Loss Gradient Gaussian Width based Generalization and Optimization Guarantees

Arindam Banerjee
Univ of Illinois Urbana-Champaign
arindamb@illinois.edu

Qiaobo Li
Univ of Illinois Urbana-Champaign
qiaobol2@illinois.edu

Yingxue Zhou
Univ of Minnesota, Twin Cities
zhou0877@umn.edu

Abstract

Generalization and optimization guarantees on the population loss often rely on uniform convergence based analysis, typically based on the Rademacher complexity of the predictors. The rich representation power of modern models has led to concerns about this approach. In this paper, we present generalization and optimization guarantees in terms of the complexity of the gradients, as measured by the Loss Gradient Gaussian Width (LGGW). First, we introduce generalization guarantees directly in terms of the LGGW under a flexible gradient domination condition. Second, we show that sample reuse in ERM does not make the empirical gradients deviate from the population gradients as long as the LGGW is small. Third, focusing on deep networks, we bound their single-sample LGGW in terms of the Gaussian width of the featurizer, i.e., the output of the last-but-one layer. To our knowledge, our generalization and optimization guarantees in terms of LGGW are the first results of its kind, and hold considerable promise towards quantitatively tight bounds for deep models.

1 INTRODUCTION

Machine learning theory typically characterizes generalization behavior in terms of uniform convergence guarantees, often in terms of Rademacher complexities of predictor function classes (Koltchinskii and Panchenko, 2000; Bartlett and Mendelson, 2002; Shalev-Shwartz and Ben-David, 2014; Mohri, 2018). However, given the rich representation power of deep models, it has been difficult to develop sharp uniform bounds, and concerns have been raised whether non-vacuous bounds following the uniform convergence route may be possible (Nagarajan and Kolter, 2019; Negrea et al., 2020; Kakade et al., 2008, 2012; Bartlett et al., 2017; Golowich et al., 2018). On the other hand, the empirical successes of deep learning has highlighted the importance of gradients, which play a critical role in learning over-parameterized models with deep representations (Bottou et al., 2018; Goodfellow et al., 2016; LeCun et al., 2015). The motivation behind this paper is to investigate generalization and optimization guarantees based on uniform convergence analysis of gradients, rather than the predictors.

Empirical work has demonstrated that gradients of overparameterized deep learning models are typically "simple," e.g., only a small fraction of entries have large values (Papyan, 2018, 2019; Li et al., 2020; Wu et al., 2020; Xie et al., 2024), gradients typically lie in a low-dimensional space (Gur-Ari et al., 2018; Ghorbani et al., 2019; Rothchild et al., 2020; Ivkin et al., 2019), using top-k components with suitable bias corrections work well in practice (Rothchild et al., 2020; Ivkin et al., 2019), etc. In this paper, we view such simplicity in terms of a generic chaining based covering of the gradients, which can be represented as the loss gradient Gaussian width (LGGW) (Talagrand, 2014; Vershynin, 2018). Based on the empirical evidence, we posit that loss gradients of modern machine learning models have small LGGW. Our main technical results in the paper present generalization and optimization guarantees in terms of the LGGW. While our main technical results hold irrespective of whether the LGGW is small or not, similar to their classical Rademacher complexity counterparts, the results imply sharper bounds if the LGGW is small. Further, while deep learning gradients form the motivation behind the work, some of the technical results we present are not restricted to deep learning and hold for general models under simple regularity conditions.

Proceedings of the 28$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

In Section 2, we present generalization bounds in terms generic chaining over gradients and, equivalently, the LGGW. Our analysis needs a mechanism to transition from losses to gradients of losses, and we assume the population loss function satisfies a flexible gradient domination (GD) condition (Foster et al., 2018), which includes the PL (Polyak-Łojasiewicz) condition (Polyak, 1963) used in modern deep learning (Liu et al., 2020a,b, 2022) as a simple special case. Then, based on recent advances in uniform convergence of vector-valued functions (Maurer, 2016; Maurer and Pontil, 2016; Foster et al., 2018), one can readily get generalization bounds based on the vector Rademacher complexity of gradients. Our main new result on generalization is to show that such vector Rademacher complexity can be bounded in terms of the LGGW, which makes the bound explicitly depend on the geometry of gradients. Our proof is based on generic chaining (GC) (Talagrand, 2014), where we first bound the vector Rademacher complexity using a hierarchical covering argument, and then bound the hierarchical covering in terms of the LGGW using the majorizing measure theorem (Talagrand, 2014). To our knowledge, this is the first generalization bound in terms of Gaussian width of loss gradients, and the results imply that models with small LGGW will generalize well.

In Section 3, we present optimization guarantees based on the LGGW. One big disconnect between the theory and practice of (stochastic) gradient descent (GD) is that the theory assumes fresh samples in each step whereas in practice one reuses samples. If one just focuses on the finite sum empirical risk minimization (ERM), sample reuse is not an issue. However, the goal of such optimization is to reduce the population loss, and sample reuse derails the standard analysis. We show that even with sample reuse, the discrepancy between the sample average and the population gradient can be bounded by the LGGW along with only a logarithmic dependence on the ambient dimensionality and the number of steps of optimization, i.e., the number of times the samples has been reused. Thus, if the LGGW is small, the sample average gradients stay close to the population gradients, even under sample reuse. We also use these results to establish population convergence results for GD with sample reuse.

In Section 4, we present the first results on bounding LGGW of deep learning gradients. We consider formal models for feed-forward and residual networks widely studied in the deep learning theory literature (Allen-Zhu et al., 2019; Du et al., 2019; Arora et al., 2019b; Liu et al., 2022; Banerjee et al., 2022) and show that their LGGW can be bounded by the Gaussian width of the featurizer, i.e., the output of the last-but-one layer. We also demonstrate benefits of architecture choices, e.g., hidden layers being wider, last-but-one layer being narrower, which yields smaller single-sample LGGW.

Notation: c_1, c_2, c_3, etc., denote constants, and they may mean different constants at different places. We use $\tilde{O}$ to hide poly-log terms.

2 GAUSSIAN WIDTH BASED GENERALIZATION BOUNDS

In this section, we formalize the notion of Loss Gradient Gaussian Width (LGGW) (Vershynin, 2018; Talagrand, 2014), and as the main result, establish generalization bounds in terms of LGGW. For parameters $\theta \in \Theta \subset \mathbb{R}^p$ and data points $z \in \mathcal{Z}$ where $z = (\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$, for any suitable loss $\ell(\theta; z) : \Theta \times \mathcal{Z} \mapsto \mathbb{R}_+$, the gradients of interest are $\xi(\theta; z) = \nabla_\theta \ell(\theta; z) \in \mathbb{R}^p$. Given a dataset of n samples, we start by defining the sets of all empirical average gradients and n-tuple of individual gradients:

Definition 1 (Gradient Sets). *Given n samples $z^{(n)} = \{z_1, \ldots, z_n\} \in \mathcal{Z}^{(n)}$, for a suitable choice of parameter set Θ and loss $\ell : \Theta \times \mathcal{Z} \mapsto \mathbb{R}_+$, let the set of all possible n-tuple of individual loss gradients and scaled stacked gradients be respectively denoted as:*

$$\hat{\Xi}^{(n)}(\Theta) := \left\{ \hat{\xi}^{(n)} = \{\xi_i, i \in [n]\} \mid \exists \theta \in \Theta, s.t. \xi_i = \nabla \ell(\theta; z_i) \right\} \tag{1}$$

$$\hat{\Xi}_n(\Theta) := \left\{ \hat{\xi}_n \in \mathbb{R}^{np} \mid \exists \theta \in \Theta, s.t. \hat{\xi}_n = \frac{1}{\sqrt{n}} \left[\nabla \ell(\theta; z_i) \right]_{i=1}^n \right\} \tag{2}$$

where $\left[\nabla \ell(\theta; z_i) \right]_{i=1}^n$ denotes the n p-dimensional gradients stacked to create $\hat{\xi}_n \in \mathbb{R}^{np}$.

The scaling of $\frac{1}{\sqrt{n}}$ in the stacked gradient vector $\hat{\xi}_n$ is maintain the L_2-norm at the same scale as individual gradient components, i.e., with $\|\nabla \ell(\theta; z_i)\|_2 = O(1)$, we will have $\|\hat{\xi}_n\|_2 = O(1)$.

For convenience, we will denote the set of n-tuples of gradients in (1) as $\hat{\Xi}^{(n)}$ and the empirical average gradient in (2) as $\hat{\Xi}_n$. Note that the sets $\hat{\Xi}^{(n)}$ and $\hat{\Xi}_n$ are conditioned on $z^{(n)}$, and the results we will present will hold for any $z^{(n)} \in \mathcal{Z}^n$, i.e., any n samples under suitable regularity conditions (see Section 4.2.1). Next, we define the Loss Gradient Gaussian Width (LGGW):

Definition 2 (Loss Gradient Gaussian Width). *Given the scaled stacked gradient set $\hat{\Xi}_n$ as in Definition 1, the LGGW is defined as:*

$$w(\hat{\Xi}_n) := \mathbb{E}_{\mathbf{g} \sim \mathcal{N}(0, \mathbb{I}_{np \times np})} \left[\sup_{\hat{\xi}_n \in \hat{\Xi}_n} \langle \hat{\xi}_n, \mathbf{g} \rangle \right] . \tag{3}$$

Gaussian width measures how much a standard Gaussian vector can align with a set, capturing its complexity (Vershynin, 2018). As a special case of generic chaining, it estimates this alignment using expected supremum over Gaussian processes (Talagrand, 2014).

2.1 Bounds based on Vector Rademacher Complexity

Our approach to generalization bounds is based on vector Rademacher complexities. Unlike the traditional approach which focus on bounds based on Rademacher complexities of the set of predictors $f(\theta; x)$ (Koltchinskii and Panchenko, 2000; Bartlett and Mendelson, 2002; Shalev-Shwartz and Ben-David, 2014; Mohri, 2018), we will develop generalization bounds based on vector Rademacher complexities (Maurer, 2016; Maurer and Pontil, 2016; Foster et al., 2018) of the set of loss gradients $\xi(\theta, z) = \nabla\ell(\theta; z)$. In particular, following Foster et al. (2018), we consider *Normed Empirical Rademacher Complexity* (NERC), which is the extension of standard Rademacher complexity to sets of vectors or vector-valued functions.

Definition 3. *In the setting of Definition 1, for the set $\hat{\Xi}^{(n)}$ of n-tuples of individual gradients, the NERC is defined as*

$$\hat{R}_n(\hat{\Xi}^{(n)}) := \frac{1}{n}\mathbb{E}_{\varepsilon^{(n)}}\left[\sup_{\hat{\xi}^{(n)}\in\hat{\Xi}^{(n)}}\left\|\sum_{i=1}^{n}\varepsilon_i\xi_i\right\|\right], \tag{4}$$

where $\varepsilon^{(n)}$ are a set of n i.i.d. Rademacher random variables, i.e., $P(\varepsilon_i = +1) = P(\varepsilon_i = -1) = \frac{1}{2}$, and $\hat{\Xi}^{(n)}$ is the set of n-tuples of gradients as in (1).

For (4), we present results for the L_2 norm $\|\cdot\|_2$, but the analyses tools and results have counterparts for general norms (Ledoux and Talagrand, 2013) and such results will be in terms of γ-functions in generic chaining (Talagrand, 2014).

Given a loss function ℓ and a distribution $\mathcal{D}$ over examples $z \in \mathcal{Z}$, we consider the minimization of the *population risk* function, that is, $\min_\theta \mathcal{L}_{\mathcal{D}}(\theta) := \mathbb{E}_{z\sim\mathcal{D}}[\ell(\theta; z)]$. The learner does not observe the distribution $\mathcal{D}$ and typically minimize the *empirical risk* $\hat{\mathcal{L}}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}\ell(\theta; z_i)$, where $z^{(n)} \sim \mathcal{D}^n$ are i.i.d. draws. While the loss functions are usually nonconvex in modern machine learning, including deep learning, they usually satisfy a form of the *gradient domination* (GD) condition (Foster et al., 2018), which includes the popular Polyak-Łojasiewicz (PL) condition as a simple special case (Polyak, 1963; Liu et al., 2020a,b, 2022). Further, we assume the gradients to be bounded.

Assumption 1 (Population Risk: Gradient Domination). *The population risk $\mathcal{L}_{\mathcal{D}}$ satisfies the $(\alpha, \bar{c}_\alpha)$-GD condition with respect to a norm $\|\cdot\|$, if there are constants $\bar{c}_\alpha > 0$, $\alpha \in [1, 2]$, such that $\forall\theta \in \Theta$ and $\theta^\star \in \operatorname{argmin}_{\theta\in\Theta}\mathcal{L}_{\mathcal{D}}(w)$, we have $\mathcal{L}_{\mathcal{D}}(\theta) - \mathcal{L}_{\mathcal{D}}(\theta^\star) \le \bar{c}_\alpha\|\nabla\mathcal{L}_{\mathcal{D}}(\theta)\|^\alpha$.*

Assumption 2 (Bounded Gradients). *The sample gradients $\xi(\theta; z) = \nabla\ell(\theta; z)$ are bounded, i.e., $\|\xi(\theta; z)\|_2 = O(1), \theta \in \Theta, z \in \mathcal{Z}$.*

Remark 2.1. To understand whether the $(\alpha, \bar{c}_\alpha)$-GD condition in Assumption 1 holds in deep learning models, in Figure 1, we plot the gradient domination ratio $\frac{\mathcal{L}_{\mathcal{D}}(\theta)}{\|\nabla\mathcal{L}_{\mathcal{D}}(\theta)\|^\alpha}$ for $\alpha = 1, 2$ for three standard deep learning models. The plots demonstrate that the condition indeed clearly holds in the optimization trajectory for $\alpha = 1$ with a small constant $\bar{c}_1 < 5$ for all models, and for $\alpha = 2$, i.e., the PL condition (Polyak, 1963; Liu et al., 2020a,b, 2022), with $\bar{c}_2 < 1$ for some models, but needs large constants $\bar{c}_2 > 100$ for for ResNet18 on CIFAR-10. We will establish the theoretical results for any $\alpha \in [1, 2]$. □

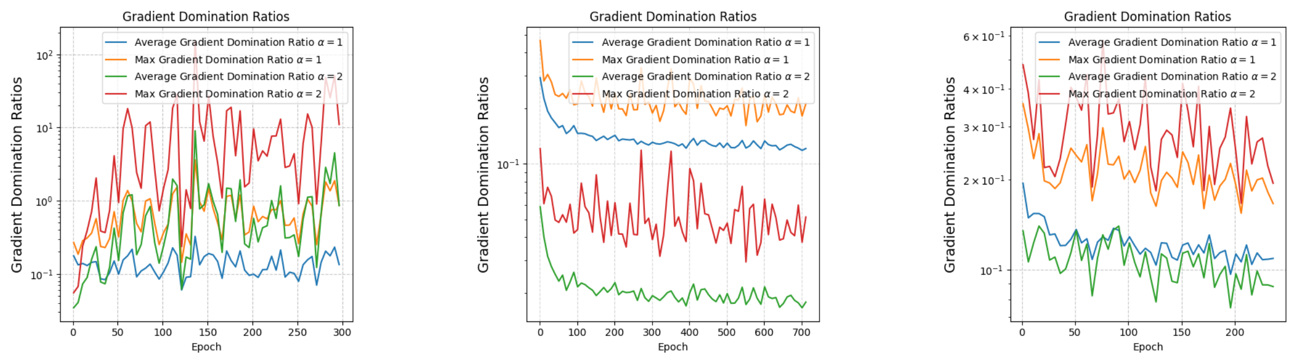
Remark 2.2. The GD condition is more flexible than the PL condition as it allows for a more relaxed relationship between function values and the gradient norm. While both conditions may fail in landscapes with multiple valleys, starting optimization from a pretrained model often positions the model within a single valley, where the GD condition is likely to hold. □

We start with a generalization bound from Foster et al. (2018) for losses satisfying gradient domination.

Proposition 1. *Under Assumptions 1 and 2, with $\theta^\star \in \operatorname{argmin}_{\theta\in\Theta}\mathcal{L}_{\mathcal{D}}(\theta)$ denoting any population loss minimizer and $\hat{R}_n(\hat{\Xi}^{(n)})$ as in Definition 3, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of the samples $z^{(n)} \sim \mathcal{D}^n$, for any $\hat{\theta} \in \Theta$ we have*

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(\hat{\theta}) - \mathcal{L}_{\mathcal{D}}(\theta^*) \le 2\bar{c}_\alpha\Bigg(&\|\nabla\hat{\mathcal{L}}_n(\hat{\theta})\|_2^\alpha \\ &+ 2\left(4\hat{R}_n(\hat{\Xi}^{(n)}) + c_3\frac{\log\frac{1}{\delta}}{n}\right)^\alpha\Bigg) + c_4\left(\frac{\log\frac{1}{\delta}}{n}\right)^{\frac{\alpha}{2}}.\end{aligned} \tag{5}$$

For the bound to be useful, the NERC $\hat{R}_n(\Xi^{(n)})$ needs to be small, e.g., $\hat{R}_n(\hat{\Xi}^{(n)}) = O(\frac{1}{\sqrt{n}})$. Foster et al. (2018) analyzed the NERC for loss functions of the form $\ell(\theta, (x, y)) = h(\psi_\theta(x))$, where $\psi_\theta(x) = \langle\theta, x\rangle$ and h is Lipschitz, so it suffices to consider the gradient of linear predictor $\psi_\theta(x) = \theta^T x$. Since $\nabla\psi_\theta(x) = x$ is independent of θ, as long as each $\|x_i\|$ is bounded (Assumption 2), we have $\frac{1}{n}\mathbb{E}_{\varepsilon^{(n)}}[\sup_{\theta\in\Theta}\|\sum_{i=1}^{n}\varepsilon_i\nabla\phi(\theta, x_i)\|] = \frac{1}{n}\mathbb{E}_{\varepsilon^{(n)}}[\|\sum_{i=1}^{n}\varepsilon_i x_i\|] = \frac{c}{\sqrt{n}}$, by Khintchine's inequality (Ledoux and Talagrand, 2013; Vershynin, 2012) for L_2 norm and under suitable assumptions on certain other norms (Kakade et al., 2008, 2012; Foster et al., 2018). The simplification here is that the $\sup_{\theta\in\Theta}$ drops out, which is only possible for linear predictors. Our analysis considers the general case where the $\sup_\theta$ does not drop out and has to be handled explicitly.



(a) ResNet18 on CIFAR-10 (b) FFN on CIFAR-10 (c) CNN on Fashion-MNIST

Figure 1: Gradient Domination (GD) Ratio for $\alpha = 1, 2$ for (a) ResNet18 on CIFAR-10, (b) FFN on CIFAR-10, and (c) CNN on Fashion-MNIST. Each experiment is repeated 5 times and we report the average and the maximum ratio. The results show that GD holds for $\alpha = 1$ with a small constant $\bar{c}_1 < 5$ for all models, and for $\alpha = 2$, with $\bar{c}_2 < 1$ for some models, but needs large constants $\bar{c}_2 > 100$ for for ResNet18.

2.2 Main Result: Gaussian Width Bound

For our main result, we consider whether the NERC $\hat{R}_n(\hat{\Xi}^{(n)})$ can somehow be related to the Gaussian width $w(\hat{\Xi}_n)$ as in (2), which arguably captures the geometry of the gradients more directly.

Theorem 1. *Based on Definitions 1, 2, and 3, with $\varepsilon^{(n)}$ denoting n i.i.d. Rademacher variables, conditioned on any $z^{(n)} \in \mathcal{Z}^{(n)}$, for any $u > 0$, with probability at least $(1 - c_0 \exp(-u^2/2))$ over the randomness of $\varepsilon^{(n)}$, we have*

$$\sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \frac{1}{\sqrt{n}} \left\| \sum_{i=1}^{n} \varepsilon_i \xi_i \right\|_2 \leq c_5 (1 + u) w(\hat{\Xi}_n) \ . \tag{6}$$

As a result, the normed empirical Rademacher complexity (NERC) satisfies

$$\hat{R}_n(\hat{\Xi}^{(n)}) = \frac{1}{n} \mathbb{E}_{\varepsilon^{(n)}} \left[\sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \left\| \sum_{i=1}^{n} \varepsilon_i \xi_i \right\|_2 \right] \leq \frac{c_6 w(\hat{\Xi}_n)}{\sqrt{n}} \ . \tag{7}$$

One can plug-in the Theorem 1 to Proposition 1 to get the desired generalization bound. For example, with Assumption 1 holding for $\alpha = 1$ (Figure 1), with probability $(1 - \delta)$, for any $\hat{\theta} \in \Theta$ we have

$$\mathcal{L}_{\mathcal{D}}(\hat{\theta}) - \mathcal{L}_{\mathcal{D}}(\theta^*) \leq 2\bar{c}_1 \Bigg(\|\nabla \hat{\mathcal{L}}_n(\hat{\theta})\|_2 + 2 \left(\frac{4 c_6 w(\hat{\Xi}_n)}{\sqrt{n}} + c_3 \frac{\log \frac{1}{\delta}}{n} \right) \Bigg) + c_4 \sqrt{\frac{\log \frac{1}{\delta}}{n}} \ . \tag{8}$$

Remark 2.3. A technically curious aspect of the result in (7) is that the expectation in $\hat{R}_n(\hat{\Xi}^{(n)})$ is from n Rademacher variables over samples whereas that in $w(\hat{\Xi}_n)$ is from p normal variables over gradient components. The shift from samples to gradient components makes the bound geometric. $\square$

Remark 2.4. The bound in (8) does not depend on the algorithm used to reach $\hat{\theta} \in \Theta$. Thus, one can use variants of stochastic gradient descent, adaptive optimization, or sharpness aware minimization (Zaheer et al., 2018; Reddi et al., 2019; Foret et al., 2020) to reach $\hat{\theta}$, and the bound would still hold. $\square$

Remark 2.5. Since the bound is with respect to the Gaussian width of the set of scaled stacked gradients $\hat{\xi}_n = \frac{1}{\sqrt{n}} \left[\nabla \ell(\theta; z_i) \right]_{i=1}^{n}$, it would depend on both the alignment of the n sample gradients $\xi_i = \nabla \ell(\theta; z_i)$ as well as the width of the set of individual sample gradients $\Xi_i = \left\{ \hat{\xi}_i \in \mathbb{R}^p \mid \exists \theta \in \Theta, \text{ s.t. } \hat{\xi}_i = \nabla \ell(\theta; z_i) \right\}$. Depending on the alignment over the n sample gradients, the width can range from $O(1)$ to $O(\sqrt{n})$, similar to Rademacher complexity (Shalev-Shwartz and Ben-David, 2014; Bartlett and Mendelson, 2002). The nature of alignment here will be based on similarity of the hierarchical covering of individual gradient sets, with similar coverings, e.g., similar distortions at each level of the hierarchy, yielding small width. For the width of individual sample gradient sets, we present bounds for two commonly used networks: FeedForward Networks and Residual Networks in Section 4. $\square$

2.3 Proof sketch of Main Result

Our analysis is based on generic chaining (GC) and needs the following definition (Talagrand, 2014).

