|-|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|\big)^2\\&\le\frac{2}{n}\sum_{i=1}^n\big(|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|^2+|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|^2\big)\max_{i\in[n]}\big(|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|-|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|\big)^2\\&\le 4r\max_{i\in[n]}\big(|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))-\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|\big)^2\\&\le 4r\beta^2\max_{i\in[n]}\max\big\{|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|^2,|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|^2\big\}\big|\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i)\big|^2,\end{aligned}$$

where in the last step we have used Eq. (7.6). Since $|\ell'(y,a)|\le G_\ell$, we further get that

$$\Big(\frac{1}{n}\sum_{i=1}^n\big(|\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))|^2-|\ell'(y_i,\Psi(\mathbf{w}';\mathbf{x}_i))|^2\big)^2\Big)^{\frac{1}{2}}\le 2\beta G_\ell\sqrt{r}\max_{i\in[n]}\big|\Psi(\mathbf{w};\mathbf{x}_i)-\Psi(\mathbf{w}';\mathbf{x}_i)\big|,$$

from which we know $\mathcal{N}_2(\epsilon, \widehat{\mathcal{L}}_r, S) \leq \mathcal{N}_\infty\big(\epsilon/(2\beta G_\ell\sqrt{r}), \mathcal{G}, S_{\mathbf{x}}\big)$ (recall that $\mathcal{G}$ is defined in Eq. (4.9)). Furthermore, for any $\mathbf{w} \in \widehat{\mathcal{W}}_r$, we know

$$\frac{1}{n}\sum_{i=1}^n |\ell'(y_i, \Psi(\mathbf{w}; \mathbf{x}_i))|^4 \leq \frac{G_\ell^2}{n}\sum_{i=1}^n |\ell'(y_i, \Psi(\mathbf{w}; \mathbf{x}_i))|^2 \leq G_\ell^2 r.$$

Define $\epsilon_0 = \max\big\{G_\ell\sqrt{r}, 6\beta G_\ell\sqrt{r}\mathfrak{R}_n(\mathcal{G})\big\}$ and $N = \big\lfloor \log_2 \frac{G_\ell\sqrt{r}}{\max\big\{6\beta G_\ell\sqrt{r}\mathfrak{R}_n(\mathcal{G}), \sqrt{r}/\sqrt{n}\big\}}\big\rfloor$. Introduce a sequence of positive numbers: $\epsilon_j = 2^{-j}\epsilon_0$, $j \in [N]$. It is clear from the definition of N that $\epsilon_j \geq 6\beta G_\ell\sqrt{r}\mathfrak{R}_n(\mathcal{G})$ for all $j = 0, 1, \ldots, N$. Lemma 14 shows

$$\log \mathcal{N}_\infty(\epsilon_j/(2\beta G_\ell\sqrt{r}), \mathcal{G}, S_{\mathbf{x}}) \leq 1 + \frac{256\beta^2 n G_\ell^2 r \mathfrak{R}_n^2(\mathcal{G})}{\epsilon_j^2}\Big(\log^2 \frac{32e\beta^2 G_\ell^2 r b^2 n}{\epsilon_j^2}\Big),$$

where we have used $|\Psi(\mathbf{w}, \mathbf{x})| \leq b$. Therefore, we can apply Lemma 13 to derive that

$$\begin{aligned}
\mathfrak{R}_S(\widehat{\mathcal{L}}_r) &\leq 2\sum_{j=1}^N \big(\epsilon_j + \epsilon_{j-1}\big)\Big(\frac{\log \mathcal{N}_2(\epsilon_j, \widehat{\mathcal{L}}_r, S)}{n}\Big)^{\frac{1}{2}} + \epsilon_N \\
&\leq 2\sum_{j=1}^N \big(\epsilon_j + \epsilon_{j-1}\big)\Big(\frac{\log \mathcal{N}_\infty(\epsilon_j/(2\beta G_\ell\sqrt{r}), \mathcal{G}, S_{\mathbf{x}})}{n}\Big)^{\frac{1}{2}} + \epsilon_N \\
&\leq 2\sum_{j=1}^N \big(\epsilon_j + \epsilon_{j-1}\big)\Big(\frac{1}{\sqrt{n}} + \frac{16\beta G_\ell\sqrt{r}\mathfrak{R}_n(\mathcal{G})\log(32e\beta^2 G_\ell^2 r b^2 n/\epsilon_j^2)}{\epsilon_j}\Big) + \epsilon_N.
\end{aligned}$$

Analyzing analogously to the proof of Theorem 2, we know that

$$\begin{aligned}
&\mathfrak{R}_S(\widehat{\mathcal{L}}_r) \leq \frac{6}{\sqrt{n}}\sum_{j=1}^N \epsilon_j + \epsilon_N + 96\beta G_\ell\sqrt{r}\mathfrak{R}_n(\mathcal{G})\sum_{j=1}^N \log(32e\beta^2 G_\ell^2 r b^2 n/\epsilon_j^2) \\
&\leq \frac{6\epsilon_0}{\sqrt{n}} + \frac{2\epsilon_0 \max\big\{6\beta G_\ell\mathfrak{R}_n(\mathcal{G}), 1/\sqrt{n}\big\}}{G_\ell} + 96\beta G_\ell\sqrt{r}\mathfrak{R}_n(\mathcal{G})\Big(N\log(32e\beta^2 G_\ell^2 r b^2 n) + 2N\log 1/\epsilon_0 + 2N^2\Big),
\end{aligned}$$

where we have used Eq. (7.10). It is clear that $N \leq \log_2(G_\ell\sqrt{n})$. Combining the above inequalities together yields

$$\begin{aligned}
\mathfrak{R}_S(\widehat{\mathcal{L}}_r) \leq \max\big\{G_\ell\sqrt{r}, 6\beta G_\ell\sqrt{r}\mathfrak{R}_n(\mathcal{G})\big\}\Big(\frac{6}{\sqrt{n}} + \frac{2\max\big\{6\beta G_\ell\mathfrak{R}_n(\mathcal{G}), 1/\sqrt{n}\big\}}{G_\ell}\Big) \\
+ 96\beta G_\ell\sqrt{r}\mathfrak{R}_n(\mathcal{G})\Big(\bar{N}\log(32e\beta^2 G_\ell^2 r b^2 n) + 2\bar{N}\log 1/(G_\ell\sqrt{r}) + 2\bar{N}^2\Big).
\end{aligned}$$

The proof is complete. □

Proof of Lemma 6. According to Lemma 16, we have the following inequality

$$\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\mathcal{W}: \frac{1}{n}\sum_{i=1}^n |\ell'(y_i, \Psi(\mathbf{w}; \mathbf{x}_i))|^2 \leq r} \frac{1}{n}\sum_{i=1}^n \sigma_i |\ell'(y_i, \Psi(\mathbf{w}; \mathbf{x}_i))|^2\Big] \leq \bar{\phi}_n(r).$$

It is clear that $\bar{\phi}_n$ is a sub-root function and the fixed point $\bar{r}_n^*$ satisfies

$$\begin{aligned}
\sqrt{\bar{r}_n^*} = \max\big\{G_\ell, 6\beta G_\ell\mathfrak{R}_n(\mathcal{G})\big\}\Big(\frac{6}{\sqrt{n}} + \frac{2\max\big\{6G_\ell\beta\mathfrak{R}_n(\mathcal{G}), 1/\sqrt{n}\big\}}{G_\ell}\Big) \\
+ 96\beta G_\ell\mathfrak{R}_n(\mathcal{G})\big(\bar{N}\log(32eb^2\beta^2 n) + 2\bar{N}^2\big).
\end{aligned}$$

Since $\mathfrak{R}_n(\mathcal{G})\beta \lesssim 1$, we know $\max\big\{G_\ell, 6\beta G_\ell\mathfrak{R}_n(\mathcal{G})\big\} \lesssim G_\ell$. Since $\mathfrak{R}_n(\mathcal{G}) \geq 1/(\beta\sqrt{n})$ and $G_\ell \gtrsim 1$, we know $\frac{1}{\sqrt{n}} + \frac{\max\{G_\ell\beta\mathfrak{R}_n(\mathcal{G}), 1/\sqrt{n}\}}{G_\ell} \lesssim \frac{1}{\sqrt{n}} + \frac{G_\ell\beta\mathfrak{R}_n(\mathcal{G})}{G_\ell} \lesssim \beta\mathfrak{R}_n(\mathcal{G})$. It then follows that

$$\sqrt{\bar{r}_n^*} \lesssim \beta G_\ell\mathfrak{R}_n(\mathcal{G}) + \beta G_\ell\mathfrak{R}_n(\mathcal{G})\big(\bar{N}\log(b^2\beta^2 n) + \bar{N}^2\big).$$

