The effect of Leaky ReLUs on the training and generalization of overparameterized networks

Yinglong Guo
School of Mathematics
University of Minnesota
Minneapolis, MN 55455

Shaohan Li
School of Mathematics
University of Minnesota
Minneapolis, MN 55455

Gilad Lerman
School of Mathematics
University of Minnesota
Minneapolis, MN 55455

Abstract

We investigate the training and generalization errors of overparameterized neural networks (NNs) with a wide class of leaky rectified linear unit (ReLU) functions. More specifically, we carefully upper bound both the convergence rate of the training error and the generalization error of such NNs and investigate the dependence of these bounds on the Leaky ReLU parameter, α. We show that $\alpha = -1$, which corresponds to the absolute value activation function, is optimal for the training error bound. Furthermore, in special settings, it is also optimal for the generalization error bound. Numerical experiments empirically support the practical choices guided by the theory.

1 INTRODUCTION

Deep neural networks (DNNs) have demonstrated remarkable success in diverse fields, including image classification and text recognition. Despite their achievements, a comprehensive understanding of these networks remains elusive. Theoretical justifications for their performance have primarily centered around the overparameterized setting and mainly considered a rectified linear unit (ReLU). This paper aims to extend and generalize insights gained from recent theoretical works to any Leaky ReLU and provide practical guidance on selecting the most suitable Leaky ReLU for overparameterized networks. By doing so, we offer valuable insights for practitioners seeking optimal performance in real-world scenarios.

To address our aim, we begin by reviewing two recent theoretical trends. The first centers around a fundamental convergence theory for the training error of overparameterized neural networks (NNs). Its pioneering work by Jacot et al. (2018) studied the training dynamics using the neural tangent kernel and showed that the training error goes to zero in the asymptotic regime where the width of the layers goes to infinity. A more reasonable regime assumes a sufficiently large lower bound on the width. In such overparameterized regime, (Goodfellow et al., 2015) empirically noticed that the corresponding NNs can avoid local minima and converge to their global optimal solutions. (Du et al., 2019) proved the convergence of gradient descent (GD) for NNs with smooth and Lipschitz continuous activation functions whose width exponentially depends on the depth of the networks and polynomially depends on the number of samples. For 2-layer NNs with a ReLU, Li and Liang (2018) proved the convergence of the training error, Oymak and Soltanolkotabi (2020) reduced the width requirement for training convergence, and Song et al. (2021) established convergence whenever the width sub-quadratically depends on the number of samples and the activation functions are sufficiently smooth.

For DNNs, it has become common to consider the polynomial regime of overparameterization, where the NN widths polynomially depend both on the numbers of samples and the depths. Allen-Zhu et al. (2019b) established the first convergence result for the training error in this polynomial regime, while assuming ReLU activation functions. They separately analyzed training by gradient descent and stochastic gradient descent (SGD). Zou and Gu (2019) improved the estimates of Allen-Zhu et al. (2019b) by enhancing the lower bound of the gradient. Chen et al. (2019) further improved the polynomial dependence of the width on the number of samples that was established in Zou and Gu (2019), but on the other hand, their polynomial dependence on the depth is worse. Banerjee et al. (2023) showed that for smooth activation functions a linear dependence of the width on the number of samples is sufficient to guarantee convergence.

Another recent progress involves bounding the generalization error of overparameterized NNs. Chizat and Bach (2020) established a generalization bound of infinitely wide two-layer NNs with homogeneous activation functions

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

for classification and showed that the probability of the misclassification bound goes to 0 as the size of the training samples increases. Arora et al. (2019) bounded the generalization error of 2-layer overparameterized NNs for classification. They also analyzed the class of functions that are learnable by two-layer NNs. Allen-Zhu et al. (2019a) studied the generalization error of two-layer and three-layer NN with a non-negative, convex, and 1-Lipschitz smooth loss function using stochastic gradient descent. They showed that overparameterization improves generalization. Cao and Gu (2020) further established the generalization error of deep NNs for classification using gradient descent. Zhu et al. (2022) extended the latter work for classification by using some other activation functions, including leaky ReLU with $\alpha \in (0,1)$.

However, these foundational and important works have not yet provided much practical guidance for designing NNs. Practitioners often use variants of ReLU for activation and this work aims to provide guidance on their choices. Leaky ReLU is widely used in DNNs for supervised learning tasks (Redmon et al., 2016; Ridnik et al., 2021) and for generative tasks (Radford et al., 2015; Chen et al., 2016; Karras et al., 2019; Wang et al., 2021). It is represented by the function $\sigma_\alpha(x)$, where $\sigma_\alpha(x)=x$ for $x>0$ and $\sigma_\alpha(x)=\alpha x$ for $x\leq 0$, with α being a parameter. ReLU is a special case of Leaky ReLU when $\alpha=0$. The Leaky ReLU function aims to prevent zero gradients for negative inputs, thus avoiding neurons from not activating. Empirical studies have demonstrated the advantage of using Leaky ReLU with small $\alpha > 0$ over ReLU (Xu et al., 2015). However, theoretical studies have primarily focused on ReLU and have not directly established the convergence theory and generalization for regression when using Leaky ReLU with any $\alpha<1$. Moreover, the optimal choice of the Leaky ReLU parameter α to expedite the training process and enhance generalization remains unclear. Therefore, a theoretical study is needed to analyze the efficacy of leaky ReLU during training and to provide guidance on selecting the parameter α in practice.

This paper studies overparameterized DNNs with a wide class of leaky ReLU activation functions and develops theories for the convergence of the training error and the upper bound of the generalization error. It builds on the proof framework and techniques introduced in previous studies, in particular, the ones of Allen-Zhu et al. (2019b), Zou and Gu (2019) and Cao and Gu (2020), but establishes the dependence of the convergence rate and the generalization error on the leaky ReLU parameter α. It reveals that the optimal convergence rate bound is achieved at $\alpha=-1$ and the optimal bound of the generalization error is achieved at $\alpha=-1$ using small training epochs as long as the NN is sufficiently deep and the dataset is sufficiently large. This means that activation by the absolute value function may outperform activation by ReLU and the commonly used leaky ReLU (with small $\alpha>0$) in terms of faster training convergence and smaller generalization error. We are not aware of any prior use of the absolute value function for activating DNNs. We are only aware of using it for activating the scattering network (Mallat, 2012) due to its help with "energy preservation" (Bruna and Mallat, 2013).

The main contributions of the current work are as follows:

1. We establish the convergence of the training errors in overparameterized NNs with any leaky ReLU using both GD and SGD. Our estimates clarify the effect of the Leaky ReLU parameter α on the convergence rate bound. In particular, $\alpha=-1$, yields the optimal bound.
2. We upper bound the generalization error for overparameterized NNs for regression with leaky ReLUs. For sufficiently large datasets, deep NNs and small training epochs, the bound is optimal at $\alpha=-1$.
3. We improve previous results for ReLU (see §4.2). In particular, we show that deep NNs achieve a similar convergence rate as a shallow NN.
4. Our predictions receive substantial support from a comprehensive set of numerical experiments

The rest of the paper unfolds as follows: §2 details the assumed setup of the NNs and the training algorithms; §3 presents the main theorems; §4 describes our technical contributions and sketches the proof of the main theorems; §5 provides extensive numerical tests supporting our predictions from the theory on synthetic and real datasets; and §6 concludes this work and discusses its limitations.

2 PROBLEM SETUP

We follow the model of Allen-Zhu et al. (2019b), while allowing a wide class of Leaky ReLU activation functions. We consider a dataset $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^n$, where $\boldsymbol{x}_i\in\mathbb{R}^p$, $\|\boldsymbol{x}_i\|=1$, $\boldsymbol{y}_i\in\mathbb{R}^d$, $\|\boldsymbol{y}_i\|\leq O(1)$ and $d<O(1)$. We focus on a NN $\mathcal{N}$: $\mathbb{R}^p\to\mathbb{R}^d$ with L hidden layers having m neurons each and linear input and output layers. Its input layer produces $\boldsymbol{h}_0=\boldsymbol{A}\boldsymbol{x}$, where ,$\boldsymbol{A}\in\mathbb{R}^{m\times p}$. For $l\in[L]:=\{1,2,...L\}$, the output of the lth hidden layer, $\boldsymbol{h}_l$, is inductively defined by

$$\boldsymbol{h}_l=\mathcal{H}_l(\boldsymbol{h}_{l-1})=\sigma_\alpha(\boldsymbol{W}_l\boldsymbol{h}_{l-1}), \tag{1}$$

where $\boldsymbol{W}_l\in\mathbb{R}^{m\times m}$ and σ_α is the leaky ReLU activation function with $\alpha<1$:

$$\sigma_\alpha(x)=\begin{cases} x, & \text{if } x\geq 0;\\ \alpha x, & \text{if } x<0.\end{cases}$$

The output layer produces $\hat{\boldsymbol{y}}=\boldsymbol{B}\boldsymbol{h}_L$, where $\boldsymbol{B}\in\mathbb{R}^{d\times m}$. Let $\boldsymbol{W}:=(\boldsymbol{W}_1,\boldsymbol{W}_2,...\,\boldsymbol{W}_L)$ store all the trainable parameters and we thus compactly denote $\hat{\boldsymbol{y}}=\mathcal{N}(\boldsymbol{x};\boldsymbol{W})$. For simplicity, we initialize $\boldsymbol{A}$ and $\boldsymbol{B}$ (see below), so they are fixed, and only train $\boldsymbol{W}_l$, $l\in[L]$.

We train the NN using the mean squared error (MSE): $\mathcal{L}(\boldsymbol{W})=\sum_{i=1}^n\|\boldsymbol{y}_i-\mathcal{N}(\boldsymbol{x}_i;\boldsymbol{W})\|^2$. We denote its gradient

by $\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}):=(\nabla_{\boldsymbol{W}_1}\mathcal{L}(\boldsymbol{W}),...\nabla_{\boldsymbol{W}_L}\mathcal{L}(\boldsymbol{W}))$. Appendix B.12 extends our theory to many other useful loss functions. We assume a specified upper bound $\epsilon>0$ on the training error and express our estimates in terms of ϵ.

When discussing generalization, we assume that the set $\{\boldsymbol{x}_i\}_{i=1}^n$ is i.i.d. drawn from an arbitrary distribution $\mathcal{D}_{\boldsymbol{X}}$ and that for $1\le i\le n$, $\boldsymbol{y}_i=F(\boldsymbol{x}_i)$ for an arbitrary measurable function F. The generalization error is thus $R(\boldsymbol{W}):=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\boldsymbol{X}}}\|F(\boldsymbol{x})-\mathcal{N}(\boldsymbol{x};\boldsymbol{W})\|^2$.

We assume the following data separation property:

Assumption 2.1. There exists $0<\delta<c_0$, where $c_0<1$, so that $\min_{i,j\in[n]}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|\ge\delta>0$.

This assumption, suggested by Allen-Zhu et al. (2019b), is reasonable. Indeed, if, on the other hand, there exists $i\neq j\in[n]$ such that $\boldsymbol{x}_i=\boldsymbol{x}_j$, then we can assume $\boldsymbol{y}_i\neq\boldsymbol{y}_i$ (otherwise we can combine these multiple instances into one single data point). It is then impossible to obtain a zero training error, which is needed for our convergence study.

Algorithm 1 Rescaled initialization

Input: Input dimension p, width of hidden layer m, output dimension d, and leaky ReLU parameter α.
Initialize:

$$\boldsymbol{A}\sim N\left(0,\frac{1}{m}\right),\ \boldsymbol{B}\sim N\left(0,\frac{1}{d}\right),$$
$$\boldsymbol{W}_l^{(0)}\sim N\left(0,\frac{2}{m}\right),\ l\in[L]$$

Activation function:

$$\tilde{\sigma}_\alpha(x)=\begin{cases}\frac{1}{\sqrt{1+\alpha^2}}x, & \text{if } x\ge0\\ \frac{\alpha}{\sqrt{1+\alpha^2}}x, & \text{if } x<0\end{cases}\tag{2}$$

Following He et al. (2015), we initialize the network parameters as follows: $\boldsymbol{A}\sim N(0,1/m)$, $\boldsymbol{B}\sim N(0,1/d)$ and $\boldsymbol{W}_l^{(0)}\sim N(0,2/(m(1+\alpha^2)))$ for $l\in[L]$. Note that the factor $1/(1+\alpha^2)$ ensures a constant variance for any choice of α. We can move the factor $1/(1+\alpha^2)$ from the weight initialization to the activation function, and equivalently initialize with Algorithm 1. The theoretical study of the latter formulation with its rescaled Leaky ReLU function, $\tilde{\sigma}_\alpha(x)$ (see (2)), turns out to be more tractable.

Algorithms 2 and 3 formulate the training procedures with simple GD and SGD, respectively.

Algorithm 2 Training (gradient descent)

Input: Learning rate η.
Initialize: Apply Algorithm 1 to obtain $\boldsymbol{A}$,$\boldsymbol{B}$ and $\boldsymbol{W}^{(0)}$
for $t=0$ **to** T **do**

$$\boldsymbol{W}^{(t+1)}=\boldsymbol{W}^{(t)}-\eta\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{(t)}).$$

end for

Algorithm 3 Training (stochastic gradient descent)

Input Learning rate η.
Initialize: Apply Algorithm 1 to obtain $\boldsymbol{A}$,$\boldsymbol{B}$ and $\boldsymbol{W}^{(0)}$
for $t=0$ **to** T **do**
Randomly select batch $B\subset[n]$ with $|B|=b$.

$$\boldsymbol{W}^{(t+1)}=\boldsymbol{W}^{(t)}-\eta\nabla_{\boldsymbol{W}}\mathcal{L}_B(\boldsymbol{W}^{(t)}),$$

where $\mathcal{L}_B(\boldsymbol{W}^{(t)}):=\sum_{i\in B}\|\boldsymbol{y}_i-\mathcal{N}(\boldsymbol{x}_i;\boldsymbol{W}^{(t)})\|^2$.
end for

3 MAIN RESULTS

The two theorems below establish the convergence of the training error for overparameterized NNs using a Leaky ReLU function with $\alpha<1$. The first theorem pertains to training with gradient descent (GD) (Algorithm 2), while the second applies to training with stochastic gradient descent (SGD) (Algorithm 3). Both theorems are formulated within the context outlined in §2. This setup includes Assumption 2.1 with a parameter δ, Algorithm 1 for the initialization of the parameters of the NN, n training points, $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^n$, where $\|\boldsymbol{x}_i\|=1$, and $\|\boldsymbol{y}_i\|\le O(1)$, output dimension d ($\boldsymbol{y}_i\in\mathbb{R}^d$), NN depth L, NN width m, Leaky ReLU parameter α, learning rate η, batch size b (for Algorithm 3) and a desired upper bound $\epsilon>0$ on the training error.

Theorem 3.1. *Assume the setup of §2, where both* $m/\ln^4 m>\frac{1+\alpha^2}{(1-\alpha)^2}\Omega(\frac{n^5L^{15}d}{\delta^4})$ *and* $m>\Omega\big(\ln\ln\epsilon^{-1}\big)$, *and the training is according to Algorithm 2 with learning rate* $\eta\le O(\frac{d}{nL^2m})$. *Then, with probability at least* $1-e^{-\Omega(\ln m)}$,

$$\mathcal{L}(\boldsymbol{W}^{(T)})<\epsilon\ \text{ and }\ \mathcal{L}(\boldsymbol{W}^{(t)})\le\gamma^t\mathcal{L}(\boldsymbol{W}^{(0)}),\ \forall t\le T,\tag{3}$$

where

$$\gamma=1-\Omega\left(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{\eta\delta m}{nd}\right),\ T=\frac{\ln\big(\epsilon/\mathcal{L}(\boldsymbol{W}^{(0)})\big)}{\ln\gamma}.\tag{4}$$

Theorem 3.2. *Assume the setup of §2, where both* $\frac{m}{\ln^4 m}>\frac{(1+\alpha^2)^4}{(1-\alpha)^8}\Omega(\frac{n^8L^{15}d}{b\delta^5})$ *and* $m\ln m>\Omega\left(\ln\ln\epsilon^{-1}\right)$ *and the NN is trained according to Algorithm 3 with* $\eta\le O(\frac{d\delta}{mn^3L^3\ln^2 m})$ *and* $t>\frac{(1+\alpha^2)^2}{(1-\alpha)^4}\Omega(\frac{n^5L^2}{b\delta^2}\ln^2 m)$. *There exists a constant* $C_0>1$ *such that*

$$\begin{gathered}\mathcal{L}(\boldsymbol{W}^{(T)})<\epsilon\ \text{and}\ \mathcal{L}(\boldsymbol{W}^{(t)})\le C_0\gamma^t\mathcal{L}(\boldsymbol{W}^{(0)})\\ \text{for all } t\le T \text{ with probability } 1-e^{-\Omega(\ln m)},\end{gathered}\tag{5}$$

where

$$\gamma=1-\Omega\left(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{\eta b\delta m}{n^2d}\right),\ T=\frac{\ln\big(\epsilon/C_0\mathcal{L}(\boldsymbol{W}^{(0)})\big)}{\ln\gamma}.\tag{6}$$

These theorems show that for any $\alpha<1$ the training error linearly converges to zero when the NN width is sufficiently large and the learning rate η is sufficiently small.