Definition 4. *For a metric space (T, d), an admissible sequence of T is a collection of subsets of T, $\Gamma = \{T_r : r \geq 0\}$, with $|T_0| = 1$ and $|T_r| \leq 2^{2^r}$ for all $r \geq 1$. For $\beta \geq 1$, the γ_β functional is defined by*

$$\gamma_\beta(T, d) := \inf_{\Gamma} \sup_{t \in T} \sum_{r=0}^{\infty} 2^{r/\beta} d(t, T_r) \ , \tag{9}$$

where the infimum is over all admissible sequences Γ.

Remark 2.6. There has been substantial developments on GC over the past two decades (Talagrand, 2014, 1996). GC lets one develop sharp upper bounds to suprema of stochastic processes indexed in a set with a metric structure in terms of γ_β functions. Typically, the bounds are in terms of γ_2 functions for sub-Gaussian processes and a mix of γ_1 and γ_2 functions for sub-exponential processes. $\square$

To get to and understand the proof of Theorem 1, we first recap the standard GC argument. For any suitable stochastic process under consideration, say $\{X_\xi\}$, the key step is to find a suitable pseudo-metric $d(\xi_1, \xi_2)$ that satisfies *increment condition* (IC) (Talagrand, 2014): for any $u > 0$

$$\mathbb{P}(|X_{\xi_1} - X_{\xi_2}| \geq u d(\xi_1, \xi_2)) \leq 2\exp(-u^2/2) . \quad (10)$$

Then, under some simple assumptions, e.g., $X_\xi = 0$ for some $\xi \in \Xi$, GC shows that $\mathbb{E}[\sup_{\xi \in \Xi} X_\xi] \leq c\gamma_2(\Xi, d)$, for some constant c, and there is a corresponding high probability version of the result (Talagrand, 2014). The analysis is typically done using the canonical distance

$$d(\xi_1, \xi_2) := (\mathbb{E}[(X_{\xi_1} - X_{\xi_2})^2])^{1/2} , \quad (11)$$

though one can use other (not canonical) distance $\bar{d}(\xi_1, \xi_2)$ which satisfies the increment condition (10).

At a high level, our proof has three parts: First, since the NERC $\hat{R}_n(\hat{\Xi}^{(n)})$ is defined in terms of the set of n-tuples $\hat{\Xi}^{(n)}$, we start with a suitable (non canonical) distance $\bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)})$ over n-tuples in $\hat{\Xi}^{(n)}$ and establish a version of Theorem 1 in terms of $\gamma_2(\hat{\Xi}^{(n)}, \bar{d}^{(n)})$, showing $\hat{R}_n(\hat{\Xi}^{(n)}) \leq c_2 \frac{\gamma_2(\hat{\Xi}^{(n)}, \bar{d}^{(n)})}{\sqrt{n}}$.

Second, for the set $\hat{\Xi}_n$ of stacked gradients $\hat{\xi}_n$, we construct a Gaussian process $X_{\hat{\xi}_n} = \langle \hat{\xi}_n, \mathbf{g} \rangle$ where $\mathbf{g} \sim \mathcal{N}(0, \mathbb{I}_{np \times np})$. We note a simple one-to-one correspondence between the set of stacked gradients $\hat{\Xi}_n$ and the set of n-tuples $\hat{\Xi}^{(n)}$ since $\hat{\xi}_n = \frac{1}{\sqrt{n}} [\xi_i]_{i=1}^n$. Further, we show that the canonical distance $d(X_{\hat{\xi}_{1,n}}, X_{\hat{\xi}_{2,n}})$ as in (11) from the Gaussian process on $\hat{\Xi}_n$ satisfies $d(X_{\hat{\xi}_{1,n}}, X_{\hat{\xi}_{2,n}}) = \bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)})$, the non-canonical distance on $\hat{\Xi}^{(n)}$. Based on the set correspondence and exact same distance measure, we have $\gamma_2(\hat{\Xi}_n, d) = \gamma_2(\hat{\Xi}^{(n)}, \bar{d}^{(n)})$, so that $\hat{R}_n(\hat{\Xi}^{(n)}) \leq c_2 \frac{\gamma_2(\hat{\Xi}_n, d)}{\sqrt{n}}$.

Third, the final step of the analysis is to show that $\gamma_2(\hat{\Xi}_n, d) \leq c_4 w(\hat{\Xi}_n)$. The result follows directly from the Majorizing Measure Theorem (Talagrand, 2014) since we are working with a Gaussian process indexed on $\hat{\Xi}_n$ and using the canonical distance d. All proofs are in Appendix A.

3 GRADIENT DESCENT WITH SAMPLE REUSE

If the goal is to make sure GD decreases the population loss $\mathcal{L}(\theta)$, then one typically assumes that fresh samples are used in every iteration t to compute $\hat{\mathcal{L}}_n(\theta_t)$ and corresponding $\nabla\hat{\mathcal{L}}_n(\theta_t)$, making sure such estimates are unbiased estimates of the population loss gradient. In practice, however, one reuses samples from the fixed training set for GD and also for mini-batch stochastic gradient descent. Such sample reuse violates the fresh sample assumption, and in turn cannot guarantee the unbiasedness of the gradient estimates (Chen and Banerjee, 2018; De Stefani and Upfal, 2019). While one can still guarantee decrease of the finite sum empirical loss from the ERM perspective (Shalev-Shwartz et al., 2010; Ghadimi and Lan, 2012; Nemirovski et al., 2009) with GD, it is unclear if the population loss should decrease with such updates. Thus, the theory and practice of (stochastic) GD and variants have diverged. If ERM optimization is done via sample reuse, one needs to then use uniform convergence (UC) to get population results (Shalev-Shwartz et al., 2010; Shalev-Shwartz and Ben-David, 2014). While this approach is well established, UC analysis over large function classes cannot quite take advantage of the geometry of individual stochastic gradients (Mei et al., 2018; Amir et al., 2022; Davis and Drusvyatskiy, 2022).

In this section, we show that empirical estimates of the gradient with sample reuse stay close to the population gradients as long as the LGGW is small. We start by consider the joint event of interest under such sequential sample reuse. The initialization parameter θ_0 for gradient descent is assumed to be independent of the samples, e.g., based on random initialization or a pre-trained model. But such a parameter can land anywhere in some Θ, so we need a uniform bound on that set. The subsequent parameters $\theta_t = \theta_{t-1} - \eta\nabla\hat{\mathcal{L}}_n(\theta_{t-1})$ are obtained sequentially and thus has dependency on all the samples, which will be reused to estimate $\nabla\hat{\mathcal{L}}_n(\theta_t)$. Our main result shows that the deviation between the empirical gradient and the population gradient jointly over the $(T+1)$ time steps can be bounded with high probability based on the LGGW and otherwise logarithmic dependencies on the ambient dimensionality p and the time horizon T.

Theorem 2. *Let* $\theta_0 \in \Theta_0$*, and* $\theta_t, t \in [T]$ *be a sequence of parameters obtained from GD by reusing a fixed set of samples* $z^{(n)} \sim \mathcal{D}^n$ *in each epoch. Let* $\Delta(\theta) := \left\| \frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta, z_i) - \nabla\mathcal{L}_D(\theta) \right\|_2$*, where the population gradient* $\nabla\mathcal{L}_D(\theta_t) = \mathbb{E}_{z\sim\mathcal{D}}[\nabla\ell(\theta_t; z)], t \in [T]$*. Under Assumption 2, with* $w(\hat{\Xi}_n^0)$ *denoting the LGGW for* Θ_0 *as in Definition 2, for any* $\delta \leq \frac{1}{2}$*, with probability* *at least* $(1-2\delta)$ *over* $z^{(n)} \sim \mathcal{D}^n$, *we have*

$$\max\left(\sup_{\theta_0\in\Theta_0} \Delta(\theta_0), \quad \max_{t\in[T]} \Delta(\theta_t)\right) \leq \frac{c_7 \max(w(\hat{\Xi}_n^0), \log p) + \sqrt{\log T + \log \frac{1}{\delta}}}{\sqrt{n}} . \tag{12}$$

Our proof does not utilize the specific form of GD and goes through for any update of the form $\theta_{t+1} = f(\{\theta_s, s \leq t\})$. Thus, Theorem 2 is an adaptive data analysis (ADA) result, but without using the traditional tools such as differential privacy (Dwork et al., 2015a; Jung et al., 2019). Our proof utilizes the conditional probability of events, and expresses conditional or adaptive events in terms of events based on fresh samples. Proofs are in the supplementary material.

Remark 3.1. There are a few key takeaways from Theorem 2. First, note that the key term in the bound depends on LGGW $w(\hat{\Xi}_n^0)$ only at initialization, since we considered a uniform bound over $\theta_0 \in \Theta_0$. Note that $w(\hat{\Xi}_n^0)$ can be bounded based on properties of gradients at random initialization or with pretrained models (see Section 4). Second, the bound has a logarithmic dependence on p, the ambient dimension, and T, the number of steps. Third, the sample dependence is $\frac{1}{\sqrt{n}}$ for GD with batch size n. Thus, the bound says that in spite of sample reuse, the sample gradients will stay close to population gradient. □

Remark 3.2. The bound in Theorem 2 does not restrict the subsequent iterates θ_t to be restricted to the initialization set Θ_0. In other words, the bound holds without a "near initialization" type restricted which has been widely used in the deep learning theory literature. Further, in case the analysis starts from a given fixed θ_0, the dependence on $w(\hat{\Xi}_n^0)$ drops out, and we are only left with dependence on $(\log p + \sqrt{\log T})$. □

Population Convergence in Optimization. The results in Theorems 2 can be readily used to straightforwardly to get convergence of population gradients, without needing an additional uniform convergence argument (Shalev-Shwartz and Ben-David, 2014).

Theorem 3. *Consider a non-convex loss* $\mathcal{L}_{\mathcal{D}}(\theta) = \mathbb{E}_{z\sim\mathcal{D}}\left[\ell(\theta, z)\right]$ *with* τ*-Lipschitz gradient. For GD using a total of* $z^{(n)} \sim \mathcal{D}^n$ *original samples and reusing the* n *samples in each step as in Theorem 2, with step-size* $\eta = \frac{1}{4\tau}$, *with probability at least* $(1-2\delta)$ *for any* $\delta > 0$ *over* $z^{(n)} \sim \mathcal{D}^n$, $\mathbb{E}_R \left\|\nabla \mathcal{L}_{\mathcal{D}}\left(\theta_R\right)\right\|^2 \leq O\left(\frac{1}{T}\right) + O\left(\frac{\max\left(w^2(\hat{\Xi}_n^0), \log^2 p\right) + \left(\log T + \log\frac{1}{\delta}\right)}{n}\right)$ *where* R *is uniformly distributed over* $\{1, \ldots, T\}$ *and the expectation is over the randomness of* R.

Remark 3.3. In the bound, the first term comes from the finite-sum optimization error, which depends on the number of iterations T; and the second term comes from the statistical error, which depends on the sample size n and the LGGW $w(\hat{\Xi}_n^0), \log p$, and $\log T$. □

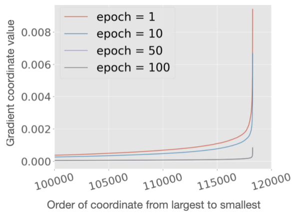
Remark 3.4. While an analysis of sample reuse based GD can be done using differential privacy based adaptive data analysis (ADA) (Dwork et al., 2015b; Jung et al., 2019) and algorithmic stability (Bousquet and Elisseeff, 2002; Hardt et al., 2016), the best-known efficient algorithm in such framework (Bassily et al., 2014) have gradient error scaling with $\sqrt{p}$. Further, without any assumption, lower bounds in ADA (Steinke and Ullman, 2015) also show that it is hard to release more than $\tilde{O}(n^2)$ statistical queries. □

4 GAUSSIAN WIDTH OF DEEP LEARNING GRADIENTS

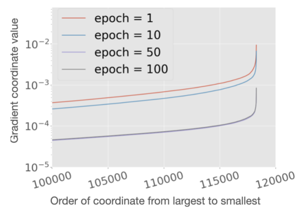
In this section, we initiate the study of loss gradient Gaussian width (LGGW) of deep learning gradients. As a warm up, we review that bounded L_2 norm, e.g., $\|\xi(\theta; z)\|_2 \leq c$ for any $\theta \in \Theta$ and $z \in \mathcal{Z}$, with no additional structure, leads to a large Gaussian width since for such $\hat{\Xi}_n$, $w(\hat{\Xi}_n) = \mathbb{E}_{\mathbf{g}}[\sup_{\hat{\xi}_n} \langle \hat{\xi}_n, \mathbf{g}\rangle] \leq c\mathbb{E}_{\mathbf{g}}[\langle \mathbf{g}/\|\mathbf{g}\|_2, \mathbf{g}\rangle] = c\mathbb{E}_{\mathbf{g}}[\|\mathbf{g}\|_2] = O(\sqrt{p})$. We posit that the Gaussian width of deep learning gradients is much smaller and present a set of technical results in support. We show that under standard assumptions, the LGGW can be bounded in terms of the Gaussian width of the featurizer, i.e., output of the last-but-one layer, for both feedforward and residual networks. The results emphasize the benefits of the stability of featurizers for deep learning.

4.1 Bounds Based on Empirical Geometry of Gradients

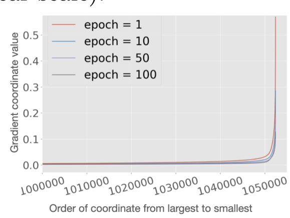
In the context of deep learning, recent work has illustrated that empirically the gradients are considerably structured (Gur-Ari et al., 2018; Papyan, 2018, 2019; Ghorbani et al., 2019; Li et al., 2020): during optimization, the gradients are "simple" in the sense that the sorted absolute gradient components decay quickly and the components have large values only for a small subset of parameters (e.g., see Figure 2), and at convergence, almost all gradients are really close to zero, the so-called interpolation condition (Belkin et al., 2019; Negrea et al., 2020; Ma et al., 2018; Bartlett et al., 2020). Such observations have formed the basis of more memory or communication-efficient algorithms, e.g., based on top-k count-sketch, truncated SGD, and heavy hitters based analysis (Rothchild et al., 2020; Ivkin et al., 2019). We augment these empirical observations with some examples of gradient sets which have small Gaussian widths by construction or by suitable composition properties.



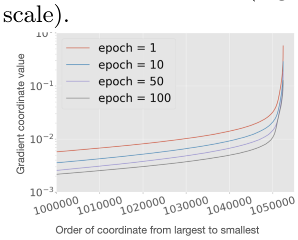
(a) Sorted gradient coordinate absolute values (linear scale).



(b) Sorted gradient coordinate absolute values (log scale).



(c) Sorted gradient coordinate absolute values (linear scale).



(d) Sorted gradient coordinate absolute values (log scale).

Figure 2: (a-b) Sorted gradient coordinate in linear scale (a) and log scale (b) for SGD over 100 epochs on CIFAR-10. Model: 4-hidden-layer ReLU network with 256 nodes on each layer, with around 1,055,000 parameters. (c-d) Sorted gradient coordinate in linear scale (c) and log scale (d) for SGD over 100 epochs on MNIST. Model: 2-layer ReLU with 128 nodes on each layer, with roughly 120,000 parameters. Y-axis is the absolute value of gradient coordinates, X-axis is sorted in increasing order.

Example 4.1 (**Ellipsoid).** Consider $\hat{\Xi}_n = \{\hat{\xi}_n \in \mathbb{R}^p \mid \sum_{j=1}^p ((\hat{\xi}_n[j])^2/a_j^2 = \tilde{O}(1), a \in \mathbb{R}^p\}$. Then, the Gaussian width $w(\hat{\Xi}_n) = \tilde{O}(\|\mathbf{a}\|_2)$ (Talagrand, 2014). Note that if $\|\mathbf{a}\|_2 = \tilde{O}(1)$, then we have $w(\hat{\Xi}_n) = \tilde{O}(1)$. For example, if the elements of $\mathbf{a}$ sorted in decreasing order satisfy $a_{(j)} \leq c_3/\sqrt{j}$, then $\|\mathbf{a}\|_2 = \tilde{O}(1)$. $\square$

Example 4.2 (**k-support norm).** The k-support norm

$$\|\xi\|_k^{sp} = \inf_{\sum_i u_i = \xi} \left\{ \sum_i \|u_i\|_2 \mid \|u_i\|_0 \leq k \right\}$$

is based on an infimum convolution over k-sparse vectors (Argyriou et al., 2012; McDonald et al., 2014; Chen and Banerjee, 2015), with $k = 1$ corresponding to the L_1 norm. Consider $\hat{\Xi}_n = \{\hat{\xi}_n \mid \|\hat{\xi}_n\|_k^{sp} \leq c_0\sqrt{k}\}$. Note that among many other things $\hat{\Xi}_n$ includes the infimum convolution of all k-sparse gradients $\hat{\xi}_n$ with each non-zero element bounded by c_0. By construction, $w(\hat{\Xi}_n) \leq c_1 k\sqrt{\log(p/k)}$ (Chen and Banerjee, 2015). $\square$

Example 4.3 (**Union of sets).** If $\hat{\Xi}_n$ is an union of p^r sets, i.e., polynomial in p, where each set has a $\tilde{O}(1)$ Gaussian width and each element has $\tilde{O}(1)$ L_2 norm, then $w(\hat{\Xi}_n) = \tilde{O}(\sqrt{r})$ (Maurer et al., 2014; Chen and Banerjee, 2015). Taking convex hull of such $\hat{\Xi}_n$ can only change its Gaussian width by a constant amount, i.e., $w(\text{conv}(\hat{\Xi}_n)) \leq c_2 w(\hat{\Xi}_n)$ (Talagrand, 2014)(Theorem 2.4.15). $\square$

Example 4.4 (**Minkowski sum of sets).** If $\hat{\Xi}_n$ is Minkowski sum of a union of p^r sets as in Example 4.3 and a $\sqrt{k}$ radius k-support norm ball as in Example 4.2, then $w(\hat{\Xi}_n) = \max(k, \sqrt{r})\tilde{O}(1)$. $\square$

Based on how the gradients empirically look for deep networks (Figures 2), it is arguably fair to assume that the gradients live in $\hat{\Xi}_n$ with small Gaussian width. We prove explicit upper bounds of single-sample LGGW for FFNs and ResNets in Sections 4.2.1 and 4.2.2 respectively.

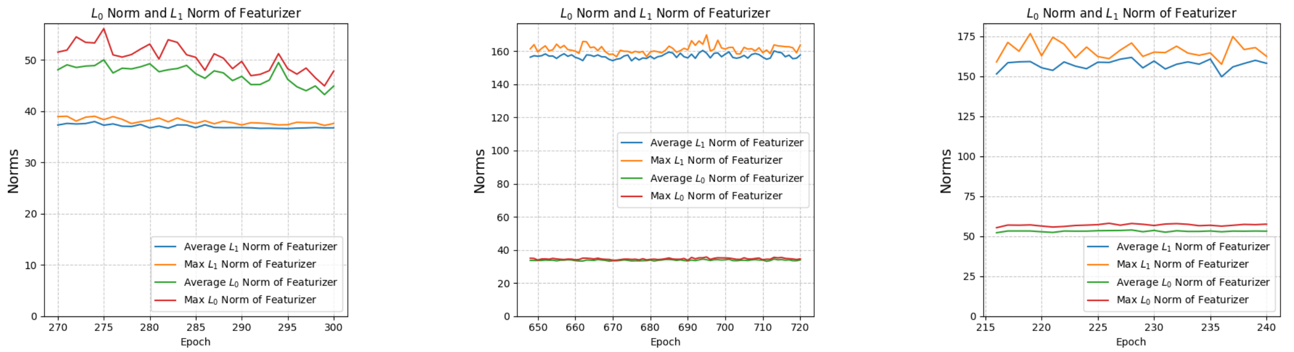
4.2 Gaussian Width Bounds for Neural Networks

Consider a training set $z^{(n)} = \{z_i, i \in [n]\}, z_i = (\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d, y_i \in \mathcal{Y} \subseteq \mathbb{R}$. For some loss function l, the goal is to minimize: $\hat{\mathcal{L}}_n(\theta) = \frac{1}{n}\sum_{i=1}^n l(y_i, \hat{y}_i) = \frac{1}{n}\sum_{i=1}^n l(y_i, f(\theta; \mathbf{x}_i))$, where the prediction $\hat{y}_i := f(\theta; \mathbf{x}_i)$ is from a neural network, and the parameter vector $\theta \in \mathbb{R}^p$. For our analysis, we will assume square loss $l(y_i, \hat{y}_i) = \frac{1}{2}(y_i - \hat{y}_i)^2$, though the analyses extends to more general losses under mild additional assumptions. We focus on feedforward networks (FFNs) and residual networks (RestNets), and consider them as L layer neural networks $f(\theta; \mathbf{x})$ with parameters θ, where each layer has width m. We make two standard assumptions (Allen-Zhu et al., 2019; Du et al., 2019; Liu et al., 2020b; Banerjee et al., 2022).

Assumption 3 (**Activation).** *The activation function ϕ is 1-Lipschitz, i.e., $|\phi'| \leq 1$ and $\phi(0) = 0$.*

Assumption 4 (**Random initialization).** *Let $\theta_0 := (\text{vec}(W_0^{(1)})^\top, \ldots, \text{vec}(W_0^{(L)})^\top, \mathbf{v}_0^\top)^\top$ denote the initial weights, and $w_{0,ij}^{(l)}$ denote i,j-th element of $W_0^{(l)}$ for $l \in [L]$. At initialization, $w_{0,ij}^{(l)} \sim \mathcal{N}(0, \sigma_0^{(l)2})$ for $l \in [L]$ where $\sigma_0^{(1)} = \frac{\sigma_1}{2\left(1+\sqrt{\frac{\log m_1}{2m_1}}\right)}$, $\sigma_0^{(l)} = \frac{\sigma_1}{1+\sqrt{\frac{m_{l-1}}{m_l}}+\sqrt{\frac{2\log m_l}{m_l}}}$, $l \geq 2$, $\sigma_1 > 0$, and $\|\mathbf{v}_0\|_2 = 1$. For any input $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, $\|\mathbf{x}\|_2 = 1$.*

Assumption 3 is satisfied by commonly used activation functions such as (non-smooth) ReLU (Allen-Zhu et al., 2019) and (smooth) GeLU (Hendrycks and Gimpel, 2016). And Assumption 4 is regarding the random initialization of the weights, variants of which are typically used in most work in deep learning theory (Jacot et al., 2018; Allen-Zhu et al., 2019; He et al., 2015; Glorot and Bengio, 2010; Sutskever et al., 2013; Arora et al., 2019a; Du et al., 2018, 2019).