Recall $\bar{N} = \lfloor \log_2(G_\ell \sqrt{n}) \rfloor$. Then, $\bar{N}\log(b^2\beta^2 n) + \bar{N}^2 \lesssim \bar{N}(\log(b\beta n) + \log(G_\ell n)) \lesssim \log(G_\ell n)\log(G_\ell b\beta n)$. It then follows that $\bar{r}_n^* \lesssim \beta^2 G_\ell^2 \mathfrak{R}_n^2(\mathcal{G}) \log^2(G_\ell n)\log^2(G_\ell b\beta n)$. The stated bound then follows from Lemma 15 with $b' = G_\ell^2$. The proof is complete. □

7.4 Proof of Theorem 3

In this section, we present the proof of Theorem 3. Our basic idea is to first develop a bound on $\sup_{\mathbf{w}\in\mathcal{W}_r} \|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\|_2$ in terms of the localization parameter r, and then apply the standard peeling argument to relate r to $L_S(\mathbf{w})$.

Lemma 17. *Let Assumption 1 and Assumption 2 hold. Assume that $\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}}) \lesssim G_\Psi$ and $\mathfrak{R}_n(\mathcal{G})\beta \leq 1$. Then, for any $\delta \in (0,1)$, with probability (w.r.t. S) at least $1-\delta$ we have*

$$\sup_{\mathbf{w}\in\mathcal{W}_r} \|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \lesssim \Big(\sqrt{r} + G_\ell \beta \mathfrak{R}_n(\mathcal{G}) \log^2(G_\ell b\beta n)\Big)\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn)\log^{\frac{1}{2}}(1/\delta) + \frac{G_\ell G_\Psi \log(1/\delta)}{n}.$$

Proof. According to Lemma 6, with probability at least $1-\delta/2$ the following inequality holds simultaneously for all $\mathbf{w} \in \mathcal{W}$

$$L_S(\mathbf{w}) - 2L(\mathbf{w}) \leq C_1 \beta^2 G_\ell^2 \mathfrak{R}_n^2(\mathcal{G}) \log^4(G_\ell b\beta n) + \frac{C_1 G_\ell^2 (\log(4/\delta) + \log\log n)}{n}, \tag{7.11}$$

where C_1 is a universal constant. Let E be the event under which Eq. (7.11) holds. Define

$$\hat{r} = 2r + C_1 \beta^2 G_\ell^2 \mathfrak{R}_n^2(\mathcal{G}) \log^4(G_\ell b\beta n) + \frac{C_1 G_\ell^2 (\log(4/\delta) + \log\log n)}{n}.$$

Under the event, we know that $L_n(\mathbf{w}) \leq \hat{r}$ for any $\mathbf{w} \in \mathcal{W}_r$. Therefore, $\mathcal{W}_r \subseteq \widehat{\mathcal{W}}_{\hat{r}}$ and

$$\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_r} \Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i \nabla f(\mathbf{w};\mathbf{z}_i)\Big\|_2\Big] \leq \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\widehat{\mathcal{W}}_{\hat{r}}} \Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i \nabla f(\mathbf{w};\mathbf{z}_i)\Big\|_2\Big].$$

Since $\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}}) \lesssim G_\Psi$, we have $\max\{G_\Psi\sqrt{r}, 3\sqrt{3r}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\} \lesssim G_\Psi\sqrt{r}$. Therefore, we can apply Theorem 2 to show that (under the event E)

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_r} \Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i \nabla f(\mathbf{w};\mathbf{z}_i)\Big\|_2\Big] &\lesssim \frac{G_\Psi\sqrt{\hat{r}}}{\sqrt{n}} + \sqrt{\hat{r}}\max\{\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}}), 1/\sqrt{n}\} + \sqrt{\hat{r}}\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn) \\
&\lesssim \frac{G_\Psi\sqrt{\hat{r}}}{\sqrt{n}} + \sqrt{\hat{r}}\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn).
\end{aligned}$$

We plug the definition of $\hat{r}$ to the above bound and get

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_r} \Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i \nabla f(\mathbf{w};\mathbf{z}_i)\Big\|_2\Big] \\
&\lesssim \Big(\sqrt{r} + \beta G_\ell \mathfrak{R}_n(\mathcal{G})\log^2(G_\ell b\beta n) + \frac{G_\ell(\log^{\frac{1}{2}}(1/\delta) + \log^{\frac{1}{2}}\log n)}{\sqrt{n}}\Big)\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn) \\
&\lesssim \Big(\sqrt{r} + \beta G_\ell \mathfrak{R}_n(\mathcal{G})\log^{\frac{1}{2}}(1/\delta)\log^2(G_\ell b\beta n)\Big)\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn),
\end{aligned} \tag{7.12}$$

where we have used $\frac{G_\Psi}{\sqrt{n}} \leq \mathfrak{R}_{2n}(\widetilde{\mathcal{F}})$ and $\frac{1}{\sqrt{n}} \leq \beta\mathfrak{R}_n(\mathcal{G})$ by the definition in Eq. (4.8). By Theorem 1, the following inequality holds with probability at least $1-\delta/2$

$$\sup_{\mathbf{w}\in\mathcal{W}_r} \|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \lesssim \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_r} \Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i \nabla f(\mathbf{w};\mathbf{z}_i)\Big\|_2\Big] + G_\Psi\Big(\frac{r\log(1/\delta)}{n}\Big)^{\frac{1}{2}} + \frac{G_\ell G_\Psi \log(1/\delta)}{n}.$$

This together with the definition of the event E shows the following inequality with probability $1-\delta$

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}_r} \big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 \\
&\lesssim \Big(\sqrt{r} + \beta G_\ell \mathfrak{R}_n(\mathcal{G}) \log^{\frac{1}{2}}(1/\delta) \log^2(G_\ell b\beta n)\Big)\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn) + G_\Psi\Big(\frac{r\log(1/\delta)}{n}\Big)^{\frac{1}{2}} + \frac{G_\ell G_\Psi \log(1/\delta)}{n} \\
&\lesssim \Big(\sqrt{r} + \beta G_\ell \mathfrak{R}_n(\mathcal{G}) \log^2(G_\ell b\beta n)\Big)\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn)\log^{\frac{1}{2}}(1/\delta) + \frac{G_\ell G_\Psi \log(1/\delta)}{n},
\end{aligned}$$

where we have used $\frac{G_\Psi}{\sqrt{n}} \le \mathfrak{R}_{2n}(\widetilde{\mathcal{F}})$ in the last step. The proof is complete. □

We now apply the peeling trick to prove Theorem 3.

Proof of Theorem 3. Let $r_k = 2^{-k}G_\ell^2, k = 0, \ldots, K$, where $K = \lceil \log_2(nG_\Psi^2)\rceil$. Define

$$B_k = \big\{\mathbf{w}\in\mathcal{W} : r_{k+1} < L(\mathbf{w}) \le r_k\big\}, \quad k = 0, \ldots, K-1,$$

and $B_K = \big\{\mathbf{w}\in\mathcal{W} : L(\mathbf{w}) \le r_K\big\}$. Since $L(\mathbf{w}) \le G_\ell^2$, we know that $\mathcal{W} = \cup_{k=0,\ldots,K} B_k$. According to Lemma 17 with $r = r_k$, we derive the following inequality with probability at least $1-\delta/(K+1)$

$$\sup_{\mathbf{w}\in B_k} \big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 \lesssim \Big(\sqrt{r_k} + \beta G_\ell \mathfrak{R}_n(\mathcal{G}) \log^2(G_\ell b\beta n)\Big)\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn)\log^{\frac{1}{2}}(K/\delta) + \frac{G_\ell G_\Psi \log(K/\delta)}{n}. \tag{7.13}$$

By the union bounds on the probability of events, we know that with probability at least $1-\delta$, Eq. (7.13) holds simultaneously for all $k = 0, 1, \ldots, K$. We denote this event by E.