Moreover, these theorems reveal the dependence of the convergence rate bound on α and this information can guide one in selecting α for optimal training speed. We note that the typical choice of the leaky ReLU parameter α (e.g., 0.01 or 0.05) does not yield a better bound for the convergence speed than ReLU (i.e., $\alpha=0$); furthermore, the negative values of α yield better results than ReLU and the optimal choice of α is -1. In §4.1 we interpret this optimality in terms of obtaining the largest derivative gap of the rescaled leaky ReLU at 0. We mathematically formulate the above observation as follows:

Corollary 3.3. *Assume the setup of §2 with either Algorithms 2 or 3 and that all parameters are chosen so that when $\alpha=0$, $\gamma<1$. Then $\alpha=-1$ minimizes the above convergence rate γ among all $\alpha<1$. Moreover, γ is decreasing in α on $(-\infty,-1)$ and increasing on $(-1,1)$.*

For $\alpha=0$, our result improves the previous analysis of both Allen-Zhu et al. (2019b) and Zou and Gu (2019). We compare our bounds with the ones of Zou and Gu (2019), since they improved the bounds of Allen-Zhu et al. (2019b). For this purpose, we examine the difference in the setups. First, Zou and Gu (2019) divides the loss function $\mathcal{L}(\boldsymbol{W})$ by n and thus we need to convert their estimate by a factor of a power of n accordingly. Second, our proof assumes that the hidden signals are separated by $\delta<O(1)$, whereas Zou and Gu (2019) assumes that $\delta<O(1/L)$. We establish this upper bound independently of L with careful mathematical estimates; therefore, our setup eliminates implicit dependence on L in the other formulas. At last, Zou and Gu (2019) enforces the initial scaled loss to be bounded by $O(1)$ (this amounts to a bound $O(n)$ on our loss) and their conclusion holds with probability at least $1-\Omega(1/n)$. On the other hand, we relax the initial unscaled loss to be bounded by $O(\sqrt{\ln m})$ and our conclusion holds with probability at least $1-e^{-\Omega(lnm)}$, which we find more natural for the overparameterized regime.

After converting to our setup, the convergence rate in Zou and Gu (2019) is $1-\Omega(\eta\delta m/(dnL))$ when using gradient descent, and our convergence rate improves to $1-\Omega(\eta\delta m/(dn))$; also, when using SGD the convergence rate in Zou and Gu (2019) is $1-\Omega(\eta\delta mb/(dn^2L))$ and we improve it to $1-\Omega(\eta\delta mb/(dn^2))$. The important finding is that in the overparameterized regime, a deeper NN does not lead to slower convergence, but rather achieves a similar convergence rate as a shallow NN. One can further note that we improved the bound of Zou and Gu (2019) on m by the factor $n^{-3}L^{-1}$ for GD and $n^{-8}L^{-2}(n/b)^{-3}\delta^3$ for SGD. Furthermore, our lower bound on the number of epochs t in Theorem 3.2 improves the one of Allen-Zhu et al. (2019b) by a factor of order $n^{-2}L^{-2}$, where there is no explicit bound in Zou and Gu (2019).

Appendix B.12 extends the above bounds to convex loss functions, which include the cross-entropy for classification and a special loss function proposed in Kumar et al. (2023). The convergence rate for these functions is different, but $\alpha=-1$ is still optimal for their bounds.

Next, we establish an upper bound of the generalization error of a NN trained using GD, where an analogous bound when using SGD is specified in Theorem B.12 in Appendix B.10. We first follow the previous analysis of generalization in overparameterized NNs by Cao and Gu (2020) and establish the corresponding bound for our setting with Leaky ReLU activation function.

Theorem 3.4. *Assume the setup of §2 with GD, where $m=\Theta(\frac{n^{10+2\tau}L^{15+2\tau}d^{1+2\tau}}{\delta^{4-2\tau}})$ for $\tau>0$ and $\eta=\Theta(\frac{d}{nL^2m})$. Assume further that m is larger than its lower bound and η is smaller than its upper bound in Theorem 3.1 (by an appropriate choice of the hidden constants in Θ and compared to the constants hidden in the lower bound of m and in the upper bound of η in Theorem 3.1). Then at a given training epoch $t\leq T$ (see (4) for T), with probability at least $1-e^{-\Omega(\ln m)}$, the generalization error is bounded as follows*

$$R(\boldsymbol{W}^{(t)})\leq\gamma^t\mathcal{L}(\boldsymbol{W}^{(0)})+\min\Bigg\{O\bigg(\frac{d^{3/2+\tau}\delta^\tau n^{1/2+\tau}}{L^{1/2-\tau}\ln m}\bigg),\ O\bigg(\frac{1-\alpha}{\sqrt{1+\alpha^2}}\frac{d^{1/3}t^{4/3}}{m^{1/6}n^{2/3}L^{2/3}}\bigg)\Bigg\}+\min\Bigg\{O\bigg(\frac{\sqrt{d\ln m}\ t}{nL}\bigg),\ O\bigg(\frac{n^{1/2+\tau}L^{2+\tau}d^{1/2+\tau}}{\delta^{1/2-\tau}\ln m}\bigg)\Bigg\}+O\bigg(d\sqrt{\frac{\ln m}{n}}\bigg). \quad (7)$$

In Appendix A, we clarify the above estimates for different regimes for the number of training epochs, t. In particular, we indicate a tradeoff between the first training term and the other NN-complexity terms (excluding the last term of data complexity) and show that we cannot make both of these kinds of terms sufficiently small. Stopping at a sufficiently small number of epochs results in a bound of the generalization error of order $O(\ln(m))$, which is also of order $O(\ln(n))$. This bound is composed of several terms. The term which contributes $O(\ln(m))$ is due to the training error and one cannot expect a better bound for it when having a small number of epochs. The rest of the terms do converge when n and L are sufficiently large and in this latter regime the overall bound is minimized when $\alpha=-1$. On the other hand, for larger numbers of epochs overfitting is observed, which results in divergent generalization error. Exploring the dependence of the generalization bound on t is advantageous to an epoch-independent bound, like the one pursued by Cao and Gu (2020) for classification instead of regression. Indeed, the bound of Cao and Gu (2020) is $\Theta(\text{poly}(n)\cdot n^{-1/2})$, which is significantly larger than $O(\log(n))$.

For very special datasets (e.g., single-layer ReLU NN separability) Cao and Gu (2020) reduced the term poly(n) so their overall bound is sufficiently small. A natural, but more complicated, extension of this to regression is to consider datasets well-approximated by L-layer leaky ReLU NNs. In Appendix B.11, we improve the convergence rate, the lower bound of m (so its dependence on n is linear) and the generalization error bound for such datasets. However, for a large number of epochs we still notice overfitting with divergent generalization error (with a smaller rate of increase to infinity than for general datasets).

At last, Kumar et al. (2023) claimed that when using the loss function discussed in (137) of Appendix B.12, minimizing a particular generalization error bound is equivalent to minimizing the latter loss function for training. Therefore, if $\alpha=-1$ is optimal for the training error, then it is also optimal for the generalization error bound. Since we verified the optimality of $\alpha=-1$ for our upper bound of the convergence rate in Appendix B.12 and experimentally demonstrated instances where this bound is comparable to the actual convergence rate in Figure 2, we get some numerical evidence that for the latter instances $\alpha=-1$ is optimal for bounding the generalization error.

4 IDEAS OF PROOFS

Our proofs follow ideas of Allen-Zhu et al. (2019b), Zou and Gu (2019) and Cao and Gu (2020) and adapt them to the general case of Leaky ReLU with $\alpha<1$. It also adapts Cao and Gu (2020) to regression. We first sketch in §4.1 the basic ideas of our proofs, while we supplement all details in the appendix. We then highlight some of the innovative ideas in §4.2.

4.1 Sketch of Proofs

We describe here a quick roadmap to verifying the theory. The proofs of Theorems 3.1 and 3.2 follow the initial framework of Allen-Zhu et al. (2019b), which was later followed by Zou and Gu (2019), but consider the effect of using any leaky RELU with $\alpha<1$.

These proofs use the following two lemmas, which are proved in §B.5 and §B.4. Let us first clarify their notation. We denote by $\|\boldsymbol{X}\|_2$ and $\|\boldsymbol{X}\|$ the spectral and Frobenius norms of a matrix $\boldsymbol{X}$. For $\boldsymbol{W}=(\boldsymbol{W}_1...\boldsymbol{W}_L)$ and $\boldsymbol{V}=(\boldsymbol{V}_1...\boldsymbol{V}_L)$, we define $\|(\boldsymbol{W}_1...\boldsymbol{W}_L)\|_F^2:=\sum_{l\in[L]}\|\boldsymbol{W}_l\|_F^2$, $\|(\boldsymbol{W}_1\ ...\ \boldsymbol{W}_L)\|_2 := \max_{l\in[L]}\|\boldsymbol{W}_l\|_2$ and $\langle\boldsymbol{W},\boldsymbol{V}\rangle := \sum_{l\in[L]}\langle\boldsymbol{W}_l,\boldsymbol{V}_l\rangle$. We denote by $\boldsymbol{W}'$ a perturbation of $\boldsymbol{W}$.

Lemma 4.1 (Semi-smoothness)**.** *Assume the setup of §2. If* $\|\boldsymbol{W}-\boldsymbol{W}^{(0)}\|_2<\omega<O\Big(\frac{1}{L^{9/2}\ln^{3/2}m}\Big)$ *and* $\|\boldsymbol{W}'\|_2<\omega$, *then with a probability at least* $1-e^{-\Omega(m)}$

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}+\boldsymbol{W}')\le&\mathcal{L}(\boldsymbol{W})+\langle\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}),\boldsymbol{W}'\rangle\\&+\frac{nL^2m}{d}O(\|\boldsymbol{W}'\|_2^2)\\&+\frac{(1-\alpha)\omega^{1/3}L^2\sqrt{mn\mathcal{L}(\boldsymbol{W})\ln m}}{\sqrt{d(1+\alpha^2)}}O(\|\boldsymbol{W}'\|_2).\end{aligned}\tag{8}$$

Lemma 4.2 (Gradient bounds)**.** *Assume the setup of §2. If* $\|\boldsymbol{W}-\boldsymbol{W}^{(0)}\|_2<\omega<O\Big(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m}\Big)$, *then with a probability at least* $1-e^{-\Omega(m\delta^2/L^3)}$

$$\|\nabla_{\boldsymbol{W}_l}\mathcal{L}(\boldsymbol{W})\|_F^2\le\mathcal{L}(\boldsymbol{W})O\Big(\frac{mn}{d}\Big),\quad \textit{for}\ \ l\in[L]\tag{9}$$

$$\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W})\|_F^2\ge\mathcal{L}(\boldsymbol{W})\Omega\Big(\frac{(1-\alpha)^2}{(1+\alpha^2)}\frac{\delta m}{nd}\Big).\tag{10}$$

We note that the factor $(1-\alpha)/\sqrt{1+\alpha^2}$ appears in both (8) and (10), where it is squared in (10). This factor is the derivative gap in Leaky ReLU, i.e., $\tilde{\sigma}'_\alpha(0+)-\tilde{\sigma}'_\alpha(0-)$, which can be viewed as a measure of nonlinearity. Its value is larger for Leaky ReLU with $\alpha<0$ than for ReLU (with $\alpha=0$) and maximized at $\alpha=-1$. Our analysis below, which combines the bounds in (8) and (10), shows that Leaky ReLU with $\alpha<0$ leads to better control of the decay of the loss function than ReLU and that the best control is at $\alpha=-1$.

Theorem 3.1 can be proved as follows. Let $\boldsymbol{W}:=\boldsymbol{W}^{(t)}$ and $\boldsymbol{W}':=-\eta\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{(t)})$ and note that by gradient descent, $\boldsymbol{W}+\boldsymbol{W}'=\boldsymbol{W}^{(t+1)}$. Denoting $\mathcal{L}^{(t)}:=\mathcal{L}(\boldsymbol{W}^{(t)})$ and applying (8) of Lemma 4.1, we can conclude that with a probability of at least $1-e^{-\Omega(m)}$, the following inequality holds

$$\mathcal{L}^{(t+1)}\le\mathcal{L}^{(t)}-\eta\langle\nabla_{\boldsymbol{W}}\mathcal{L}^{(t)},\nabla_{\boldsymbol{W}}\mathcal{L}^{(t)}\rangle\tag{11}$$

$$+\frac{\eta(1-\alpha)\omega^{\frac{1}{3}}L^2\sqrt{mn\mathcal{L}^{(t)}\ln m}}{\sqrt{d(1+\alpha^2)}}O\Big(\|\nabla_{\boldsymbol{W}}\mathcal{L}^{(t)}\|_2\Big)\tag{12}$$

$$+\frac{\eta^2nL^2m}{d}O\Big(\|\nabla_{\boldsymbol{W}}\mathcal{L}^{(t)}\|_2^2\Big).\tag{13}$$

Using (10) we bound $\sqrt{\mathcal{L}^{(t)}}$ as follows with probability at least $1-e^{-\Omega(m\delta^2/L^3)}$:

$$\sqrt{\mathcal{L}^{(t)}}\le\frac{\sqrt{1+\alpha^2}}{1-\alpha}O\left(\sqrt{\frac{nd}{\delta m}}\right)\|\nabla_{\boldsymbol{W}}\mathcal{L}^{(t)}\|_F.\tag{14}$$

Applying (14), we control the term in (12), with probability at least $1-e^{-\Omega(m\delta^2/L^3)}$, by

$$\eta\omega^{1/3}nL^2(\sqrt{\ln m}/\sqrt{\delta})\ O\Big(\|\nabla_{\boldsymbol{W}}\mathcal{L}^{(t)}\|_F^2\Big).\tag{15}$$

Using $\omega<O(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m})$, which is required by Lemma 4.2, we reduce (15) to $\eta\|\nabla_{\boldsymbol{W}}\mathcal{L}^{(t)}\|_F^2/3$. Using

$\eta < O(d/(nL^2m))$, which is required in Theorem 3.1, we reduce the bound in (13) to $\eta\|\nabla_{\boldsymbol{W}}\mathcal{L}^{(t)}\|_F^2/3$.

Next, we apply these bounds to the respective terms in (11) and use the identity $\langle \boldsymbol{X},\boldsymbol{X}\rangle = \|\boldsymbol{X}\|_F^2$ for a vector of matrices $\boldsymbol{X}=(\boldsymbol{X}_1,...,\boldsymbol{X}_L)$ to reduce (11) to

$$\mathcal{L}^{(t+1)} \leq \mathcal{L}^{(t)} - 1/3\eta\|\nabla_{\boldsymbol{W}}\mathcal{L}^{(t)}\|_F^2. \tag{16}$$

Further application of the lower bound in (10) to the above equation results in $\mathcal{L}^{(t+1)} \leq \gamma\mathcal{L}^{(t)}$ with γ specified in (4) and we consequently conclude (3) of Theorem 3.1.

The above argument holds for one training step with probability at least $1-e^{-\Omega(m)}$. This argument extends to T steps with probability at least $1-Te^{-\Omega(m)}$. We note that the number of epochs T can be bounded using the bound ϵ on the training error, the convergence rate in (4) and the estimate $\mathcal{L}(\boldsymbol{W}^{(0)}) \leq O(n\sqrt{\ln m})$, which is shown in Appendix B.6, as follows:

$$\begin{aligned} T &= \ln(\epsilon/\mathcal{L}(\boldsymbol{W}^{(0)}))/\ln\gamma \leq \Theta(\ln(\epsilon/n\sqrt{\ln m})/\ln\gamma) \\ &\leq O\bigg(\frac{nd}{\eta\delta m}(\ln\epsilon^{-1} + \ln(n\sqrt{\ln m}))\bigg). \end{aligned}$$

Thus the total probability to ensure T-steps training with training error lower than ϵ is at least $1-O(\frac{nd}{\eta\delta m}(\ln\epsilon^{-1} + \ln(n\sqrt{\ln m})))e^{-\Omega(m)}$. Given that $m > \Omega(\mathrm{poly}(n,L,d,\delta^{-1}))$ and $m > \Omega(\ln\ln\epsilon^{-1})$, this probability is of order $1-e^{-\Omega(m)}$.

In Appendix B.6, we demonstrate that the inequality $\|\boldsymbol{W}^{(t)} - \boldsymbol{W}^{(0)}\|_2 < \omega < O(\delta^{3/2}/(n^{3/2}L^{15/2}\ln^{3/2}m))$ holds with probability at least $1-e^{-\Omega(\ln m)}$. Note that the latter bound implies the conditions for both Lemmas 4.1 and 4.2 and thus concludes the proof of Theorem 3.1

The proof of Theorem 3.2 is detailed in §B.7. We briefly describe the proof idea as follows. First, we use a similar argument as in the proof of Theorem 3.1 to bound the expectations of the loss functions at each step. Second, we use (9) to find an absolute upper bound of the loss functions. By combining these two bounds and using Azuma's inequality, we derive the decay of the loss function in (5) with the convergence rate in (6) in Theorem 3.2. Finally, we verify that the conditions for Lemma 4.1 and Lemma 4.2 are satisfied when the NN width satisfies $m/\ln^4 m > (1+\alpha^2)^4/(1-\alpha)^8\Omega(n^8L^{15}d/(b\delta^5))$ and thus conclude the theorem.

The proof of Theorem 3.4, which appears in §B.9, relies on the following lemma that bounds the generalization error for a class of NNs whose parameters are close to $\boldsymbol{W}^{(0)}$.

Lemma 4.3 (Generalization error with perturbation)**.** *Assume the setup of §2, where α is the leaky ReLU parameter. If $\|\boldsymbol{W} - \boldsymbol{W}^{(0)}\| < \omega < O\left(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m}\right)$, then with probability at least $1-e^{-\Omega(\ln m)}$*

$$\begin{aligned} R(\boldsymbol{W}) \leq \frac{1}{n}\mathcal{L}(\boldsymbol{W}) + \frac{1-\alpha}{\sqrt{1+\alpha^2}}O(d(\ln m)\sqrt{m}L^2\omega^{4/3}) + \\ O(d\sqrt{m(\ln m)/n}L\omega) + O\Big(d\sqrt{\ln m/n}\Big). \end{aligned}$$

The proof of Lemma 4.3, which appears in Appendix B.8, follows similar ideas as those of Cao and Gu (2020) but adapted to the different task of regression. Theorem 3.4 is a consequence of this lemma and two different estimates of the size of ω during training. The first estimate controls ω during the entire training with GD, regardless of how large the training epoch is, and is expressed in Lemma B.9. The second estimate uses direct bounds of the learning steps and provides a better upper bound of ω when the training epoch is small.