(a) ResNet18 on CIFAR-10 (b) FFN on CIFAR-10 (c) CNN on Fashion-MNIST

Figure 3: L_0 norm and L_1 norm of featurizers for (a) ResNet18 on CIFAR-10, (b) FFN on CIFAR-10, and (c) CNN on Fashion-MNIST of the last 10% epochs close to convergence. We plot both the average and maximum norms among 5 repetitions for each experiment. The figures demonstrate that for these models, as approaching convergence, the featurizers has relatively low L_0 norm and L_1 norm, which further indicates a small LGGW.

4.2.1 Bounds for FeedForward Networks (FFNs)

We consider f to be an FFN given by $f(\theta;x) = \mathbf{v}^\top \phi(\frac{1}{\sqrt{m_L}} W^{(L)} \phi(\cdots \phi(\frac{1}{\sqrt{m_1}} W^{(1)}\mathbf{x}))))$, where $W^{(1)} \in \mathbb{R}^{m \times d}, W^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}, l \in \{2,\ldots,L\}$ are the layer-wise weight matrices, $\mathbf{v} \in \mathbb{R}^{m_L}$ is the last layer vector, $\phi(\cdot)$ is the (pointwise) activation function, and the total set of parameters $\theta \in \mathbb{R}^p, p = \sum_{k=1}^L m_k m_{k-1} + m_L$ is given by

$$\theta = (\mathrm{vec}(W^{(1)})^\top, \ldots, \mathrm{vec}(W^{(L)})^\top, \mathbf{v}^\top)^\top. \tag{13}$$

in which $m_0 = d$. Our analysis for both FFNs and ResNets will be for gradients over all parameters in a fixed radius *spectral* norm ball around the initialization θ_0, which is more flexible than Frobenius norm balls typically used in the literature (Woodworth et al., 2020; Chizat et al., 2019; Zou et al., 2020; Banerjee et al., 2022):

$$\begin{aligned} B^{\mathrm{Spec}}_{\rho,\rho_1}(\theta_0) := \big\{ \theta \in \mathbb{R}^p \text{ as in (13) } \mid \|\mathbf{v} - \mathbf{v}_0\|_2 \\ \leq \rho_1, \|W^{(\ell)} - W_0^{(\ell)}\|_2 \leq \rho, \ell \in [L] \big\} . \end{aligned} \tag{14}$$

Let $B^{\mathrm{Spec}}_{\rho}(\theta_0) \in \mathbb{R}^{p-m}$ be the corresponding set of weight matrices $W = \{W^{(l)}, \ell \in [L]\}$. The main implications of our results and analyses go through as long as the layerwise spectral radius $\rho = o(\sqrt{m_l})$ for all $l \in [L]$, and last layer radius $\rho_1 = O(1)$, which are arguably practical.

Let $h^{(L)}(W,\mathbf{x}) := \phi\left(\frac{1}{\sqrt{m_L}} W^{(L)} \phi(\cdots \phi(\frac{1}{\sqrt{m_1}} W^{(1)}\mathbf{x})))\right) \in \mathbb{R}^{m_L}$ be the "featurizer," i.e., output of the last but one layer, so that $f(\theta;\mathbf{x}) = \mathbf{v}^\top h^{(L)}(W,\mathbf{x})$ with $\theta = (W,\mathbf{v})$. Consider the set of all featurizer values for a single sample:

$$A^{(L)}(\mathbf{x}) = \left\{ h^{(L)}(W,\mathbf{x}) \mid W \in B^{\mathrm{Spec}}_{\rho}(\theta_0) \right\}. \tag{15}$$

We now establish an upper bound on the single-sample LGGW of FFN gradient set Ξ^{ffn} corresponding to parameters in $B^{\mathrm{Spec}}_{\rho,\rho_1}(\theta_0)$. In particular, we show that the single-sample LGGW can effectively be bounded by the Gaussian width of the sample average of the featurizer.

Theorem 4 (LGGW: FFNs). *Under Assumptions 3 and 4, with $\beta_l := \sigma_1 + \frac{\rho}{\sqrt{m_l}}$, $l \in [L]$, with probability at least $\left(1 - \sum_{l=1}^L \frac{2}{m_l}\right)$ over the randomness of the initialization, we have*

$$w(\Xi^{\mathrm{ffn}}) \leq c_1 w(A^{(L)}) + c_2(1+\rho_1)\sqrt{m_L}\left(\prod_{l=1}^L \beta_l\right) \sum_{l=1}^L \frac{1}{\beta_l\sqrt{m_l}}.$$

Remark 4.1. For the simplest case, with $m_l = m$ for all $l \in [L]$, we have that $w(\Xi^{\mathrm{ffn}}) \leq c_1 w(A^{(L)}) + c_2 L(1+\rho_1)\beta^{L-1}$, where $\beta = \sigma_1 + \frac{\rho}{\sqrt{m}}$. Note that choosing $\sigma_1 < 1 - \frac{\rho}{\sqrt{m}}$, i.e., mildly small initialization variance, satisfies $\beta < 1$. When $\beta < 1$, we have $w(\Xi^{\mathrm{ffn}}) = O(1)$ as long as $L = \Omega(\frac{\log(\sqrt{m}+L(1+\rho_1))}{\log\frac{1}{\beta}})$, which holds for moderate depth L. To see this, note that we have $w(\Xi^{\mathrm{ffn}}) = O(w(A^{(L)}) + L(1+\rho_1)\beta^{L-1})$. Further, we have whp $\|\alpha^{(L)}\|_2 \leq \beta^L$ (Lemma 7 in Appendix C), so that $w(A^{(L)}) \leq c\beta^L\sqrt{m}$. Then, for $\beta < 1$, the bound $\beta^L\sqrt{m} + \beta^{L-1}L(1+\rho_1) = O(1)$ if $L = \Omega(\frac{\log(\sqrt{m}+L(1+\rho_1))}{\log\frac{1}{\beta}})$. Interestingly, sufficient depth L helps control the Gaussian width. □

Remark 4.2. In practice, the dimensions of the hidden layers $m_l, l \in [L-1]$ are always much larger than the dimension of the output layer m_L, i.e., $m_l \gg m_L$, and our bound is even better under such situations. To see this, first note that $1 - \sum_{l=1}^L \frac{2}{m_l} \gg 1 - \frac{2L}{m_L}$. Second, for the second term in the bound which depends on parts other than the featurizer, $\sqrt{m_L}\prod_{l=1}^L \beta_l \sum_{l=1}^L \frac{1}{\beta_l\sqrt{m_l}} \ll L\beta_L^{L-1}$, since $\beta_L > \beta_l$.

Therefore, the probability is higher and the bound is smaller when $m_l \gg m_L$, as in real neural networks. $\square$

Remark 4.3. The result implies that under suitable conditions, the Gaussian width of gradients over all $p = \sum_{k=1}^{d} m_k m_{k-1} + m_L$ parameters can be reduced to the Gaussian width of the featurizer, which is m_L dimensional. If the featurizer does feature selection, e.g., as in variants of invariant risk minimization (IRM) (Arjovsky et al., 2019; Zhou et al., 2022) or sharpness aware minimization (SAM) (Foret et al., 2020; Andriushchenko et al., 2023), then the Gaussian width of the featurizer will be small. On the other hand, if the featurizer has many spurious and/or random features (Rosenfeld et al., 2020; Zhou et al., 2022), then the Gaussian width will be large. $\square$

4.2.2 Bounds for Residual Networks (ResNets)

We consider f to be a ResNet given by $\alpha^{(l)}(\mathbf{x}) = \alpha^{(l-1)}(\mathbf{x}) + \phi\left(\frac{1}{L\sqrt{m_l}} W^{(l)} \alpha^{(l-1)}(\mathbf{x})\right), l \in [L]$, and $f(\theta; \mathbf{x}) = \alpha^{(L+1)}(\mathbf{x}) = \mathbf{v}^\top \alpha^{(L)}(\mathbf{x})$, where $\alpha^{(0)}(\mathbf{x}) = \mathbf{x}$, and $W^{(l)}, l \in [L], \mathbf{v}, \phi$ are as in FFNs. Let $h^{(L)}(W, \mathbf{x}) := \alpha^{(L)}(\mathbf{x}) \in \mathbb{R}^{m_L}$ be the "featurizer," i.e., output of the last layer, so that $f(\theta; \mathbf{x}) = \mathbf{v}^\top h^{(L)}(W, \mathbf{x})$ with $\theta = (W, \mathbf{v})$. Note that the scaling of the residual layers is assumed to have a smaller scaling factor $\frac{1}{L}$ following standard theoretical analysis of ResNets (Du et al., 2019) and also aligns with practice where smaller initialization variance yields state-of-the-art performance (Zhang et al., 2021). With the featurizer and associated sets defined as in (15) for FFNs, we now establish an upper bound on the single-sample LGGW of ResNet gradients in terms of that of the featurizer.

Theorem 5 (**LGGW: ResNets**). *Under Assumptions 3 and 4, with $\beta_l := \sigma_1 + \frac{\rho}{\sqrt{m_l}}$, $l \in [L]$, with probability at least $\left(1 - \sum_{l=1}^{L} \frac{2}{m_l}\right)$ over the randomness of the initialization, we have*

$$
\begin{aligned}
w(\Xi^{\mathrm{rn}}) &\leq c_1 w(A^{(L)}) \\
&\quad + c_2 \frac{1+\rho_1}{L} \sqrt{m_L} \prod_{l=1}^{L} \left(1 + \frac{\beta_l}{L}\right) \sum_{l=1}^{L} \frac{1}{\left(1 + \frac{\beta_l}{L}\right)\sqrt{m_l}}.
\end{aligned}
$$

Remark 4.4. For the simplest case, with $m_l = m$ for all $l \in [L]$, we have that $w(\Xi^{\mathrm{rn}}) \leq c_1 w(A^{(L)}) + c_2 (1+\rho_1) \left(1 + \frac{\beta}{L}\right)^{L-1} \leq c_1 w(A^{(L)}) + c_2 (1+\rho_1) e^{\beta}$. in which $\beta = \sigma_1 + \frac{\rho}{\sqrt{m}}$, and the second inequality holds since $\left(1 + \frac{\beta}{L}\right)^{L-1} \leq e^{\beta}$. As for FFNs, choosing $\sigma_1 < 1 - \frac{\rho}{\sqrt{m}}$, i.e., mildly small initialization variance, satisfies $\beta < 1$. When $\beta < 1$, we have $w(\Xi^{\mathrm{rn}}) = O(w(A^{(L)}))$ since $(1+\rho_1)e^{\beta} = O(1)$. Thus, for ResNets, the LGGW on all parameters reduces to the Gaussian width of the featurizer. $\square$

Remark 4.5. As for FFNs, our ResNet bound is also much sharper when the dimensions in the hidden layers are much larger than the output dimension, which is the case in commonly used ResNet models. $\square$

Remark 4.6. To get a sense of the Gaussian width of the featurizer, which will be small if the featurizer is sparse (Vershynin, 2018), in Figure 3, we plot the average and max (over 5 runs) of the L_0 and L_1 norm of the featurizer for the last 10% epochs close to convergence for three standard deep learning models: ResNet18 on CIFAR-10 (512 dimension featurizer), FFN on CIFAR-10 (256 dimension featurizer), and CNN on Fashion-MNIST (128 dimension featurizer). The small L_0 and L_1 norm of the featurizer, indicating both feature selection and being in small L_1 ball, imply small Gaussian width of the featurizer which in turn implies small LGGW based on Theorems 4 and 5. $\square$

5 CONCLUSIONS

In this work, we have introduced a new approach to analyzing the generalization and optimization of learning models based on the loss gradient Gaussian width (LGGW). For loss satisfying the gradient domination condition, which includes the popular PL condition as a special case, we establish generalization bounds in terms of LGGW. In the context of optimization, we show that gradient descent with sample reuse does not go haywire and the gradient estimates stay close to the population counterparts as long as the LGGW at initialization is small. We also consider the LGGW of deep learning models, especially feedforward networks and residual networks. In these settings, we show that the LGGW with single sample can be bounded by the Gaussian width of the featurizer and additional terms which are small under mild conditions. Overall, our work demonstrates the benefits of small Gaussian width in the context of learning, and will hopefully serve as motivation for further study of this geometric perspective especially for modern learning models.

Acknowledgements. The work was supported by grants from the National Science Foundation (NSF) through awards IIS 21-31335, OAC 21-30835, DBI 20-21898, IIS-2002540 as well as a C3.ai research award.

References

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

Idan Amir, Roi Livni, and Nati Srebro. Thinking outside the ball: Optimal learning with gradient descent for generalized linear stochastic convex optimization. *Advances in Neural Information Processing Systems*, 35:23539–23550, 2022.

Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. *Advances in Neural Information Processing Systems*, 36:47032–47051, 2023.

Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. *Advances in Neural Information Processing Systems*, 25, 2012.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

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

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019b.

Arindam Banerjee, Pedro Cisneros-Velarde, Libin Zhu, and Mikhail Belkin. Restricted strong convexity of deep learning models with smooth activations. *arXiv preprint arXiv:2209.15106*, 2022.

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

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE, 2014.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116 (32):15849–15854, 2019.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, February 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Sheng Chen and Arindam Banerjee. Structured estimation with atomic norms: General bounds and applications. *Advances in Neural Information Processing Systems*, 28, 2015.

Sheng Chen and Arindam Banerjee. An improved analysis of alternating minimization for structured multi-response regression. *Advances in Neural Information Processing Systems*, 31, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

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

Lorenzo De Stefani and Eli Upfal. A rademacher complexity based method for controlling power and confidence level in adaptive statistical analysis. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 71–80. IEEE, 2019.

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

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *Advances in neural information processing systems*, 28, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126, 2015b.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. *Advances in neural information processing systems*, 31, 2018.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

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

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Ion Stoica, Raman Arora, et al. Communication-efficient distributed sgd with sketching. *Advances in Neural Information Processing Systems*, 32, 2019.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. *arXiv preprint arXiv:1909.03577*, 2019.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13(1):1865–1890, 2012.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Xinyan Li, Qilong Gu, Yingxue Zhou, Tiancong Chen, and Arindam Banerjee. Hessian based analysis of sgd for deep nets: Dynamics and generalization. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 190–198. SIAM, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *arXiv preprint arXiv:2003.00307*, 7, 2020a.

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

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

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016,* *Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

Andreas Maurer and Massimiliano Pontil. Bounds for vector-valued function estimation. *arXiv preprint arXiv:1606.01487*, 2016.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. An inequality with applications to structured sparsity and multitask dictionary learning. In *Conference on Learning Theory*, pages 440–460. PMLR, 2014.

Andrew M McDonald, Massimiliano Pontil, and Dimitris Stamos. Spectral k-support norm regularization. *Advances in neural information processing systems*, 27, 2014.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Mehryar Mohri. Foundations of machine learning, 2018.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, pages 7263–7272. PMLR, 2020.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Vardan Papyan. The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size. *arXiv preprint arXiv:1811.07062*, 2018.

Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians. In *International Conference on Machine Learning*, pages 5012–5021. PMLR, 2019.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora. Fetchsgd: Communication-efficient federated learning with sketching. In *International Conference on Machine Learning*, pages 8253–8265. PMLR, 2020.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Conference on learning theory*, pages 1588–1628. PMLR, 2015.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

Michel Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103, 1996.

Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60. Springer, 2014.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing*, pages 210–268, 2012.

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

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

Yikai Wu, Xingyu Zhu, Chenwei Wu, Annie Wang, and Rong Ge. Dissecting hessian: Understanding common structure of hessian in neural networks. *arXiv preprint arXiv:2010.04261*, 2020.

Zeke Xie, Qian-Yuan Tang, Mingming Sun, and Ping Li. On the overlooked structure of stochastic gradients. *Advances in Neural Information Processing Systems*, 36, 2024.

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

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

Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. Sparse invariant risk minimization. In *International Conference on Machine Learning*, pages 27222–27244. PMLR, 2022.

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**/No/Not Applicable]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

Loss Gradient Gaussian Width Based Generalization and Optimization Guarantees: Supplementary Materials

A Generalization Bounds: Proofs for Section 2

In this section, we first provide a proof sketch of the main results. Then we provide detail proofs for results in Section 2.

Proof sketch of Main Result We provide additional details and associated Lemma for the first part of the analysis. Consider the stochastic process $X_{\hat{\xi}^{(n)}} := \frac{1}{\sqrt{n}} \|\sum_{i=1}^n \varepsilon_i \xi_i\|_2$, by triangle inequality, we have

$$\left| X_{\hat{\xi}_1^{(n)}} - X_{\hat{\xi}_2^{(n)}} \right| \leq \frac{1}{\sqrt{n}} \left\| \sum_{i=1}^n \varepsilon_i (\xi_{i,1} - \xi_{i,2}) \right\|_2 \ . \tag{16}$$

Thus, for a suitable (non canonical) metric $\bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)})$, it suffices to show an increment condition (IC) of the form: for any $u > 0$, with probability at least $1 - 2\exp(-u^2/2)$ over the randomness of $\varepsilon^{(n)}$, we have

$$\frac{1}{\sqrt{n}} \left\| \sum_{i=1}^n \varepsilon_i (\xi_{i,1} - \xi_{i,2}) \right\|_2 \leq u \bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)}) \ . \tag{17}$$

The in-expectation version of such a result is usually easy to establish using Jensen's inequality, e.g., see (26.16) in Shalev-Shwartz and Ben-David (2014). The high probability version can be more tricky. We focus on a variant of the increment condition which allows for a constant shift, i.e., right hand side of the form $(\mu + u)\bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)})$ where $\mu = O(1)$. First, we consider the non-cannnical distance

$$\bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)}) := \left(\frac{1}{n} \sum_{i=1}^n \|\xi_{i,1} - \xi_{i,2}\|_2^2 \right)^{1/2} \tag{18}$$

which is greater or equal to the canonical distance as in (11), due to Jensen's inequality in (16). With our choice of distance $\bar{d}^{(n)}$, note that $\mathbf{v}_i = \frac{\xi_{i,1} - \xi_{i,2}}{\bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)})}$ satisfies $\sum_{i=1}^n \|\mathbf{v}_i\|_2^2 = n$. Then, the following result is effectively a *shifted increment condition* (SIC), a shifted variant of (17) with a constant shift of $\mu \leq 1$:

Lemma 1. *Let $\mathbf{v}_i \in \mathbb{R}^p, i = 1, \ldots, n$ be a set of vectors such that $\sum_{i=1}^n \|\mathbf{v}_i\|_2^2 \leq n$. Let ε_i be a set of i.i.d. Rademacher random variables. Then, for any $u > 0$,*

$$\mathbb{P}_{\varepsilon^{(n)}} \left(\left| \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \mathbf{v}_i \right\|_2 - \mu \right| \geq u \right) \leq 2 \exp\left(-\frac{u^2}{2} \right), \tag{19}$$

where $\mu \leq 1$ is a positive constant.

Next, we show that such constant shifts can be gracefully handled by GC:

Lemma 2. *Consider a stochastic process $\{X_{\hat{\xi}^{(n)}}\}, \hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}$ which satisfies: for any $u > 0$, with probability at least $1 - c_1 \exp(-u^2/2)$, we have*

$$\left| X_{\hat{\xi}_1^{(n)}} - X_{\hat{\xi}_2^{(n)}} \right| \leq c_0 (\mu + u) \bar{d}^{(n)} \left(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)} \right) \ . \tag{20}$$

Further, assume that $X_{\hat{\xi}_0^{(n)}} = 0$ for some $\hat{\xi}_0^{(n)} \in \hat{\Xi}^{(n)}$. Then, for any $u > 0$, with probability at least $1 - c_3 \exp(-u^2/2)$, we have

$$\sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \left| X_{\hat{\xi}^{(n)}} \right| \leq c_2 (\mu + u) \gamma_2 \left(\hat{\Xi}^{(n)}, \bar{d}^{(n)} \right) \ . \tag{21}$$

Remark A.1. The $X_{\hat{\xi}_0^{(n)}} = 0$ is easy to satisfy for gradients as long as the gradients are all zero for some θ_0, e.g., minima of the loss, interpolation condition, etc. That condition is not necessary, and a more general result can be established, e.g., see Talagrand (2014)[Theorem 2.4.12]. $\square$

The SIC in Lemma 1 and the shifted variant of GC in Lemma 2 can now be used to establish the first part of the proof of Theorem 1. The technicalities for the other parts are more straightforward and are directly handled in the proof of Theorem 1.

A.1 Generalization bound in terms of NERC

We start with a variation of Proposition 1 of Foster et al. (2018), which will be convenient for our purposes.

Proposition 2. *Under Assumptions 1 and 2, with $\theta^\star \in \arg\min_{\theta\in\Theta} L_{\mathcal{D}}(w)$ denoting any population loss minimizer and $\hat{R}_n(\hat{\Xi}^{(n)})$ as in Definition 3, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of the samples $z^{(n)} \sim \mathcal{D}^n$, for any $\hat{\theta} \in \Theta$ we have*

$$\mathcal{L}_{\mathcal{D}}(\hat{\theta}) - \mathcal{L}(\theta^*) \leq 2c_0 \left(\|\nabla \hat{\mathcal{L}}_n(\hat{\theta})\|^\alpha + 2\left(\mathbb{E} \sup_{\theta\in\Theta} \|\nabla \hat{\mathcal{L}}_n(\theta) - \nabla \mathcal{L}_D(\theta)\| \right)^\alpha + c_2 \left(\frac{\log \frac{1}{\delta}}{n} \right)^{\frac{\alpha}{2}} \right) . \tag{22}$$

Proof. Since $\mathcal{L}_{\mathcal{D}}$ satisfies the (α, c_0)-GD condition, since $\alpha \in [1,2]$, we have

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(\hat{\theta}) - \mathcal{L}(\theta^*) &\leq c_0 \|\nabla \mathcal{L}_{\mathcal{D}}(\hat{\theta})\|^\alpha \leq 2c_0 \left(\|\nabla \hat{\mathcal{L}}_n(\hat{\theta})\|^\alpha + \|\nabla \hat{\mathcal{L}}_n(\hat{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\hat{\theta})\|^\alpha \right) \\ &\leq 2c_0 \left(\|\nabla \hat{\mathcal{L}}_n(\hat{\theta})\|^\alpha + \left(\sup_{\theta\in\Theta} \|\nabla \hat{\mathcal{L}}_n(\theta) - \nabla \mathcal{L}_{\mathcal{D}}(\theta)\| \right)^\alpha \right)\end{aligned}$$

Now, by a direct application of McDiarmid's inequality, since $z^{(n)} \sim \mathcal{D}^n$, with probability at least $(1-\delta)$ over the draw of $z^{(n)}$, we have

$$\sup_{\theta\in\Theta} \left\| \frac{1}{n} \sum_{i=1}^n \nabla \ell(\theta; z_i) - \mathbb{E}_z[\nabla \ell(\theta; z)] \right\| \leq \mathbb{E} \sup_{\theta\in\Theta} \left\| \frac{1}{n} \sum_{i=1}^n \nabla \ell(\theta; z_i) - \mathbb{E}_z[\nabla \ell(\theta; z)] \right\| + c \sup_{\substack{\theta\in\Theta \\ z\in\mathcal{Z}}} \|\nabla \ell(\theta; z)\| \sqrt{\frac{\log \frac{1}{\delta}}{n}} .$$

Noting that $\|\nabla \ell(\theta; z)\|_2 = O(1)$ for $\theta \in \Theta, z \in \mathcal{Z}$ completes the proof. $\square$

Next, we effectively restate Proposition 2 of Foster et al. (2018), in our notation.