For any $\mathbf{w}\in B_k$ with $k < K$, the definition of B_k implies that $r_k \le 2L(\mathbf{w})$. Then, under the event E, the following inequality holds uniformly for all $\mathbf{w}\in\cup_{k=0,\ldots,K-1} B_k$

$$\big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 \lesssim \Big(\sqrt{2L(\mathbf{w})} + \beta G_\ell \mathfrak{R}_n(\mathcal{G}) \log^2(G_\ell b\beta n)\Big)\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn)\log^{\frac{1}{2}}(K/\delta) + \frac{G_\ell G_\Psi \log(K/\delta)}{n}. \tag{7.14}$$

Furthermore, for any $\mathbf{w}\in B_K$, we have $L(\mathbf{w}) \le r_K \lesssim G_\ell^2/(nG_\Psi^2)$ and therefore Eq. (7.13) implies

$$\begin{aligned}
&\sup_{\mathbf{w}\in B_K} \big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 \\
&\lesssim \Big(G_\ell/(G_\Psi\sqrt{n}) + \beta G_\ell \mathfrak{R}_n(\mathcal{G}) \log^2(G_\ell b\beta n)\Big)\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn)\log^{\frac{1}{2}}(K/\delta) + \frac{G_\ell G_\Psi \log(K/\delta)}{n} \\
&\lesssim \beta G_\ell \mathfrak{R}_n(\mathcal{G})\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn)\log^2(G_\ell b\beta n)\log^{\frac{1}{2}}(K/\delta) + \frac{G_\ell G_\Psi \log(K/\delta)}{n},
\end{aligned}$$

where we have used the fact $\frac{1}{\sqrt{n}} \le \beta\mathfrak{R}_n(\mathcal{G})$ by Eq. (4.8) and $G_\Psi \ge 1$. We combine this inequality and Eq. (7.14) together, and know that Eq. (7.14) holds uniformly for all $\mathbf{w}\in\mathcal{W} = \cup_{k=0,\ldots,K} B_k$. Furthermore, by Lemma 6, we know the following inequality with probability at least $1-\delta$

$$L(\mathbf{w}) - 2L_S(\mathbf{w}) \lesssim \beta^2 G_\ell^2 \mathfrak{R}_n^2(\mathcal{G}) \log^4(G_\ell b\beta n) + \frac{G_\ell^2(\log(1/\delta) + \log\log n)}{n},$$

which implies that

$$\sqrt{L(\mathbf{w})} \lesssim L_S^{\frac{1}{2}}(\mathbf{w}) + \beta G_\ell \mathfrak{R}_n(\mathcal{G}) \log^2(G_\ell b\beta n) + \frac{G_\ell(\log^{\frac{1}{2}}(1/\delta) + \log^{\frac{1}{2}}\log n)}{\sqrt{n}}.$$

We plug this inequality back into Eq. (7.14), and derive

$$\begin{aligned}
\big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 &\lesssim \frac{G_\ell G_\Psi \log(K/\delta)}{n} + \\
&\mathfrak{R}_{2n}(\widetilde{\mathcal{F}})\widetilde{N}\log(Bn)\log^{\frac{1}{2}}(K/\delta)\Big(L_S^{\frac{1}{2}}(\mathbf{w}) + \beta G_\ell \mathfrak{R}_n(\mathcal{G})\log^2(G_\ell b\beta n) + \frac{G_\ell(\log^{\frac{1}{2}}(1/\delta) + \log^{\frac{1}{2}}\log n)}{\sqrt{n}}\Big).
\end{aligned}$$

The proof is complete by noting $1/\sqrt{n} \le \beta\mathfrak{R}_n(\mathcal{G})$ and $\frac{G_\Psi}{\sqrt{n}} \le \mathfrak{R}_{2n}(\widetilde{\mathcal{F}})$. □

Proof of Lemma 4. Fix any $(\tilde{\mathbf{x}}_1, \tilde{k}_1), \ldots, (\tilde{\mathbf{x}}_{2n}, \tilde{k}_{2n}) \in \widetilde{X}$, and define $I_1 = \big\{i \in [2n] : \tilde{k}_i = 1\big\}$ and $I_2 = \big\{i \in [2n] : \tilde{k}_i = 2\big\}$. By the definition of $\tilde{f}_{\mathbf{w},\mathbf{v}}$, we know

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\tilde{f}_{\mathbf{w},\mathbf{v}} \in \widetilde{\mathcal{F}}} \sum_{i=1}^{2n} \sigma_i \tilde{f}_{\mathbf{w},\mathbf{v}}(\tilde{\mathbf{x}}_i, \tilde{k}_i) &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\tilde{f}_{\mathbf{w},\mathbf{v}} \in \widetilde{\mathcal{F}}} \Big(\sum_{i \in I_1} \sigma_i \tilde{f}_{\mathbf{w},\mathbf{v}}(\tilde{\mathbf{x}}_i, 1) + \sum_{i \in I_2} \sigma_i \tilde{f}_{\mathbf{w},\mathbf{v}}(\tilde{\mathbf{x}}_i, 2) \Big) \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\tilde{f}_{\mathbf{w},\mathbf{v}} \in \widetilde{\mathcal{F}}} \sum_{i \in I_1} \sigma_i \tilde{f}_{\mathbf{w},\mathbf{v}}(\tilde{\mathbf{x}}_i, 1) + \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\tilde{f}_{\mathbf{w},\mathbf{v}} \in \widetilde{\mathcal{F}}} \sum_{i \in I_2} \sigma_i \tilde{f}_{\mathbf{w},\mathbf{v}}(\tilde{\mathbf{x}}_i, 2) \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{(\mathbf{w},\mathbf{v}) \in \mathcal{W} \times \mathcal{V}} \sum_{i \in I_1} \sigma_i \langle \nabla \Psi(\mathbf{w}, \tilde{\mathbf{x}}_i), \mathbf{v} \rangle + 2 G_\Psi \beta \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{w} \in \mathcal{W}} \sum_{i \in I_2} \sigma_i \Psi(\mathbf{w}, \tilde{\mathbf{x}}_i) \\
&\leq |I_1| \mathfrak{R}_{S_{\mathbf{x}}, |I_1|}(\widetilde{\mathcal{G}}) + 2 G_\Psi \beta |I_2| \mathfrak{R}_{S_{\mathbf{x}}, |I_2|}(\mathcal{G}).
\end{aligned}$$

If we take the supremum over all $(\tilde{\mathbf{x}}_1, \tilde{k}_1), \ldots, (\tilde{\mathbf{x}}_{2n}, \tilde{k}_{2n}) \in \widetilde{X}$, we get

$$\mathfrak{R}_{\widetilde{X}, 2n}(\widetilde{\mathcal{F}}) \leq \max_{I_1, I_2} \Big(\frac{|I_1| \mathfrak{R}_{S_{\mathbf{x}}, |I_1|}(\widetilde{\mathcal{G}})}{2n} + \frac{2 G_\Psi \beta |I_2| \mathfrak{R}_{S_{\mathbf{x}}, |I_2|}(\mathcal{G})}{2n} \Big).$$

The stated bound then follows directly. □

8 Conclusions

We provide a general result on relating the uniform convergence of gradients to the Rademacher complexity of two real-valued functions (i.e., $\mathcal{G}$ and $\widetilde{\mathcal{F}}$), which are relatively easy to control. Our uniform convergence rate is dimension-free and can be fast in an interpolation setting. Our key observation is that the Rademacher complexity of a local function space parameterized by r scales as a square root function of r. We justify the effectiveness of our result by applying it to develop optimistic bounds for SNNs. In particular, we develop almost width-independent Rademacher complexity bounds for SNNs with an initialization-dependent space for Lipschitz and smooth activation functions under a Frobenius-norm constraint.

Acknowledgements

The work is partially supported by the Research Grants Council of Hong Kong [Project No. 17302624] and [Project No. 17305425].

Appendix

A Table of Notations

We collect in Table 2 the notations of performance measures used in this paper. We denote $\mathbb{I}_{[\cdot]}$ the indicator function, i.e., returning the value 1 if the argument is true and 0 otherwise.