4.2 Discussion of Innovation

While we followed, extended and improved an existing proof framework, we would like to emphasize some innovation in our proof techniques. To begin with, it is difficult to directly extend the previous methods to any leaky ReLU with $\alpha < 1$. Our idea of rescaling the leaky ReLU activation function, along with the observation that, with rescaled initialization, it is equivalent to using the unscaled leaky ReLU, helped tremendously simplify our initial technical and complex effort. This allowed us to elegantly use the previous ideas and further improve them. Additional technical steps that are required to address the case $\alpha \neq 0$ can be noticed in the proofs of Lemmas B.1, B.2, B.4 and B.7.

We have also made notable improvements to previous estimates. In particular, we improved the lower bound for the gradient established by Zou and Gu (2019) by a factor of L. We also eliminated the previous dependence of the convergence rate on a negative power of L, which was undesirable as it implied that deeper networks might experience slower convergence. This demonstrates that the convergence rate of deep neural networks is at least comparable to that of shallow neural networks. Specifically, the later estimates can be found in the proof of Lemma 4.2 in Appendix B.4. They are motivated by a suggestion from Allen-Zhu et al. (2019b) to incorporate gradients from all layers' parameters, departing from previous estimates that solely relied on the gradients of parameters from the last layer. More specifically, improved lower bounds for the gradients from all layers' parameters can be found in Lemma B.7 in Appendix B.4. We also obtained a tighter bound for the spectral norm of $\boldsymbol{W}^{(t)} - \boldsymbol{W}^{(0)}$ when using SGD. This improved the lower bound on the width m for training convergence by a factor of order $n^{-8}L^{-2}(n/b)^{-3}\delta^3$.

Additionally, a more careful and fresh look helped improve the interpretation of the results. In particular, noting the effect on the number of epochs t on the generalization error, while developing tighter bounds when t was sufficiently small, helped with a meaningful bound on the generalization error. Another example includes making all the probabilities dependent on m, a choice we deemed more suitable for the overparameterized regime. Furthermore, to avoid the hidden dependence of δ on L in the previous works, we had to develop some careful mathematical estimates (see (29) in the appendix),

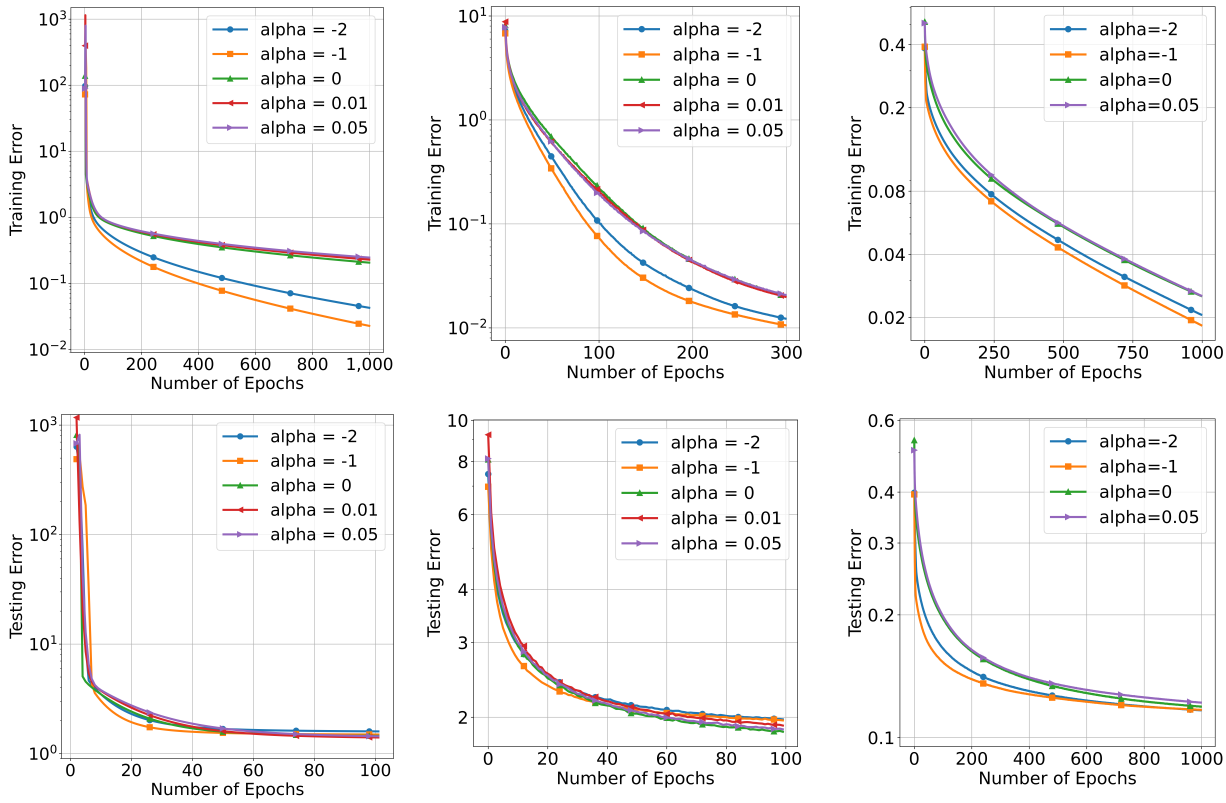

Figure 1: Log-scale training and testing errors using different datasets and different α's. From left to right: synthetic dataset, F-MNIST and CIFAR-10. Top row: training errors. Bottom row: testing errors.

so we could explicitly identify the dependence on L and relax the previous assumption $\delta<O(1/L)$ to $\delta<O(1)$.

5 NUMERICAL EXPERIMENTS

As our theory deals only with upper bounds, we conduct numerical experiments to examine the dependence of the actual training convergence rate and generalization error, particularly at an early epoch, on the parameter α. Our main goal is to determine whether $\alpha=-1$ is the optimal choice for convergence and generalization in overparameterized NNs with LeakyReLU activation functions. Appendix C provides additional experiments.

5.1 Setup

We summarize our implementation for the following datasets. We provide additional details in §C.1.

Synthetic dataset: We simulate a dataset which contains 1,000 data points in $\mathbb{R}^5$ i.i.d. sampled from a normalized Gaussian distribution, $N(0,\boldsymbol{I}_5)$. We verified that Assumption 2.1 holds for the generated dataset with $\delta=0.21$. We generate real-valued labels, y, by the following noisy nonlinear function of $\boldsymbol{x}$:

$$y=\sin(10x_1+20x_2^3)+\cos(3x_3+5x_4^2)\\ +2/(1+\mathrm{ReLU}(0.05+x_5))^{1/2}+2x_1x_5+\varepsilon,$$

where $\varepsilon\sim N(0,0.01)$. We construct NNs with five hidden layers, $m=5,000$ and leaky ReLUs with $\alpha\in\{-2,-1,0,0.01,0.05\}$. We initialize the NNs by Algorithm 1 and train them with GD using the MSE loss.

F-MNIST: This standard grayscale image classification benchmark consists of ten classes (Xiao et al., 2017). We build NNs with two hidden layers and width $m=2{,}000$. We use leaky ReLUs with $\alpha\in\{-2,-1,0,0.01,0.05\}$. We initialize the NNs by Algorithm 1 and train them using SGD with batch size 64 and the cross entropy loss.

CIFAR-10: This is another standard dataset for image classification (Krizhevsky et al., 2009). It consists of ten classes of RGB natural images. We modify the architecture of VGG19 (Simonyan and Zisserman, 2014) with four convolutional layers (width 512) and two linear layers (width 512) using Leaky ReLUs with $\alpha\in\{-2,-1,0,0.05\}$. We use Algorithm 1 to initialize the NNs and train them using SGD with batch size 64 and cross entropy loss.

To ensure that m is sufficiently large with respect to n, we randomly sample subsets of F-MNIST and CIFAR-10 (see more details in Appendix C.1).

5.2 Results

Figure 1 demonstrates both training errors (top) and testing errors (bottom) for the synthetic dataset, F-MNIST and CIFAR-10 (from left to right) for different αs. We

Table 1: Training and testing errors for the three main datasets. The first three rows report the training error at the last epoch. The next ones report the testing error at an early epoch ($t=30$ for synthetic, $t=20$ for F-MNIST and $t=200$ for CIFAR-10).

Metric	Dataset	$\alpha=-2$	$\alpha=-1$	$\alpha=0$	$\alpha=0.05$
Final training error	Synthetic	0.039±0.002	**0.022**±0.002	0.197±0.013	0.245±0.022
	F-MNIST	0.096±0.009	**0.076**±0.008	0.211±0.018	0.229±0.032
	CIFAR-10	0.019±0.001	**0.018**±0.001	0.024±0.001	0.027±0.001
Early Epoch testing error	Synthetic	1.914±0.067	**1.775**±0.065	2.086±0.173	2.313±0.218
	F-MNIST	2.371±0.103	**2.362**±0.053	2.442±0.067	2.470±0.092
	CIFAR-10	0.146±0.004	**0.143**±0.005	0.169±0.012	0.173±0.007

remark that we use the testing error as an approximation of the generalization error. Observing the training errors in the top row we note that the convergence is fastest for the NN with $\alpha=-1$ and the ranking of α from fastest to slowest convergence corresponds to the one predicted by our theory; that is, if α obtains a lower estimate for γ in (4) than α', then it results in faster convergence in our experiments. Observing the testing errors, we note that around a small training epoch (e.g., 30 for the synthetic dataset, 20 for F-MNIST, and 200 for CIFAR-10), the testing error is smallest when $\alpha=-1$. However, at larger training epochs the gaps of the testing errors are small for most of the αs.

To get a better quantitative idea, Table 1 summarizes for the different data sets the training error at the last epoch and the testing error at an early epoch. We ran the experiments 10 times and reported the mean and standard deviations (std's). We note that the std's are small and for better visualization we did not include them in Figure 1. We observe that choosing $\alpha=-1$ gives the least final training error in all datasets. Compared to ordinary ReLU, our choice of $\alpha=-1$ reduces the final training error by at least 22% (CIFAR-10) and at most 91% (synthetic). At early training epoch, compared to ordinary ReLU, the choice of $\alpha=-1$ reduces the testing error by at least 4% (F-MNIST) and at most 15% (CIFAR-10). This correlates with the predictions we made by our theory that the optimal bounds of the convergence rate and generalization error (at a sufficiently small epoch) are achieved with $\alpha=-1$.

Lastly, we compare the theoretically predicted upper bounds of the convergence rate and the empirical convergence rates with different αs. For this purpose, we ran experiments using the synthetic dataset and California housing (see its detailed description in Appendix C.1) with choices of α from [−10,0.5]. We approximate the convergence rate for each α using the training errors from the experiments at time steps 100 (i.e., $\mathcal{L}^{(100)}$) and 1,000 (i.e., $\mathcal{L}^{(1000)}$). The empirical convergence rate is calculated as

$$\hat{\gamma}(\alpha) := (\mathcal{L}^{(1000)}/\mathcal{L}^{(100)})^{1/900}.$$

To simplify our upper bound, we denote the constant $\Omega\left(\frac{\eta\delta m}{nd}\right)$ in (4) by C_γ and estimate its value based on the calculated convergence rate at $\alpha=0$ as

$$C_\gamma := C_0(1-\hat{\gamma}(0)), \tag{17}$$

where we choose $C_0 = 1$ for the synthetic dataset and $C_0=0.5$ for California housing. Consequently, we obtain our theoretical upper bounds of the convergence rates

$$\gamma(\alpha)=1-0.00143\frac{(1-\alpha)^2}{1+\alpha^2} \quad \text{for the synthetic dataset,}$$
$$\gamma(\alpha)=1-0.000537\frac{(1-\alpha)^2}{1+\alpha^2} \quad \text{for California housing.}$$

Figure 2 compares the theoretical upper bound of the convergence rate, $\gamma(\alpha)$, with the experimental convergence rate $\hat{\gamma}(\alpha)$ for the tested values of αs. It is interesting to note that the predicted upper bound dependence on α correlates very well with both numerical experiments.

Appendix C.2 includes additional details and numerical results. In particular, it performs experiments similar to the ones reported in Figure 1, while incorporating the datasets MNIST, California housing and IMDb movie reviews; the architectures of recurrent NNs and transformer NNs; and another loss function for regression. It also demonstrates how the training and testing errors depend on the NN hyperparameters (e.g., depth and width).

All codes are available at `https://github.com/sli743/leakyReLU`.

6 DISCUSSION

We established a mathematical theory that clarifies the impact of the Leaky ReLU parameter on bounds of both the training error convergence rate and the generalization error for overparameterized NNs. We showed that the absolute value function yields the optimal convergence rate bound for the training error and also the optimal generalization error bound when the training epoch is sufficiently small, with a sufficiently large dataset and a deep NN. Our extensive empirical tests support using the absolute value function for effective training of overparameterized NNs and for effective generalization with sufficiently small epochs and sufficiently large datasets and deep overparameterized NNs.

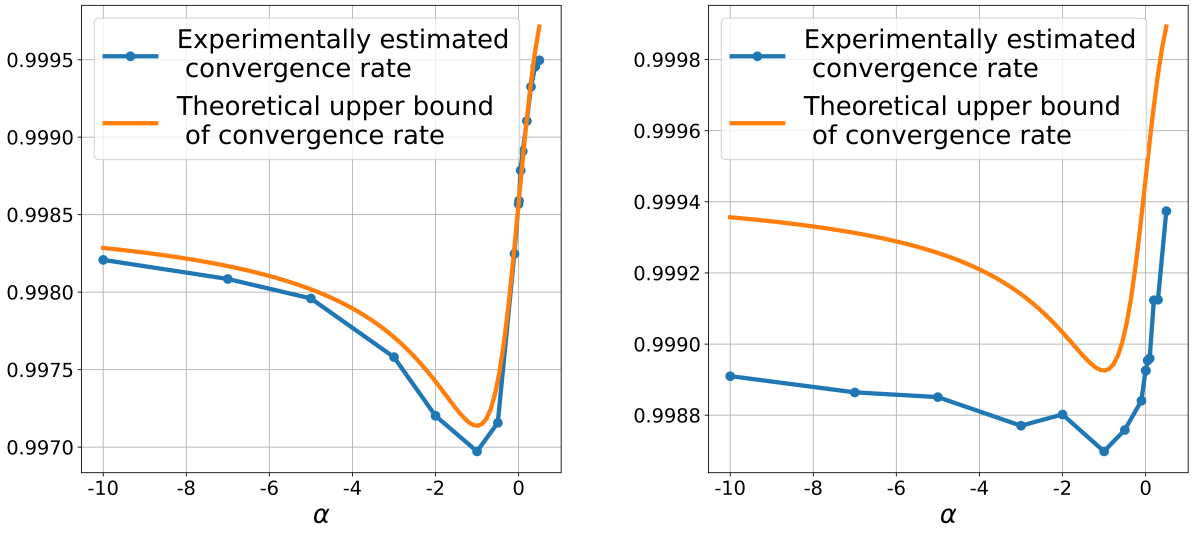

Figure 2: Comparison of the "shape" of the theoretical upper bound of the training convergence rate (orange line) with the calculated convergence rate (blue dots). We used the synthetic dataset (left) and California housing dataset (right) with different values of α's.

There are different possible extensions of our theory. For example, it is useful to extend it to other structured NNs, such as convolutional NNs (CNNs), while allowing any Leaky ReLU. Allen-Zhu et al. (2019b) established convergence for overparameterized CNNs with ReLU and one can directly extend their analysis to any Leaky ReLU. Nevertheless, it still remains open to extend the generalization theory to other structured NNs. Furthermore, it is useful to study the training convergence and generalization for larger classes of activation functions, such as the Gaussian error linear unit (Hendrycks and Gimpel, 2016).

Our work has three major limitations. First, our generalization error bound is not sufficiently small. Nevertheless, we believe it still indicates some interesting and relevant phenomena, in particular, the behavior when stopping at an early epoch. We further improved our estimates for a special class of datasets, although we observed that it was not sufficiently small in general. This is likely due to the fact that the regression setting poses greater challenges than classification. We also highlighted the possible implications of Kumar et al. (2023) to a generalization estimate given tight training error bounds.

Second, the lower bound that we require on the width, m, is generally unrealistically large and we thus find it important to extend our theory to lower values of m. Developing such a theory seems to require a careful analysis of nonlinear dynamical systems, given that current methods aim to linearize the underlying dynamical system. Nevertheless, for the special class of datasets discussed in Appendix B.11, we were able to provide a satisfying linear dependence of the lower bound of m on n.

Lastly, to theoretically guarantee the use of $\alpha = -1$, we need to develop respective lower bounds. We are not aware of useful and generic lower bounds and we find it rather difficult to develop them. Nevertheless, we still believe that making predictions based on the carefully developed upper bound and empirically testing them is valuable for practitioners. Indeed, our numerical results indicate the optimality of $\alpha = -1$ in many scenarios of overparameterized networks. On the other hand, we are unaware of much practical guidance that stems from the many other important and fundamental estimates in the study of overparameterized NNs. Additionally, Figure 2 shows cases where our upper bound for the convergence rate aligns with the observed convergence rate.