Proposition 3. *Let $\hat{\Xi}^{(n)}$ be the set of all n-tuple gradients as in (1) and $\hat{R}_n(\hat{\Xi}^{(n)})$ be the NERC (normed empirical Rademacher complexity) as in Definition 3. For any $\delta > 0$, with probability at least $1-\delta$ over the draw of $z^{(n)} \sim \mathcal{D}^n$, we have*

$$\mathbb{E} \sup_{\theta\in\Theta} \left\| \nabla \widehat{L}_n(\theta) - \nabla L_{\mathcal{D}}(\theta) \right\| \leq 4 \cdot \hat{R}_n(\hat{\Xi}^{(n)}) + c_1 \left(\frac{\log\left(\frac{1}{\delta}\right)}{n} \right) . \tag{23}$$

Proof. The proof follows from that of Proposition 2 of Foster et al. (2018). $\square$

Combining Propositions 2 and 3, we now get the following result.

Proposition 1. *Under Assumptions 1 and 2, with $\theta^\star \in \operatorname{argmin}_{\theta\in\Theta} \mathcal{L}_{\mathcal{D}}(\theta)$ denoting any population loss minimizer and $\hat{R}_n(\hat{\Xi}^{(n)})$ as in Definition 3, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of the samples $z^{(n)} \sim \mathcal{D}^n$, for any $\hat{\theta} \in \Theta$ we have*

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(\hat{\theta}) - \mathcal{L}_{\mathcal{D}}(\theta^*) \leq 2\bar{c}_\alpha \Bigg(&\|\nabla \hat{\mathcal{L}}_n(\hat{\theta})\|_2^\alpha \\ &+ 2\left(4\hat{R}_n(\hat{\Xi}^{(n)}) + c_3 \frac{\log \frac{1}{\delta}}{n} \right)^\alpha \Bigg) + c_4 \left(\frac{\log \frac{1}{\delta}}{n} \right)^{\frac{\alpha}{2}} .\end{aligned} \tag{5}$$

A.2 Bounding NERC with Gaussian width

Lemma 1. *Let $\mathbf{v}_i \in \mathbb{R}^p, i = 1, \dots, n$ be a set of vectors such that $\sum_{i=1}^n \|\mathbf{v}_i\|_2^2 \leq n$. Let ε_i be a set of i.i.d. Rademacher random variables. Then, for any $u > 0$,*

$$\mathbb{P}_{\varepsilon^{(n)}}\left(\left| \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \mathbf{v}_i \right\|_2 - \mu \right| \geq u \right) \leq 2 \exp\left(-\frac{u^2}{2} \right), \tag{19}$$

where $\mu \leq 1$ is a positive constant.

We establish the result by an application of the entropy method, and the specific application is just the bounded difference inequality (Boucheron et al., 2013). Recall that a function $f : \mathcal{X}^n \mapsto \mathbb{R}$ is said to have the bounded difference property is for some non-negative constants $c_1, \dots, c_n$, we have for $1 \leq i \leq n$

$$\sup_{\substack{x_1,\dots,x_n \\ x_i' \in \mathcal{X}}} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \leq c_i \ . \tag{24}$$

If f satisfies the bounded difference property, then an application of the Efron-Stein inequality (Boucheron et al., 2013)[Theorem 3.1] implies that $Z = f(X_1, \dots, X_n)$ satisfies the variance $\mathrm{var}(Z) \leq \frac{1}{4} \sum_{i=1}^n c_i^2$ (Boucheron et al., 2013)[Corollary 3.2]. The bounded difference inequality is an application of the entropy method (Boucheron et al., 2013)[Theorem 6.2] which shows that such Z satisfies a sub-Gaussian tail inequality where the role of the sub-Gaussian norm (Vershynin, 2012, 2018) is played by the variance $v = \frac{1}{4} \sum_i c_i^2$, so that

$$\mathbb{P}(|Z - \mathbb{E}[Z]| > u) \leq 2 \exp(-u^2/2v) \ . \tag{25}$$

We now focus on the specific result we need and prove it using the bounded difference inequality.

Proof. Since $\mathbf{v}_i$ are fixed, let

$$f(\varepsilon_i, \dots, \varepsilon_n) := \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \mathbf{v}_i \right\|_2 \ .$$

By triangle inequality, for any $j \in \{1, \dots, n\}$, we have

$$\begin{aligned}
&\sup_{\substack{\varepsilon_1,\dots,\varepsilon_n \\ \varepsilon_j' \in \{-1,+1\}}} |f(\varepsilon_1, \dots, \varepsilon_n) - f(\varepsilon_1, \dots, \varepsilon_{j-1}, \varepsilon_j', \varepsilon_{j+1}, \dots, \varepsilon_n)| \\
&\leq \sup_{\substack{\varepsilon_1,\dots,\varepsilon_n \\ \varepsilon_j' \in \{-1,+1\}}} \left| \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \mathbf{v}_i \right\|_2 - \left\| \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n \varepsilon_i \mathbf{v}_i - \varepsilon_j \mathbf{v}_j + \varepsilon_j' \mathbf{v}_j \right) \right\|_2 \right| \\
&\leq \sup_{\varepsilon_j, \varepsilon_j' \in \{-1,+1\}} \frac{1}{\sqrt{n}} \left\| \varepsilon_j \mathbf{v}_j - \varepsilon_j' \mathbf{v}_j \right\|_2 \\
&\leq \frac{2}{\sqrt{n}} \|\mathbf{v}_j\|_2 \ .
\end{aligned}$$

As a result, the varinace bound $v = \frac{1}{4} \sum_{i=1}^n \frac{2^2}{n} \|\mathbf{v}_i\|_2^2 \leq 1$, since $\sum_{i=1}^n \|\mathbf{v}_i\|_2^2 \leq n$.

Further, we have $\mathbb{E}[f(\varepsilon_1, \dots, \varepsilon_n)] \leq 1$. To see this, first note that

$$\mathbb{E}[f^2(\varepsilon_1, \dots, \varepsilon_n)] = \mathbb{E}\left[\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \mathbf{v}_i \right\|_2^2 \right] = \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^n \|\mathbf{v}_i\|_2^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^n \langle \varepsilon_i \mathbf{v}_i, \varepsilon_j \mathbf{v}_j \rangle \right] \overset{(a)}{\leq} 1$$

where (a) follows since $\sum_{i=1}^n \|\mathbf{v}_i\|_2^2 \leq n$ and by independence, for $i \neq j$, $\mathbb{E}[\langle \varepsilon_i \mathbf{v}_i, \varepsilon_j \mathbf{v}_j \rangle] = \langle \mathbb{E}[\varepsilon_i \mathbf{v}_i], \mathbb{E}[\varepsilon_j \mathbf{v}_j] \rangle = 0$ since $\mathbb{E}[\varepsilon_i] = 0$. Then, by Jensen's inequality, we have

$$\mathbb{E}[f((\varepsilon_1, \dots, \varepsilon_n)] = \mathbb{E}\left[\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \mathbf{v}_i \right\|_2 \right] \leq \sqrt{ \mathbb{E}\left[\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \mathbf{v}_i \right\|_2^2 \right] } \leq 1 \ .$$

Then, an application of the bounded difference inequality (25) (Boucheron et al., 2013)[Theorem 6.2] completes the proof. $\square$

Next we focus on the proof of Lemma 2. The proof follows the standard generic chaining analysis (Talagrand, 2014) with suitable adjustments to handle the constant shift μ.

Lemma 2. *Consider a stochastic process $\{X_{\hat{\xi}^{(n)}}\}, \hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}$ which satisfies: for any $u > 0$, with probability at least $1 - c_1 \exp(-u^2/2)$, we have*

$$\left| X_{\hat{\xi}_1^{(n)}} - X_{\hat{\xi}_2^{(n)}} \right| \leq c_0(\mu + u)\bar{d}^{(n)} \left(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)} \right) . \tag{20}$$

Further, assume that $X_{\hat{\xi}_0^{(n)}} = 0$ for some $\hat{\xi}_0^{(n)} \in \hat{\Xi}^{(n)}$. Then, for any $u > 0$, with probability at least $1 - c_3 \exp(-u^2/2)$, we have

$$\sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \left| X_{\hat{\xi}^{(n)}} \right| \leq c_2(\mu + u)\gamma_2 \left(\hat{\Xi}^{(n)}, \bar{d}^{(n)} \right) . \tag{21}$$

Proof. We consider an optimal admissible sequence $\{\Gamma_r^{(n)}\}$, i.e., sequence of subsets $\Gamma_r^{(n)}$ of $\hat{\Xi}^{(n)}$ with $|\Gamma_r^{(n)}| \leq N_r$, where $N_0 = 1, N_r = 2^{2^r}$ if $r \geq 1$, and

$$\gamma_2(\hat{\Xi}^{(n)}, \bar{d}^{(n)}) = \sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \sum_{r \geq 0} 2^{r/2} \bar{d}^{(n)}(\hat{\Xi}^{(n)}, \Gamma_r^{(n)}) . \tag{26}$$

For any $\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}$, let $\pi_r(\hat{\xi}^{(n)}) \in \Gamma_r^{(n)}$ be the "projection" of $\hat{\xi}^{(n)}$ to $\Gamma_r^{(n)}$, i.e.,

$$\pi_r(\hat{\xi}^{(n)}) = \operatorname*{argmin}_{\hat{\xi}_r^{(n)} \in \Gamma_r^{(n)}} \bar{d}^{(n)}(\hat{\xi}^{(n)}, \hat{\xi}_r^{(n)}) . \tag{27}$$

Further, let $\hat{\xi}_0^{(n)} \in \hat{\Xi}^{(n)}$ be such that $X_{\hat{\xi}_0^{(n)}} = 0$. For any $\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}$, we can decompose $X_{\hat{\xi}^{(n)}}$ as

$$X_{\hat{\xi}^{(n)}} = \sum_{r \geq 1} \left(X_{\pi_r(\hat{\xi}^{(n)})} - X_{\pi_{r-1}(\hat{\xi}^{(n)})} \right) + X_{\hat{\xi}_0^{(n)}} , \tag{28}$$

which holds provided we have chosen the sequence of sets $\{\Gamma_r^{(n)}\}$ such that $\pi_R(\hat{\xi}^{(n)}) = \hat{\xi}^{(n)}$ for R large enough. Now, based on the shifted increment condition (20), for any $u > 0$, we have

$$\mathbb{P}(\left| X_{\pi_r(\hat{\xi}^{(n)})} - X_{\pi_{r-1}(\hat{\xi}^{(n)})} \right| > c_0(\mu + u 2^{r/2})\bar{d}^{(n)}(\pi_r(\hat{\xi}^{(n)}), \pi_{r-1}(\hat{\xi}^{(n)})) \leq c_1 \exp(-u^2 2^r/2) . \tag{29}$$

Over all $\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}$, the number of possible pairs $\left(\pi_r(\hat{\xi}^{(n)}), \pi_{r-1}(\hat{\xi}^{(n)}) \right)$ is

$$|\Gamma_r^{(n)}| \cdot |\Gamma_{r-1}^{(n)}| \leq N_r N_{r-1} \leq N_{r+1} = 2^{2^{r+1}} .$$

Applying union bound over all the possible pairs of $\left(\pi_r(\hat{\xi}^{(n)}), \pi_{r-1}(\hat{\xi}^{(n)}) \right)$, we have

$$\forall r \geq 1, \forall \hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}, \qquad \left| X_{\pi_r(\hat{\xi}^{(n)})} - X_{\pi_{r-1}(\hat{\xi}^{(n)})} \right| \leq c_0(\mu + u 2^{r/2})\bar{d}^{(n)}(\pi_r(\hat{\xi}^{(n)}), \pi_{r-1}(\hat{\xi}^{(n)})) \tag{30}$$

with probability at least

$$1 - \sum_{r \geq 1} 2^{2^{r+1}} \cdot c_1 \exp(-u^2 2^r/2) \geq 1 - c_3 \exp(-u^2/2) , \tag{31}$$

where c_3 is a positive constant. From (28), (30), and (31), using the fact that $X_{\hat{\xi}_0^{(n)}} = 0$, we have

$$\mathbb{P} \left(\sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \left| X_{\hat{\xi}^{(n)}} \right| > \sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \sum_{r \geq 1} c_0(\mu + u 2^{r/2})\bar{d}^{(n)}(\pi_r(\hat{\xi}^{(n)}), \pi_{r-1}(\hat{\xi}^{(n)})) \right) \leq c_3 \exp\left(-\frac{u^2}{2} \right) .$$

By triangle inequality, we have

$$\bar{d}^{(n)}(\pi_r(\hat{\xi}^{(n)}), \pi_{r-1}(\hat{\xi}^{(n)})) \leq \bar{d}^{(n)}(\hat{\xi}^{(n)}, \pi_r(\hat{\xi}^{(n)})) + \bar{d}^{(n)}(\hat{\xi}^{(n)}, \pi_{r-1}(\hat{\xi}^{(n)})) \ ,$$

so that

$$\begin{aligned}\sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \sum_{r \geq 1} (\mu + u 2^{r/2}) \bar{d}^{(n)}(\pi_r(\hat{\xi}^{(n)}), \pi_{r-1}(\hat{\xi}^{(n)})) &\leq \sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} 2 \sum_{r \geq 0} (\mu + u 2^{r/2}) \bar{d}^{(n)}(\hat{\xi}^{(n)}, \pi_r(\hat{\xi}^{(n)})) \\ &\leq 2(\mu + u) \sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \sum_{r \geq 0} 2^{r/2} \bar{d}^{(n)}(\hat{\xi}^{(n)}, \pi_r(\hat{\xi}^{(n)})) \\ &\leq 2(\mu + u) \gamma_2(\hat{\Xi}^{(n)}, \bar{d}^{(n)}) \ .\end{aligned}$$

As a result, we have

$$\mathrm{P}\left(\sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \left| X_{\hat{\xi}^{(n)}} \right| > 2c_0(\mu + u)\gamma_2(\hat{\Xi}^{(n)}, \bar{d}^{(n)}) \right) \leq c_3 \exp\left(-\frac{u^2}{2} \right). \tag{32}$$

That completes the proof. □

Using Lemma 1 and 2, we can now prove Theorem 1.

Theorem 1. *Based on Definitions 1, 2, and 3, with $\varepsilon^{(n)}$ denoting n i.i.d. Rademacher variables, conditioned on any $z^{(n)} \in \mathcal{Z}^{(n)}$, for any $u > 0$, with probability at least $(1 - c_0 \exp(-u^2/2))$ over the randomness of $\varepsilon^{(n)}$, we have*

$$\sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \frac{1}{\sqrt{n}} \left\| \sum_{i=1}^n \varepsilon_i \xi_i \right\|_2 \leq c_5(1 + u) w(\hat{\Xi}_n) \ . \tag{6}$$

As a result, the normed empirical Rademacher complexity (NERC) satisfies

$$\hat{R}_n(\hat{\Xi}^{(n)}) = \frac{1}{n} \mathbb{E}_{\varepsilon^{(n)}} \left[\sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \left\| \sum_{i=1}^n \varepsilon_i \xi_i \right\|_2 \right] \leq \frac{c_6 w(\hat{\Xi}_n)}{\sqrt{n}} \ . \tag{7}$$

Proof. For any $\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)} \in \hat{\Xi}^{(n)}$, for any $u > 0$, by triangle inequality,

$$\begin{aligned}&\mathbb{P}_{\varepsilon^{(n)}} \left(\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \xi_{i,1} \right\|_2 - \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \xi_{i,2} \right\|_2 \geq u \bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)}) \right) \\ &\quad \leq \mathbb{P}_{\varepsilon^{(n)}} \left(\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (\xi_{i,1} - \xi_{i,2}) \right\|_2 \geq u \bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)}) \right) \ .\end{aligned}$$

Similarly,

$$\begin{aligned}&\mathbb{P}_{\varepsilon^{(n)}} \left(\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \xi_{i,2} \right\|_2 - \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \xi_{i,1} \right\|_2 \geq u \bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)}) \right) \\ &\quad \leq \mathbb{P}_{\varepsilon^{(n)}} \left(\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (\xi_{i,1} - \xi_{i,2}) \right\|_2 \geq u \bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)}) \right) \ .\end{aligned}$$

Let $\mathbf{v}_i = \frac{\xi_{i,1} - \xi_{i,2}}{\bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)})}$, so that $\sum_{i=1}^n \|\mathbf{v}_i\|_2^2 = n$. Then, from Lemma 1, for some $\mu \leq 1$, for any $u > 0$ we have

$$\mathbb{P}_{\varepsilon^{(n)}} \left(\left| \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \mathbf{v}_i \right\|_2 - \mu \right| \geq u \right) \leq 2 \exp\left(-\frac{u^2}{2} \right) \ .$$

so that

$$\mathbb{P}_{\varepsilon^{(n)}}\left(\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(\xi_{i,1}-\xi_{i,2})\right\|_2 \geq (\mu+u)\bar{d}^{(n)}(\hat{\xi}_1^{(n)},\hat{\xi}_2^{(n)})\right) \leq 2\exp\left(-\frac{u^2}{2}\right)$$

Then,

$$\begin{aligned}
&\mathbb{P}_{\varepsilon^{(n)}}\left(\left|\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\xi_{i,1}\right\|_2-\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\xi_{i,2}\right\|_2\right| \geq (\mu+u)\bar{d}^{(n)}(\hat{\xi}_1^{(n)},\hat{\xi}_2^{(n)})\right)\\
&\leq \mathbb{P}_{\varepsilon^{(n)}}\left(\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\xi_{i,1}\right\|_2-\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\xi_{i,2}\right\|_2 \geq (\mu+u)\bar{d}^{(n)}(\hat{\xi}_1^{(n)},\hat{\xi}_2^{(n)})\right)\\
&\qquad + \mathbb{P}_{\varepsilon^{(n)}}\left(\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\xi_{i,1}\right\|_2-\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\xi_{i,2}\right\|_2 \geq (\mu+u)\bar{d}^{(n)}(\hat{\xi}_1^{(n)},\hat{\xi}_2^{(n)})\right)\\
&\leq 2\mathbb{P}_{\varepsilon^{(n)}}\left(\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(\xi_{i,1}-\xi_{i,2})\right\|_2 \geq (\mu+u)\bar{d}^{(n)}(\hat{\xi}_1^{(n)},\hat{\xi}_2^{(n)})\right)\\
&\leq 4\exp(-u^2/2)\ .
\end{aligned}$$

Then, applying Lemma 2 with constants, for any $u > 0$ we have

$$\mathbb{P}_{\varepsilon^{(n)}}\left[\sup_{\hat{\xi}^{(n)}\in\hat{\Xi}^{(n)}}\frac{1}{\sqrt{n}}\left\|\sum_{i=1}^{n}\varepsilon_i\xi_i\right\|_2 \geq c_1(1+u)\gamma_2(\hat{\Xi}^{(n)},\bar{d}^{(n)})\right] \leq c_0\exp(-u^2/2)\ . \tag{33}$$

Let

$$Y = \frac{\sup_{\hat{\xi}^{(n)}\in\hat{\Xi}^{(n)}}\frac{1}{\sqrt{n}}\|\sum_{i=1}^{n}\varepsilon_i\xi_i\|_2}{c_1\gamma_2(\hat{\Xi}^{(n)},\bar{d}^{(n)})}\ .$$

For any constant $u_0 > 0$, i.e., $u_0 = O(1)$, we have

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^\infty \mathbb{P}(Y\geq u)\mathrm{d}u\\
&= \int_0^{1+u_0}\mathbb{P}(Y\geq u)\mathrm{d}u + \int_{u_0+2}^{\infty}\mathbb{P}(Y\geq u)\mathrm{d}u\\
&\leq 1+u_0+\int_{u_0}^{\infty}\mathbb{P}(Y\geq 2+u)\mathrm{d}u\\
&= 1+u_0+\frac{c_0}{u_0}\exp\left(-\frac{u_0^2}{2}\right)\\
&\leq 1+c_3,
\end{aligned}$$

where c_3 is a constant. Hence,

$$\begin{aligned}
&\mathbb{E}_{\varepsilon^{(n)}}\left[\sup_{\hat{\xi}^{(n)}\in\hat{\Xi}^{(n)}}\frac{1}{\sqrt{n}}\left\|\sum_{i=1}^{n}\varepsilon_i\xi_i\right\|_2\right] \leq c_1(1+c_3)\gamma_2(\hat{\Xi}^{(n)},\bar{d}^{(n)})\\
\Rightarrow\qquad &\frac{1}{n}\mathbb{E}_{\varepsilon^{(n)}}\left[\sup_{\hat{\xi}^{(n)}\in\hat{\Xi}^{(n)}}\left\|\sum_{i=1}^{n}\varepsilon_i\xi_i\right\|_2\right] \leq c_2\frac{\gamma_2(\hat{\Xi}^{(n)},\bar{d}^{(n)})}{\sqrt{n}}\ ,
\end{aligned}$$

for a suitable constant c_2.

Now that we have established a version of both parts of Theorem 1 in terms of $\gamma_2(\hat{\Xi}^{(n)},\bar{d}^{(n)})$, to complete the proof by showing that $\gamma_2(\hat{\Xi}^{(n)},\bar{d}^{(n)}) \leq c_4 w(\hat{\Xi}_n)$ for some constant c_4. We do this in two steps:

- first showing that $\gamma_2(\hat{\Xi}^{(n)}, \bar{d}^{(n)}) = \gamma_2(\hat{\Xi}_n, d)$; and
- then showing $\gamma_2(\hat{\Xi}_n, d) \leq c_4 w(\hat{\Xi}_n)$.