B Examples of Loss Functions

In this section, we will show that the logistic loss and polynomially tailed losses [30, 31] satisfy Assumption 1 and the Lipschitz continuity assumption. It is also direct to show that sub-exponential tailed losses satisfy these assumptions. We omit the discussions for brevity.

Notation	Meaning	Reference
$F_S(\mathbf{w})$	$\frac{1}{n}\sum_{i=1}^n f(\mathbf{w};\mathbf{z}_i)$	Eq. (3.1)
$F(\mathbf{w})$	$\mathbb{E}_{\mathbf{z}}[f(\mathbf{w};\mathbf{z})]$	Eq. (3.1)
$L_S(\mathbf{w})$	$\frac{1}{n}\sum_{i=1}^n \lvert\ell'(y_i,\Psi(\mathbf{w};\mathbf{x}_i))\rvert^2$	Eq. (3.2)
$L(\mathbf{w})$	$\mathbb{E}_{\mathbf{z}}[\lvert\ell'(y,\Psi(\mathbf{w};\mathbf{x}))\rvert^2]$	Eq. (3.2)
$\mathcal{W}_r$	$\{\mathbf{w}:\mathbf{w}\in\mathcal{W}, L(\mathbf{w})\le r\}$	Eq. (4.2)
$\widehat{\mathcal{W}}_r$	$\{\mathbf{w}:\mathbf{w}\in\mathcal{W}, L_S(\mathbf{w})\le r\}$	Eq. (4.3)
$\mathcal{V}$	$\{\mathbf{v}\in\mathbb{R}^d:\lVert\mathbf{v}\rVert_2\le 1\}$	Eq. (4.4)
$\widetilde{X}$	$\{(\mathbf{x}_1,1),(\mathbf{x}_1,2),(\mathbf{x}_2,1),(\mathbf{x}_2,2),\ldots,(\mathbf{x}_n,1),(\mathbf{x}_n,2)\}$	Eq. (4.5)
$\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x},k)$	$\langle\nabla\Psi(\mathbf{w},\mathbf{x}),\mathbf{v}\rangle\mathbb{I}_{[k=1]}+2G_\Psi\beta\Psi(\mathbf{w};\mathbf{x})\mathbb{I}_{[k=2]}$	Eq. (4.6)
$\widetilde{\mathcal{F}}$	$\{(\mathbf{x},k)\mapsto\tilde{f}_{\mathbf{w},\mathbf{v}}(\mathbf{x},k):(\mathbf{w},\mathbf{v})\in\mathcal{W}\times\mathcal{V}\}$	Eq. (4.7)
$\mathcal{G}$	$\{\mathbf{x}\mapsto\Psi(\mathbf{w};\mathbf{x}):\mathbf{w}\in\mathcal{W}\}$	Eq. (4.9)
$\widetilde{\mathcal{G}}$	$\{\mathbf{x}\mapsto\langle\nabla\Psi(\mathbf{w};\mathbf{x}),\mathbf{v}\rangle:(\mathbf{w},\mathbf{v})\in\mathcal{W}\times\mathcal{V}\}$	Eq. (4.10)
$g_{\mathbf{w},\mathbf{v}}(\mathbf{z})$	$\langle\nabla f(\mathbf{w};\mathbf{z}),\mathbf{v}\rangle$	Eq. (5.1)
$\widehat{\mathcal{G}}_r$	$\{\mathbf{z}\mapsto\langle\nabla f(\mathbf{w};\mathbf{z}),\mathbf{v}\rangle:(\mathbf{w},\mathbf{v})\in\widehat{\mathcal{W}}_r\times\mathcal{V}\}$	Eq. (5.3)

Table 2: Table of Notations.

B.1 Logistic Loss

Assume $y\in\{-1,+1\}$. The logistic loss $\ell(y,a)=\log(1+\exp(-ya))$ is a popular loss function for classification. It can be checked that (recall ℓ' means the derivative of ℓ w.r.t. the second argument)

$$\ell'(y,a)=\frac{-y\exp(-ya)}{1+\exp(-ya)}\quad\text{and}\quad\ell''(y,a)=\frac{\exp(-ya)}{(1+\exp(-ya))^2}.$$

It is clear that $\ell''(y,a)\ge 0$ and therefore ℓ is convex. Then, we know $\ell'(y,\cdot)$ is a monotonic function and therefore for any $a\in[a_1,a_2]$ we know

$$\min\{\ell'(y,a_1),\ell'(y,a_2)\}\le\ell'(y,a)\le\max\{\ell'(y,a_1),\ell'(y,a_2)\},$$

which implies that

$$|\ell'(y,a)|\le\max\{|\ell'(y,a_1)|,|\ell'(y,a_2)|\}. \tag{B.1}$$

It is also clear that $|\ell'(y,a)|\le 1$ and $\ell''(y,a)\le|\ell'(y,a)|$. Therefore, ℓ is 1-Lipschitz continuous and it satisfies Assumption 1 with $\beta=1$.

B.2 Polynomially tailed Losses

Assume $y\in\{-1,+1\}$. Consider the following loss function with polynomially-decaying tails [30]

$$\ell(y,a)=\begin{cases}(1+ya)^{-\alpha}, & \text{if } ya\ge 0\\ 1-\alpha ya, & \text{otherwise,}\end{cases}$$

where $\alpha>0$ is a parameter. Then, we know

$$\ell'(y,a)=\begin{cases}-\alpha y(1+ya)^{-\alpha-1}, & \text{if } ya\ge 0\\ -\alpha y, & \text{otherwise,}\end{cases}\quad\text{and}\quad\ell''(y,a)=\begin{cases}\alpha(\alpha+1)(1+ya)^{-(\alpha+2)}, & \text{if } ya\ge 0\\ 0, & \text{otherwise.}\end{cases}$$

Then, it is clear that $|\ell'(y,a)|\le\alpha$. Note $\ell''(y,a)\ge 0$ and therefore Eq. (B.1) holds. Furthermore, if $ya\ge 0$ we know

$$\frac{\ell''(y,a)}{|\ell'(y,a)|}=\frac{\alpha(\alpha+1)(1+ya)^{-(\alpha+2)}}{\alpha(1+ya)^{-\alpha-1}}=\frac{1+\alpha}{1+ya}\le 1+\alpha$$

and $\ell''(y,a)\le|\ell'(y,a)|$ if $a<0$. Therefore, Assumption 1 holds with $\beta=1+\alpha$.

C Proofs on Shallow Neural Networks

C.1 Proofs on SNNs with smooth activations

This section presents the proofs of results on SNNs. To this aim, we first introduce some lemmas. The following lemma gives a contraction principle for Rademacher complexity [3].

Lemma 18 (Contraction Lemma [4]). *Suppose $\psi : \mathbb{R} \mapsto \mathbb{R}$ is G-Lipschitz. Then we have $\mathbb{E}_{\boldsymbol{\sigma}} \sup_{f\in\widetilde{\mathcal{F}}} \sum_{i=1}^n \sigma_i \psi\big(f(x_i)\big)$ $G\mathbb{E}_{\boldsymbol{\sigma}} \sup_{f\in\widetilde{\mathcal{F}}} \sum_{i=1}^n \sigma_i f(x_i)$.*

The following lemma is a comparison inequality for Gaussian processes, see, e.g., Theorem 7.2.11 in [40]. Let $\mathbf{g} = (g_1, \ldots, g_n)$ be independent standard Gaussian random variables.

Lemma 19. *Let $\{\mathfrak{X}_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$ and $\{\mathfrak{Y}_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$ be two mean-zero separable Gaussian processes indexed by the same set $\mathcal{W}$ and suppose that*

$$\mathbb{E}_{\boldsymbol{g}}[(\mathfrak{X}_{\mathbf{w}} - \mathfrak{X}_{\mathbf{w}'})^2] \leq \mathbb{E}_{\boldsymbol{g}}[(\mathfrak{Y}_{\mathbf{w}} - \mathfrak{Y}_{\mathbf{w}'})^2], \quad \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}. \tag{C.1}$$

Then $\mathbb{E}_{\boldsymbol{g}}[\sup_{\mathbf{w}\in\mathcal{W}} \mathfrak{X}_{\mathbf{w}}] \leq \mathbb{E}_{\boldsymbol{g}}[\sup_{\mathbf{w}\in\mathcal{W}} \mathfrak{Y}_{\mathbf{w}}]$.