Acknowledgements

This work was partially supported by NSF award DMS 2124913.

References

Allen-Zhu, Z., Li, Y., and Liang, Y. (2019a). Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32.

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

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. (2019).

Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR.

Banerjee, A., Cisneros-Velarde, P., Zhu, L., and Belkin, M. (2023). Neural tangent kernel at initialization: linear width suffices. In *Uncertainty in Artificial Intelligence*, pages 110–118. PMLR.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2022). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21.

Bruna, J. and Mallat, S. (2013). Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886.

Cao, Y. and Gu, Q. (2020). Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3349–3356.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29.

Chen, Z., Cao, Y., Zou, D., and Gu, Q. (2019). How much over-parameterization is sufficient to learn deep relu networks? *ArXiv*, abs/1911.12360.

Chizat, L. and Bach, F. (2020). Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR.

Goodfellow, I., Vinyals, O., and Saxe, A. (2015). Qualitatively characterizing neural network optimization problems. In *International Conference on Learning Representations*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

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

Higham, C. F. and Higham, D. J. (2019). Deep learning: An introduction for applied mathematicians. *SIAM review*, 61(4):860–891.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. Available at `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Kumar, R., Majmundar, K., Nagaraj, D., and Suggala, A. S. (2023). Stochastic re-weighted gradient descent via distributionally robust optimization. *arXiv preprint arXiv:2306.09222*.

Li, Y. and Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Mallat, S. (2012). Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press, 2nd edition.

Oymak, S. and Soltanolkotabi, M. (2020). Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105.

Pace, R. K. and Barry, R. (1997). Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional

generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Ridnik, T., Lawen, H., Noy, A., Ben Baruch, E., Sharir, G., and Friedman, I. (2021). Tresnet: High performance GPU-dedicated architecture. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1400–1409.

Shamir, O. (2011). A variant of Azuma's inequality for martingales with subgaussian tails. *arXiv preprint arXiv:1110.2392*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition.

Song, C., Ramezani-Kebrya, A., Pethick, T., Eftekhari, A., and Cevher, V. (2021). Subquadratic overparameterization for shallow neural networks. *Advances in Neural Information Processing Systems*, 34:11247–11259.

Wang, X., Li, Y., Zhang, H., and Shan, Y. (2021). Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747.

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Zhu, Z., Liu, F., Chrysos, G., and Cevher, V. (2022). Generalization properties of NAS under activation and skip connection search. *Advances in Neural Information Processing Systems*, 35:23551–23565.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. (2020). Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109(3):467–492.

Zou, D. and Gu, Q. (2019). An improved analysis of training over-parameterized deep neural networks. *Advances in neural information processing systems*, 32.

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] **Yes.**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] **Not Applicable: Our theoretical analysis and recommended choice of hyperparameters do not affect the computational complexity.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] **Yes.**
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] **Yes.**
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] **Yes, see Appendix B.**
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] **Yes.**
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] **Yes.**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] **Yes.**
   (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] **Yes.**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] **Not Applicable: The computing power is not of a concern.**
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] **Yes.**
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable] **Yes.**
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] **Not Applicable: We do not produce new assets.**
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable] **Not Applicable.**

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] **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] **Not Applicable.**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] **Not Applicable.**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] **Not Applicable.**

Appendix

Section A discusses the generalization error bound, established in Theorem 3.4, under different regimes for the number of training epochs. Section B completes the proofs of the theorems stated in the main text and establishes four additional theorems: Theorem B.12, which bounds the generalization error when applying SGD; Theorem B.16, which bounds the convergence rate when using another loss function for regression; and Theorems B.14 and B.15, which bound the convergence rate and generalization error, respectively, for a special class of datasets. Section C describes additional numerical experiments and the full details of implementation for both the previous and the new experiments.

A Discussion of the Generalization Error Bound

In this section, we clarify the estimates for generalization error in (7) for different regimes of the number of training epochs, t.

We first note that the last term in (7) can be sufficiently small for a sufficiently large sample size n, so we may ignore it. The first bounding term in (7) reflects the training error and the middle two bounding terms represent the NN complexity. There is a tradeoff between the training and NN-complexity terms, as we explain below; in particular, we cannot make both of them sufficiently small. We remark that the closest bound on the generalization error for overparameterized deep NNs was established in the context of classification using GD in Cao and Gu (2020). Their generalization bound is independent of the training epoch. Instead, their bound is of order $\Theta(\text{poly}(n)\cdot n^{-1/2})$ and is typically not small even for arbitrarily large n. For very special cases (e.g., linear separability) they reduced the term $\text{poly}(n)$ so their overall bound is sufficiently small. In this work, we investigate the dependence of the generalization bound on t for regression without making assumptions about the data distribution. Nevertheless, one may consider similar special assumptions as in Cao and Gu (2020) and apply them to our theory in order to better control our generalization bound.

To better understand the bound in (7), we apply the bound on γ from Theorem 3.1 and our choice of m. We first quickly show that T is at order of $\Theta((nL)^2)$, from §4, we know that

$$T=\ln(\epsilon/\mathcal{L}(\boldsymbol{W}^{(0)}))/\ln\gamma\leq\Theta(\ln(\epsilon/n\sqrt{\ln m})/\ln\gamma),$$

by using (3) and $\eta=\Theta(d/(nmL^2))$, this upper bound is $\Theta((nL)^2)$, and when n is large, a lower bound with the same order can be achieved. We observe two different regions of $t\leq\Theta((nL)^2)$ (in §4, we show that $\Theta((nL)^2)$ approximates T). When $t=\Theta\big((nL)^{1-\kappa}\big)$, where $0<\kappa<1$, the first 3 terms of $R(\boldsymbol{W}^{(t)})$ are bounded by

$$\exp\bigg(-\Omega\bigg(\frac{(1-\alpha)^2\delta}{(1+\alpha^2)(nL)^{1+\kappa}}\bigg)\bigg)O(\ln m)+\frac{(1-\alpha)}{\sqrt{1+\alpha^2}}O\bigg(\frac{d^{1/6}\delta^{2/3}}{nL^{11/6}(nL)^{4\kappa/3}}\bigg)+O\bigg(\frac{\sqrt{d\ln m}}{(nL)^{\kappa}}\bigg).$$

The last two terms above are sufficiently small for sufficiently large n or L and the first training term is of the order $O(\ln m)$ and is thus the dominant one. In practice, it can be reduced through careful initialization. We note that this dominant term is minimized at $\alpha=-1$. When n and L are not sufficiently large and the second bounding term is comparable to the first term, then the bound is minimized at a certain α between -1 and 1. If, on the other hand, $t=\Omega((nL)^{(1+\kappa)})$, where $0<\kappa\leq 1$, then the order of the NN-complexity terms of (7) is $O(n^{\min\{\kappa,1/2+\tau\}}L^{\min\{\kappa,2+\tau\}})$, which becomes extremely large when n and L grow. This illustrates the overfitting phenomenon in neural network training, where the generalization error bound increases significantly as the training error approaches zero. Overall, we note that a smaller bound is obtained when $t=\Theta((nL)^{1-\kappa})$ and moreover overfitting occurs when $t=\Theta((nL)^{1+\kappa})$. These observations support the benefit of early stopping. We remark that when $t=T$, which is roughly at $\Theta((nL)^2)$, we can express the upper bound in (7), excluding its last term, in terms of ϵ as follows:

$$\begin{aligned}\epsilon+\min\bigg\{\bigg(\frac{(1-\alpha)^{11/3}}{(1+\alpha^2)^{11/6}}\bigg)O\bigg(\frac{d^{1/3}\delta^{4/3}}{m^{1/6}n^{10/3}L^{10/3}}\ln^{4/3}(n\sqrt{\ln m}/\epsilon)\bigg),O\bigg(\frac{d^{3/2+\tau}\delta^{\tau}n^{1/2+\tau}}{L^{1/2-\tau}\ln m}\bigg)\bigg\}\\ +\min\bigg\{\bigg(\frac{(1-\alpha)^2}{1+\alpha^2}\bigg)O\bigg(\frac{d^{1/2}\delta\sqrt{\ln m}}{n^3L^3}\bigg)\ln(n\sqrt{\ln m}/\epsilon),O\bigg(\frac{n^{1/2+\tau}L^{2+\tau}d^{1/2+\tau}}{\delta^{1/2-\tau}\ln m}\bigg)\bigg\}.\end{aligned}$$

The examination of our above theoretical results on generalization error bounds reveals two weaknesses when compared to the convergence theorems, that is, Theorems 3.1 and 3.2. Firstly, unlike the convergence rate that guarantees the training error's convergence, the generalization error bound doesn't assure a convergence to zero. Consequently, this bound may not offer a precise guideline about the optimal choice of α, especially when the number of epochs is large.

Secondly, $\alpha=-1$ is the optimal choice for the generalization error bound when training terminates early and both n and L are sufficiently large. In contrast, the convergence theorem asserts that $\alpha=-1$ consistently ensures the fastest convergence. Numerical results align with these observations.

B Proofs

We detail the proofs of Lemmas 4.1, 4.2 and 4.3 and the conclusion of Theorems 3.1, 3.2 and 3.4 from these lemmas. Moreover, we formulate and prove some the following additional theorems: a theorem that bounds the generalization error when using SGD, which is the analog of Theorem 3.4 for SGD instead of GD; theorems that improve our estimates for for a special class of datasets; and a theorem for the convergence theory when using a different loss function. Section B.1 introduces notation needed for the proof, § B.2 quantifies the bounds for the initial weights, § B.3 extends the latter bounds to weights within a small perturbation around the initialization, § B.4 proves the lower and upper bounds for the gradient at initial weight and within a small perturbation (Lemma 4.2), § B.5 shows the proof of semi-smoothness (Lemma 4.1), § B.6 and § B.7 conclude the main theorem for gradient descent and stochastic gradient descent (Theorem 3.1 and 3.2), §B.8 proves the upper bound of the generalization error for a class of NN functions (Lemma 4.3), §B.9 concludes the generalization error bound for GD (Theorem 3.4), §B.10 formulates and clarifies an upper bound of the generalization error for SGD, §B.11 introduces a special dataset and establishes theorems on the convergence rate bound and generalization error bound using this dataset, and §B.12 extends Theorem 3.1 and provides bounds of the convergence rate for a special loss function.

For the study of training convergence, we follow the notation and proof framework of Allen-Zhu et al. (2019b), while incorporating the improvements suggested by Zou and Gu (2019) and some additional ones. For the study of the generalization error, we follow the proof framework of Cao and Gu (2020) while extending the latter work to the task of regression. Whenever previous ideas require adaptation to Leaky ReLUs or to some of our technical contributions (summarized in §4.2), we prefer to repeat and even add more details so the reader can fully follow the current text and will not need to switch between references. However, when we feel that the ideas of previous works directly extend to our setting we formulate the analogous lemmas without proving them.

B.1 Notation

Throughout this appendix, we denote the entries of a vector $\boldsymbol{x}\in\mathbb{R}^m$ by x_j or $(\boldsymbol{x})_j$, $j\in[m]$. We denote the entries of a matrix $\boldsymbol{A}\in\mathbb{R}^{m\times m}$ by A_{ij} or $(\boldsymbol{A})_{ij}$, $i,j\in[m]$. For $i\in[m]$, the ith row vector of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}_{i,\cdot}$ and its ith column vector is denoted by $\boldsymbol{A}_{\cdot,i}$. The default norm $\|\cdot\|$ is the ℓ_2 norm. We denote by 1_E the indicator function of the event E, which equals 1 when E occurs and 0 otherwise. We denote by $\mathcal{B}_1^m$ the unit ball in $\mathbb{R}^m$.

We use the rescaled leaky ReLU introduced in (2) as the activation function of the neural networks under consideration. When acting on each coordinate of a vector $\boldsymbol{x}\in\mathbb{R}^p$ we express its action using the following diagonal matrix $\boldsymbol{D}_{\boldsymbol{x}}$:

$$\tilde{\sigma}_\alpha(\boldsymbol{x})=\boldsymbol{D}_{\boldsymbol{x}}\boldsymbol{x},\ \text{ where } (\boldsymbol{D}_{\boldsymbol{x}})_{jj}=\frac{1_{x_j\geq 0}}{\sqrt{1+\alpha^2}}+\frac{\alpha 1_{x_j<0}}{\sqrt{1+\alpha^2}}\quad\text{and for } k\neq j\ (D_{\boldsymbol{x}})_{kj}=0. \tag{18}$$

For $i\in[n]$ and a data point $\boldsymbol{x}_i\in\mathbb{R}^p$, We inductively define

$$\boldsymbol{g}_{i,l}:=\boldsymbol{W}_l\boldsymbol{h}_{i,l-1},\ \boldsymbol{h}_{i,l}:=\tilde{\sigma}_\alpha(\boldsymbol{W}_l\boldsymbol{h}_{i,l-1})\equiv\tilde{\sigma}_\alpha(\boldsymbol{g}_{i,l}),\ \boldsymbol{h}_{i,0}=\boldsymbol{A}\boldsymbol{x}_i \tag{19}$$

and use the notation $h_{i,l,k}:=(\boldsymbol{h}_{i,l})_k$ and $g_{i,l,k}:=(\boldsymbol{g}_{i,l})_k$. We denote

$$\boldsymbol{D}_{i,l}:=\boldsymbol{D}_{\boldsymbol{g}_{i,l}}\ \text{ and } D_{i,l,jj}:=(\boldsymbol{D}_{i,l})_{jj}\equiv\frac{1_{g_{i,l,j}\geq 0}}{\sqrt{1+\alpha^2}}+\frac{\alpha 1_{g_{i,l,j}<0}}{\sqrt{1+\alpha^2}}.$$

We further denote $\boldsymbol{D}_0:=\boldsymbol{I}$ and use the new notation to express the outputs of all hidden layers via matrix products (where according to the notation of §2 $\boldsymbol{W}_0\equiv\boldsymbol{A}$ and $\boldsymbol{W}_{L+1}\equiv\boldsymbol{B}$:

$$\begin{aligned}
&\boldsymbol{g}_{i,0}=\boldsymbol{h}_{i,0}=\boldsymbol{A}\boldsymbol{x}_i,\\
&\boldsymbol{g}_{i,l}=\boldsymbol{W}_l\boldsymbol{D}_{i,l-1}\boldsymbol{W}_{l-1}...\boldsymbol{W}_2\boldsymbol{D}_{i,1}\boldsymbol{W}_1\boldsymbol{A}\boldsymbol{x}_i,\\
&\boldsymbol{h}_{i,l}=\boldsymbol{D}_{i,l}\boldsymbol{W}_l\boldsymbol{D}_{i,l-1}\boldsymbol{W}_{l-1}...\boldsymbol{W}_2\boldsymbol{D}_{i,1}\boldsymbol{W}_1\boldsymbol{A}\boldsymbol{x}_i,\\
&\boldsymbol{g}_{i,L+1}:=\boldsymbol{B}\boldsymbol{h}_{i,L}\equiv\boldsymbol{B}\boldsymbol{D}_{i,L}\boldsymbol{W}_L\boldsymbol{D}_{i,L-1}\boldsymbol{W}_{L-1}...\boldsymbol{W}_2\boldsymbol{D}_{i,1}\boldsymbol{W}_1\boldsymbol{A}\boldsymbol{x}_i.
\end{aligned}$$

We denote the residual and its elements by

$$\boldsymbol{e}_i:=\boldsymbol{g}_{i,L+1}-\boldsymbol{y}_i,\ e_{i,j}=(\boldsymbol{e}_i)_j$$

and the loss function by

$$\mathcal{L}(\boldsymbol{W}):=\sum_{i=1}^{n}\text{loss}(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W}):=\sum_{i=1}^{n}\frac{1}{2}\|\boldsymbol{y}_i-\boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i;\boldsymbol{W})\|^2\equiv\frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{e}_i\|^2.$$

Section 5 in Higham and Higham (2019) presents a comprehensive derivation for the gradient of the loss function in a neural network. In our case, the activation function derivative can be written as

$$\frac{\partial h_{i,l,j}}{\partial g_{i,l,k}}=\delta_{jk}\cdot\left(\frac{1_{g_{i,l,k}\geq 0}}{\sqrt{1+\alpha^2}}+\frac{\alpha 1_{g_{i,l,k}<0}}{\sqrt{1+\alpha^2}}\right)\equiv D_{i,l,jk},\text{ for } l\in[L].$$

Denoting $\mathbf{Back}_{i,L+1}:=\boldsymbol{B}$ and $\mathbf{Back}_{i,l}:=\boldsymbol{B}\boldsymbol{D}_{i,L}\boldsymbol{W}_L...\boldsymbol{W}_l$ (this is the backpropagation operator) we can express the derivative of the loss with respect to the rt entry of $\boldsymbol{W}_l$, where $r,t\in[m]$, as

$$\nabla_{(\boldsymbol{W}_l)_{rt}}\text{loss}(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W})=(\mathbf{Back}_{i,l+1}^T\boldsymbol{e}_i)_r D_{i,l,rr}\boldsymbol{h}_{i,l-1,t}.$$

Similarly, the gradient of the loss according to the matrix $\boldsymbol{W}_l$ and according to its kth row vector, $(\boldsymbol{W}_l)_{k,\cdot}$, can be expressed as

$$\begin{aligned}\nabla_{\boldsymbol{W}_l}\text{loss}(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W})&=\boldsymbol{D}_{i,l}\mathbf{Back}_{i,l+1}^T\boldsymbol{e}_i\boldsymbol{h}_{l-1}^T(\boldsymbol{x}_i),\\ \nabla_{(\boldsymbol{W}_l)_{k,\cdot}}\text{loss}(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W})&=D_{i,l,kk}\langle(\mathbf{Back}_{i,l+1})_{\cdot,k},\boldsymbol{e}_i\rangle\boldsymbol{h}_{l-1}(\boldsymbol{x}_i).\end{aligned}$$

For a vector $\boldsymbol{v}\in\mathbb{R}^p$, we denote its ℓ_2 norm by $\|\boldsymbol{v}\|_2$ (where $\|\boldsymbol{v}\|_2^2=\sum_{j\in[p]}v_j^2$), ℓ_∞ norm by $\|\boldsymbol{v}\|_\infty=\max_{j\in[p]}|v_j|$, and ℓ_0 "size" by $\|\boldsymbol{v}\|_0=|\{j\in[p]:v_j\neq 0\}|$. For a matrix $\boldsymbol{X}\in\mathbb{R}^{m\times m}$, we denote its spectral norm by $\|\boldsymbol{X}\|_2=\max_{j\in[m]}|\lambda_j(\boldsymbol{X})|$, Frobenius norm by $\|\boldsymbol{W}\|_F=\sqrt{\sum_{i,j\in[m]}W_{ij}^2}$, and ℓ_0 "size" by $\|\boldsymbol{D}\|_0=|\{(i,j)\in[m]^2:D_{ij}\neq 0\}|$. For a vector of matrices $\boldsymbol{W}=(\boldsymbol{W}_1,...,\boldsymbol{W}_l)$, where $\boldsymbol{W}_1$, ..., $\boldsymbol{W}_l\in\mathbb{R}^{m\times m}$, we define its ℓ_2 norm by $\|\boldsymbol{W}\|_2:=\max_{l\in[L]}\|\boldsymbol{W}_l\|_2$ and Frobenius norm by $\|\boldsymbol{W}\|_F:=\sqrt{\sum_{l=1}^{L}\|\boldsymbol{W}_l\|_F^2}$. For simplicity of notation we use $\|\cdot\|$ instead of $\|\cdot\|_2$ for vectors, matrices and vectors of matrices.