For the first step, note that there is a correspondence between $\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}$ and $\hat{\xi}_n \in \hat{\Xi}_n$ as $\hat{\xi}_n$ is a scaled stacked version of the n-tuples. Moreover, for any pair $\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)} \in \hat{\Xi}^{(n)}$ respectively corresponding to $\hat{\xi}_{1,n}, \hat{\xi}_{2,n} \in \hat{\Xi}_n$, by definition of $\bar{d}^{(n)}$ as in (18) and d as in (11), we have

$$\begin{aligned}
d(\hat{\xi}_{1,n}, \hat{\xi}_{2,n}) &= \mathbb{E}\left[\left(X_{\hat{\xi}_{1,n}} - X_{\hat{\xi}_{2,n}}\right)^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\left(\langle \hat{\xi}_{1,n}, g\rangle - \langle \hat{\xi}_{2,n}, g\rangle\right)^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\langle \hat{\xi}_{1,n} - \hat{\xi}_{2,n}, g\rangle^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \|\xi_{i,1} - \xi_{i,2}\|_2^2 \\
&= \bar{d}^{(n)}(\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)}) \ .
\end{aligned}$$

Further, because of the correspondence between the elements of $\hat{\Xi}^{(n)}$ and $\hat{\Xi}_n$, there is also a correspondence between admissible sequences $\Gamma^{(n)}$ of $\hat{\Xi}^{(n)}$ and Γ of $\hat{\Xi}_n$. As a result, we have

$$\begin{aligned}
\gamma_2(\hat{\Xi}^{(n)}, \bar{d}^{(n)}) &= \inf_{\Gamma^{(n)}} \sup_{\hat{\xi}^{(n)} \in \hat{\Xi}^{(n)}} \sum_{r=0}^{\infty} 2^{r/2} \bar{d}^{(n)}(\hat{\xi}^{(n)}, \Gamma_r^{(n)}) && (34) \\
&= \inf_{\Gamma} \sup_{\hat{\xi}_n \in \hat{\Xi}_n} \sum_{r=0}^{\infty} 2^{r/2} d(\hat{\xi}_n, \Gamma_r) && (35) \\
&= \gamma_2(\hat{\Xi}_n, d) \ . && (36)
\end{aligned}$$

For the second step, since for the Gaussian process $X_{\hat{\xi}_n} = \langle \hat{\xi}_n, \mathbf{g}\rangle$ on $\hat{\Xi}_n$, the canonical distance is

$$\left(\mathbb{E}\left[\left(X_{\hat{\xi}_{1,n}} - X_{\hat{\xi}_{2,n}}\right)^2\right]\right)^{1/2} = \left(\frac{1}{n}\sum_{i=1}^{n} \|\xi_{i,1} - \xi_{i,2}\|_2^2\right)^{1/2} , \tag{37}$$

from the majorizing measure theorem (Talagrand, 2014), for some constant c_4, we have

$$\gamma_2(\hat{\Xi}_n, d) \leq c_4 w(\hat{\Xi}_n) \tag{38}$$

That completes the proof. $\square$

B SGD with Sample Reuse: Proofs for Section 3

In this section, we provide proofs for results in Section 3.

B.1 GD with sample reuse

Theorem 2. *Let $\theta_0 \in \Theta_0$, and $\theta_t, t \in [T]$ be a sequence of parameters obtained from GD by reusing a fixed set of samples $z^{(n)} \sim \mathcal{D}^n$ in each epoch. Let $\Delta(\theta) := \left\| \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(\theta, z_i) - \nabla \mathcal{L}_D(\theta) \right\|_2$, where the population gradient $\nabla \mathcal{L}_D(\theta_t) = \mathbb{E}_{z\sim\mathcal{D}}[\nabla \ell(\theta_t; z)], t \in [T]$. Under Assumption 2, with $w(\hat{\Xi}_n^0)$ denoting the LGGW for Θ_0 as in Definition 2, for any $\delta \leq \frac{1}{2}$, with probability at least $(1-2\delta)$ over $z^{(n)} \sim \mathcal{D}^n$, we have*

$$\begin{aligned}
&\max\left(\sup_{\theta_0 \in \Theta_0} \Delta(\theta_0), \quad \max_{t \in [T]} \Delta(\theta_t)\right) \\
&\qquad \leq \frac{c_7 \max(w(\hat{\Xi}_n^0), \log p) + \sqrt{\log T + \log \frac{1}{\delta}}}{\sqrt{n}} \ .
\end{aligned} \tag{12}$$

Proof. Remembering that the population gradient $\nabla \mathcal{L}_D(\theta_t) = \mathbb{E}_{z\sim\mathcal{D}}[\nabla \ell(\theta_t; z)]$, for $t \in [T]$ consider the sequence of events

$$\Lambda_0 \triangleq \left\{ \sup_{\theta_0 \in \Theta_0} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(\theta_0, z_i) - \nabla \mathcal{L}_D(\theta_0) \right\|_2 \leq \epsilon \right\} \ , \ \text{and} \ \ \bar{\Lambda}_0 \triangleq \left\{ \sup_{\theta_0 \in \Theta_0} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(\theta_0, z_i) - \nabla \mathcal{L}_D(\theta_0) \right\|_2 > \epsilon \right\} \ , \tag{39}$$

$$\Lambda_t \triangleq \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(\theta_t, z_i) - \nabla \mathcal{L}_D(\theta_t) \right\|_2 \leq \epsilon \right\} \ , \ \text{and} \ \ \bar{\Lambda}_t \triangleq \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(\theta_t, z_i) - \nabla \mathcal{L}_D(\theta_t) \right\|_2 > \epsilon \right\} \ . \tag{40}$$

We define a filtration based these sequence of events Λ_t which considers conditional events $\lambda|\{\Lambda_{\tau\leq t}\}$with corresponding conditional probability distributions:

$$\mathcal{D}_0 = \mathcal{D} \ , \tag{41}$$

$$\mathcal{D}_t = \mathcal{D} \mid \{\Lambda_{\tau<t}\} \ . \tag{42}$$

To make the analysis precise, we use $\mathbb{P}_{\mathcal{D}_t}[\lambda]$ to denote probability of events λ in the t-th stage of the filtration and $\mathbb{P}_{\mathcal{D}}[\lambda]$ to denote probability of suitable events λ according to the original distribution $\mathcal{D}$.

Note that (posterior) probability of any event λ_1 at stage 1, i.e., conditioned on Λ_1 of the filtration can be written in terms of probabilities at the previous stage based on the definition of conditional probability as

$$\mathbb{P}_{\mathcal{D}_1}[\lambda_1] = \mathbb{P}_{\mathcal{D}|\{\Lambda_0\}}[\lambda_1] = \frac{\mathbb{P}_{\mathcal{D}}[\lambda_1, \Lambda_0]}{\mathbb{P}_{\mathcal{D}}[\Lambda_0]} \ . \tag{43}$$

More generally, by the definition of conditional probability

$$\mathbb{P}_{\mathcal{D}_t}[\lambda] = \mathbb{P}_{\mathcal{D}|\{\Lambda_{\tau<t}\}}[\lambda_t] = \frac{\mathbb{P}_{\mathcal{D}}[\lambda_t, \{\Lambda_{\tau<t}\}]}{\mathbb{P}_{\mathcal{D}}[\{\Lambda_{\tau<t}\}]} \ , \tag{44}$$

where $\{\Lambda_{\tau<t}\}$ denotes the joint event $\{\Lambda_0, \Lambda_1, \ldots, \Lambda_{t-1}\}$ where individual events Λ_τ are as in (39) and (40). While seemingly straightforward, (44) shows a way to express probabilities $\mathbb{P}_{\mathcal{D}_t}[\cdot]$ of events in the sequence directly in terms $\mathbb{P}_{\mathcal{D}}[\cdot]$, the original distribution.

For our analysis, instead of working with the full distributions, we will be working on the finite sample versions based on $z^{(n)} \sim \mathcal{D}^n$. In particular, let $\mathcal{D}_0(z^{(n)})$ be the empirical distribution based on the samples, i.e., $\mathcal{D}_0(z^{(n)}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{z_i}$, where $\mathbf{1}$ is the indicator function. As before, the (posterior) probability of any event λ_1 at any stage 1 of the filtration can be written in terms of probabilities at the previous stage based on the definition of conditional probability:

$$\mathbb{P}_{\mathcal{D}_1(z^{(n)})}[\lambda_1] = \mathbb{P}_{\mathcal{D}_0(z^{(n)})|\{\Lambda_0\}}[\lambda_1] = \frac{\mathbb{P}_{\mathcal{D}_0(z^{(n)})}[\lambda_1, \Lambda_0]}{\mathbb{P}_{\mathcal{D}_0(z^{(n)})}[\Lambda_0]} \ . \tag{45}$$

More generally, by the definition of conditional probability:

$$\mathbb{P}_{\mathcal{D}_t(z^{(n)})}[\lambda_t] = \mathbb{P}_{\mathcal{D}_0(z^{(n)})|\{\Lambda_{\tau\leq t}\}}[\lambda_t] = \frac{\mathbb{P}_{\mathcal{D}_0(z^{(n)})}[\lambda_t, \{\Lambda_{\tau<t}\}]}{\mathbb{P}_{\mathcal{D}_0(z^{(n)})}[\{\Lambda_{\tau<t}\}]} \ . \tag{46}$$

Note that events, Λ_t, of interest are defined in terms of the samples $z^{(n)}$, and the definition of conditional probability provides a way to express the probabilities of such events under $\mathcal{D}_t(z^{(n)})$ in terms of events under $\mathcal{D}_0(z^{(n)})$. Since $\mathcal{D}_0(z^{(n)})$ is the empirical distribution of the samples $z^{(n)} \sim \mathcal{D}^n$, for convenience, we will denote $\mathbb{P}_{\mathcal{D}_0(z^{(n)})}[\cdot]$ as $\mathbb{P}_{z^{(n)}\sim\mathcal{D}^n}[\cdot]$ and even more briefly as $\mathbb{P}_{z^{(n)}}[\cdot]$.

To simplify the notation, consider the random variable

$$X_0 \triangleq \sup_{\theta_0 \in \Theta_0} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(\theta_0, z_i) - \nabla \mathcal{L}_D(\theta_0) \right\|_2 \ , \tag{47}$$

$$X_t \triangleq \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(\theta_t, z_i) - \nabla \mathcal{L}_D(\theta_t) \right\|_2 \ . \tag{48}$$

Then, from (39) and (40), we have $\Lambda_t = \{X_t \leq \epsilon\}$ and $\bar{\Lambda}_t = \{X_t > \epsilon\}$. Then, we have

$$
\begin{aligned}
&\mathbb{P}\left[\max\left(\sup_{\theta_0\in\Theta_0}\left\|\frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta_0,z_i) - \nabla\mathcal{L}_D(\theta_0)\right\|_2 , \quad \max_{t\in[T]}\left\|\frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta_t,z_i) - \nabla\mathcal{L}_D(\theta_t)\right\|_2\right) > \epsilon\right] \qquad (49)\\
&\leq \mathbb{P}_{z^{(n)}}\left[\sup_{\theta_0\in\Theta_0}\left\|\frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta_0,z_i) - \nabla\mathcal{L}_D(\theta_0)\right\|_2 > \epsilon\right] + \sum_{t=1}^T \mathbb{P}_{\mathcal{D}_t(z^{(n)})}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta_t,z_i) - \nabla\mathcal{L}_D(\theta_t)\right\|_2 > \epsilon\right]\\
&= \sum_{t=0}^T \mathbb{P}_{\mathcal{D}_t(z^{(n)})}\left[X_t > \epsilon\right]\\
&= \sum_{t=0}^T \mathbb{P}_{\mathcal{D}_t(z^{(n)})}[\bar{\Lambda}_t] \ . \qquad (50)
\end{aligned}
$$

Now, with $\{\Lambda_{0:\tau}\} := (\Lambda_0, \Lambda_1, \ldots, \Lambda_\tau)$ and $\{0:\tau\} := \{0, 1, \ldots, \tau\}$, we have

$$
\begin{aligned}
\sum_{t=0}^T \mathbb{P}_{\mathcal{D}_t(z^{(n)})}[\bar{\Lambda}_t] &= \mathbb{P}_{\mathcal{D}_0(z^{(n)})}[\bar{\Lambda}_0] + \mathbb{P}_{\mathcal{D}_1(z^{(n)})}[\bar{\Lambda}_1] + \ldots + \mathbb{P}_{\mathcal{D}_T(z^{(n)})}[\bar{\Lambda}_T]\\
&= \mathbb{P}_{z^{(n)}}[\bar{\Lambda}_0] + \frac{\mathbb{P}_{z^{(n)}}[\Lambda_0, \bar{\Lambda}_1]}{\mathbb{P}_{z^{(n)}}[\Lambda_0]} + \ldots + \frac{\mathbb{P}_{z^{(n)}}[\{\Lambda_{0:(T-1)}\}, \bar{\Lambda}_T]}{\mathbb{P}_{z^{(n)}}[\{\Lambda_{0:(T-1)}\}]}\\
&\overset{(a)}{\leq} \frac{\mathbb{P}_{z^{(n)}}[\bar{\Lambda}_0]}{\mathbb{P}_{z^{(n)}}[\{\Lambda_{0:(T-1)}\}]} + \frac{\mathbb{P}_{z^{(n)}}[\Lambda_0, \bar{\Lambda}_1]}{\mathbb{P}_{z^{(n)}}[\{\Lambda_{0:(T-1)}\}]} + \ldots + \frac{\mathbb{P}_{z^{(n)}}[\{\Lambda_{0:(T-1)}\}, \bar{\Lambda}_T]}{\mathbb{P}_{z^{(n)}}[\{\Lambda_{0:(T-1)}\}]}\\
&\overset{(b)}{\leq} \frac{1 - \mathbb{P}_{z^{(n)}}[\{\Lambda_{0:T}\}]}{\mathbb{P}_{z^{(n)}}[\{\Lambda_{0:(T-1)}\}]}\\
&= \frac{1 - \mathbb{P}_{z^{(n)}}\left[\max_{t\in\{0:T\}} X_t \leq \epsilon\right]}{\mathbb{P}_{z^{(n)}}\left[\max_{t\in\{0:T-1\}} X_t \leq \epsilon\right]}\\
&= \frac{\mathbb{P}_{z^{(n)}}\left[\max_{t\in\{0:T\}} X_t > \epsilon\right]}{\mathbb{P}_{z^{(n)}}\left[\max_{t\in\{0:T-1\}} X_t \leq \epsilon\right]}\\
&\overset{(c)}{\leq} \frac{\mathbb{P}_{z^{(n)}}\left[\max_{t\in\{0:T\}} X_t > \epsilon\right]}{\mathbb{P}_{z^{(n)}}\left[\max_{t\in\{0:T\}} X_t \leq \epsilon\right]} \ . \qquad (51)
\end{aligned}
$$

where (a) follows since $\mathbb{P}_{z^{(n)}}[\{\Lambda_{0:(T-1)}\}] \leq \mathbb{P}_{z^{(n)}}[\{\Lambda_{0:\tau}\}] \leq 1$ for all $\tau \in \{0:T\}$, (b) follows since the counter-event of $\{\Lambda_{0:T}\}$ covers the events in (a), and (c) follows since $\mathbb{P}_{z^{(n)}}\left[\max_{t\in\{0:T\}} X_t \leq \epsilon\right] \leq \mathbb{P}_{z^{(n)}}\left[\max_{t\in\{0:T-1\}} X_t \leq \epsilon\right]$. Thus, it suffices to focus on an upper bound on $\mathbb{P}_{z^{(n)}\sim\mathcal{D}^n}[\max_{t\in\{0:T\}} X_t > \epsilon]$ which corresponds to the same 'bad' event $\{\max_{t\in\{0:T\}} X_t > \epsilon\}$ but in the non-adaptive setting, since the probability is w.r.t. $\mathbb{P}_{z^{(n)}\sim\mathcal{D}^n}[\cdot]$.

From Lemma 3, for any $u > 0$, we have

$$
\mathbb{P}_{z^{(n)}}\left[\sup_{\theta_0\in\Theta_0}\left\|\frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta,z_i) - \nabla\mathcal{L}_{\mathcal{D}}(\theta)\right\|_2 > \frac{c_1 w(\hat{\Xi}_n^0) + u}{\sqrt{n}}\right] \leq \exp(-2u^2) \ . \qquad (52)
$$

Further, from Lemma 5, for any $u > 0$, we have

$$
\mathbb{P}_{z^{(n)}}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta,z_i) - \nabla\mathcal{L}_{\mathcal{D}}(\theta)\right\|_2 > \frac{c_1\left(\log p + u\right)}{\sqrt{n}}\right] \leq \exp(-u^2) \ . \qquad (53)
$$

Choosing $\epsilon = \frac{c_1 \max(w(\hat{\Xi}_n), \log p) + u}{\sqrt{n}}$ and $u = \log(T+1) + \log\frac{1}{\delta}$, we have

$$
\begin{aligned}
\mathbb{P}_{z^{(n)}}\left[\max_{t\in\{0:T\}} X_t > \epsilon\right] &\overset{(a)}{\leq} \sum_{t=0}^T \mathbb{P}_{z^{(n)}}\left[X_t > \epsilon\right] = \mathbb{P}_{z^{(n)}}\left[X_0 > \epsilon\right] + \sum_{t=1}^T \mathbb{P}_{z^{(n)}}\left[X_t > \epsilon\right]\\
&\overset{(b)}{\leq} \exp(-2u^2) + T\exp(-u^2)
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(c)}{\leq} (T+1)\exp(-u^2) \\
&\overset{(d)}{\leq} (T+1)\exp\left(-\left(\log(T+1)+\log\frac{1}{\delta}\right)\right)=\delta,
\end{aligned}
$$

where (a) follows by union bound, (b) follows from 52 and 53, (c) follows since $u>1$ for $T\geq 2$, and (d) follows from the form of u. Then, for $\delta \leq 1/2$ we have

$$
\mathbb{P}_{\mathcal{D}_T(z^{(n)})}\left[\max_{t\in\{0:T\}} X_t > \epsilon\right] \leq \frac{\mathbb{P}_{z^{(n)}}\left[\max_{t\in\{0:T\}} X_t > \epsilon\right]}{\mathbb{P}_{z^{(n)}}\left[\max_{t\in\{0:T\}} X_t \leq \epsilon\right]} \leq \frac{\delta}{1-\delta} \leq 2\delta \ .
$$

That completes the proof. □

Lemma 3. *With $\hat{\Xi}_n, w(\hat{\Xi}_n)$ respectively denoting the set of empirical gradients and its Gaussian width as in Definition 2, for any $u>0$, we have*

$$
\mathbb{P}_{z^{(n)}}\left[\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta,z_i)-\nabla\mathcal{L}_{\mathcal{D}}(\theta)\right\|_2 > \frac{c_1 w(\hat{\Xi}_n)+u}{\sqrt{n}}\right] \leq \exp(-2u^2) \ . \tag{54}
$$

Proof. From Foster et al. (2018), with $\nabla\hat{\mathcal{L}}_n(\theta)=\frac{1}{n}\sum_{i=1}^n\nabla\ell(\theta,z_i)$, for any $\delta>0$, with probability at least $1-\delta$ over the data $z^{(n)}$, we have

$$
\mathbb{E}_{z^{(n)}}\sup_{\theta\in\Theta}\|\nabla\hat{\mathcal{L}}_n(\theta)-\nabla\mathcal{L}_{\mathcal{D}}(\theta)\| \leq 4\hat{R}_n(\hat{\Xi}^{(n)})+\frac{\log\frac{1}{\delta}}{n} \ . \tag{55}
$$

Using the bounded difference inequality (Boucheron et al., 2013), for any $\epsilon>0$ and $\|\nabla\ell(\theta;z)\|_2=O(1)$ for $\theta\in\Theta, z\in\mathcal{Z}$, we have

$$
\mathbb{P}_{z^{(n)}}\left[\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta,z_i)-\nabla\mathcal{L}_{\mathcal{D}}(\theta)\right\|_2 > \mathbb{E}_{z^{(n)}}\sup_{\theta\in\Theta}\|\nabla\hat{\mathcal{L}}_n(\theta)-\nabla\mathcal{L}_{\mathcal{D}}(\theta)\|+\epsilon\right] \leq \exp(-2n\epsilon^2) \ . \tag{56}
$$

From Section 2 and Theorem 1, we have

$$
\hat{R}_n(\hat{\Xi}^{(n)}) \leq c_2\frac{w(\hat{\Xi}_n)}{\sqrt{n}} \ . \tag{57}
$$

Combining the above results, with $\epsilon=\frac{u}{\sqrt{n}}$, for any $\delta>0$, with probability at least $1-\delta$ over the data $z^{(n)}$, we have

$$
\mathbb{P}_{z^{(n)}}\left[\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta,z_i)-\nabla\mathcal{L}_{\mathcal{D}}(\theta)\right\|_2 > \frac{c_1 w(\hat{\Xi}_n)+u}{\sqrt{n}}\right] \leq \exp(-2u^2) \ , \tag{58}
$$

where $c_1=4c_2$. That completes the proof. □

Lemma 4 (Matrix Bernstein Inequality). *(Tropp et al., 2015) Let $S_1,\ldots,S_n$ be independent, centered random matrices with common dimension $d_1\times d_2$, and assume that each one is uniformly bounded*

$$
\mathbb{E}S_k=0, \|S_k\|\leq L, \forall k\in[n].
$$

Introduce the sum

$$
Z=\sum_{k=1}^n S_k,
$$

and let $v(Z)$ denote the matrix variance statistic of the sum

$$
v(Z)=\max\left\{\|\mathbb{E}\left(ZZ^*\right)\|, \|\mathbb{E}\left(Z^*Z\right)\|\right\}
$$

$$= \max \left\{ \left\| \sum_{k=1}^{n} \mathbb{E}\left(S_k S_k^*\right) \right\|, \left\| \sum_{k=1}^{n} \mathbb{E}\left(S_k^* S_k\right) \right\| \right\}.$$

Then

$$\mathbb{P}\left\{\|Z\| \geq t\right\} \leq (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{v(Z) + Lt/3}\right), \forall t \geq 0.$$

Furthermore,

$$\mathbb{E}\,\|Z\| \leq \sqrt{2v(Z)\log(d_1+d_2)} + \frac{1}{3} L \log(d_1+d_2).$$

Lemma 5. *Under Assumption 2, for any $u > 0$, we have*

$$\mathbb{P}_{z^{(n)}}\left[\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(\theta, z_i) - \nabla \mathcal{L}_{\mathcal{D}}(\theta) \right\|_2 > \frac{c_1 (\log p + u)}{\sqrt{n}}\right] \leq \exp(-u^2)\ . \tag{59}$$