Our analysis of SNNs requires the following Efron-Stein inequality, which controls the variance of a general function.

Lemma 20 (Efron-Stein Inequality). *Consider $g : \mathcal{Z}^n \mapsto \mathbb{R}$. Let $Z_1, \ldots, Z_n$ be independent random variables taking values in $\mathcal{Z}$ and $\widetilde{Z} = g(Z_1, \ldots, Z_n)$. Then $\mathrm{Var}(\widetilde{Z}) \leq \sum_{i=1}^n \mathbb{E}\big[(\widetilde{Z} - \mathbb{E}_i[\widetilde{Z}])^2\big]$, where $\mathbb{E}_i[\widetilde{Z}] = \mathbb{E}\big[\widetilde{Z}|Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n\big]$.*

Proof of Lemma 7. The bound on $\mathfrak{R}_S(\mathcal{G})$ was proved in the work [24] (note we fix μ_j here and therefore the complexity bound there does not involve the spectral norm of $\mathbf{w}_0$). We only control $\mathfrak{R}_S(\widetilde{\mathcal{G}})$ here. According to Eq. (6.1), we know ($\mathbf{v} = (\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(m)})$ with $\mathbf{v}^{(k)} \in \mathbb{R}^d$)

$$\begin{aligned}
n\mathfrak{R}_S(\widetilde{\mathcal{G}}) &= \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{(\mathbf{w},\mathbf{v})\in\mathcal{W}\times\mathcal{V}} \sum_{i=1}^n \sigma_i \langle \nabla\Psi(\mathbf{w}; \mathbf{x}_i), \mathbf{v}\rangle\Big] = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{(\mathbf{w},\mathbf{v})\in\mathcal{W}\times\mathcal{V}} \sum_{i=1}^n \sigma_i \sum_{j=1}^m \Big\langle \mu_j \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i, \mathbf{v}^{(j)}\Big\rangle\Big] \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{(\mathbf{w},\mathbf{v})\in\mathcal{W}\times\mathcal{V}} \Big(\sum_{j=1}^m \|\mu_j \mathbf{v}^{(j)}\|_2^2\Big)^{\frac{1}{2}} \Big(\sum_{j=1}^m \Big\|\sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i\Big\|_2^2\Big)^{\frac{1}{2}}\Big] \\
&\leq \frac{1}{\sqrt{m}} \mathbb{E}_{\boldsymbol{\sigma}}\Big[\Big(\sup_{\mathbf{w}\in\mathcal{W}} \sum_{j=1}^m \Big\|\sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i\Big\|_2^2\Big)^{\frac{1}{2}}\Big] \leq \frac{1}{\sqrt{m}} \mathbb{E}_{\boldsymbol{\sigma}}\Big[\Big(\sum_{j=1}^m \sup_{\mathbf{w}\in\mathcal{W}} \Big\|\sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i\Big\|_2^2\Big)^{\frac{1}{2}}\Big],
\end{aligned}$$

where we have used the Cauchy-Schwarz's inequality, $\|\mathbf{v}\|_2^2 = \sum_{j=1}^m \|\mathbf{v}^{(j)}\|_2^2 \leq 1$ and $\mu_j \in \{-1/\sqrt{m}, +1/\sqrt{m}\}$. It then follows from the concavity of $x \mapsto \sqrt{x}$ that

$$\mathfrak{R}_S(\widetilde{\mathcal{G}}) \leq \frac{1}{\sqrt{mn}} \Big(\sum_{j=1}^m \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\mathcal{W}} \Big\|\sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i\Big\|_2^2\Big]\Big)^{\frac{1}{2}}. \tag{C.2}$$

Define $g : \mathbb{R}^n \mapsto \mathbb{R}$ as $g(\sigma_1, \ldots, \sigma_n) = \sup_{\mathbf{w}\in\mathcal{W}} \big\|\sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i\big\|_2$ and $\widetilde{Z} = g(\sigma_1, \ldots, \sigma_n)$. We know

$$\begin{aligned}
\widetilde{Z} - \mathbb{E}_k[\widetilde{Z}] &= \sup_{\mathbf{w}\in\mathcal{W}} \Big\|\sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i\Big\|_2 - \mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{W}} \Big\|\sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i\Big\|_2 | \sigma_j : j \neq k\Big] \\
&= \frac{1}{2}\Big(\sup_{\mathbf{w}\in\mathcal{W}} \Big\|\sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i\Big\|_2 - \sup_{\mathbf{w}\in\mathcal{W}} \Big\|\sum_{i\in[n]:i\neq k} \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i + \gamma'(\mathbf{x}_k^\top \mathbf{w}^{(j)})\mathbf{x}_k\Big\|_2\Big) \\
&\quad + \frac{1}{2}\Big(\sup_{\mathbf{w}\in\mathcal{W}} \Big\|\sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i\Big\|_2 - \sup_{\mathbf{w}\in\mathcal{W}} \Big\|\sum_{i\in[n]:i\neq k} \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i - \gamma'(\mathbf{x}_k^\top \mathbf{w}^{(j)})\mathbf{x}_k\Big\|_2\Big).
\end{aligned}$$

If $\sigma_k = 1$, then it follows that

$$\begin{aligned}
\big|\widetilde{Z} - \mathbb{E}_k[\widetilde{Z}]\big| &= \frac{1}{2}\Big| \sup_{\mathbf{w}\in\mathcal{W}} \Big\| \sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i \Big\|_2 - \sup_{\mathbf{w}\in\mathcal{W}} \Big\| \sum_{i\in[n]:i\neq k} \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i - \gamma'(\mathbf{x}_k^\top \mathbf{w}^{(j)})\mathbf{x}_k \Big\|_2 \Big| \\
&\leq \frac{1}{2} \sup_{\mathbf{w}\in\mathcal{W}} \Big\| \sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i - \sum_{i\in[n]:i\neq k} \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i + \gamma'(\mathbf{x}_k^\top \mathbf{w}^{(j)})\mathbf{x}_k \Big\|_2 \\
&= \frac{1}{2} \sup_{\mathbf{w}\in\mathcal{W}} \Big\| \gamma'(\mathbf{x}_k^\top \mathbf{w}^{(j)})\mathbf{x}_k + \gamma'(\mathbf{x}_k^\top \mathbf{w}^{(j)})\mathbf{x}_k \Big\|_2 \leq G_\gamma \|\mathbf{x}_k\|_2,
\end{aligned}$$

where we have used the elementary inequality $\big|\sup_{\mathbf{w}} h(\mathbf{w}) - \sup_{\mathbf{w}} \tilde{h}(\mathbf{w})\big| \leq \sup_{\mathbf{w}} |h(\mathbf{w}) - \tilde{h}(\mathbf{w})|$ for any $h, \tilde{h} : \mathcal{W} \mapsto \mathbb{R}$. If $\sigma_k = -1$, we can derive a similar inequality. Therefore, we always have $\big|\widetilde{Z} - \mathbb{E}_k[\widetilde{Z}]\big| \leq G_\gamma \|\mathbf{x}_k\|_2$, where $\mathbb{E}_k[\cdot]$ is defined in Lemma 20. It then follows from Lemma 20 that $\mathbb{E}[\widetilde{Z}^2] - \big(\mathbb{E}[\widetilde{Z}]\big)^2 \leq G_\gamma^2 \sum_{k=1}^n \|\mathbf{x}_k\|_2^2$. That is, (for any $R > 0$ and $\tilde{\mathbf{v}} \in \mathbb{R}^{\bar{d}}$, we introduce $\mathcal{B}_{R,\tilde{\mathbf{v}}} := \{\tilde{\mathbf{w}} \in \mathbb{R}^{d'} : \|\tilde{\mathbf{w}} - \tilde{\mathbf{v}}\|_2 \leq R\}$)

$$\begin{aligned}
\mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{W}} \Big\| \sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i \Big\|_2^2 \Big] &\leq \Big(\mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{W}} \Big\| \sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i \Big\|_2 \Big] \Big)^2 + G_\gamma^2 \sum_{i=1}^n \|\mathbf{x}_i\|_2^2 \\
&\leq \Big(\mathbb{E}\Big[\sup_{\tilde{\mathbf{w}}\in\mathcal{B}_{R,\mathbf{w}_0^{(j)}}} \Big\| \sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\mathbf{x}_i \Big\|_2 \Big] \Big)^2 + G_\gamma^2 \sum_{i=1}^n \|\mathbf{x}_i\|_2^2, \qquad \text{(C.3)}
\end{aligned}$$

where we have used the inequality $\|\mathbf{w}^{(j)} - \mathbf{w}_0^{(j)}\|_2 \leq R$ due to the constraint $\sum_{j=1}^m \|\mathbf{w}^{(j)} - \mathbf{w}_0^{(j)}\|_2^2 \leq R^2$. We now control $\mathbb{E}\big[\sup_{\mathbf{w}\in\mathcal{B}_{R,\mathbf{w}_0^{(j)}}} \big\| \sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\mathbf{x}_i \big\|_2\big]$. We only consider the case $j = 1$ for simplicity as the discussion is the same for all $j \in [m]$. By the relationship between Rademacher and Gaussian complexities [3], we know