Throughout this appendix, we apply Algorithm 1 to initialize the weights $\boldsymbol{W}$, $\boldsymbol{A}$, $\boldsymbol{B}$ for the neural network.

We use the big O, Ω and Θ notation. That is, $f=O(N)$ or $f=\Omega(N)$ if there exists $C>0$ and $N_0\in\mathbb{N}$ such that $f\leq CN$ or $f\geq CN$, respectively, for all $N>N_0$. Also, $f=\Theta(N)$ if and only if $f=O(N)$ and $f=\Omega(N)$.

Throughout this appendix, we may neglect the subscript i or superscripts (t) or (0) when there is no confusion.

B.2 Initialization

In this section, we focus on properties of the weights initialized by Algorithm 1 without training. We thus denote $\boldsymbol{W}:=\boldsymbol{W}^{(0)}$ and for any input vector $\boldsymbol{x}\in\mathbb{R}^p$ and $l\in[L]$

$$\begin{aligned}&\boldsymbol{g}_0=\boldsymbol{h}_0:=\boldsymbol{A}\boldsymbol{x},\\ &\boldsymbol{g}_l:=\boldsymbol{W}_l^{(0)}\boldsymbol{D}_{\boldsymbol{g}_{l-1}}...\boldsymbol{W}_2^{(0)}\boldsymbol{D}_{\boldsymbol{g}_1}\boldsymbol{W}_1^{(0)}\boldsymbol{A}\boldsymbol{x},\\ &\boldsymbol{h}_l:=\boldsymbol{D}_{\boldsymbol{g}_l}\boldsymbol{g}_l.\end{aligned}$$

For simplicity, we denote $\boldsymbol{D}_l:=\boldsymbol{D}_{\boldsymbol{g}_l}$.

We first establish Lemma B.1 which controls the norms of the outputs of the hidden layers with high probability. We then establish Lemma B.2 that upper bounds $\max_{i\neq j\in[n]}\langle\boldsymbol{h}_{i,l}/\|\boldsymbol{h}_{i,l}\|,\boldsymbol{h}_{j,l}/\|\boldsymbol{h}_{j,l}\|\rangle$ for all $l\in[L]$. Lastly, Lemma B.3 summarizes useful bounds of the norms of some relevant matrices.

We remark that the proof of Lemma B.1 adapts ideas of Allen-Zhu et al. (2019b) to the setting of Leaky ReLUs. The proof of Lemma B.2 follows ideas of Zou and Gu (2019), while assuming that $\delta<O(1)$ instead of $\delta<O(1/L)$ and applying minor adaptation to Leaky ReLUs. At last, Lemma B.3 directly follows the same proof argument in Allen-Zhu et al. (2019b) (while using the conclusion of Lemma B.1) and we thus omit its proof.

Lemma B.1. *Assume the setup of §2 and the above notation. If $\boldsymbol{x}\in\mathbb{R}^p$, $\|\boldsymbol{x}\|=1$ and ϵ is a fixed number in $(\Omega(\frac{L}{m}),1)$, then*

$$\|\boldsymbol{h}_l\|\in[1-\epsilon,1+\epsilon]\quad\textit{for all}\quad l\in\{0\}\cup[L]\quad\textit{with probability at least}\ 1-e^{-\Omega(m\epsilon^2/L)}.$$

Proof. We first prove the lemma for $l = 0$. Due to the initialization of the input layer by Algorithm 1, $\boldsymbol{h}_0 = \boldsymbol{A}\boldsymbol{x} \sim N(0, \|\boldsymbol{x}\|^2/m) = N(0,1/m)$. Therefore, $m\|\boldsymbol{h}_0\|^2 \sim \chi^2(m)$, where $\chi^2(m)$ denotes the chi-square distribution with m degrees of freedom. Using the tail bound for this sub-Gaussian distribution

$$\mathbb{P}\Big(\big|\|\boldsymbol{h}_0\|^2 - 1\big| > \frac{\epsilon}{2}\Big) \leq 2e^{-m\epsilon^2/32} \leq e^{-\Omega(m\epsilon^2)}. \tag{20}$$

We next prove the lemma for $l \geq 1$. For each layer l, we analyze the distribution of each entry of $\boldsymbol{h}_l$, and denote by $h_{l,j} := (\boldsymbol{h}_l)_j$, $j \in [m]$, conditioned on the output from the former layer $\boldsymbol{h}_{l-1}$. We note that the randomness of $\boldsymbol{h}_l$ comes from $\boldsymbol{W}_l$ given the fixed $\boldsymbol{h}_{l-1}$.

We note the following expression for $h_{l,j}$, which follows from (18) and (19):

$$h_{l,j} = \tilde{\sigma}_\alpha(g_{l,j}) = 1_{g_{l,j}>0}\frac{g_{l,j}}{\sqrt{1+\alpha^2}} + 1_{g_{l,j}\leq 0}\frac{\alpha g_{l,j}}{\sqrt{1+\alpha^2}}. \tag{21}$$

We remark that unlike previous analyses (Allen-Zhu et al., 2019b; Zou and Gu, 2019), we need to deal with two different terms in the sum in order to address Leaky ReLU and note just ReLU. We observe that due to the initialization of $\boldsymbol{W}_l$ and (19), $g_{l,j} \sim N(0, 2\sum h_{l-1,k}^2/m) = N(0, 2\|\boldsymbol{h}_{l-1}\|^2/m)$. By the symmetry of the normal distribution, $g_{l,j}$ is positive with probability 0.5. Thefore, the random variable

$$B_j := 1_{g_{l,j}>0}$$

is Bernoulli with probability 0.5, that is, $B_j \sim \mathrm{B}(0.5)$. We further note that $B_j g_{l,j} = B_j g_{l,j} | g_{l,j} > 0$. We thus rewrite (21) as

$$h_{l,j} = \frac{1}{\sqrt{1+\alpha^2}} B_j g_{l,j} | \{g_{l,j} > 0\} - \frac{\alpha}{\sqrt{1+\alpha^2}}(1-B_j)(-g_{l,j}) | \{g_{l,j} \leq 0\}. \tag{22}$$

Conditioning on the event $g_{l,j} > 0$, $g_{l,j} \overset{d}{=} |X|$, where $X \sim N(0, 2\|\boldsymbol{h}_{l-1}\|^2/m)$. Therefore,

$$g_{l,j}\Big|(g_{l,j} > 0) \ \sim |N(0, 2\|\boldsymbol{h}_{l-1}\|^2/m)|.$$

Similarly,

$$-g_{l,j}\Big|(g_{l,j} \leq 0) \ \sim |N(0, 2\|\boldsymbol{h}_{l-1}\|^2/m)|.$$

Therefore, (22) and the above two equations imply the following distribution law for h_{ij}:

$$h_{l,j} \overset{d}{=} \frac{1}{\sqrt{1+\alpha^2}} B_j V_{j,1} - \frac{\alpha}{\sqrt{1+\alpha^2}}(1-B_j) V_{j,2},$$

where $V_{j,1}$, $V_{j,2} \sim |N(0, 2\|\boldsymbol{h}_{l-1}\|^2/m)|$, $B_j \sim B(0, \frac{1}{2})$ and $V_{j,1}$, $V_{j,2}$ and B_j are independent. We further claim that if the former layer $\boldsymbol{h}_{l-1}$ is given, then $V_{j,1}$ and $V_{j,2}$ are independent for $j \in [m]$. Indeed, We first observe that conditioned on $\boldsymbol{h}_{l-1}$ the entries $h_{l,j}$, $j \in [m]$, are independent. Indeed, they depend on different rows in $\boldsymbol{W}_l$ and due to Algorithm 1 for the initialization of the lth layer these rows are independent. We also note that $V_{j,1}$ and $V_{j,2}$ only rely on $h_{l,j}$, and thus conditioned on $\boldsymbol{h}_{l-1}$ they are independent for $j \in [m]$.

We next derive an expression that clarifies the distribution of $\|\boldsymbol{h}_l\|^2$ conditioned on $\boldsymbol{h}_{l-1}$. We denote

$$P_l := \{j \in [m] : \ g_{l,j} > 0\}, \qquad K_l := |P_l|,$$
$$H_{l,1} := \frac{m}{2\|\boldsymbol{h}_{l-1}\|^2} \sum_{j \in P_l} V_{j,1}^2 | \boldsymbol{h}_{l-1}, \qquad H_{l,2} := \frac{m}{2\|\boldsymbol{h}_{l-1}\|^2} \sum_{j \in [m], j \notin P_j} V_{j,2}^2 | \boldsymbol{h}_{l-1}.$$

We note that K_l is Bernoulli with m trials and probability 0.5, i.e.,

$$K_l \sim B(m, 0.5).$$

The above observations imply that conditioning on $\boldsymbol{h}_{l-1}$ and P_l, $H_{l,1} \sim \chi^2(K_l)$ and $H_{l,2} \sim \chi^2(m - K_l)$. Therefore, $\|\boldsymbol{h}_l\|^2$ conditioned on $\boldsymbol{h}_{l-1}$ is given by

$$\|\boldsymbol{h}_l\|^2 | \boldsymbol{h}_{l-1} \overset{d}{=} \frac{2\|\boldsymbol{h}_{l-1}\|^2}{(1+\alpha^2)m} H_{l,1} + \frac{2\alpha^2\|\boldsymbol{h}_{l-1}\|^2}{(1+\alpha^2)m} H_{l,2}. \tag{23}$$

Note that the indices used by $H_{l,1}$ and indices used by $H_{l,2}$ do not overlap and thus form a partition of $[m]$. This partition is determined by P_l and $H_{l,1}$ and $H_{l,2}$ are conditionally independent given P_l.

We denote $\Delta_l := \frac{\|\boldsymbol{h}_l\|^2}{\|\boldsymbol{h}_{l-1}\|^2}$ and rewrite $\|\boldsymbol{h}_b\|^2$ (fixing $l=b$) as follows

$$\ln\|\boldsymbol{h}_b\|^2 = \ln\|\boldsymbol{h}_0\|^2 + \sum_{l=1}^{b} \ln\Delta_l. \tag{24}$$

Using the distribution of $\|\boldsymbol{h}_l\|^2$ conditioning on $\boldsymbol{h}_{l-1}$, where $1\le l\le b$, we first derive upper and lower bounds of the expectation $\mathbb{E}(\ln\Delta_l|\boldsymbol{h}_{l-1})$. We then show that given $\boldsymbol{h}_{l-1}$ and other information, $\ln\Delta_l$ is an $O(m^{-1})$ sub-Gaussian random variable. With these two properties we conclude the lemma by applying a variant of Azuma's inequality for sub-Gaussian random variables on $\sum_{l=1}^{b}\ln\Delta_l$.

Bounds on the expectation of $\ln\Delta_l|\boldsymbol{h}_{l-1}$: We note that $\mathbb{E}(H_{l,1}|P_l) = K_l$, $\mathbb{E}(H_{l,1}|P_l) = m - K_l$ and thus $\mathbb{E}(H_{l,1}) = \mathbb{E}(\mathbb{E}(H_{l,1}|P_l)) = \mathbb{E}(K_l) = 0.5m$. Similarly, $\mathbb{E}(H_{l,2}) = 0.5m$ and therefore $\mathbb{E}(H_{l,1}) = \mathbb{E}(H_{l,2})$. Using the latter observation and (23) we obtain

$$\mathbb{E}(\Delta_l|\boldsymbol{h}_{l-1}) = \frac{2}{m(1+\alpha^2)}\big(\mathbb{E}(H_{l,1}) + \alpha^2\mathbb{E}(H_{l,1})\big) = 1. \tag{25}$$

Applying the concavity of the log function, Jensen's inequality and then (23) and (25) yields

$$\mathbb{E}\left(\frac{1}{1+\alpha^2}\ln\frac{2}{m}H_{l,1} + \frac{\alpha^2}{1+\alpha^2}\ln\frac{2}{m}H_{l,2}\right) \le \mathbb{E}\ln(\Delta_l|\boldsymbol{h}_{l-1}) \le \ln\mathbb{E}(\Delta_l|\boldsymbol{h}_{l-1}) = 0. \tag{26}$$

Using the Chernoff bound for the binomial distribution, we note that

$$K_l \in [0.4m, 0.6m], \text{ or equivalently } m - K_l \in [0.4m, 0.6m], \text{ with probability } 1 - e^{-\Omega(m)}. \tag{27}$$

We next use the property that if $H \sim \chi^2(K)$ and $K \in [0.4m, 0.6m]$, then $\mathbb{E}\ln\frac{2}{m}H \ge -\frac{4}{m}$ (see page 13 in the proof of Lemma 7.1 in Allen-Zhu et al. (2019b)). This property and (26) imply

$$\mathbb{E}(\ln(\Delta_l)|\boldsymbol{h}_{l-1}) \in \left[-\frac{4}{m}, 0\right]. \tag{28}$$

Conditional sub-Gaussianity of $\ln\Delta_l$: We derive a tail bound for $\ln\Delta_l|\boldsymbol{h}_{l-1}$ and consequently conclude its sub-Gaussianity. We denote

$$E_l := \{|P_l| \in [0.4m, 0.6m]\}.$$

The combination of (23), basic probabilistic manipulations and the conditional independence of $H_{l,1}$ and $H_{l,2}$ yields

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\frac{m}{2}\Delta_l - \frac{m}{2}\Big| < t \Big| \boldsymbol{h}_{l-1}, E_l, P_l\Big) \\
&= \mathbb{P}\left(\left|\frac{1}{1+\alpha^2}H_{l,1} + \frac{\alpha^2}{1+\alpha^2}H_{l,2} - \frac{m}{2}\right| < t \Big| E_l, P_l\right) \\
&\ge \mathbb{P}\left(\left|\frac{1}{1+\alpha^2}H_{l,1} - \frac{1}{1+\alpha^2}\frac{m}{2}\right| < t/2 \text{ and } \left|\frac{\alpha^2}{1+\alpha^2}H_{l,2} - \frac{\alpha^2}{1+\alpha^2}\frac{m}{2}\right| < t/2 \Big| E_l, P_l\right) \\
&\ge \mathbb{P}\left(\left|\frac{1}{1+\alpha^2}H_{l,1} - \frac{1}{1+\alpha^2}\frac{m}{2}\right| < t/2 \Big| E_l, P_l\right)\mathbb{P}\left(\left|\frac{\alpha^2}{1+\alpha^2}H_{l,2} - \frac{\alpha^2}{1+\alpha^2}\frac{m}{2}\right| < t/2 \Big| E_l, P_l\right) \\
&\ge \mathbb{P}\Big(\Big|H_{l,1} - \frac{m}{2}\Big| < t/2 \Big| E_l, P_l\Big)\mathbb{P}\Big(\Big|H_{l,2} - \frac{m}{2}\Big| < t/2 \Big| E_l, P_l\Big).
\end{aligned}$$

Recall that given P_l, $H_{l,1}$ and $H_{l,2}$ are $\chi^2(|P_l|)$ and $\chi^2(m-|P_l|)$, respectively. We thus apply the corresponding tail bounds of $H_{l,1}$ and $H_{l,2}$ and (27) to the bound above and obtain that

$$\mathbb{P}\Big(\Big|\frac{m}{2}\Delta_l - \frac{m}{2}\Big| < t \Big| \boldsymbol{h}_{l-1}\Big) \ge \Big(1 - e^{-\Omega(t^2/m)}\Big)^2 \ge 1 - \Omega\Big(e^{-\Omega(t^2/m)}\Big).$$

Consequently,

$$\mathbb{P}\left(\ln|\Delta_l|<\frac{t}{m}\Big|\boldsymbol{h}_{l-1}\right)\geq 1-e^{-\Omega((\frac{t}{m})^2 m)} \quad \text{for} \quad t\in(0,m/4].$$

Therefore, $\ln\Delta_l$ conditioned on $\boldsymbol{h}_{l-1}$ and $K_l\in[0.4m,0.6m]$ is $O(m^{-1})$-sub-Gaussian.