Proof. Following Assumption 2, denote the L_2 norm bound of the gradient is G. There are two different cases:

- If $\log p + u \geq \sqrt{n}$, then applying the bounded difference inequality (Boucheron et al., 2013), for any $\epsilon > 0$,

$$\mathbb{P}_{z^{(n)}}\left[\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla \ell\left(\theta, z_i\right) - \nabla \mathcal{L}_{\mathcal{D}}\left(\theta\right) \right\|_2 > \mathbb{E}_{z^{(n)}} \left\| \nabla \hat{\mathcal{L}}_n\left(\theta\right) - \nabla \mathcal{L}_{\mathcal{D}}\left(\theta\right) \right\| + \epsilon \right] \leq \exp\left(-\frac{n\epsilon^2}{2G^2}\right)$$

  Since

$$\mathbb{E}_{z^{(n)}} \left\| \nabla \hat{\mathcal{L}}_n\left(\theta\right) - \nabla \mathcal{L}_{\mathcal{D}}\left(\theta\right) \right\| \leq 2G = \frac{2\sqrt{n}G}{\sqrt{n}} \leq \frac{2G\left(\log p + u\right)}{\sqrt{n}}$$

  then taking $\epsilon = \frac{\sqrt{2}Gu}{\sqrt{n}}$, we have

$$\begin{aligned}
&\mathbb{P}_{z^{(n)}}\left[\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla \ell\left(\theta, z_i\right) - \nabla \mathcal{L}_{\mathcal{D}}\left(\theta\right) \right\|_2 > \frac{2G\left(\log p + u\right) + \sqrt{2}Gu}{\sqrt{n}}\right] \\
&\leq \mathbb{P}_{z^{(n)}}\left[\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla \ell\left(\theta, z_i\right) - \nabla \mathcal{L}_{\mathcal{D}}\left(\theta\right) \right\|_2 > \mathbb{E}_{z^{(n)}} \left\| \nabla \hat{\mathcal{L}}_n\left(\theta\right) - \nabla \mathcal{L}_{\mathcal{D}}\left(\theta\right) \right\| + \frac{\sqrt{2}Gu}{\sqrt{n}}\right] \\
&\leq \exp(-u^2)
\end{aligned}$$

  Therefore,

$$\mathbb{P}_{z^{(n)}}\left[\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla \ell\left(\theta, z_i\right) - \nabla \mathcal{L}_{\mathcal{D}}\left(\theta\right) \right\|_2 > \frac{2G\left(\log p + u\right) + \sqrt{2}Gu}{\sqrt{n}}\right] \leq \exp(-u^2)$$

- If $\log p + u < \sqrt{n}$, then with $S_i = \nabla \ell\left(\theta_t, z_i\right) - \mathbb{E}\nabla \ell\left(\theta_t, z_i\right)$ and $Z = \sum_{i=1}^{n} S_i$, we have $\mathbb{E}S_i = 0$, and

$$\|S_i\| = \left\|\nabla \ell\left(\theta_t, z_i\right) - \mathbb{E}\nabla \ell\left(\theta_t, z_i\right)\right\| \leq \left\|\nabla \ell\left(\theta_t, z_i\right)\right\| + \left\|\mathbb{E}\nabla \ell\left(\theta_t, z_i\right)\right\| \leq 2G.$$

  so we can take $L = 2G$ as an upper bound of $\|S_i\|$. Besides,

$$v(Z) = \max \left\{ \left\| \sum_{i=1}^{n} \mathbb{E}\left(S_i S_i^*\right) \right\|, \left\| \sum_{i=1}^{n} \mathbb{E}\left(S_i^* S_i\right) \right\| \right\} \leq n \max_{i \in [n]} \|S_i\|^2 \leq 4nG^2$$

  Then directly apply Lemma 4 with $d_1 = 1$, $d_2 = p$, we can get that

$$\mathbb{P}\left\{\|Z\| \geq t\right\} \leq (p+1) \cdot \exp\left(-\frac{t^2/2}{v(Z) + Lt/3}\right)$$

Substituting $L = 2G$ and $v(Z) \leq 4nG^2$, we have

$$\mathbb{P}\left\{\|Z\| \geq t\right\} \leq (p+1) \cdot \exp\left(-\frac{t^2/2}{4nG^2 + 2Gt/3}\right) = (p+1) \cdot \exp\left(\frac{-3t^2}{24nG^2 + 4Gt}\right)$$

Taking $t = 4\sqrt{n}G\left(\log p + u\right)$, we have

$$\begin{aligned}
\mathbb{P}\left\{\|Z\| \geq 4\sqrt{n}G\left(\log p + u\right)\right\} &\leq (p+1) \cdot \exp\left(-\frac{48nG^2\left(\log p + u\right)^2}{24nG^2 + 16\sqrt{n}G^2\left(\log p + u\right)}\right) \\
&< (p+1) \cdot \exp\left(-\frac{48nG^2\left(\log p + u\right)^2}{24nG^2 + 16\sqrt{n}G^2 \cdot \sqrt{n}}\right) \\
&< (p+1) \cdot \exp\left(-\left(\log p + u\right)^2\right) < (p+1) \cdot \exp\left(-\left(\left(\log p\right)^2 + u^2\right)\right) \\
&\leq \exp\left(-u^2\right)
\end{aligned}$$

So we have

$$\begin{aligned}
&\mathbb{P}_{z^{(n)}}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla\ell\left(\theta, z_i\right) - \nabla\mathcal{L}_{\mathcal{D}}\left(\theta\right)\right\|_2 > \frac{4G\left(\log p + u\right)}{\sqrt{n}}\right] \\
&= \mathbb{P}\left\{\frac{1}{n}\|Z\| \geq \frac{4G\left(\log p + u\right)}{\sqrt{n}}\right\} = \mathbb{P}\left\{\|Z\| \geq 4\sqrt{n}G\left(\log p + u\right)\right\} \\
&\leq \exp(-u^2)
\end{aligned}$$

Therefore,

$$\mathbb{P}_{z^{(n)}}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla\ell\left(\theta, z_i\right) - \nabla\mathcal{L}_{\mathcal{D}}\left(\theta\right)\right\|_2 > \frac{4G\left(\log p + u\right)}{\sqrt{n}}\right] \leq \exp(-u^2)$$

Combining these two cases, there exist c_3 such that

$$\mathbb{P}_{z^{(n)}}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla\ell\left(\theta, z_i\right) - \nabla\mathcal{L}_{\mathcal{D}}\left(\theta\right)\right\|_2 > \frac{c_3G\left(\log p + u\right)}{\sqrt{n}}\right] \leq \exp(-u^2)$$

Since $G = O(1)$ according to Assumption 2, by taking $c_1 = Gc_3$, we complete the proof.

B.2 Population Convergence in Optimization

Theorem 3. *Consider a non-convex loss $\mathcal{L}_{\mathcal{D}}(\theta) = \mathbb{E}_{z\sim\mathcal{D}}\left[\ell(\theta, z)\right]$ with τ-Lipschitz gradient. For GD using a total of $z^{(n)} \sim \mathcal{D}^n$ original samples and re-using the n samples in each step as in Theorem 2, with step-size $\eta = \frac{1}{4\tau}$, with probability at least $(1-2\delta)$ for any $\delta > 0$ over $z^{(n)} \sim \mathcal{D}^n$, $\mathbb{E}_R\left\|\nabla\mathcal{L}_{\mathcal{D}}\left(\theta_R\right)\right\|^2 \leq O\left(\frac{1}{T}\right) + O\left(\frac{\max\left(w^2(\hat{\Xi}_n^0), \log^2 p\right) + \left(\log T + \log\frac{1}{\delta}\right)}{n}\right)$ where R is uniformly distributed over $\{1, \ldots, T\}$ and the expectation is over the randomness of R.*

Proof. With the condition that $\mathcal{L}_{\mathcal{D}}(\theta)$ has τ-Lipschitz gradient, i.e., for any $\theta, \theta' \in \Theta$

$$\|\nabla\mathcal{L}_{\mathcal{D}}(\theta) - \nabla\mathcal{L}_{\mathcal{D}}(\theta')\|_2 \leq \tau\|\theta - \theta'\|, \tag{60}$$

we have

$$\mathcal{L}_{\mathcal{D}}(\theta) \leq \mathcal{L}_{\mathcal{D}}(\theta') + \langle\theta - \theta', \mathcal{L}_{\mathcal{D}}(\theta)\rangle + \frac{\tau}{2} \cdot \|\theta - \theta'\|_2^2. \tag{61}$$

At iteration t, to simplify the notation, we use the $g(\theta_t) = \frac{1}{n}\sum_{i=1}^{n}\nabla\ell\left(\theta_t, z_i\right)$ to be the sample gradient estimate and $\nabla\mathcal{L}_{\mathcal{D}}(\theta_t)$ is the population gradient. Let $\Delta_t = g(\theta_t) - \nabla\mathcal{L}_{\mathcal{D}}(\theta_t)$.

Given the update of SGD to be $\theta_{t+1} = \theta_t - \eta_t g(\theta_t)$, we have

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}(\theta_{t+1}) &\le \mathcal{L}_{\mathcal{D}}(\theta_t) + \langle \nabla \mathcal{L}_{\mathcal{D}}(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\tau \eta_t^2}{2} \|\theta_{t+1} - \theta_t\|_2^2 \\
&\le \mathcal{L}_{\mathcal{D}}(\theta_t) - \eta_t \langle \nabla \mathcal{L}_{\mathcal{D}}(\theta_t), \nabla \mathcal{L}_{\mathcal{D}}(\theta_t) + \Delta_t \rangle + \frac{\tau \eta_t^2}{2} \|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t) + \Delta_t\|_2^2 \\
&\le \mathcal{L}_{\mathcal{D}}(\theta_t) - \frac{\eta_t}{2} \|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t)\|_2^2 + \frac{\eta_t}{2} \|\Delta_t\|_2^2 + \tau \eta_t^2 \|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t)\|^2 + \tau \eta_t^2 \|\Delta_t\|_2^2 \\
&= \mathcal{L}_{\mathcal{D}}(\theta_t) - \left(\frac{\eta_t}{2} - \tau \eta_t^2\right) \|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t)\|_2^2 + \left(\frac{\eta_t}{2} + \tau \eta_t^2\right) \|\Delta_t\|_2^2. \qquad (62)
\end{aligned}$$

From Theorem 2, with probability at least $1 - 2\delta$, for all $t \in \{0 : T\}$ we have

$$\|\Delta_t\|_2^2 \le \frac{\left(c_1 \max(w(\hat{\Xi}_n^0), \log p)) + \sqrt{\log T + \log \frac{1}{\delta}}\right)^2}{n} \le \frac{2c_1^2 \max\left(w^2(\hat{\Xi}_n^0), \log^2 p\right) + 2\left(\log T + \log \frac{1}{\delta}\right)}{n}. \qquad (63)$$

Bring this to the (62) and sum over iteration from $t = 1$ to $t = T$, we have

$$\frac{1}{T} \sum_{t=1}^{T} \left(\frac{\eta_t}{2} - \tau \eta_t^2\right) \|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t)\|_2^2 \le \frac{\mathcal{L}_{\mathcal{D}}(\theta_t) - L_{\mathcal{D}}^{\star}}{T} + \frac{1}{T} \sum_{t=1}^{T} \left(\frac{\eta_t}{2} + \tau \eta_t^2\right) \|\Delta_t\|_2^2, \qquad (64)$$

where $\mathcal{L}_{\mathcal{D}}(\theta_{T+1})$ is lower bounded by $L_{\mathcal{D}}^{\star}$. Choosing $\eta_t = \frac{1}{4\tau}$, we have

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} \|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t)\|_2^2 &\le \frac{16\tau \left(\mathcal{L}_{\mathcal{D}}(\theta_T) - L_{\mathcal{D}}^{\star}\right)}{T} + \frac{1}{T} \sum_{t=1}^{T} 3 \|\Delta_t\|_2^2 \qquad (65) \\
&\le \frac{16\tau \left(\mathcal{L}_{\mathcal{D}}(\theta_T) - L_{\mathcal{D}}^{\star}\right)}{T} + 3 \frac{2c_1^2 \max\left(w^2(\hat{\Xi}_n^0), \log^2 p\right) + 2\left(\log T + \log \frac{1}{\delta}\right)}{n} \qquad (66)
\end{aligned}$$

Given that θ_R is uniformly sampled from $\{\theta_1, ..., \theta_T\}$, we have

$$\mathbb{E}[\|\nabla L_{\mathcal{D}}(\theta_R)\|_2^2] = \frac{1}{T} \sum_{t=1}^{T} \|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t)\|_2^2 \qquad (67)$$

That completes the proof. □

C Gaussian Width of Gradients: Proofs for Section 4

C.1 Gaussian Width bounds for Feed-Forward Networks (FFNs)

We consider f to be a FFN with given by

$$f(\theta; \mathbf{x}) = \mathbf{v}^\top \phi\Big(\frac{1}{\sqrt{m_L}} W^{(L)} \phi(\cdots \phi(\frac{1}{\sqrt{m_1}} W^{(1)} \mathbf{x}))\Big)\ , \qquad (68)$$

where $W^{(1)} \in \mathbb{R}^{m_1 \times d}$, $W^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}, l \in \{2, ..., L\}$ are layer-wise weight matrices, $\mathbf{v} \in \mathbb{R}^{m_L}$ is the last layer vector, $\phi(\cdot)$ is the smooth (pointwise) activation function, and the total set of parameters

$$\theta := \left(\operatorname{vec}\left(W^{(1)}\right)^\top, ..., \operatorname{vec}\left(W^{(L)}\right)^\top, v^\top\right)^\top \in \mathbb{R}^{\sum_{k=1}^{L} m_k m_{k-1} + m_L} \qquad (69)$$

with $m_0 = d$. For convenience, we write the model in terms of the layerwise outputs or features as:

$$\alpha^{(0)}(\mathbf{x}) = \mathbf{x}\ , \qquad (70)$$

$$\alpha^{(l)} = \phi\left(\frac{1}{\sqrt{m_l}} W^{(l)} \alpha^{(l-1)}(\mathbf{x})\right) \tag{71}$$

$$f(\theta;\mathbf{x}) = \mathbf{v}^\top \alpha^{(L)}(\mathbf{x}) \ . \tag{72}$$

Our analysis for both FFNs and ResNets will be for gradients over all parameters in a fixed radius spectral norm ball around the initialization θ_0:

$$B^{\mathrm{Spec}}_{\rho,\rho_1}(\theta_0) := \left\{\theta \in \mathbb{R}^p \text{ as in (69) } \mid \|\mathbf{v}-\mathbf{v}_0\|_2 \le \rho_1,\ \|W^{(\ell)} - W_0^{(\ell)}\|_2 \le \rho, \ell \in [L]\right\} \ . \tag{73}$$

As is standard, we assume $\|\mathbf{x}\|_2 = 1$ according to Assumption 4.

We start with the following standard consequence of the Gaussian random initialization, variants of which is widely used in practice (Allen-Zhu et al., 2019; Du et al., 2019; Arora et al., 2019a; Banerjee et al., 2022).

Lemma 6. *(Banerjee et al., 2022) Under Assumption 4, for $\theta \in B^{Spec}_{\rho,\rho_1}(\theta_0)$, with probability at least $1 - \frac{2}{m_l}$,*

$$\left\|W^{(l)}\right\|_2 \le \left(\sigma_1 + \frac{\rho}{\sqrt{m_l}}\right)\sqrt{m_l} = \beta_l \sqrt{m_l}$$

with $\beta_l = \sigma_1 + \frac{\rho}{\sqrt{m_l}}$.

Proof. For a $(m_l \times m_{l-1})$ random matrix $W_0^{(l)}$ with i.i.d. entries $w^{(l)}_{0,ij} \sim \mathcal{N}(0, \sigma_0^{(l)2})$, with probability at least $1 - 2\exp(-t^2/2\sigma_0^{(l)2})$, the largest singular value of W_0 is bounded by

$$\sigma_{\max}\left(W_0^{(l)}\right) \le \sqrt{m_l} + \sqrt{m_{l-1}} + t$$

Let us choose $t = \sigma_0^l \sqrt{2\log m_l}$ so that the inequality holds with probability at least $1 - \frac{2}{m_l}$. Then, there are two cases:

- **Case 1:** $l = 1$. Since $m_0 = d$ and $m_1 \ge d$, with probability at least $1 - \frac{2}{m_1}$,

$$\left\|W_0^{(1)}\right\|_2 \le \sigma_0^{(1)}\left(\sqrt{m_1} + \sqrt{d} + \sqrt{2\log m_1}\right) \le \sigma_0^{(1)}\left(2\sqrt{m_1} + \sqrt{2\log m_1}\right) = \sigma_1\sqrt{m_1}$$

- **Case 2:** $2 \le l \le L$. With probability at least $1 - \frac{2}{m_l}$,

$$\left\|W_0^{(l)}\right\|_2 \le \sigma_0^{(l)}\left(\sqrt{m_l} + \sqrt{m_{l-1}} + \sqrt{2\log m_l}\right) = \sigma_1\sqrt{m_l}$$

Then, by triangle inequality, for $\theta \in B^{\mathrm{Spec}}_{\rho,\rho_1}(\theta_0)$,

$$\left\|W^{(l)}\right\|_2 \le \left\|W_0^{(l)}\right\|_2 + \left\|W^{(l)} - W_0^{(l)}\right\|_2 \overset{(a)}{\le} \sigma_1\sqrt{m_l} + \rho = \beta_l\sqrt{m_l}$$

Lemma 7. *Consider any $l \in [L]$. Under Assumption 3 and 4, for $\theta \in B^{Spec}_{\rho,\rho_1}(\theta_0)$, with probability at least $1 - \sum_{k=1}^{l} \frac{2}{m_k}$, we have*

$$\left\|\alpha^{(l)}\right\|_2 \le \prod_{k=1}^{l}\left(\sigma_1 + \frac{\rho}{\sqrt{m_k}}\right) = \prod_{k=1}^{l} \beta_k$$

with $\beta_k = \sigma_1 + \frac{\rho}{\sqrt{m_k}}$.

Proof. Following Allen-Zhu et al. (2019); Liu et al. (2020b), we prove the result by recursion. First, recall that since $\|\mathbf{x}\|_2^2 = 1$, we have $\|\alpha^{(0)}\|_2 = 1$. Then, since ϕ is 1-Lipschitz and $\phi(0) = 0$,

$$\left\|\phi\left(\frac{1}{\sqrt{m_1}}W^{(1)}\alpha^{(0)}\right)\right\|_2 - \|\phi(\mathbf{0})\|_2 \le \left\|\phi\left(\frac{1}{\sqrt{m_1}}W^{(1)}\alpha^{(0)}\right) - \phi(\mathbf{0})\right\|_2 \le \left\|\frac{1}{\sqrt{m_1}}W^{(1)}\alpha^{(0)}\right\|_2 \ ,$$

so that

$$\begin{aligned}\|\alpha^{(1)}\|_2 = \left\|\phi\left(\frac{1}{\sqrt{m_1}}W^{(1)}\alpha^{(0)}\right)\right\|_2 &\leq \left\|\frac{1}{\sqrt{m_1}}W^{(1)}\alpha^{(0)}\right\|_2 + \|\phi(\mathbf{0})\|_2 \\ &\leq \frac{1}{\sqrt{m_1}}\|W^{(1)}\|_2\|\alpha^{(0)}\|_2 \\ &\leq \left(\sigma_1 + \frac{\rho}{\sqrt{m_1}}\right) ,\end{aligned}$$

where we used Lemma 6 in the last inequality, which holds with probability at least $1 - \frac{2}{m_1}$. For the inductive step, we assume that for some $l-1$, we have

$$\|\alpha^{(l-1)}\|_2 \leq \prod_{k=1}^{l-1}\left(\sigma_1 + \frac{\rho}{\sqrt{m_k}}\right) ,$$

which holds with the probability at least $1 - \sum_{k=1}^{l-1}\frac{2}{m_k}$. Since ϕ is 1-Lipschitz, for layer l, we have

$$\begin{aligned}\left\|\phi\left(\frac{1}{\sqrt{m_l}}W^{(l)}\alpha^{(l-1)}\right)\right\|_2 - \|\phi(\mathbf{0})\|_2 &\leq \left\|\phi\left(\frac{1}{\sqrt{m_l}}W^{(l)}\alpha^{(l-1)}\right) - \phi(\mathbf{0})\right\|_2 \\ &\leq \left\|\frac{1}{\sqrt{m_l}}W^{(l)}\alpha^{(l-1)}\right\|_2 ,\end{aligned}$$

so that

$$\begin{aligned}\|\alpha^{(l)}\|_2 = \left\|\phi\left(\frac{1}{\sqrt{m_l}}W^{(l)}\alpha^{(l-1)}\right)\right\|_2 &\leq \left\|\frac{1}{\sqrt{m_l}}W^{(l)}\alpha^{(l-1)}\right\|_2 + \|\phi(\mathbf{0})\|_2 \\ &\leq \frac{1}{\sqrt{m_l}}\|W^{(l)}\|_2\|\alpha^{(l-1)}\|_2 \\ &\overset{(a)}{\leq} \left(\sigma_1 + \frac{\rho}{\sqrt{m_l}}\right)\|\alpha^{(l-1)}\|_2 \\ &\overset{(b)}{\leq} \prod_{k=1}^{l}\left(\sigma_1 + \frac{\rho}{\sqrt{m_k}}\right) ,\end{aligned}$$

where (a) follows from Lemma 6 and (b) from the inductive step. Since we have used Lemma 6 l times, after a union bound, our result would hold with probability at least $1 - \sum_{k=1}^{l}\frac{2}{m_k}$. This completes the proof. $\square$

Lemma 8. *Consider any $l \in \{2, \ldots, L\}$. Under Assumptions 3 and 4, for $\theta \in B^{\mathrm{Spec}}_{\rho,\rho_1}(\theta_0)$, with probability at least $\left(1 - \frac{2}{m_l}\right)$,*

$$\left\|\frac{\partial\alpha^{(l)}}{\partial\alpha^{(l-1)}}\right\|_2^2 \leq \left(\sigma_1 + \frac{\rho}{\sqrt{m_l}}\right)^2 = \beta_l^2 . \tag{74}$$

Proof. By definition, we have

$$\left[\frac{\partial\alpha^{(l)}}{\partial\alpha^{(l-1)}}\right]_{i,j} = \frac{1}{\sqrt{m_l}}\phi'(\tilde{\alpha}_i^{(l)})W_{ij}^{(l)} . \tag{75}$$