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\tilde{\mathbf{w}}\in\mathcal{B}_{R,\mathbf{w}_0^{(1)}}} \Big\| \sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\mathbf{x}_i \Big\|_2 &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\tilde{\mathbf{w}}\in\mathcal{B}_{R,\mathbf{w}_0^{(1)}}, \tilde{\mathbf{v}}\in\mathbb{R}^{d'}:\|\tilde{\mathbf{v}}\|_2\leq 1} \Big\langle \sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\mathbf{x}_i, \tilde{\mathbf{v}} \Big\rangle \\
&\leq \Big(\frac{\pi}{2}\Big)^{\frac{1}{2}} \mathbb{E}_{\boldsymbol{g}} \sup_{\tilde{\mathbf{w}}\in\mathcal{B}_{R,\mathbf{w}_0^{(1)}}, \tilde{\mathbf{v}}\in\mathbb{R}^{d'}:\|\tilde{\mathbf{v}}\|_2\leq 1} \Big\langle \sum_{i=1}^n g_i \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\mathbf{x}_i, \tilde{\mathbf{v}} \Big\rangle, \qquad \text{(C.4)}
\end{aligned}$$

where $g_1, \ldots, g_n$ are independent standard Gaussian random variables. Define two Gaussian processes as follows

$$\mathfrak{X}_{\tilde{\mathbf{w}},\tilde{\mathbf{v}}} = \sum_{i=1}^n g_i \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\langle \mathbf{x}_i, \tilde{\mathbf{v}} \rangle, \quad \mathfrak{Y}_{\tilde{\mathbf{w}},\tilde{\mathbf{v}}} = \sqrt{2} G_\gamma \sum_{i=1}^n g_{i,1} \langle \mathbf{x}_i, \tilde{\mathbf{v}} \rangle + \sqrt{2} R_{\mathbf{x}} L_\gamma \sum_{i=1}^n g_{i,2} \mathbf{x}_i^\top \tilde{\mathbf{w}}, \qquad \text{(C.5)}$$

where $g_{i,1}, g_{i,2}, i \in [n]$ are independent standard Gaussian random variables. Then it is clear that

$$\mathbb{E}_{\boldsymbol{g}}\Big[\big(\mathfrak{X}_{\tilde{\mathbf{w}},\tilde{\mathbf{v}}} - \mathfrak{X}_{\tilde{\mathbf{w}}',\tilde{\mathbf{v}}'}\big)^2\Big] = \mathbb{E}_{\boldsymbol{g}}\Big(\sum_{i=1}^n g_i \big(\gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\langle \mathbf{x}_i, \tilde{\mathbf{v}} \rangle - \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}}')\langle \mathbf{x}_i, \tilde{\mathbf{v}}' \rangle \big) \Big)^2 = \sum_{i=1}^n \Big(\gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\langle \mathbf{x}_i, \tilde{\mathbf{v}} \rangle - \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}}')\langle \mathbf{x}_i, \tilde{\mathbf{v}}' \rangle \Big)^2.$$

It then follows from the standard inequality $(a+b)^2 \leq 2a^2 + 2b^2$ that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}}\Big[\Big(\mathfrak{X}_{\tilde{\mathbf{w}},\tilde{\mathbf{v}}} - \mathfrak{X}_{\tilde{\mathbf{w}}',\tilde{\mathbf{v}}'}\Big)^2\Big] \\
&\leq 2\sum_{i=1}^n \Big(\gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\langle \mathbf{x}_i, \tilde{\mathbf{v}} \rangle - \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\langle \mathbf{x}_i, \tilde{\mathbf{v}}' \rangle \Big)^2 + 2\sum_{i=1}^n \Big(\gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\langle \mathbf{x}_i, \tilde{\mathbf{v}}' \rangle - \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}}')\langle \mathbf{x}_i, \tilde{\mathbf{v}}' \rangle \Big)^2 \\
&\leq 2G_\gamma^2 \sum_{i=1}^n \big(\langle \mathbf{x}_i, \tilde{\mathbf{v}} \rangle - \langle \mathbf{x}_i, \tilde{\mathbf{v}}' \rangle \big)^2 + 2\sum_{i=1}^n \|\mathbf{x}_i\|_2^2 \big(\gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}}) - \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}}') \big)^2 \qquad \text{(C.6)} \\
&\leq 2G_\gamma^2 \sum_{i=1}^n \big(\langle \mathbf{x}_i, \tilde{\mathbf{v}} \rangle - \langle \mathbf{x}_i, \tilde{\mathbf{v}}' \rangle \big)^2 + 2R_{\mathbf{x}}^2 L_\gamma^2 \sum_{i=1}^n \big(\mathbf{x}_i^\top \tilde{\mathbf{w}} - \mathbf{x}_i^\top \tilde{\mathbf{w}}' \big)^2 = \mathbb{E}_{\boldsymbol{g}}\big[\big(\mathfrak{Y}_{\tilde{\mathbf{w}},\tilde{\mathbf{v}}} - \mathfrak{Y}_{\tilde{\mathbf{w}}',\tilde{\mathbf{v}}'} \big)^2 \big], \qquad \text{(C.7)}
\end{aligned}$$

where we have used the L_γ-smoothness of γ in the last inequality. It then follows from Lemma 19 that

$$\mathbb{E}_{\boldsymbol{g}} \sup_{(\tilde{\mathbf{w}},\tilde{\mathbf{v}})} \sum_{i=1}^n g_i \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}}) \langle \mathbf{x}_i, \tilde{\mathbf{v}} \rangle \le \mathbb{E}_{\boldsymbol{g}} \sup_{(\tilde{\mathbf{w}},\tilde{\mathbf{v}})} \Big(\sqrt{2} G_\gamma \sum_{i=1}^n g_{i,1} \langle \mathbf{x}_i, \tilde{\mathbf{v}} \rangle + \sqrt{2} L_\gamma R_{\mathbf{x}} \sum_{i=1}^n g_{i,2} \mathbf{x}_i^\top \tilde{\mathbf{w}} \Big)$$

$$\le \sqrt{2} G_\gamma \mathbb{E}_{\boldsymbol{g}} \sup_{\tilde{\mathbf{v}}:\|\tilde{\mathbf{v}}\|_2 \le 1} \sum_{i=1}^n g_i \langle \mathbf{x}_i, \tilde{\mathbf{v}} \rangle + \sqrt{2} L_\gamma R_{\mathbf{x}} \mathbb{E}_{\boldsymbol{g}} \sup_{\tilde{\mathbf{w}} \in \mathcal{B}_{R,\mathbf{w}_0^{(1)}}} \sum_{i=1}^n g_i \mathbf{x}_i^\top \tilde{\mathbf{w}}$$