Conclusion of the proof of the lemma: We define a new variable $\tilde{\Delta}_l$, where $\tilde{\Delta}_l=\Delta_l$ if $K_l\in[0.4m,0.6m]$ and $\tilde{\Delta}_l=1$, otherwise. From the tail probability of $\ln\Delta_l$ and the definition, it is clear that $\ln\tilde{\Delta}_l|\boldsymbol{h}_{l-1}$ is $O(m^{-1})$-sub-Gaussian. It follows from (27) that with overwhelming probability $\Delta=\tilde{\Delta}$. We consider the sequence of the following random variables $\{(\ln\tilde{\Delta}_l-\mathbb{E}\ln\tilde{\Delta}_l)|\boldsymbol{h}_{l-1}\}_{l=1}^b$. By Azuma's inequality for sub-Gaussian variables (see Theorem 2 with $c=m$ in Shamir (2011))

$$\mathbb{P}\left(\left|\sum_{l=1}^b \ln\Delta_l-\mathbb{E}(\ln\Delta_l|\boldsymbol{h}_{l-1})\right|>b\epsilon\right)<e^{-\Omega(b\epsilon^2 m)}.$$

Applying (28) to the above inequality yields

$$\mathbb{P}\left(\left|\sum_{l=1}^b \ln\Delta_l\right|>\epsilon+O\left(\frac{b}{m}\right)\right)<e^{-\Omega(\epsilon^2 m/b)}.$$

We can choose $\epsilon>\Omega(\frac{L}{m})$ such that

$$\mathbb{P}\left(\left|\sum_{l=1}^b \ln\Delta_l\right|>\epsilon/2\right)<e^{-\Omega(\epsilon^2 m/b)}$$

Combining (20), (24) and the above equation we obtain that

$$\mathbb{P}\big(\big|\|\boldsymbol{h}_b\|^2-1\big|>\epsilon_0\big)<e^{-\Omega(m\epsilon^2/L)}, \text{ for } \ b\in[L].$$

□

Lemma B.2. *Assume the setup of §2 and the notation introduced in this section. If $\delta<O(1)$ and $m>\Omega(\ln nL^4)$, then*

$$\max_{i\neq j\in[n]}\left\langle\frac{\boldsymbol{h}_{i,l}}{\|\boldsymbol{h}_{i,l}\|},\frac{\boldsymbol{h}_{j,l}}{\|\boldsymbol{h}_{j,l}\|}\right\rangle^2\leq 1-\Omega\left(\frac{\delta^2}{L^2}\right) \quad \textit{with probability at least } 1-e^{-\Omega(\delta^4 m/L^4)}. \tag{29}$$

Proof. We separate the proof of this lemma into three parts. The first one establishes a useful upper bound of the expectation of the multiplication of two leaky ReLUs of certain inner products (see (30) below). Given this upper bound, the second part shows that with high probability,

$$\min_{i\neq j\in[n]}\|\boldsymbol{h}_{i,l}-\boldsymbol{h}_{j,l}\|\geq\Omega(\delta/L), \text{ for any } l\in[L].$$

The third part uses the result to conclude this Lemma.

Part 1. We verify the following probabilistic estimate:

$$\mathbb{E}\tilde{\sigma}_\alpha(\boldsymbol{u}^T\boldsymbol{h}_i)\tilde{\sigma}_\alpha(\boldsymbol{u}^T\boldsymbol{h}_j)\leq\frac{1}{2}\left(1-\frac{1}{2}\theta^2\right)+\frac{(1-\alpha)^2}{(1+\alpha^2)}O(\theta^3), \tag{30}$$
$$\text{where } \boldsymbol{h}_i,\boldsymbol{h}_j\in\mathbb{R}^p, \text{ for } \theta>0,\langle\boldsymbol{h}_i,\boldsymbol{h}_j\rangle\leq 1-\frac{1}{2}\theta^2, \text{ and } \boldsymbol{u}\sim N(0,\boldsymbol{I})\in\mathbb{R}^p.$$

Since $\boldsymbol{u}\sim N(0,\boldsymbol{I})$, $\mathbb{E}u_k u_{k'}=0$ whenever $k\neq k'$. We denote $\boldsymbol{u}:=(u_1,u_2\cdots u_p)^T$, $\boldsymbol{h}_i:=(h_{i,1},h_{i,2},\cdots h_{i,p})^T$ and $\boldsymbol{h}_j:=(h_{j,1},h_{j,2},\cdots h_{j,p})^T$. We first note that

$$\mathbb{E}(\boldsymbol{u}^T\boldsymbol{h}_i)(\boldsymbol{u}^T\boldsymbol{h}_j)=\mathbb{E}\left(\sum_{k=1}^p u_k h_{i,k}\right)\left(\sum_{k'=1}^p u_{k'}h_{j,k'}\right)=\mathbb{E}\sum_{k=1}^p u_k^2 h_{i,k}h_{j,k}=\boldsymbol{h}_i^T\boldsymbol{h}_j\mathbb{E}\boldsymbol{u}^T\boldsymbol{u}\leq 1-\frac{1}{2}\theta^2.$$

For simplicity, we denote $Z_i := \boldsymbol{u}^T \boldsymbol{h}_i$ and $Z_j := \boldsymbol{u}^T \boldsymbol{h}_j$ and thus express the above equation as

$$\mathbb{E}(Z_i Z_j) \leq 1 - \frac{1}{2}\theta^2. \tag{31}$$

Using the symmetry of normal distribution, we obtain that

$$\mathbb{E}(Z_i Z_j | Z_i, Z_j \geq 0) = \mathbb{E}(Z_i Z_j | Z_i, Z_j < 0)$$

and

$$\mathbb{E}(Z_i Z_j | Z_i < 0, Z_j \geq 0) = \mathbb{E}(Z_i Z_j | Z_i \geq 0, Z_j < 0).$$

Consequently, the expectation of $\tilde{\sigma}_\alpha(Z_i)\tilde{\sigma}_\alpha(Z_j)$ can be rewritten as

$$\begin{aligned}\mathbb{E}\tilde{\sigma}_\alpha(Z_i)\tilde{\sigma}_\alpha(Z_j) = \frac{1}{1+\alpha^2}\Big(&\mathbb{E}(Z_i Z_j | Z_i, Z_j \geq 0) + \alpha \mathbb{E}(Z_i Z_j | Z_i \geq 0, Z_j < 0) \\ &+ \alpha \mathbb{E}(Z_i Z_j | Z_i < 0, Z_j \geq 0) + \alpha^2 \mathbb{E}(Z_i Z_j | Z_i, Z_j < 0)\Big) \\ = &\mathbb{E}(Z_i Z_j | Z_i, Z_j \geq 0) + \frac{2\alpha}{1+\alpha^2}\mathbb{E}(Z_i Z_j | Z_i \geq 0, Z_j < 0).\end{aligned} \tag{32}$$

Similarly, we express $\mathbb{E}Z_i Z_j$ as follows: (32)

$$\begin{aligned}\mathbb{E}Z_i Z_j &= 2\mathbb{E}(Z_i Z_j | Z_i, Z_j \geq 0) + 2\mathbb{E}(Z_i Z_j | Z_i \geq 0, Z_j \leq 0) \\ &= 2\mathbb{E}\tilde{\sigma}_\alpha(Z_1)\tilde{\sigma}_\alpha(Z_2) + \left(2 - \frac{4\alpha}{1+\alpha^2}\right)\mathbb{E}(Z_1 Z_2 | Z_1 \geq 0, Z_2 \leq 0).\end{aligned}$$

Rearranging the above equation yields

$$\mathbb{E}\tilde{\sigma}_\alpha(Z_1)\tilde{\sigma}_\alpha(Z_2) = \frac{1}{2}\mathbb{E}Z_1 Z_2 - \frac{(1-\alpha)^2}{1+\alpha^2}\mathbb{E}(Z_1 Z_2 | Z_1 \geq 0, Z_2 < 0). \tag{33}$$

Noting that $\mathbb{E}Z_1 Z_2 \leq 1 - \frac{1}{2}\theta^2$ and using the proof of Lemma A.3 of Zou et al. (2020) result in

$$|\mathbb{E}(Z_1 Z_2 | Z_1 \geq 0, Z_2 < 0)| \leq O(\theta^3).$$

The application of both (31) and the above estimate to (33) results in(30) and thus concludes this part.

Part 2. For $l = 0, \ldots, L$ and $\delta_l := \frac{\delta}{2(l+1)}$ we prove by induction:

$$\min_{i \neq j \in [n]} \|\boldsymbol{h}_{i,l} - \boldsymbol{h}_{j,l}\| \geq \delta_l \text{ with probability at least } 1 - e^{-\Omega(\delta^4 m / L^4)}. \tag{34}$$

We first prove (34) when $l = 0$. Recall that $\boldsymbol{h}_{i,0} = \boldsymbol{A}\boldsymbol{x}_i$ and note that for any $i, j \in [n]$,

$$\begin{aligned}\mathbb{E}\left(\|\boldsymbol{A}\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{x}_j\|^2\right) &= \mathbb{E}\langle \boldsymbol{A}\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{x}_j, \boldsymbol{A}\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{x}_j \rangle = \mathbb{E}\|\boldsymbol{A}\boldsymbol{x}_i\|^2 + \mathbb{E}\|\boldsymbol{A}\boldsymbol{x}_j\|^2 - 2\mathbb{E}\langle \boldsymbol{A}\boldsymbol{x}_i, \boldsymbol{A}\boldsymbol{x}_j \rangle \\ &= 2 - 2\mathbb{E}\sum_{k=1}^{m}\sum_{s,t} A_{ks} x_{i,s} A_{kt} x_{j,t} = 2 - 2\mathbb{E}\sum_{k,s=1}^{m} x_{i,s} x_{j,s} A_{ks}^2 \\ &= 2 - 2\sum_{s,k} x_{i,s} x_{j,s} \mathbb{E} A_{ks}^2 = 2 - 2\sum_{s,k} x_{i,s} x_{j,s} \frac{1}{m} \\ &= 2 - 2\boldsymbol{x}_i^T \boldsymbol{x}_j.\end{aligned} \tag{35}$$

Recall that Assumption 2.1 implies that $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \geq \delta$ and thus clearly

$$\boldsymbol{x}_i^T \boldsymbol{x}_j \leq 1 - \delta^2/2.$$

Applying this estimate in (35) yields the that

$$\mathbb{E}(\|\boldsymbol{A}\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{x}_j\|^2) \geq \delta^2. \tag{36}$$

Due to the random initialization, $m\|\boldsymbol{A}\boldsymbol{x}\|^2\sim\chi^2(m)$ and therefore $\|\boldsymbol{A}\boldsymbol{x}\|^2$ is $(O(1/m,4)$ sub-exponential. Since

$$\mathbb{P}(\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j\|^2>s)\leq\mathbb{P}(\|\boldsymbol{A}\boldsymbol{x}_i\|^2>s/2)+\mathbb{P}(\|\boldsymbol{A}\boldsymbol{x}_j\|^2>s/2),$$

the tail probability of $\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j\|^2$ is of the same order as the tail probabilities of $\|\boldsymbol{A}\boldsymbol{x}_i\|^2$ and $\|\boldsymbol{A}\boldsymbol{x}_j\|^2$. Therefore, we conclude that $\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j\|^2$ is also $(O(1/m),4)$ sub-exponential. Using the assumption $\delta<c_0$, where c_0 can be appropriately chosen (here we assume that $c_0\delta<3/4$), (36) and the fact that $\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j\|^2$ is $(O(1/m),4)$ sub-exponential) we conclude that

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j\|^2<\frac{\delta^2}{4}\right)&\leq\mathbb{P}\big(\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j\|^2<\delta^2(1-\delta)\big)\\&\leq\mathbb{P}\big(\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j\|^2<(1-\delta)\mathbb{E}\|\boldsymbol{A}\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_j\|^2\big)\\&\leq O(e^{-\delta^4 m}).\end{aligned}$$

Applying a union bound over all distinct $i,\,j\in[n]$, we conclude that with probability at least $1-n^2e^{-\Omega(\delta^4 m)}$,

$$\min_{i\neq j\in[n]}\|\boldsymbol{h}_{i,0}-\boldsymbol{h}_{j,0}\|\geq\frac{\delta}{2}\equiv\delta_0.$$

Next, we fix $l\in[L]$, assume that (34) holds for all $k\in[0,l-1]$ and verify (34) for l. Using the fact that $\mathbb{E}(\|\boldsymbol{h}_{i,l}\|^2|\boldsymbol{h}_{i,l-1})=\|\boldsymbol{h}_{i,l-1}\|^2$ and the definition of $\boldsymbol{h}_{i,l}$ we obtain

$$\begin{aligned}&\mathbb{E}(\|\boldsymbol{h}_{i,l}-\boldsymbol{h}_{j,l}\|^2|\boldsymbol{h}_{l-1})\\&=\mathbb{E}(\|\boldsymbol{h}_{i,l}\|^2|\boldsymbol{h}_{i,l-1})+\mathbb{E}(\|\boldsymbol{h}_{j,l}\|^2|\boldsymbol{h}_{j,l-1})-2\mathbb{E}(\langle\boldsymbol{h}_{i,l},\boldsymbol{h}_{j,l}\rangle|\boldsymbol{h}_{l-1})\\&=\|\boldsymbol{h}_{i,l-1}\|^2+\|\boldsymbol{h}_{j,l-1}\|^2-2\mathbb{E}(\langle\tilde{\sigma}_\alpha(\boldsymbol{W}_l\boldsymbol{h}_{i,l-1}),\tilde{\sigma}_\alpha(\boldsymbol{W}_l\boldsymbol{h}_{j,l-1})\rangle|\boldsymbol{h}_{l-1}).\end{aligned}\tag{37}$$

Applying the induction assumption (i.e., (30) with $\theta=\delta_{l-1}$) and the fact that $(\boldsymbol{W}_l)_{k,\cdot}\sim N\big(0,\frac{2}{m}\boldsymbol{I}\big)$ and denoting by $\boldsymbol{u}$ a random variable such that $\boldsymbol{u}\sim N(0,\boldsymbol{I})$ so $(\boldsymbol{W}_l)_{k,\cdot}^T\overset{d}{=}2\boldsymbol{u}/m$ result in

$$\begin{aligned}\mathbb{E}(\langle\tilde{\sigma}_\alpha(\boldsymbol{W}_l\boldsymbol{h}_{i,l-1}),\tilde{\sigma}_\alpha(\boldsymbol{W}_l\boldsymbol{h}_{j,l-1})\rangle|\boldsymbol{h}_{l-1})&=\sum_k\mathbb{E}(\tilde{\sigma}_\alpha((\boldsymbol{W}_l)_{k,\cdot}^T\boldsymbol{h}_{i,l-1})\tilde{\sigma}_\alpha((\boldsymbol{W}_l)_{k,\cdot}^T\boldsymbol{h}_{j,l-1})|\boldsymbol{h}_{l-1})\\&=\frac{2m}{m}\mathbb{E}(\tilde{\sigma}_\alpha(\boldsymbol{u}^T\boldsymbol{h}_{i,l-1})\tilde{\sigma}_\alpha(\boldsymbol{u}^T\boldsymbol{h}_{j,l-1})|\boldsymbol{h}_{l-1})\\&\leq1-\frac{1}{2}\delta_{l-1}^2+\frac{(1-\alpha)^2}{1+\alpha^2}O(\delta_{l-1}^3).\end{aligned}$$

Using Lemma B.1, we note for any $i\in[n]$, $\|\boldsymbol{h}_{i,l}\|^2\in(1-O(\delta_{l-1}^3),1+O(\delta_{l-1}^3))$ with probability at least $1-ne^{-\Omega(\delta_{l-1}^3 m)}$. Combining this observation with (37) yields for a constant $C>0$

$$\mathbb{E}(\|\boldsymbol{h}_{i,l}-\boldsymbol{h}_{j,l}\|^2|\boldsymbol{h}_{l-1})\geq\delta_{l-1}^2\left(1-C\frac{(1-\alpha)^2}{1+\alpha^2}\delta_{l-1}\right)+O(\delta_{l-1}^3).$$

It follows from (23) and the fact that $H_{l,1}\sim\chi^2(K_l)$ and $H_{l,2}\sim\chi^2(m-K_l)$ that $\|\boldsymbol{h}_{i,l}\|^2|\boldsymbol{h}_{l-1}$ is $(O(1/m),4)$ sub-exponential and thus $\|\boldsymbol{h}_{i,l}-\boldsymbol{h}_{j,l}\|^2|\boldsymbol{h}_{l-1}$ is also $(O(1/m),4)$ sub-exponential. Thus for $i\neq j\in[n]$

$$\mathbb{P}\left(\|\boldsymbol{h}_{i,l}-\boldsymbol{h}_{j,l}\|^2\leq\delta_{l-1}^2\left(1-\left(C\frac{(1-\alpha)^2}{1+\alpha^2}\right)\delta_{l-1}\right)(1-\delta_{l-1})\middle|\boldsymbol{h}_{l-1}\right)\leq O(\exp(-\delta_{l-1}^4 m)).$$

Applying a union bound for all $n(n-1)/2$ pairs yields

$$\begin{aligned}\mathbb{P}\left(\min_{i\neq j\in[n]}\|\boldsymbol{h}_{i,l}-\boldsymbol{h}_{j,l}\|^2\leq\delta_{l-1}^2\left(1-\left(C\frac{(1-\alpha)^2}{1+\alpha^2}\right)\delta_{l-1}\right)(1-\delta_{l-1})\middle|\boldsymbol{h}_{l-1}\right)&\\\leq n(n-1)/2O(\exp(-\delta_{l-1}^4 m)).&\end{aligned}\tag{38}$$