Since $\|A\|_2 = \sup_{\|\mathbf{u}\|_2=1}\|A\mathbf{u}\|_2$, so that $\|A\|_2^2 = \sup_{\|\mathbf{u}\|_2=1}\sum_i\langle\mathbf{a}_i, \mathbf{u}\rangle^2$, we have that for $2 \leq l \leq L$,

$$\left\|\frac{\partial\alpha^{(l)}}{\partial\alpha^{(l-1)}}\right\|_2^2 = \sup_{\|\mathbf{u}\|_2=1}\frac{1}{m_l}\sum_{i=1}^{m_l}\left(\phi'(\tilde{\alpha}_i^{(l)})\sum_{j=1}^{m_{l-1}}W_{ij}^{(l)}u_j\right)^2$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \sup_{\|\mathbf{u}\|_2=1} \frac{1}{m_l} \|W^{(l)}\mathbf{u}\|_2^2 \\
&= \frac{1}{m_l} \|W^{(l)}\|_2^2 \\
&\overset{(b)}{\leq} \beta_l^2 \ ,
\end{aligned}
$$

where (a) follows from ϕ being 1-Lipschitz by Assumption 3 and (b) from Lemma 6. This completes the proof. $\square$

Lemma 9. *Consider any $l \in [L]$. Denote the parameter vector $\mathbf{w}^{(l)} = vec\left(W^{(l)}\right)$. Under Assumption 3 and 4, for $\theta \in B^{Spec}_{\rho,\rho_1}(\theta_0)$ and $\beta_k = \sigma_1 + \frac{\rho}{\sqrt{m_k}}$, with probability at least $1 - \sum_{k=1}^{l-1} \frac{2}{m_k}$,*

$$
\left\| \frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}} \right\|_2^2 \leq \frac{\prod_{k=1}^{l-1} \beta_k^2}{m_l} \ .
$$

Proof. Note that the parameter vector $\mathbf{w}^{(l)} = \text{vec}(W^{(l)})$ and can be indexed with $j \in [m_1]$ and $j' \in [d]$ when $l = 1$, $j \in [m_l]$ and $j' \in [m_{l-1}]$ when $l \geq 2$. Then, we have

$$
\left[\frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}}\right]_{i,jj'} = \left[\frac{\partial \alpha^{(l)}}{\partial W^{(l)}}\right]_{i,jj'} = \frac{1}{\sqrt{m_l}} \phi'(\tilde{\alpha}_i^{(l)}) \alpha_{j'}^{(l-1)} \mathbf{1}_{[i=j]} \ . \tag{76}
$$

For $l \in \{2, \ldots, L\}$, noting that $\frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}} \in \mathbb{R}^{m_l \times m_l m_{l-1}}$ and $\|V\|_F = \|\text{vec}(V)\|_2$ for any matrix V, we have

$$
\begin{aligned}
\left\| \frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}} \right\|_2^2 &= \sup_{\|V\|_F=1} \frac{1}{m_l} \sum_{i=1}^{m_l} \left(\phi'(\tilde{\alpha}_i^{(l)}) \sum_{j=1}^{m_l} \sum_{j'=1}^{m_{l-1}} \alpha_{j'}^{(l-1)} \mathbf{1}_{[i=j]} V_{jj'} \right)^2 \\
&\leq \sup_{\|V\|_F=1} \frac{1}{m_l} \|V\alpha^{(l-1)}\|_2^2 \\
&\leq \frac{1}{m_l} \sup_{\|V\|_F=1} \|V\|_2^2 \|\alpha^{(l-1)}\|_2^2 \\
&\overset{(a)}{\leq} \frac{1}{m_l} \|\alpha^{(l-1)}\|_2^2 \\
&\overset{(b)}{\leq} \frac{1}{m_l} \prod_{k=1}^{l-1} \beta_k^2 \ ,
\end{aligned}
$$

where (a) follows from $\|V\|_2^2 \leq \|V\|_F^2$ for any matrix V, and (b) from Lemma 7.

The $l = 1$ case follows in a similar manner:

$$
\left\| \frac{\partial \alpha^{(1)}}{\partial \mathbf{w}^{(1)}} \right\|_2^2 \leq \frac{1}{m_1} \|\alpha^{(0)}\|_2^2 = \frac{1}{m_1} \|\mathbf{x}\|_2^2 = \frac{1}{m_1} \ ,
$$

which satisfies the form for $l = 1$. That completes the proof. $\square$

Before getting into the Gaussian width analysis, we need two results which will be used in our main proofs.

Lemma 10. *Let $\mathcal{A} \in \mathbb{R}^{d_1+d_2}$. For any $\mathbf{a} \in \mathcal{A}$, let $\mathbf{a}_1 = \Pi_{\mathbb{R}^{d_1}}(\mathbf{a})$, i.e., projection of $\mathbf{a}$ on the first d_1 coordinates, and let $\mathbf{a}_2 = \Pi_{\mathbb{R}^{d_2}}(\mathbf{a})$, i.e., projection of $\mathbf{a}$ on the latter d_2 coordinates. Let $\mathcal{A}_1 := \{\mathbf{a}_1 | \mathbf{a} \in \mathcal{A}\}$ and $A_2 = \{\mathbf{a}_2 | \mathbf{a} \in \mathcal{A}\}$. Then,*

$$
w(\mathcal{A}) \leq w(\mathcal{A}_1) + w(\mathcal{A}_2) \ . \tag{77}
$$

Proof. Note that $\mathcal{A} \subset \mathcal{A}_1 + \mathcal{A}_2$, where $+$ here denotes the Minkowski sum, i.e., all $\mathbf{a}' = \mathbf{a}_1 + \mathbf{a}_2$ for $\mathbf{a}_1 \in \mathcal{A}_1, \mathbf{a}_2 \in \mathcal{A}_2$. Then,

$$
w(A) \leq w(\mathcal{A}_1 + \mathcal{A}_2)
$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathbf{g}_{d_1+d_2}}\left[\sup_{\mathbf{a}_1+\mathbf{a}_2\in\mathcal{A}_1+\mathcal{A}_2}\langle\mathbf{a}_1+\mathbf{a}_2,\mathbf{g}_{d_1+d_2}\rangle\right] \\
&\leq \mathbb{E}_{\mathbf{g}_{d_1+d_2}}\left[\sup_{\mathbf{a}_1\in\mathcal{A}_1}\langle\mathbf{a}_1,\mathbf{g}_{d_1}\rangle+\sup_{\mathbf{a}_2\in\mathcal{A}_2}\langle\mathbf{a}_2,\mathbf{g}_{d_2}\rangle\right] \\
&= \mathbb{E}_{\mathbf{g}_{d_1}}\left[\sup_{\mathbf{a}_1\in\mathcal{A}_1}\langle\mathbf{a}_1,\mathbf{g}_{d_1}\rangle\right]+\mathbb{E}_{\mathbf{g}_{d_2}}\left[\sup_{\mathbf{a}_2\in\mathcal{A}_2}\langle\mathbf{a}_2,\mathbf{g}_{d_2}\rangle\right] \\
&= w(\mathcal{A}_1)+w(\mathcal{A}_2)\ .
\end{aligned}
$$

That completes the proof. □

Lemma 11. *Consider $\mathcal{A}=\{\mathbf{a}\}$ with $\mathbf{a}\in\mathbb{R}^D$ such that each $\mathbf{a}$ can be written as $\mathbf{a}=B\mathbf{v}$ for some $B\in\mathcal{B}\subset\mathbb{R}^{D\times d},\mathbf{v}\in\mathcal{V}\subset\mathbb{R}^d$. Then, there exists a constant C such that*

$$
w(\mathcal{A})\leq C\sup_{B\in\mathcal{B}}\|B\|_2 w(\mathcal{V})\ . \tag{78}
$$

Proof. According to the definition of γ_2 function, we have

$$
\begin{aligned}
\gamma_2(\mathcal{A},d) &= \inf_{\Gamma}\sup_{\mathbf{a}\in A}\sum_{k=0}^{\infty}2^{k/2}d(\mathbf{a},\mathcal{A}_k) \\
\gamma_2(\mathcal{V},d) &= \inf_{\Lambda}\sup_{\mathbf{v}\in\mathcal{V}}\sum_{k=0}^{\infty}2^{k/2}d(\mathbf{v},\mathcal{V}_k)
\end{aligned}
$$

where the infimum is taken with respect to all admissible sequences Γ of A and Λ of $\mathcal{V}$, and the distance d is the ℓ_2 distance for both spaces:

$$
\begin{aligned}
d(\mathbf{a}_1,\mathbf{a}_2) &:= \|\mathbf{a}_1-\mathbf{a}_2\|_2 \\
d(\mathbf{v}_1,\mathbf{v}_2) &:= \|\mathbf{v}_1-\mathbf{v}_2\|_2
\end{aligned}
$$

For any fixed $\mathbf{a}\in\mathcal{A}$, we can write it as $\mathbf{a}=B\mathbf{v}$. For any admissible sequence $\Lambda=\{\mathcal{V}_k\}_{k=1}^{\infty}$ of $\mathcal{V}$, we consider the following sequence:

$$
B\Lambda := \{B\mathcal{V}_k\}_{k=1}^{\infty}
$$

in which $B\mathcal{V}_k:=\{B\mathbf{v}:\mathbf{v}\in\mathcal{V}_k\}$. Since for each k, $B\mathcal{V}_k$ is a subset of $\mathcal{A}$, and $|B\mathcal{V}_k|\leq|\mathcal{V}_k|\leq 2^{2^k}$, so $B\Lambda$ is a admissible sequence of $\mathcal{A}$. Now in $\mathcal{V}$, we walk from the original point $\mathbf{v}_0$ to $\mathbf{v}$ along the chain

$$
\mathbf{v}_0=\pi_0(\mathbf{v})\to\pi_1(\mathbf{v})\to\cdots\to\pi_K(\mathbf{v})\to\cdots
$$

in which $\pi_k(\mathbf{v})\in\mathcal{V}_k$ is the best approximation to $\mathbf{v}$ in $\mathcal{V}_k$, i.e.

$$
d(\mathbf{v},\mathcal{V}_k)=d(\mathbf{v},\pi_k(\mathbf{v}))
$$

then obviously $d(\mathbf{a},B\mathcal{V}_k)\leq d(\mathbf{a},B\pi_k(\mathbf{v}))$ and $B\pi_k(\mathbf{v})\in B\mathcal{V}_k$. Therefore,

$$
\begin{aligned}
\sum_{k=0}^{\infty}2^{k/2}d(\mathbf{a},B\mathcal{V}_k) &\leq \sum_{k=0}^{\infty}2^{k/2}d(B\mathbf{v},B\pi_k(\mathbf{v}))=\sum_{k=0}^{\infty}2^{k/2}\|B\mathbf{v}-B\pi_k(\mathbf{v})\|_2 \\
&\leq \sum_{k=0}^{\infty}2^{k/2}\|B\|_2\|\mathbf{v}-\pi_k(\mathbf{v})\|_2\leq\sup_{B\in\mathcal{B}}\|B\|_2\sum_{k=0}^{\infty}2^{k/2}\|\mathbf{v}-\pi_k(\mathbf{v})\|_2 \\
&= \sup_{B\in\mathcal{B}}\|B\|_2\sum_{k=0}^{\infty}2^{k/2}d(\mathbf{v},\pi_k(\mathbf{v}))=\sup_{B\in\mathcal{B}}\|B\|_2\sum_{k=0}^{\infty}2^{k/2}d(\mathbf{v},\mathcal{V}_k) \\
&\leq \sup_{B\in\mathcal{B}}\|B\|_2\sup_{\mathbf{v}\in\mathcal{V}}\sum_{k=0}^{\infty}2^{k/2}d(\mathbf{v},\mathcal{V}_k)
\end{aligned}
$$

By taking supreme over $\mathbf{a} \in \mathcal{A}$, we can get that

$$\sup_{\mathbf{a}\in\mathcal{A}} \sum_{k=0}^{\infty} 2^{k/2} d(\mathbf{a}, B\mathcal{V}_k) \leq \sup_{B\in\mathcal{B}} \|B\|_2 \sup_{\mathbf{v}\in\mathcal{V}} \sum_{k=0}^{\infty} 2^{k/2} d\left(\mathbf{v}, \mathcal{V}_k\right)$$

Since all $B\Lambda$ are admissible sequences of $\mathcal{A}$, then by taking infimum over all admissible sequences Λ of $\mathcal{V}$, we have

$$\begin{aligned}
\gamma_2(\mathcal{A}, d) &= \inf_{\Gamma} \sup_{\mathbf{a}\in\mathcal{A}} \sum_{k=0}^{\infty} 2^{k/2} d(\mathbf{a}, \mathcal{A}_k) \leq \inf_{\Lambda} \sup_{\mathbf{a}\in\mathcal{A}} \sum_{k=0}^{\infty} 2^{k/2} d(\mathbf{a}, B\mathcal{V}_k) \\
&\leq \inf_{\Lambda} \sup_{B\in\mathcal{B}} \|B\|_2 \sup_{\mathbf{v}\in\mathcal{V}} \sum_{k=0}^{\infty} 2^{k/2} d\left(\mathbf{v}, \mathcal{V}_k\right) = \sup_{B\in\mathcal{B}} \|B\|_2 \inf_{\Lambda} \sup_{\mathbf{v}\in\mathcal{V}} \sum_{k=0}^{\infty} 2^{k/2} d\left(\mathbf{v}, \mathcal{V}_k\right) \\
&= \sup_{B\in\mathcal{B}} \|B\|_2 \, \gamma_2(\mathcal{V}, d)
\end{aligned}$$

For the mean-zero Gaussian process $X_{\mathbf{a}} = \langle \mathbf{a}, g \rangle$ on $\mathcal{A}$, in which g is a standard Gaussian vector in $\mathbb{R}^D$, we have

$$\|X_{\mathbf{a}_1} - X_{\mathbf{a}_2}\|_2^2 = \mathbb{E}_g \left|\langle \mathbf{a}_1, g\rangle - \langle \mathbf{a}_2, g\rangle\right|^2 = \mathbb{E}_g \left\langle \mathbf{a}_1 - \mathbf{a}_2, g\right\rangle^2 = \|\mathbf{a}_1 - \mathbf{a}_2\|_2^2$$

so $d\left(\mathbf{a}_1, \mathbf{a}_2\right) = \|\mathbf{a}_1 - \mathbf{a}_2\|_2 = \|X_{\mathbf{a}_1} - X_{\mathbf{a}_2}\|_2$ is the canonical distance on $\mathcal{A}$. Similarly, for the mean-zero Gaussian process $Y_{\mathbf{v}} = \langle \mathbf{v}, h \rangle$ on $\mathcal{V}$, in which h is a standard Gaussian vector in $\mathbb{R}^d$,

$$d\left(\mathbf{v}_1, \mathbf{v}_2\right) = \|\mathbf{v}_1 - \mathbf{v}_2\|_2 = \|Y_{\mathbf{v}_1} - Y_{\mathbf{v}_2}\|_2$$

is also the canonical distance on $\mathcal{V}$. Then by Talagrand's majorizing measure theorem, there are absolute constants c, C such that

$$\begin{aligned}
w\left(\mathcal{A}\right) &= \mathbb{E}_g \sup_{\mathbf{a}\in\mathcal{A}} X_{\mathbf{a}} \leq c\gamma_2\left(\mathcal{A}, d\right) \leq c \sup_{B\in\mathcal{B}} \|B\|_2 \, \gamma_2(\mathcal{V}, d) \\
&\leq C \sup_{B\in\mathcal{B}} \|B\|_2 \, \mathbb{E}_h \sup_{\mathbf{v}\in\mathcal{V}} Y_{\mathbf{v}} = C \sup_{B\in\mathcal{B}} \|B\|_2 \, w\left(\mathcal{V}\right)
\end{aligned}$$

so the inequality holds. □

We now restate and prove the main result for FFNs.

Theorem 4 (LGGW: FFNs). *Under Assumptions 3 and 4, with* $\beta_l := \sigma_1 + \frac{\rho}{\sqrt{m_l}}$, $l \in [L]$, *with probability at least* $\left(1 - \sum_{l=1}^{L} \frac{2}{m_l}\right)$ *over the randomness of the initialization, we have*

$$w(\Xi^{\text{ffn}}) \leq c_1 w(A^{(L)}) + c_2(1+\rho_1)\sqrt{m_L} \left(\prod_{l=1}^{L} \beta_l\right) \sum_{l=1}^{L} \frac{1}{\beta_l \sqrt{m_l}}.$$

Proof. The loss at any input $(\mathbf{x}, y)$ is given by $\ell(y, f(\theta; \mathbf{x}))$ with $\theta \in B^{\text{Spec}}_{\rho,\rho_1}(\theta_0)$. With $\hat{y} = f(\theta; \mathbf{x})$, let $\ell' := \frac{d\ell(y,\hat{y})}{d\hat{y}}$. Then, the gradient of the loss

$$\nabla_\theta \ell(y, f(\theta; \mathbf{x})) = \ell' \nabla_\theta f(\theta, \mathbf{x}) \ . \tag{79}$$

With $\ell' = O(1)$, we focus our analysis on the Gaussian width of the gradient of the predictor $\nabla_\theta f(\theta, \mathbf{x})$, and the Gaussian width of the loss gradient set will be bounded by a constant times the Gaussian width of predictor gradient set.

Recall that

$$f(\theta; \mathbf{x}) = \mathbf{v}^\top \phi(\frac{1}{\sqrt{m_L}} W^{(L)} \phi(\cdots \phi(\frac{1}{\sqrt{m_1}} W^{(1)} \mathbf{x})))) \ , \tag{80}$$

where

$$\theta = (\text{vec}(W^{(1)})^\top, \ldots, \text{vec}(W^{(L)})^\top, \mathbf{v}^\top)^\top \ . \tag{81}$$

For convenience, we write the model in terms of the layerwise outputs or features as:

$$\alpha^{(0)}(\mathbf{x}) = \mathbf{x} \ , \tag{82}$$

$$\alpha^{(l)} = \phi\left(\frac{1}{\sqrt{m_l}} W^{(l)} \alpha^{(l-1)}(\mathbf{x})\right) \tag{83}$$

$$f(\theta; \mathbf{x}) = \mathbf{v}^\top \alpha^{(L)}(\mathbf{x}) \ . \tag{84}$$

By Lemma 10, we bound the Gaussian width of the overall gradient by the sum of the width of the gradients of the last layer parameters $\mathbf{v} \in \mathbb{R}^m$ and that of all the intermediate parameters $W^{(l)}, \ell \in [L]$.

Starting with the last layer parameter $\mathbf{v} \in \mathbb{R}^m$, the gradient is in $\mathbb{R}^m$ and we have

$$\frac{\partial f}{\partial \mathbf{v}} = \alpha^{(L)}(\mathbf{x}) \ , \tag{85}$$

which is the output of the last layer or the so-called featurizer, i.e., $h^{(L)}(W, \mathbf{x}) = \alpha^{(L)}(\mathbf{x})$. Then, the Gaussian width is simply $w(\hat{A}_n^{(L)})$.

For any hidden layer parameter $\mathbf{w}^{(l)} \in \mathbb{R}^{m_l m_{l-1}}$ with $\mathbf{w}^{(l)} = \mathrm{vec}(W^{(l)})$, the gradient is in $\mathbb{R}^{m_l m_{l-1}}$ and we have

$$\begin{aligned} \frac{\partial f}{\partial \mathbf{w}^{(l)}} &= \frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}} \frac{\partial f}{\partial \alpha^{(l)}} \\ &= \frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}} \left(\prod_{l'=l+1}^{L} \frac{\partial \alpha^{(l')}}{\partial \alpha^{(l'-1)}} \right) \frac{\partial f}{\partial \alpha^{(L)}} \end{aligned} \tag{86}$$

Define $\mathcal{Z}^{(l)} = \left\{ Z^{(l)} \in \mathbb{R}^{m_l \times m_l m_{l-1}} : \left\| Z^{(l)} \right\|_2 \leq \frac{\prod_{k=1}^{l-1} \beta_k}{\sqrt{m_l}} \right\}$, then according to Lemma 9, $\frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}} \in \mathcal{Z}^{(l)}$ with probability $1 - \sum_{k=1}^{l-1} \frac{2}{m_k}$. Similarly, define $\mathcal{B}^{(l)} = \left\{ B^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}} : \left\| B^{(l)} \right\|_2 \leq \beta_l \right\}$, then according to Lemma 8, $\frac{\partial \alpha^{(l)}}{\partial \alpha^{(l-1)}} \in \mathcal{B}$ with probability $1 - \frac{2}{m_l}$. Besides, define $\mathcal{V}$ the set of $\mathbf{v}$, then according to (73), $\mathbf{v} \in \mathbb{R}^{m_L}$ and $\|\mathbf{v}\|_2 \leq 1 + \rho_1$, so $w(\mathcal{V}) \leq c \sup_{\mathbf{v} \in \mathcal{V}} \|\mathbf{v}\|_2 \sqrt{m_L} \leq c(1 + \rho_1)\sqrt{m_L}$.

Therefore, with $\mathcal{H}^{(l)} = \left\{ \frac{\partial f}{\partial \mathbf{w}^{(l)}} \mid W \in B_\rho^{\mathrm{Spec}}(\theta_0) \right\}$, according to (86), we have

$$\begin{aligned} w(\mathcal{H}^{(l)}) &\overset{(a)}{\leq} C \sup_{Z^{(l)} \in \mathcal{Z}^{(l)}} \|Z^{(l)}\|_2 \prod_{l'=l+1}^{L} \sup_{B^{(l')} \in \mathcal{B}^{l'}} \|B^{(l')}\|_2 \ w(\mathcal{V}) \\ &\overset{(b)}{\leq} C \frac{\prod_{k=1}^{l-1} \beta_k}{\sqrt{m_l}} \prod_{k=l+1}^{L} \beta_k \cdot c\,(1 + \rho_1)\sqrt{m_L} \\ &= C_0 \sqrt{\frac{m_L}{m_l}} (1 + \rho_1) \prod_{k \neq l} \beta_k \ . \end{aligned}$$

where (a) follows from Lemma 11 and (b) from the definition of $\mathcal{Z}^{(l)}$ and $\mathcal{B}^{(l)}$.