$$= \sqrt{2} G_\gamma \mathbb{E}_{\boldsymbol{g}} \sup_{\tilde{\mathbf{v}}:\|\tilde{\mathbf{v}}\|_2 \le 1} \sum_{i=1}^n g_i \langle \mathbf{x}_i, \tilde{\mathbf{v}} \rangle + \sqrt{2} L_\gamma R_{\mathbf{x}} \mathbb{E}_{\boldsymbol{g}} \sup_{\tilde{\mathbf{w}} \in \mathcal{B}_{R,\mathbf{w}_0^{(1)}}} \sum_{i=1}^n g_i \mathbf{x}_i^\top (\tilde{\mathbf{w}} - \mathbf{w}_0^{(1)}) \tag{C.8}$$

$$= \sqrt{2} G_\gamma \mathbb{E}_{\boldsymbol{g}} \Big\| \sum_{i=1}^n g_i \mathbf{x}_i \Big\|_2 + \sqrt{2} L_\gamma R_{\mathbf{x}} R \mathbb{E}_{\boldsymbol{g}} \Big\| \sum_{i=1}^n g_i \mathbf{x}_i \Big\|_2 \le \sqrt{2} \big(G_\gamma + L_\gamma R_{\mathbf{x}} R \big) \Big(\sum_{i=1}^n \|\mathbf{x}_i\|_2^2 \Big)^{\frac{1}{2}}, \tag{C.9}$$

where we have used the following inequality due to the concavity of $t \mapsto \sqrt{t}$

$$\mathbb{E}_{\boldsymbol{g}} \Big\| \sum_{i=1}^n g_i \mathbf{x}_i \Big\|_2 \le \Big(\mathbb{E}_{\boldsymbol{g}} \Big\| \sum_{i=1}^n g_i \mathbf{x}_i \Big\|_2^2 \Big)^{\frac{1}{2}} = \Big(\mathbb{E}_{\boldsymbol{g}} \sum_{i=1}^n \sum_{j=1}^n \langle g_i \mathbf{x}_i, g_j \mathbf{x}_j \rangle \Big)^{\frac{1}{2}} = \big(\sum_{i=1}^n \|\mathbf{x}_i\|_2^2 \big)^{\frac{1}{2}}.$$

We combine the above inequality and Eq. (C.4) together, and derive

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\tilde{\mathbf{w}} \in \mathcal{B}_{R,\mathbf{w}_0^{(1)}}} \Big\| \sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}}) \mathbf{x}_i \Big\|_2 \le \sqrt{\pi} \big(G_\gamma + L_\gamma R_{\mathbf{x}} R \big) \Big(\sum_{i=1}^n \|\mathbf{x}_i\|_2^2 \Big)^{\frac{1}{2}}.$$

We plug this bound back into Eq. (C.3), and derive

$$\mathbb{E}_{\boldsymbol{\sigma}} \Big[\sup_{\mathbf{w} \in \mathcal{W}} \Big\| \sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)}) \mathbf{x}_i \Big\|_2^2 \Big] \le \pi \big(G_\gamma + L_\gamma R_{\mathbf{x}} R \big)^2 \sum_{i=1}^n \|\mathbf{x}_i\|_2^2 + G_\gamma^2 \sum_{i=1}^n \|\mathbf{x}_i\|_2^2.$$

We plug this inequality back into Eq. (C.2), and derive

$$\mathfrak{R}_S(\widetilde{\mathcal{G}}) \le \frac{1}{\sqrt{m}n} \Big(\sum_{j=1}^m \Big(\pi \big(G_\gamma + L_\gamma R_{\mathbf{x}} R \big)^2 + G_\gamma^2 \Big) \sum_{i=1}^n \|\mathbf{x}_i\|_2^2 \Big)^{\frac{1}{2}} \le \frac{\sqrt{\pi}(G_\gamma + L_\gamma R_{\mathbf{x}} R) + G_\gamma}{n} \Big(\sum_{i=1}^n \|\mathbf{x}_i\|_2^2 \Big)^{\frac{1}{2}}.$$

The proof is complete. □

Proof of Theorem 8. By Lemma 4 and Lemma 7, we know

$$\mathfrak{R}_{\widetilde{X},2n}(\widetilde{\mathcal{F}}) \le \max_{I \subseteq [2n]} \Big(\frac{|I| \mathfrak{R}_{S_{\mathbf{x}},|I|}(\widetilde{\mathcal{G}})}{2n} + \frac{G_\Psi \beta |\bar{I}| \mathfrak{R}_{S_{\mathbf{x}},|\bar{I}|}(\mathcal{G})}{n} \Big) = \widetilde{O}\Big(\frac{R_{\mathbf{x}}}{n} \max_{I \subseteq [2n]} \Big(\big(G_\gamma + L_\gamma R_{\mathbf{x}} R \big) |I|^{\frac{1}{2}} + G_\Psi \beta |\bar{I}|^{\frac{1}{2}} G_\gamma R \Big) \Big)$$

$$= \widetilde{O}\Big(\frac{R_{\mathbf{x}}}{n} \Big((G_\gamma + L_\gamma R_{\mathbf{x}} R)^2 + G_\Psi^2 \beta^2 G_\gamma^2 R^2 \Big)^{\frac{1}{2}} \max_{I \subseteq [2n]} \big(|\bar{I}| + |I| \big)^{\frac{1}{2}} \Big) = \widetilde{O}\Big(\frac{R_{\mathbf{x}}}{\sqrt{n}} \big(G_\gamma + L_\gamma R_{\mathbf{x}} R + G_\Psi \beta G_\gamma R \big) \Big),$$

where we have used the Cauchy-Schwarz's inequality. We then plug the above inequality and Lemma 7 back into Theorem 3, and derive the stated inequality. □

C.2 Proofs on SNNs with ReLU Activation

We now turn to the proofs for SNNs with the ReLU activation function. To this aim, we introduce the definition of VC dimension and its connection to Gaussian complexity.

Definition 6 (VC dimension)**.** Let $\mathcal{H} = \{h : \mathcal{X} \mapsto \{0,1\}\}$. For any set $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\} \subset \mathcal{X}$ of cardinality n, define $\mathcal{H}_A = \{(h(\mathbf{a}_1), \ldots, h(\mathbf{a}_n)) : h \in \mathcal{H}\}$. We say that the VC dimension of $\mathcal{H}$ is $d_{\rm vc}$ if there exists a set $A \subset \mathcal{X}$ of cardinality $d_{\rm vc}$ such that $|\mathcal{H}_A| = 2^{d_{\rm vc}}$, and for any set A of cardinality $k > d_{\rm vc}$ we have $|\mathcal{H}_A| < 2^k$.

The following lemma shows the connection between Gaussian complexity and VC dimension, e.g., Theorem 8.3.23 in [40].

Lemma 21. *$\mathcal{H} = \{h : \mathcal{X} \mapsto \{0,1\}\}$ and $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\} \subset \mathcal{X}$. If the VC dimension of $\mathcal{H}$ is $d_{\rm vc}$, then we have $\frac{1}{n}\sup_{h\in\mathcal{H}} \mathbb{E}_{\mathbf{g}} \sum_{i=1}^n g_i h(\mathbf{a}_i) \lesssim \big(\frac{d_{\rm vc}}{n}\big)^{\frac{1}{2}}$.*

Proof of Lemma 9. Note Eq. (C.3) and Eq. (C.6) do not use the smoothness of γ and therefore still hold for the ReLU activation function. Define $\mathfrak{X}_{\tilde{\mathbf{w}},\tilde{\mathbf{v}}}$ by Eq. (C.5) and $\mathfrak{Y}_{\tilde{\mathbf{w}},\tilde{\mathbf{v}}} = \sqrt{2}G_\gamma \sum_{i=1}^n g_{i,1}\langle \mathbf{x}_i, \tilde{\mathbf{v}}\rangle + \sqrt{2}R_{\mathbf{x}} \sum_{i=1}^n g_{i,2}\gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})$, where $g_{i,1}, g_{i,2}, i \in [n]$ are independent standard Gaussian random variables. Then, Eq. (C.6) shows that $\mathbb{E}_{\boldsymbol{g}}\big[\big(\mathfrak{X}_{\tilde{\mathbf{w}},\tilde{\mathbf{v}}} - \mathfrak{X}_{\tilde{\mathbf{w}}',\tilde{\mathbf{v}}'}\big)^2\big] \le \mathbb{E}_{\boldsymbol{g}}\big[\big(\mathfrak{Y}_{\tilde{\mathbf{w}},\tilde{\mathbf{v}}} - \mathfrak{Y}_{\tilde{\mathbf{w}}',\tilde{\mathbf{v}}'}\big)^2\big]$. By Eq. (C.4) and similar to Eq. (C.8), we can show that