Consequently,

$$\begin{aligned}
&1-n^2\Omega(\exp(-\Omega(\delta_{l-1}^4 m)))\\
&\le \mathbb{P}\left(\min_{i\neq j\in[n]}\|\boldsymbol{h}_{i,l}-\boldsymbol{h}_{j,l}\|^2 \ge \delta_{l-1}^2\left(1-\left(C\frac{(1-\alpha)^2}{1+\alpha^2}\right)\delta_{l-1}\right)(1-\delta_{l-1})\middle|\boldsymbol{h}_{l-1}\right)\\
&= \mathbb{P}\left(\min_{i\neq j\in[n]}\|\boldsymbol{h}_{i,l}-\boldsymbol{h}_{j,l}\|^2 \ge \delta_{l-1}^2\left(1-\left(C\frac{(1-\alpha)^2}{1+\alpha^2}+1\right)\delta_{l-1}\right)+C\frac{(1-\alpha)^2}{1+\alpha^2}\delta_{l-1}^4\middle|\boldsymbol{h}_{l-1}\right)\\
&\le \mathbb{P}\left(\min_{i\neq j\in[n]}\|\boldsymbol{h}_{i,l}-\boldsymbol{h}_{j,l}\|^2 \ge \delta_{l-1}^2\left(1-\left(C\frac{(1-\alpha)^2}{1+\alpha^2}+1\right)\delta_{l-1}\right)\middle|\boldsymbol{h}_{l-1}\right). \qquad (39)
\end{aligned}$$

Next, we verify that $1-(C(1-\alpha)^2/(1+\alpha^2)+1)\delta_{l-1}\ge l^2/(l+1)^2$ for a sufficiently small c_0 (recall that $\delta<c_0$). We first note that for $l\ge 1$,

$$\frac{l^2}{(l+1)^2}=1-\frac{2l+1}{(l+1)^2}\le 1-\frac{l+1}{(l+1)^2}=1-\frac{1}{l+1}\le 1-\frac{1}{2l}.$$

Therefore, if $\delta<1/(C(1-\alpha)^2/(1+\alpha^2)+1)<c_0$, then for any $l\in[L]$

$$1-\left(C\frac{(1-\alpha)^2}{1+\alpha^2}+1\right)\delta_{l-1}=1-\left(C\frac{(1-\alpha)^2}{1+\alpha^2}+1\right)\frac{\delta}{2l}\ge 1-\frac{1}{2l}\ge\frac{l^2}{(l+1)^2}.$$

Thus (39) implies $\min_{i\neq j\in[n]}\|\boldsymbol{h}_{i,l}-\boldsymbol{h}_{j,l}\|^2\ge\delta_{l-1}^2\frac{l^2}{(l+1)^2}\equiv\delta_l^2$ with probability $1-n^2e^{-\Omega(\delta^4 m/L^4)}$. When $m>\Omega(\ln nL^4)$, the latter probability can be written as $1-e^{-\Omega(\delta^4 m/L^4)}$, which concludes (34).

Part 3. We conclude the lemma as follows. We recall that Lemma B.1 implies that with probability at least $1-e^{-\Omega(m\delta_l^3/L)}$: $\|\boldsymbol{h}_{i,l}\|^2\in[1-O(\delta_l^3),1+O(\delta_l^3)]$. Applying this conclusion and (34) we conclude that for any $i\neq j\in[n]$

$$\begin{aligned}
\left\|\frac{1}{\|\boldsymbol{h}_{j,l}\|}\boldsymbol{h}_{j,l}-\frac{1}{\|\boldsymbol{h}_{i,l}\|}\boldsymbol{h}_{i,l}\right\| &= \left\|\frac{1}{\|\boldsymbol{h}_{j,l}\|}\boldsymbol{h}_{j,l}-\frac{1}{\|\boldsymbol{h}_{j,l}\|}\boldsymbol{h}_{i,l}+\frac{1}{\|\boldsymbol{h}_{j,l}\|}\boldsymbol{h}_{i,l}-\frac{1}{\|\boldsymbol{h}_{i,l}\|}\boldsymbol{h}_{i,l}\right\|\\
&\ge \frac{1}{\|\boldsymbol{h}_{j,l}\|}\|\boldsymbol{h}_{j,l}-\boldsymbol{h}_{i,l}\|-\left|\frac{1}{\|\boldsymbol{h}_{j,l}\|}-\frac{1}{\|\boldsymbol{h}_{i,l}\|}\right|\|\boldsymbol{h}_{i,l}\|\\
&\ge \delta_l(1-\delta_l^{1/2}) \ \text{ with probability at least } 1-2e^{-\Omega(m\delta^4/L^4)}.
\end{aligned}$$

We note that for $\delta<c_0<1/2$, $\delta_l<\frac{1}{4}$ and thus $\delta_l(1-\delta_l^{1/2})\ge\frac{1}{2}\delta_l$. Consequently,

$$\begin{aligned}
&\left\langle\frac{1}{\|\boldsymbol{h}_{j,l}\|}\boldsymbol{h}_{j,l},\frac{1}{\|\boldsymbol{h}_{i,l}\|}\boldsymbol{h}_{i,l}\right\rangle\\
&=\frac{1}{2}\left(\frac{\|\boldsymbol{h}_{j,l}\|^2}{\|\boldsymbol{h}_{j,l}\|^2}+\frac{\|\boldsymbol{h}_{i,l}\|^2}{\|\boldsymbol{h}_{i,l}\|^2}-\left\|\frac{1}{\|\boldsymbol{h}_{j,l}\|}\boldsymbol{h}_{j,l}-\frac{1}{\|\boldsymbol{h}_{i,l}\|}\boldsymbol{h}_{i,l}\right\|^2\right)\\
&=1-\frac{1}{2}\left\|\frac{1}{\|\boldsymbol{h}_{j,l}\|}\boldsymbol{h}_{j,l}-\frac{1}{\|\boldsymbol{h}_{i,l}\|}\boldsymbol{h}_{i,l}\right\|^2\le 1-\frac{1}{8}\delta_l^2 \ \text{ with probability at least } 1-2e^{-\Omega(m\delta^4/L^4)}.
\end{aligned}$$

Therefore, if $\delta_l^2<8$, then

$$\left\langle\frac{1}{\|\boldsymbol{h}_{j,l}\|}\boldsymbol{h}_{j,l},\frac{1}{\|\boldsymbol{h}_{i,l}\|}\boldsymbol{h}_{i,l}\right\rangle^2\le\left(1-\frac{1}{8}\delta_l^2\right)^2\le 1-\frac{1}{8}\delta_l^2 \ \text{ with probability at least } 1-2e^{-\Omega(m\delta^4/L^4)}.$$

Finally, we apply a union bound on all the distinct i, j pairs to obtain

$$\max_{i,j\in[n]}\left\langle\frac{1}{\|\boldsymbol{h}_{j,l}\|}\boldsymbol{h}_{j,l},\frac{1}{\|\boldsymbol{h}_{i,l}\|}\boldsymbol{h}_{i,l}\right\rangle^2\le 1-\frac{1}{8}\delta_l^2 \ \text{ with probability at least } 1-n^2e^{-\Omega(m\delta^4/L^4)}.$$

The proof of the lemma is concluded by the above bound and the following two immediate observations: $\delta_l\equiv\delta/2(l+1)\ge\Omega(\delta/L)$ and when $m>\Omega(\ln n)L^4$ the above probability can be expressed as $1-e^{-\Omega(\delta^4 m/L^4)}$.

□

Lemma B.3. *Assume the setup of §2 and the notation introduced in this section. If $0\le a<b\le L$, then with probability at least $1-e^{-\Omega(m/L)}$ the following statements hold:*

1. $\|\boldsymbol{W}_{b+1}\boldsymbol{D}_b\boldsymbol{W}_b...\boldsymbol{D}_a\|\le O(\sqrt{L})$.
2. *If $d<O(\frac{m}{L\ln m})$, then $\|\boldsymbol{Back}_a\|\equiv\|\boldsymbol{B}\boldsymbol{D}_L\boldsymbol{W}_L...\boldsymbol{D}_a\boldsymbol{W}_a\|\le O(\sqrt{\frac{m}{d}})$.*
3. *If $\boldsymbol{v}\in\mathbb{R}^m$ and $\|\boldsymbol{v}\|_0\le O\big(\frac{m}{L\ln m}\big)$, then $\|\boldsymbol{W}_b\boldsymbol{D}_{b-1}...\boldsymbol{D}_a\boldsymbol{W}_a\boldsymbol{v}\|\le 2\|\boldsymbol{v}\|$.*

For $s<O(m/L\ln m)$ and $d<O(\frac{m}{L\ln m})$, with probability at least $1-\exp(-\Omega(s\log m))$, the following statement holds:

4. *For any vector $\boldsymbol{u}\in\boldsymbol{R}^d$, $\boldsymbol{v}\in\boldsymbol{R}^m$ such that $\|\boldsymbol{v}\|_0\le s$, then $|\boldsymbol{u}^T\boldsymbol{B}\boldsymbol{D}_L\boldsymbol{W}_L\cdots\boldsymbol{D}_a\boldsymbol{W}_a\boldsymbol{v}|\le O(\sqrt{s\ln m/d}\|\boldsymbol{v}\|\|\boldsymbol{u}\|)$.*

The proof of the lemma follows the same argument of the proof of Lemma 7.3 (a), (b) and Lemma 7.4 (a), (b) in Allen-Zhu et al. (2019b) and is not directly affected by our use of Leaky ReLU. We remark though that it requires applying Lemma B.1, which was formulated for any Leaky ReLU function instead of Lemma 7.1 of Allen-Zhu et al. (2019b).

B.3 Perturbation

We establish Lemma B.4 which quantifies the effect of a small perturbation of the randomly initialized parameters $\boldsymbol{W}^{(0)}$ on the output of the hidden layers. Lemma B.5 uses the former lemma to bound the norms of the perturbed matrices and the perturbations themselves. The proof of Lemma B.4 directly follows ideas of Lemma 8.2 of Allen-Zhu et al. (2019b), but adapts them to the setting of Leaky ReLUs. The final conclusion of this lemma is independent of α since the leading terms turn out to be independent of α. For completeness, we find it useful to include all these details. Lemma B.5 directly follows arguments of Allen-Zhu et al. (2019b) and we thus omit its proof.

We denote the perturbation matrix by $\boldsymbol{W}'$ and the perturbed matrix of parameters by $\boldsymbol{W}:=\boldsymbol{W}^{(0)}+\boldsymbol{W}'$. Given an input vector $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|=1$, we denote as follows the variables at the initialization (in first column), the variables after perturbation (in middle column) and the perturbation themselves (in last column):

$$
\begin{array}{lll}
\boldsymbol{h}_0^{(0)}=\boldsymbol{A}\boldsymbol{x} & \boldsymbol{h}_0=\boldsymbol{A}\boldsymbol{x} & \boldsymbol{h}_0'=\boldsymbol{0}\\
\boldsymbol{g}_l^{(0)}=\boldsymbol{W}_l^{(0)}\boldsymbol{h}_{l-1}^{(0)}, & \boldsymbol{g}_l=\boldsymbol{W}_l\boldsymbol{h}_{l-1} & \boldsymbol{g}_l'=\boldsymbol{g}_l-\boldsymbol{g}_l^{(0)}\\
(\boldsymbol{D}_l)_{jj}^{(0)}=\dfrac{1_{(\boldsymbol{g}_l^{(0)})_j\ge 0}+\alpha 1_{(\boldsymbol{g}_l^{(0)})_j<0}}{\sqrt{1+\alpha^2}}, & (\boldsymbol{D}_l)_{jj}=\dfrac{1_{(\boldsymbol{g}_l)_j\ge 0}+\alpha 1_{(\boldsymbol{g}_l)_j<0}}{\sqrt{1+\alpha^2}}, & \boldsymbol{D}_l'=\boldsymbol{D}_l-\boldsymbol{D}_l^{(0)}\\
\boldsymbol{h}_l^{(0)}=\tilde{\sigma}_\alpha(\boldsymbol{W}_l^{(0)}\boldsymbol{h}_{l-1}^{(0)})\equiv\tilde{\sigma}_\alpha(\boldsymbol{g}_l^{(0)}), & \boldsymbol{h}_l=\tilde{\sigma}_\alpha(\boldsymbol{W}_l\boldsymbol{h}_{l-1})\equiv\tilde{\sigma}_\alpha(\boldsymbol{g}_l), & \boldsymbol{h}_l'=\boldsymbol{h}_l-\boldsymbol{h}_l^{(0)}.
\end{array}
$$

Since we fix $\boldsymbol{A}$ and $\boldsymbol{W}_{L+1}\equiv\boldsymbol{B}$ in the training, $\boldsymbol{B}^{(0)}:=\boldsymbol{B}$ and $\boldsymbol{A}^{(0)}:=\boldsymbol{A}$.

Lemma B.4. *If $\|\boldsymbol{W}'\|_2=\omega<O(\frac{1}{L^{9/2}\ln^{3/2}m})$ and $m\ge\Omega(L^2)$, then the following events hold with probability at least $1-e^{-\Omega\left(\frac{m^{1/2}}{\ln m}\right)}$*

1. *$\|\boldsymbol{D}_l'\|_0<O(m\omega^{2/3}L)$ and $\|\boldsymbol{D}_l'\boldsymbol{g}_l\|<\frac{1-\alpha}{\sqrt{1+\alpha^2}}O(\omega L^{3/2})$*
2. *there exist vectors $\boldsymbol{g}_{l,1}'$ and $\boldsymbol{g}_{l,2}'$ such that $\boldsymbol{g}_l'=\boldsymbol{g}_{l,1}'+\boldsymbol{g}_{l,2}'$, and $\|\boldsymbol{g}_{l,1}'\|=O(\omega L^{3/2})$ and $\|\boldsymbol{g}_{l,2}'\|_\infty=O\Big(\frac{\omega L^{5/2}\sqrt{\ln m}}{\sqrt{m}}\Big)$,*
3. *$\|\boldsymbol{g}_l'\|,\|\boldsymbol{h}_l'\|<O\Big(\omega L^{5/2}\sqrt{\ln m}\Big)$.*

Proof. We divide the proof into two steps. First, we show that statements 2 and 3 of the lemma imply statement 1 . We then prove statements 2 and 3 of the lemma using an induction argument for $l\in\{0,1,...L\}$.

Statements 2 and 3 imply statement 1. We fix $l\in\{0,1,...L\}$. In view of Lemma B.1 and the focus on the lth layer, we assume that $\boldsymbol{h}_{l-1}^{(0)}$ is a fixed vector such that $\|\boldsymbol{h}_{l-1}^{(0)}\|\in[0.5,1.5]$. More precisely, we can condition on $\boldsymbol{h}_{l-1}^{(0)}$ and we know that with overwhelming probability $\|\boldsymbol{h}_{l-1}^{(0)}\|\in[0.5,1.5]$. We denote $g_{l,j}^{(0)}:=(\boldsymbol{g}_l^{(0)})_j$ (note the difference between the vector notation $\boldsymbol{g}_{i,l}$ and the scalar notation $g_{l,j}^{(0)}$). We recall that

$$
\boldsymbol{g}_l^{(0)}=\boldsymbol{W}_l^{(0)}\boldsymbol{h}_{l-1}^{(0)}\sim N\left(0,\frac{2\|\boldsymbol{h}_{l-1}^{(0)}\|^2}{m}\boldsymbol{I}\right)\text{ and thus }\ g_{l,j}^{(0)}\sim N\left(0,\frac{2\|\boldsymbol{h}_{l-1}^{(0)}\|^2}{m}\right)\text{ for }j\in[m].
$$

We define the following vector $\boldsymbol{d}$ and express it using the decomposition $\boldsymbol{g}_l'=\boldsymbol{g}_{l,1}'+\boldsymbol{g}_{l,2}'$ in statement 2 of this lemma:

$$
\boldsymbol{d}:=\boldsymbol{D}_l'(\boldsymbol{W}_l^{(0)}\boldsymbol{h}_l^{(0)}+\boldsymbol{g}_l')=\boldsymbol{D}_l'(\boldsymbol{W}_l^{(0)}\boldsymbol{h}_l^{(0)}+\boldsymbol{g}_{l,1}'+\boldsymbol{g}_{l,2}').
$$

We denote $D'_{l,jj}:=(\boldsymbol{D}'_l)_{jj}$, $g'_{l,1,j}:=(\boldsymbol{g}'_{l,1})_j$ and $g'_{l,2,j}:=(\boldsymbol{g}'_{l,2})_j$.

To estimate $\|\boldsymbol{d}\|$ and $\|\boldsymbol{d}\|_0$ we define the following auxiliary sets that partition $\{j\in[m]:d_j\neq 0\}$, S_1 and S_2. To do this we arbitrarily choose a positive number $\xi>2\|\boldsymbol{g}'_{l,2}\|_\infty$ and define

$$S_1:=\{j\in[m]:|g^{(0)}_{l,j}|<\xi,\, d_j\neq 0\}$$

and

$$S_2:=\{j:j\in[m]/S_1, d_j\neq 0\}.$$

In the rest of the proof we bound $|S_1|$, $\sum_{j\in S_1}d_j^2$, $|S_2|$ and $\sum_{j\in S_2}d_j^2$. We then use these estimates to bound $\|\boldsymbol{d}\|$ and $\|\boldsymbol{d}\|_0$.