Define $\hat{\mathcal{H}}_n^{(l)} = \left\{ \frac{1}{n} \sum_{z \in z^{(n)}} \frac{\partial f}{\partial \mathbf{w}^{(l)}}(z) \mid W \in B_\rho^{\mathrm{Spec}}(\theta_0) \right\}$, we have

$$w(\hat{\mathcal{H}}_n^{(l)}) \leq w(\mathcal{H}^{(l)}) \leq C_0 \sqrt{\frac{m_L}{m_l}} (1 + \rho_1) \prod_{k \neq l} \beta_k$$

Since

$$\theta = (\mathrm{vec}(W^{(1)})^\top, \ldots, \mathrm{vec}(W^{(L)})^\top, \mathbf{v}^\top)^\top = \left(\mathbf{w}^{(1)\top}, ..., \mathbf{w}^{(L)\top}, \mathbf{v}^\top \right)^\top$$

According to Lemma 10, the Gaussian width of the average predictor gradient set is bounded by

$$\sum_{l=1}^{L} w(\hat{\mathcal{H}}_n^{(l)}) + w(\hat{A}_n^{(L)}) \leq \sum_{l=1}^{L} C_0 \sqrt{\frac{m_L}{m_l}} (1 + \rho_1) \prod_{k \neq l} \beta_k + w(\hat{A}_n^{(L)})$$

$$= C_0(1+\rho_1)\sqrt{m_L}\prod_{l=1}^{L}\beta_l\sum_{l=1}^{L}\frac{1}{\beta_l\sqrt{m_l}} + w(\hat{A}_n^{(L)})$$

then there exists constant c_1, c_2 such that

$$w(\hat{\Xi}_n^{\text{ffn}}) \leq c_1 w(\hat{A}_n^{(L)}) + c_2(1+\rho_1)\sqrt{m_L}\prod_{l=1}^{L}\beta_l\sum_{l=1}^{L}\frac{1}{\beta_l\sqrt{m_l}}$$

Since the spectral norm bound for each layer except the last layer holds with probability $(1-\frac{2}{m_l})$ according to Lemma 6, and we use it L times in the proof, the Gaussian width bound holds with probability $(1-\sum_{l=1}^{L}\frac{2}{m_l})$. $\square$

C.2 Gaussian Width bounds for Residual Networks (ResNets)

We now consider the setting where f is a residual network (ResNet) with depth L and widths $m_l, l \in [L] := \{1,\dots,L\}$ given by

$$\begin{aligned}
\alpha^{(0)}(\mathbf{x}) &= \mathbf{x} \ , \\
\alpha^{(l)}(\mathbf{x}) &= \alpha^{(l-1)}(\mathbf{x}) + \phi\left(\frac{1}{L\sqrt{m_l}}W^{(l)}\alpha^{(l-1)}(\mathbf{x})\right) \ , \quad l = 1,\dots,L \ , \\
f(\theta;\mathbf{x}) = \alpha^{(L+1)}(\mathbf{x}) &= \mathbf{v}^\top\alpha^{(L)}(\mathbf{x}) \ ,
\end{aligned} \tag{87}$$

where $W^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}, l \in [L]$ are layer-wise weight matrices, $\mathbf{v} \in \mathbb{R}^{m_L}$ is the last layer vector, $\phi(\cdot)$ is the smooth (pointwise) activation function, and the total set of parameters

$$\theta := (\text{vec}(W^{(1)})^\top, \dots, \text{vec}(W^{(L)})^\top, \mathbf{v}^\top)^\top \in \mathbb{R}^{\sum_{k=1}^{L} m_k m_{k-1} + m_L} \ , \tag{88}$$

with $m_0 = d$.

We start by bounding the norm of the output $\alpha^{(l)}$ of layer l.

Lemma 12. *Consider any $l \in [L]$ and let $\beta := \sigma_1 + \frac{\rho}{\sqrt{m}}$. Under Assumptions 3 and 4, for $\theta \in B_{\rho,\rho_1}^{\text{Spec}}(\theta_0)$, with probability at least $\left(1-\sum_{k=1}^{l}\frac{2}{m_k}\right)$, we have*

$$\|\alpha^{(l)}\|_2 \leq \prod_{k=1}^{l}\left(1+\frac{\beta_k}{L}\right) \ . \tag{89}$$

Proof. Following Allen-Zhu et al. (2019); Liu et al. (2020b), we prove the result by recursion. First, recall that since $\|\mathbf{x}\|_2^2 = 1$, we have $\|\alpha^{(0)}\|_2 = 1$. Then, ϕ is 1-Lipschitz,

$$\left\|\phi\left(\frac{1}{L\sqrt{m_1}}W^{(1)}\alpha^{(0)}\right)\right\|_2 - \|\phi(\mathbf{0})\|_2 \leq \left\|\phi\left(\frac{1}{L\sqrt{m_1}}W^{(1)}\alpha^{(0)}\right) - \phi(\mathbf{0})\right\|_2 \leq \left\|\frac{1}{L\sqrt{m_1}}W^{(1)}\alpha^{(0)}\right\|_2 \ ,$$

so that, using $\phi(0) = 0$, we have

$$\begin{aligned}
\|\alpha^{(1)}\|_2 &\leq \|\alpha^{(0)}\|_2 + \left\|\phi\left(\frac{1}{L\sqrt{m_1}}W^{(1)}\alpha^{(0)}\right)\right\|_2 \\
&\leq 1 + \left\|\frac{1}{L\sqrt{m_1}}W^{(1)}\alpha^{(0)}\right\|_2 \\
&\leq 1 + \frac{1}{L\sqrt{m_1}}\|W^{(1)}\|_2\|\alpha^{(0)}\|_2 \\
&\overset{(a)}{\leq} 1 + \frac{\beta_1}{L} \ ,
\end{aligned}$$

where (a) follows from Lemma 6 which holds with probability at least $1 - \frac{2}{m_1}$.

For the inductive step, we assume that for some $(l-1)$, we have

$$\|\alpha^{(l-1)}\|_2 \leq \prod_{k=1}^{l-1}\left(1+\frac{\beta_k}{L}\right) ,$$

which holds with the probability at least $1-\sum_{k=1}^{l-1}\frac{2}{m_k}$. Since ϕ is 1-Lipschitz, for layer l, we have

$$\left\|\phi\left(\frac{1}{L\sqrt{m_l}}W^{(l)}\alpha^{(l-1)}\right)\right\|_2 - \|\phi(\mathbf{0})\|_2 \leq \left\|\phi\left(\frac{1}{L\sqrt{m_l}}W^{(l)}\alpha^{(l-1)}\right)-\phi(\mathbf{0})\right\|_2 \leq \left\|\frac{1}{L\sqrt{m_l}}W^{(l)}\alpha^{(l-1)}\right\|_2 ,$$

so that, using $\phi(0)=0$, we have

$$\begin{aligned}
\|\alpha^{(l)}\|_2 &\leq \|\alpha^{(l-1)}\|_2 + \left\|\phi\left(\frac{1}{L\sqrt{m_l}}W^{(l)}\alpha^{(l-1)}\right)\right\|_2 \\
&\leq \|\alpha^{(l-1)}\|_2 + \left\|\frac{1}{L\sqrt{m_l}}W^{(l)}\alpha^{(l-1)}\right\|_2 \\
&\leq \|\alpha^{(l-1)}\|_2 + \frac{1}{L\sqrt{m_l}}\|W^{(l)}\|_2\|\alpha^{(l-1)}\|_2 \\
&\overset{(a)}{\leq} \|\alpha^{(l-1)}\|_2 + \frac{\beta_l}{L}\|\alpha^{(l-1)}\|_2 \\
&= \left(1+\frac{\beta_l}{L}\right)\|\alpha^{((l-1)}\|_2 \\
&\overset{(b)}{\leq} \prod_{k=1}^{l}\left(1+\frac{\beta_k}{L}\right) ,
\end{aligned}$$

where (a) follows from Lemma 6 and (b) follows from the inductive step. Since we have used Lemma 6 l times, after a union bound, our result would hold with probability at least $1-\sum_{k=1}^{l}\frac{2}{m_k}$. That completes the proof. □

Recall that in our setup, the layerwise outputs and pre-activations are respectively given by:

$$\alpha^{(l)} = \alpha^{(l-1)} + \phi\left(\tilde{\alpha}^{(l)}\right) , \quad \tilde{\alpha}^{(l)} := \frac{1}{L\sqrt{m_l}}W^{(l)}\alpha^{(l-1)} . \tag{90}$$

Lemma 13. *Consider any $l \in \{2,\ldots,L\}$ and let $\beta_l := \sigma_1 + \frac{\rho}{\sqrt{m_l}}$. Under Assumptions 3 and 4, for $\theta \in B^{\mathrm{Spec}}_{\rho,\rho_1}(\theta_0)$, with probability at least $\left(1-\frac{2}{m_l}\right)$,*

$$\left\|\frac{\partial\alpha^{(l)}}{\partial\alpha^{(l-1)}}\right\|_2 \leq 1 + \frac{\beta_l}{L} . \tag{91}$$

Proof. By definition, we have

$$\left[\frac{\partial\alpha^{(l)}}{\partial\alpha^{(l-1)}}\right]_{i,j} = \mathbf{1}_{[i=j]} + \frac{1}{L\sqrt{m_l}}\phi'(\tilde{\alpha}_i^{(l)})W_{ij}^{(l)} . \tag{92}$$

so that with $C^{(l)} = [c_{ij}^{(l)}], c_{ij}^{(l)} = \frac{1}{L\sqrt{m}}\phi'(\tilde{\alpha}_i^{(l)})W_{ij}^{(l)}$, we have

$$\left\|\frac{\partial\alpha^{(l)}}{\partial\alpha^{(l-1)}}\right\|_2 = \left\|\mathbb{I}+C^{(l)}\right\|_2 \leq 1 + \|C^{(l)}\|_2 . \tag{93}$$

Since $\|C^{(l)}\|_2 = \sup_{\|\mathbf{u}\|_2=1}\|C^{(l)}\mathbf{u}\|_2$, so that $\|C^{(l)}\|_2^2 = \sup_{\|\mathbf{u}\|_2=1}\sum_i\langle\mathbf{c}_i^{(l)},\mathbf{u}\rangle^2$, we have that for $2 \leq l \leq L$,

$$\left\|C^{(l)}\right\|_2^2 = \sup_{\|\mathbf{u}\|_2=1}\frac{1}{L^2 m_l}\sum_{i=1}^{m_l}\left(\phi'(\tilde{\alpha}_i^{(l)})\sum_{j=1}^{m_{l-1}}W_{ij}^{(l)}u_j\right)^2$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \sup_{\|\mathbf{u}\|_2=1} \frac{1}{L^2 m_l} \|W^{(l)}\mathbf{u}\|_2^2 \\
&= \frac{1}{L^2 m_l} \|W^{(l)}\|_2^2 \\
&\overset{(b)}{\leq} \left(\frac{\beta_l}{L}\right)^2 ,
\end{aligned}
$$

where (a) follows from ϕ being 1-Lipschitz by Assumption 3 and (b) from Lemma 6. Putting the bound back in (93) completes the proof. □

Lemma 14. *Consider any $l \in [L]$. Under Assumptions 3 and 4, for $\theta \in B^{\mathrm{Spec}}_{\rho,\rho_1}(\theta_0)$, with probability at least $\left(1 - \sum_{k=1}^{l} \frac{2}{m_k}\right)$,*

$$
\left\| \frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}} \right\|_2 \leq \frac{1}{L\sqrt{m_l}} \prod_{k=1}^{l-1} \left(1 + \frac{\beta_k}{L}\right) . \tag{94}
$$

Proof. Note that the parameter vector $\mathbf{w}^{(l)} = \mathrm{vec}(W^{(l)})$ and can be indexed with $j \in [m_1]$ and $j' \in [d]$ when $l = 1$, $j \in [m_l]$ and $j' \in [m_{l-1}]$ when $l \geq 2$. Then, we have

$$
\left[\frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}}\right]_{i,jj'} = \left[\frac{\partial \alpha^{(l)}}{\partial W^{(l)}}\right]_{i,jj'} = \frac{1}{L\sqrt{m_l}} \phi'(\tilde{\alpha}_i^{(l)}) \alpha_{j'}^{(l-1)} \mathbf{1}_{[i=j]} . \tag{95}
$$

For $l \in \{2, \ldots, L\}$, noting that $\frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}} \in \mathbb{R}^{m_l \times m_l m_{k-1}}$ and $\|V\|_F = \|\mathrm{vec}(V)\|_2$ for any matrix V, we have

$$
\begin{aligned}
\left\| \frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}} \right\|_2^2 &= \sup_{\|V\|_F=1} \frac{1}{L^2 m_l} \sum_{i=1}^{m_l} \left(\phi'(\tilde{\alpha}_i^{(l)}) \sum_{j=1}^{m_l} \sum_{j'=1}^{m_{l-1}} \alpha_{j'}^{(l-1)} \mathbf{1}_{[i=j]} V_{jj'} \right)^2 \\
&\leq \sup_{\|V\|_F=1} \frac{1}{L^2 m_l} \|V\alpha^{(l-1)}\|_2^2 \\
&\leq \frac{1}{L^2 m_l} \sup_{\|V\|_F=1} \|V\|_2^2 \|\alpha^{(l-1)}\|_2^2 \\
&\overset{(a)}{\leq} \frac{1}{L^2 m_l} \|\alpha^{(l-1)}\|_2^2 \\
&\overset{(b)}{\leq} \frac{1}{L^2 m_l} \prod_{k=1}^{l-1} \left(1 + \frac{\beta_k}{L}\right)^2
\end{aligned}
$$

where (a) follows from $\|V\|_2^2 \leq \|V\|_F^2$ for any matrix V, and (b) from Lemma 12.

The $l = 1$ case follows in a similar manner:

$$
\left\| \frac{\partial \alpha^{(1)}}{\partial \mathbf{w}^{(1)}} \right\|_2^2 \leq \frac{1}{L^2 m_1} \|\alpha^{(0)}\|_2^2 = \frac{1}{L^2 m_1} \|\mathbf{x}\|_2^2 = \frac{1}{L^2 m_1} ,
$$

which satisfies the form for $l = 1$. That completes the proof. □

We now restate and prove the main result for ResNets.

Theorem 5 (LGGW: ResNets). *Under Assumptions 3 and 4, with $\beta_l := \sigma_1 + \frac{\rho}{\sqrt{m_l}}$, $l \in [L]$, with probability at least $\left(1 - \sum_{l=1}^{L} \frac{2}{m_l}\right)$ over the randomness of the initialization, we have*

$$
\begin{aligned}
w(\Xi^{\mathrm{rn}}) \leq\ & c_1 w(A^{(L)}) \\
& + c_2 \frac{1 + \rho_1}{L} \sqrt{m_L} \prod_{l=1}^{L} \left(1 + \frac{\beta_l}{L}\right) \sum_{l=1}^{L} \frac{1}{\left(1 + \frac{\beta_l}{L}\right)\sqrt{m_l}} .
\end{aligned}
$$

Proof. The loss at any input $(\mathbf{x}, y)$ is given by $\ell(y, f(\theta; \mathbf{x}))$ with $\theta \in B^{\text{Spec}}_{\rho,\rho_1}(\theta_0)$. With $\hat{y} = f(\theta; \mathbf{x})$, let $\ell' := \frac{d\ell(y,\hat{y})}{d\hat{y}}$. Then, the gradient of the loss

$$\nabla_\theta \ell(y, f(\theta; \mathbf{x})) = \ell' \nabla_\theta f(\theta, \mathbf{x}) \ . \tag{96}$$

With $\ell' = O(1)$, we focus our analysis on the Gaussian width of the gradient of the predictor $\nabla_\theta f(\theta, \mathbf{x})$, and the Gaussian width of the loss gradient set will be bounded by a constant times the Gaussian width of predictor gradient set.

Recall that

$$f(\theta; x) = \mathbf{v}^\top \phi(\frac{1}{\sqrt{m_L}} W^{(L)} \phi(\cdots \phi(\frac{1}{\sqrt{m_1}} W^{(1)} \mathbf{x})))) \ , \tag{97}$$

where

$$\theta = (\text{vec}(W^{(1)})^\top, \ldots, \text{vec}(W^{(L)})^\top, \mathbf{v}^\top)^\top \ . \tag{98}$$

For convenience, we write the model in terms of the layerwise outputs or features as:

$$\alpha^{(0)}(\mathbf{x}) = \mathbf{x} \ , \tag{99}$$

$$\alpha^{(l)} = \alpha^{(l-1)} + \phi\left(\frac{1}{L\sqrt{m_l}} W^{(l)} \alpha^{(l-1)}(\mathbf{x})\right) \tag{100}$$

$$f(\theta; \mathbf{x}) = \mathbf{v}^\top \alpha^{(L)}(\mathbf{x}) \ . \tag{101}$$

By Lemma 10, we bound the Gaussian width of the overall gradient by the sum of the width of the gradients of the last layer parameters $\mathbf{v} \in \mathbb{R}^{m_L}$ and that of all the intermediate parameters $W^{(l)}, \ell \in [L]$.

Starting with the last layer parameter $\mathbf{v} \in \mathbb{R}^m$, the gradient is in $\mathbb{R}^m$ and we have

$$\frac{\partial f}{\partial \mathbf{v}} = \alpha^{(L)}(\mathbf{x}) \ , \tag{102}$$

which is the output of the last layer or the so-called featurizer. Then, the Gaussian width of the average set is simply $w(\hat{A}^{(L)}_n)$

For any hidden layer parameter $\mathbf{w}^{(l)} \in \mathbb{R}^{m_l m_{l-1}}$ with $\mathbf{w}^{(l)} = \text{vec}(W^{(l)})$, the gradient is in $\mathbb{R}^{m_l m_{l-1}}$ and we have

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{w}^{(l)}} &= \frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}} \frac{\partial f}{\partial \alpha^{(l)}} \\ &= \frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}} \left(\prod_{l'=l+1}^{L} \frac{\partial \alpha^{(l')}}{\partial \alpha^{(l'-1)}}\right) \frac{\partial f}{\partial \alpha^{(l)}}\end{aligned} \tag{103}$$

Define $\mathcal{Z}^{(l)} = \left\{Z^{(l)} \in \mathbb{R}^{m_l \times m_l m_{l-1}} : \left\|Z^{(l)}\right\|_2 \le \frac{1}{L\sqrt{m_l}} \prod_{k=1}^{l-1}\left(1 + \frac{\beta_k}{L}\right)\right\}$, then according to Lemma 14, $\frac{\partial \alpha^{(l)}}{\partial \mathbf{w}^{(l)}} \in \mathcal{Z}^{(l)}$ with probability $1 - \sum_{k=1}^{l-1} \frac{2}{m_k}$. Similarly, define $\mathcal{B}^{(l)} = \left\{B^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}} : \left\|B^{(l)}\right\|_2 \le 1 + \frac{\beta_l}{L}\right\}$, then according to Lemma 13, $\frac{\partial \alpha^{(l)}}{\partial \alpha^{(l-1)}} \in \mathcal{B}^{(l)}$ with probability $1 - \frac{2}{m_l}$. Besides, define $\mathcal{V}$ the set of $\mathbf{v}$, then according to (73), $\mathbf{v} \in \mathbb{R}^{m_L}$ and $\|\mathbf{v}\|_2 \le 1 + \rho_1$, so $w(\mathcal{V}) \le c \sup_{\mathbf{v} \in \mathcal{V}} \|\mathbf{v}\|_2 \sqrt{m_L} \le c(1+\rho_1)\sqrt{m_L}$.

Therefore, with $\mathcal{H}^{(l)} = \left\{\frac{\partial f}{\partial \mathbf{w}^{(l)}} \mid W \in B^{\text{Spec}}_\rho(\theta_0)\right\}$, according to (103), we have

$$\begin{aligned}w(\mathcal{H}^{(l)}) &\overset{(a)}{\le} C \sup_{Z^{(l)} \in \mathcal{Z}^{(l)}} \|Z^{(l)}\|_2 \prod_{l'=l+1}^{L} \sup_{B^{(l)} \in \mathcal{B}^{(l)}} \|B^{(l)}\|_2 \, w(\mathcal{V}) \\ &\overset{(b)}{\le} C \frac{1}{L\sqrt{m_l}} \prod_{k=1}^{l-1}\left(1 + \frac{\beta_k}{L}\right) \prod_{k=l+1}^{L}\left(1 + \frac{\beta_k}{L}\right) \cdot c\,(1+\rho_1)\sqrt{m_L} \\ &= C_0 \frac{1+\rho_1}{L} \sqrt{\frac{m_L}{m_l}} \prod_{k \ne l}\left(1 + \frac{\beta_k}{L}\right)\end{aligned}$$

where (a) follows from Lemma 11, (b) from the definition of $\mathcal{Z}^{(l)}$ and $\mathcal{B}^{(l)}$. Define $\hat{\mathcal{H}}_n^{(l)} = \left\{ \frac{1}{n} \sum_{z \in z^{(n)}} \frac{\partial f}{\partial \mathbf{w}^{(l)}}(z) \mid W \in B_\rho^{\text{Spec}}(\theta_0) \right\}$, we have

$$w(\hat{\mathcal{H}}_n^{(l)}) \leq w(\mathcal{H}^{(l)}) \leq C_0 \frac{1+\rho_1}{L} \sqrt{\frac{m_L}{m_l}} \prod_{k \neq l} \left(1 + \frac{\beta_k}{L}\right)$$

Since

$$\theta = (\text{vec}(W^{(1)})^\top, \ldots, \text{vec}(W^{(L)})^\top, \mathbf{v}^\top)^\top = \left(\mathbf{w}^{(1)\top}, ..., \mathbf{w}^{(L)\top}, \mathbf{v}^\top\right)^\top$$

According to Lemma 10, the Gaussian width of the average predictor gradient set is bounded by

$$\begin{aligned}
\sum_{l=1}^{L} w(\hat{\mathcal{H}}_n^{(l)}) + w(\hat{A}_n^{(L)}) &\leq \sum_{l=1}^{L} C_0 \frac{1+\rho_1}{L} \sqrt{\frac{m_L}{m_l}} \prod_{k \neq l} \left(1 + \frac{\beta_k}{L}\right) + w(\hat{A}_n^{(L)}) \\
&= C_0 \frac{1+\rho_1}{L} \sqrt{m_L} \prod_{l=1}^{L} \left(1 + \frac{\beta_l}{L}\right) \sum_{l=1}^{L} \frac{1}{\left(1 + \frac{\beta_l}{L}\right)\sqrt{m_l}} + w(\hat{A}_n^{(L)})
\end{aligned}$$

then there exists constant c_1, c_2 such that

$$w(\hat{\Xi}_n^{\text{rn}}) \leq c_1 w(\hat{A}_n^{(L)}) + c_2 \frac{1+\rho_1}{L} \sqrt{m_L} \prod_{l=1}^{L} \left(1 + \frac{\beta_l}{L}\right) \sum_{l=1}^{L} \frac{1}{\left(1 + \frac{\beta_l}{L}\right)\sqrt{m_l}}$$

Since the spectral norm bound for each layer except the last layer holds with probability $(1 - \frac{2}{m_l})$ according to Lemma 6, and we use it L times in the proof, the Gaussian width bound holds with probability $(1 - \sum_{l=1}^{L} \frac{2}{m_l})$. $\square$