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\tilde{\mathbf{w}}\in\mathcal{B}_{R,\mathbf{w}_0^{(1)}}} \Big\| \sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\mathbf{x}_i \Big\|_2 \le \Big(\frac{\pi}{2}\Big)^{\frac{1}{2}} \mathbb{E}_{\boldsymbol{g}} \sup_{\tilde{\mathbf{w}}\in\mathcal{B}_{R,\mathbf{w}_0^{(1)}}, \tilde{\mathbf{v}}\in\mathbb{R}^{d'}:\|\tilde{\mathbf{v}}\|_2\le 1} \sum_{i=1}^n g_i \gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\big\langle \mathbf{x}_i, \tilde{\mathbf{v}}\big\rangle \\
&\le \mathbb{E}_{\boldsymbol{g}} \sup_{\tilde{\mathbf{w}}\in\mathcal{B}_{R,\mathbf{w}_0^{(1)}}, \tilde{\mathbf{v}}\in\mathbb{R}^{d'}:\|\tilde{\mathbf{v}}\|_2\le 1} \Big(\sqrt{\pi}G_\gamma \sum_{i=1}^n g_{i,1}\langle \mathbf{x}_i, \tilde{\mathbf{v}}\rangle + \sqrt{\pi}R_{\mathbf{x}} \sum_{i=1}^n g_{i,2}\gamma'(\mathbf{x}_i^\top \tilde{\mathbf{w}})\Big) \\
&\le \sqrt{\pi}G_\gamma \Big(\sum_{i=1}^n \|\mathbf{x}_i\|_2^2\Big)^{\frac{1}{2}} + \sqrt{\pi}R_{\mathbf{x}} \mathbb{E}_{\boldsymbol{g}} \sup_{\tilde{\mathbf{w}}\in\mathcal{B}_{R,\mathbf{w}_0^{(1)}}} \sum_{i=1}^n g_{i,2}\mathbb{I}_{[\mathbf{x}_i^\top \tilde{\mathbf{w}}\ge 0]} \lesssim R_{\mathbf{x}}\sqrt{n} + R_{\mathbf{x}}\big(n\tilde{d}\big)^{\frac{1}{2}},
\end{aligned}
$$

where we have used the fact that the VC dimension of linear function class is $\{\mathbf{x} \mapsto \mathbb{I}_{[\mathbf{x}^\top\tilde{\mathbf{w}}\ge 0]} : \tilde{\mathbf{w}} \in \mathbb{R}^{\tilde{d}}\}$ is $\tilde{d}$, e.g., Chapter 8.3 in [40], Lemma 21 and $G_\gamma = 1$. It then follows from Eq. (C.3) that

$$
\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\mathbf{w}\in\mathcal{W}} \Big\| \sum_{i=1}^n \sigma_i \gamma'(\mathbf{x}_i^\top \mathbf{w}^{(j)})\mathbf{x}_i \Big\|_2^2\Big] \lesssim \big(R_{\mathbf{x}}\big(n\tilde{d}\big)^{\frac{1}{2}}\big)^2 + \sum_{i=1}^n \|\mathbf{x}_i\|_2^2.
$$

This together with Eq. (C.2) implies the stated bound. The proof is complete. □

References

[1] R. Arora, R. Bassily, T. González, C. A. Guzmán, M. Menart, and E. Ullah. Faster rates of convergence to stationary points in differentially private optimization. In *International Conference on Machine Learning*, pages 1060–1092. PMLR, 2023.

[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. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4): 1497–1537, 2005.

[5] 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.

[6] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[7] O. Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, École Polytechnique: Department of Applied Mathematics Paris, France, 2002.

[8] A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[9] D. Davis and D. Drusvyatskiy. Graphical convergence of subgradients in nonconvex optimization and learning. *Mathematics of Operations Research*, 47(1):209–231, 2022.

[10] J. Fan and Y. Lei. High-probability generalization bounds for pointwise uniformly stable algorithms. *Applied and Computational Harmonic Analysis*, 70:101632, 2024.

[11] C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018.

[12] D. J. Foster, A. Sekhari, and K. Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, pages 8759–8770, 2018.

[13] H. Fu, Y. Chi, and Y. Liang. Guaranteed recovery of one-hidden-layer neural networks via cross entropy. *IEEE transactions on Signal Processing*, 68:3225–3235, 2020.

[14] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[15] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[16] N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. *Information and Inference: A Journal of the IMA*, 9(2):473–504, 2020.

[17] P. Gopalani and A. Mukherjee. Global convergence of sgd on two layer neural nets. *Information and Inference: A Journal of the IMA*, 14(1):iaae035, 2025.

[18] Y. Guermeur. l_p-norm sauer–shelah lemma for margin multi-category classifiers. *Journal of Computer and System Sciences*, 89:450–473, 2017.

[19] F. Hellström, G. Durisi, B. Guedj, and M. Raginsky. Generalization bounds: Perspectives from information theory and pac-bayes. *Foundations and Trends® in Machine Learning*, 18(1):1–223, 2025.

[20] 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. PMLR, 2025.

[21] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM*, 68(2):1–29, 2021.

[22] A. Ledent, R. Alves, and Y. Lei. Generalization bounds for rank-sparse neural networks. In *Advances in Neural Information Processing Systems*, volume 38, pages 147927–147996, 2025.

[23] Y. Lei and K. Tang. Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4505–4511, 2021.

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

[25] Y. Lei and Y. Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819, 2020.

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

[27] S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[28] H. Reeve and A. Kaban. Optimistic bounds for multi-output learning. In *International Conference on Machine Learning*, pages 8030–8040. PMLR, 2020.

[29] 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.

[30] 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.

[31] M. Schliserman and T. Koren. Tight risk bounds for gradient descent on separable data. In *Advances in Neural Information Processing Systems*, volume 36, pages 68749–68759, 2023.

[32] E. Schnoor, A. Behboodi, and H. Rauhut. Generalization error bounds for iterative recovery algorithms unfolded as neural networks. *Information and Inference: A Journal of the IMA*, 12(3):2267–2299, 2023.

[33] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

[34] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge university press, 2004.

[35] 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.

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

[37] 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.

[38] H. Taheri, C. Thrampoulidis, and A. Mazumdar. Sharper guarantees for learning neural network classifiers with gradient methods. In *International Conference on Learning Representations*, 2025.

[39] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pages 11–30. Springer, 2015.

[40] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[41] A. Watkins, E. Ullah, T. Nguyen-Tang, and R. Arora. Optimistic rates for multi-task representation learning. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[42] Y. Xu. *Algorithm Design and Localization Analysis in Sequential and Statistical Learning*. Columbia University, 2023.

[43] Y. Xu and A. Zeevi. Towards optimal problem dependent generalization error bounds in statistical learning theory. *Mathematics of Operations Research*, 2024.

[44] Z. Yang, Y. Bai, and S. Mei. Exact gap between generalization error and uniform convergence in random feature models. In *International Conference on Machine Learning*, pages 11704–11715. PMLR, 2021.

[45] L. Zhang, T. Yang, and R. Jin. Empirical risk minimization for stochastic convex optimization: $\mathrm{O}(1/n)$- and $\mathrm{O}(1/n^2)$-type of risk bounds. In *Conference on Learning Theory*, pages 1954–1979, 2017.

[46] S. Zhang, Y. Hu, L. Zhang, and N. He. Generalization bounds of nonconvex-(strongly)-concave stochastic minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 694–702. PMLR, 2024.

[47] T. Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

[48] X. Zhang, Y. Yu, L. Wang, and Q. Gu. Learning one-hidden-layer relu networks via gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 1524–1534, 2019.

[49] P. Zhao, Y.-J. Zhang, L. Zhang, and Z.-H. Zhou. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *Journal of Machine Learning Research*, 25(98):1–52, 2024.

[50] B. Zhu, S. Li, M. Yi, and Y. Liu. Stability and sharper risk bounds with convergence rate $\widetilde{O}(1/n^2)$. In *Advances in Neural Information Processing Systems*, volume 38, pages 17961–17997, 2025.