In order to bound $|S_1|$, we first note that

$$\mathbb{P}(|g^{(0)}_{l,j}|<\xi, d_j\neq 0)\leq\mathbb{P}(|g^{(0)}_{l,j}|<\xi)\leq\Theta\left(\xi\sqrt{\frac{m}{\|\boldsymbol{h}^{(0)}_{l-1}\|^2}}\right)=\Theta(\xi\sqrt{m}).$$

Combining a Chernoff bound for the binomial distribution with the above estimate yields

$$|S_1|<O(\xi m^{3/2}) \text{ with probability at least } 1-e^{-\Omega(m^{3/2}\xi)}. \tag{40}$$

For $j\in S_1$, we upper bound the coordinate d_j of $\boldsymbol{d}$:

$$|d_j|\leq\left|\frac{1-\alpha}{\sqrt{1+\alpha^2}}\right||g^{(0)}_{l,j}+g'_{l,1,j}+g'_{l,2,j}|\leq\left|\frac{1-\alpha}{\sqrt{1+\alpha^2}}\right|(\xi+\|\boldsymbol{g}'_2\|_\infty+|g'_{l,1,j}|).$$

For each index $j\in[m]$ such as $D'_{l,jj}\neq 0$ we note from the definition of $\boldsymbol{D}'$ that $|D'_{l,jj}|=(1-\alpha)/\sqrt{1+\alpha^2}$. By squaring both sides of the above inequality, summing over the indices in S_1 and applying (40), we conclude that with probability at least $1-e^{-\Omega(m^{3/2}\xi)}$

$$\begin{aligned}\sum_{j\in S_1}|d_j|^2&\leq 3\sum_{j\in S_1}\frac{(1-\alpha)^2}{1+\alpha^2}(\xi^2+\|\boldsymbol{g}'_{l,2}\|_\infty^2+|g'_{l,1,j}|^2)\\&\leq\frac{3(1-\alpha)^2}{1+\alpha^2}|S_1|(\xi^2+\|\boldsymbol{g}'_{l,2}\|_\infty^2)+\frac{3(1-\alpha)^2}{1+\alpha^2}\|\boldsymbol{g}'_{l,1}\|^2\\&\leq\frac{3(1-\alpha)^2}{1+\alpha^2}O\Big(\xi m^{3/2}\Big)(\xi^2+\|\boldsymbol{g}'_{l,2}\|_\infty^2)+\frac{3(1-\alpha)^2}{1+\alpha^2}\|\boldsymbol{g}'_{l,1}\|^2.\end{aligned} \tag{41}$$

We next estimate $|S_2|$. The definitions of the diagonal matrices $\boldsymbol{D}_l$, $\boldsymbol{D}_l^{(0)}$ and $\boldsymbol{D}'_l$ imply that if $D'_{jj}\neq 0$, then $g^{(0)}_{l,j}$ and $g_{l,j}$ have opposite signs, or equivalently, $g^{(0)}_{l,j}+g'_{l,j}$ and $g^{(0)}_{i,l}$ have opposite signs, which further implies that $|g'_{l,j}|\geq|g^{(0)}_{l,j}|$. We further note that by the triangle inequality $|g'_{l,j}|\leq|g'_{l,1,j}|+|g'_{l,2,j}|$. Combining these two observation and then applying additional basic estimates, we obtain

$$|g'_{l,1,j}|\geq|g^{(0)}_{l,j}|-|g'_{l,2,j}|\geq\xi-\|\boldsymbol{g}'_{l,2}\|_\infty \text{ for } j\in S_2.$$

This bound clearly implies

$$\|\boldsymbol{g}'_{l,1}\|^2\geq\sum_{j\in S_2}|g'_{l,1,j}|^2\geq|S_2|(\xi-\|\boldsymbol{g}'_{l,2}\|_\infty)^2$$

and consequently

$$|S_2|\leq\frac{\|\boldsymbol{g}'_{l,1}\|^2}{(\xi-\|\boldsymbol{g}'_{l,2}\|_\infty)^2}. \tag{42}$$

For $j\in S_2$, we note as above that $g^{(0)}_{l,j}$ and $g'_{l,j}$ have opposite signs and $|g'_{l,j}|>|g^{(0)}_{l,j}|$. The combination of both of these observations imply $|g^{(0)}_{l,j}+g'_{l,j}|\leq|g'_{l,j}|$. The later observation and the partition of $\boldsymbol{g}_l$ according to the second statement

of the lemma yield the following bound for $j\in S_2$:

$$|d_j|=\frac{|1-\alpha|}{\sqrt{1+\alpha^2}}|g_{l,j}^{(0)}+g'_{l,j}|\le\frac{|1-\alpha|}{\sqrt{1+\alpha^2}}|g'_{l,j}| \tag{43}$$

$$\le\frac{|1-\alpha|}{\sqrt{1+\alpha^2}}(|g'_{l,1,j}|+\|\boldsymbol{g}'_{l,2}\|_\infty). \tag{44}$$

Squaring both sides of (44), summing over $j\in S_2$ and applying (42) yield

$$\sum_{j\in S_2}|d_j|^2\le 2\frac{(1-\alpha)^2}{1+\alpha^2}\sum_{j\in S_2}(|g'_{l,1,j}|^2+\|\boldsymbol{g}'_{l,2}\|_\infty^2)\le 2\frac{(1-\alpha)^2}{1+\alpha^2}(\|\boldsymbol{g}'_{l,1}\|^2+|S_2|\|\boldsymbol{g}'_{l,2}\|_\infty^2)$$

$$\le 2\frac{(1-\alpha)^2}{1+\alpha^2}\left(\|\boldsymbol{g}'_{l,1}\|^2+\frac{\|\boldsymbol{g}'_{l,2}\|_\infty^2\|\boldsymbol{g}'_{l,1}\|^2}{(\xi-\|\boldsymbol{g}'_{l,2}\|_\infty)^2}\right). \tag{45}$$

Obtaining these four different estimates we conclude with bounds on $\|\boldsymbol{d}\|_0$ and $\|\boldsymbol{d}\|$. We first note that (40) and (42) yield

$$\|\boldsymbol{d}\|_0\le|S_1|+|S_2|\le\Theta(\xi m^{3/2})+\frac{\|\boldsymbol{g}'_{l,1}\|^2}{(\xi-\|\boldsymbol{g}'_{l,2}\|_\infty)^2}\quad\text{with probability at least}\quad 1-e^{-\Omega(m^{3/2}\xi)}.$$

Since $\xi>2\|\boldsymbol{g}'_{l,2}\|_\infty$, we can obtain the following bound:

$$\|\boldsymbol{d}\|_0\le\Theta(\xi m^{3/2})+\frac{4\|\boldsymbol{g}'_{l,1}\|^2}{\xi^2}.$$

In order to tighten the above bound, we minimize the right hand side term with respect to ξ and note that its minimal value is $m\|\boldsymbol{g}'_{l,1}\|^{2/3}$ and is obtained at $\xi_{\min}=\Theta\Big(\|\boldsymbol{g}'_{l,1}\|^{2/3}/m^{1/2}\Big)$. We note that the assumed conditions: $\omega<O\big(L^{-9/2}(\ln m)^{-3/2}\big)$, $\|\boldsymbol{g}'_{l,1}\|=O(\omega L^{3/2})$ and $\|\boldsymbol{g}'_{l,2}\|_\infty<O(\omega L^{5/2}\sqrt{\ln m}/\sqrt{m})$ imply that $\xi_{\min}>2\|\boldsymbol{g}'_{l,2}\|_\infty$ so that the minimum is achieved. Thus, an upper bound of $\|\boldsymbol{d}\|_0$ is obtained as

$$\|\boldsymbol{d}\|_0\le O(m\|\boldsymbol{g}'_{l,1}\|^{2/3})\le O(m\omega^{2/3}L).$$

Combining (41) and (45) yields

$$\|\boldsymbol{d}\|^2=\sum_{j=1}^m d_j^2=\sum_{j\in S_1}d_j^2+\sum_{j\in S_2}d_j^2$$

$$\le\frac{3(1-\alpha)^2}{1+\alpha^2}O\Big(\xi m^{3/2}\Big)(\xi^2+\|\boldsymbol{g}_{l,2}\|_\infty^2)+\frac{5(1-\alpha)^2}{1+\alpha^2}\|\boldsymbol{g}'_{l,1}\|^2+2\frac{(1-\alpha)^2}{1+\alpha^2}\frac{\|\boldsymbol{g}'_{l,1}\|^2\|\boldsymbol{g}'_{l,2}\|_\infty^2}{(\xi-\|\boldsymbol{g}'_{l,2}\|_\infty)^2}$$

$$\le C\frac{(1-\alpha)^2}{1+\alpha^2}(\xi^3m^{3/2}+\|\boldsymbol{g}'_{l,1}\|_2^2).$$

Plugging in $\xi=\xi_{\min}$ to the above equation and applying the second statement of this lemma result in

$$\|\boldsymbol{d}\|^2\le O\left(\frac{(1-\alpha)^2}{1+\alpha^2}\|\boldsymbol{g}'_{l,1}\|^2\right)\le\frac{1-\alpha}{\sqrt{1+\alpha^2}}O(\omega^2L^3). \tag{46}$$

Consequently, our bounds for $\|\boldsymbol{D}_l\|_0$ and $\|\boldsymbol{d}\|=\|\boldsymbol{D}'_l\boldsymbol{g}_l\|$ are

$$\|\boldsymbol{D}\|_0\le\|\boldsymbol{d}\|_0\le O(m(\omega L^{3/2})^{2/3})=O(m\omega^{2/3}L), \tag{47}$$

$$\|\boldsymbol{D}'_l\boldsymbol{g}_l\|=\|\boldsymbol{d}\|\le O(\omega L^{3/2}). \tag{48}$$

Proof of Statements 2 and 3. We prove statements 2 and 3 of Lemma B.4 by induction on $l\in\{0,1,\cdots L\}$. These statements clearly hold at $l=0$ because there is no perturbation at $l=0$ and $\boldsymbol{g}'_0=\boldsymbol{h}'_0=\boldsymbol{0}$. In view of the previous

part of the proof, we assume the lemma holds for layers $0\le j\le l-1$ and prove that the second and third statements of the lemma hold at layer l.

Following the given definitions, we expand $\boldsymbol{g}'_l$ as follows

$$\begin{aligned}\boldsymbol{g}'_l&=\boldsymbol{W}_l\boldsymbol{D}_{l-1}\boldsymbol{g}_{l-1}-\boldsymbol{W}_l^{(0)}\boldsymbol{D}_{l-1}^{(0)}\boldsymbol{g}_{l-1}^{(0)}\\&=(\boldsymbol{W}_l^{(0)}+\boldsymbol{W}'_l)(\boldsymbol{D}_{l-1}^{(0)}+\boldsymbol{D}'_{l-1})(\boldsymbol{g}_{l-1}^{(0)}+\boldsymbol{g}'_{l-1})-\boldsymbol{W}_l^{(0)}\boldsymbol{D}_{l-1}^{(0)}\boldsymbol{g}_{l-1}^{(0)}\\&=\boldsymbol{W}'_l(\boldsymbol{D}_{l-1}^{(0)}+\boldsymbol{D}'_{l-1})(\boldsymbol{g}_{l-1}^{(0)}+\boldsymbol{g}'_{l-1})+\boldsymbol{W}_l^{(0)}\boldsymbol{D}'_{l-1}(\boldsymbol{g}_{l-1}^{(0)}+\boldsymbol{g}'_{l-1})+\boldsymbol{W}_l^{(0)}\boldsymbol{D}_{l-1}^{(0)}\boldsymbol{g}'_{l-1}.\end{aligned}\tag{49}$$

We first expand $\boldsymbol{g}'_{l-1}$ in the last term of the above equation. Similarly, we then iteratively expand $\boldsymbol{g}'_{l-2}$, ..., $\boldsymbol{g}'_1$ and obtain the following expression:

$$\begin{aligned}\boldsymbol{g}'_l&=\boldsymbol{W}'_l(\boldsymbol{D}_{l-1}^{(0)}+\boldsymbol{D}'_{l-1})(\boldsymbol{g}_{l-1}^{(0)}+\boldsymbol{g}'_{l-1})+\boldsymbol{W}_l^{(0)}\boldsymbol{D}'_{l-1}(\boldsymbol{g}_{l-1}^{(0)}+\boldsymbol{g}'_{l-1})\\&\quad+\boldsymbol{W}_l^{(0)}\boldsymbol{D}_{l-1}^{(0)}\big(\boldsymbol{W}'_{l-1}(\boldsymbol{D}_{l-2}^{(0)}+\boldsymbol{D}'_{l-2})(\boldsymbol{g}_{l-2}^{(0)}+\boldsymbol{g}'_{l-2})+\boldsymbol{W}_{l-1}^{(0)}\boldsymbol{D}'_{l-2}(\boldsymbol{g}_{l-2}^{(0)}+\boldsymbol{g}'_{l-2})\big)\\&\quad+\boldsymbol{W}_l^{(0)}\boldsymbol{D}_{l-1}^{(0)}\boldsymbol{W}_{l-1}^{(0)}\boldsymbol{D}_{l-2}^{(0)}\boldsymbol{g}'_{l-2}\\&=...\\&=\sum_{k=0}^{l-1}\left(\prod_{j=1}^{k}\boldsymbol{W}_{l-j+1}^{(0)}\boldsymbol{D}_{l-j}^{(0)}\right)\Big(\boldsymbol{W}'_{l-k}(\boldsymbol{D}_{l-k-1}^{(0)}+\boldsymbol{D}'_{l-k-1})(\boldsymbol{g}_{l-k-1}^{(0)}+\boldsymbol{g}'_{l-k-1})\\&\quad+\boldsymbol{W}_{l-k}^{(0)}\boldsymbol{D}'_{l-k-1}(\boldsymbol{g}_{l-k-1}^{(0)}+\boldsymbol{g}'_{l-k-1})\Big)+\left(\prod_{j=1}^{l-1}\boldsymbol{W}_{l-j+1}^{(0)}\boldsymbol{D}_{l-j}^{(0)}\right)\boldsymbol{g}'_0.\end{aligned}$$

Since $\boldsymbol{g}'_0=\boldsymbol{0}$, the last term is $\boldsymbol{0}$. We consequently express $\boldsymbol{g}'_l$ as a sum of the following two terms:

$$\boldsymbol{g}'_l=\sum_{k=0}^{l-1}\left(\prod_{j=1}^{k}\boldsymbol{W}_{l-j+1}^{(0)}\boldsymbol{D}_{l-j}^{(0)}\right)\Big(\boldsymbol{W}'_{l-k}(\boldsymbol{D}_{l-k-1}^{(0)}+\boldsymbol{D}'_{l-k-1})(\boldsymbol{g}_{l-k-1}^{(0)}+\boldsymbol{g}'_{l-k-1})\Big)\tag{50}$$

$$+\sum_{k=0}^{l-1}\left(\prod_{j=1}^{k}\boldsymbol{W}_{l-j+1}^{(0)}\boldsymbol{D}_{l-j}^{(0)}\right)\Big(\boldsymbol{W}_{l-k}^{(0)}\boldsymbol{D}'_{l-k-1}(\boldsymbol{g}_{l-k-1}^{(0)}+\boldsymbol{g}'_{l-k-1})\Big).\tag{51}$$

We estimate with high probability the above first term (right hand side in (50)) by using the assumption $\|\boldsymbol{W}'\|<\omega$ and the first statement in Lemma B.3 (to bound $\|\prod_{j=1}^{k}\boldsymbol{W}_{l-j+1}^{(0)}\boldsymbol{D}_{l-j}^{(0)}\|$, $k=0,1,...l-1$). We thus obtain with probability at least $1-Le^{\Omega(m/L)}$

$$\begin{aligned}&\left\|\sum_{k=0}^{l-1}\left(\prod_{j=1}^{k}\boldsymbol{W}_{l-j+1}^{(0)}\boldsymbol{D}_{l-j}^{(0)}\right)(\boldsymbol{W}'_{l-k}(\boldsymbol{D}_{l-k-1}^{(0)}+\boldsymbol{D}'_{l-k-1})(\boldsymbol{g}_{l-k-1}^{(0)}+\boldsymbol{g}'_{l-k-1}))\right\|\\&\le L\max_k\left\|\prod_{j=1}^{k}\boldsymbol{W}_{l-j+1}^{(0)}\boldsymbol{D}_{l-j}^{(0)}\right\|\left\|\boldsymbol{W}'_{l-k}(\boldsymbol{D}_{l-k-1}^{(0)}+\boldsymbol{D}'_{l-k-1})(\boldsymbol{g}_{l-k-1}^{(0)}+\boldsymbol{g}'_{l-k-1})\right\|\\&\le L\cdot O(\sqrt{L})\cdot\max_k\|\boldsymbol{W}'_{l-k}\|\cdot\|\boldsymbol{D}_{l-k-1}\|\cdot\|\boldsymbol{g}_{l-k-1}^{(0)}+\boldsymbol{g}'_{l-k-1}\|\\&\le L\cdot O(\sqrt{L})\cdot\omega\cdot\frac{\max(|\alpha|,1)}{\sqrt{1+\alpha^2}}\cdot\max_k\|\boldsymbol{g}_{l-k-1}^{(0)}+\boldsymbol{g}'_{l-k-1}\|\\&\le O(\omega L^{3/2})\max_k\|\boldsymbol{g}_{l-k-1}^{(0)}+\boldsymbol{g}'_{l-k-1}\|.\end{aligned}$$

We further use Lemma B.1 to bound $\|\boldsymbol{g}_{l-k-1}^{(0)}\|$, $k\in\{0,1,...l-1\}$, by a constant and use the induction assumption to bound $\|\boldsymbol{g}'_{l-k-1}\|$, $k\in\{0,1,...l-1\}$, by $O(\omega L^{5/2}\sqrt{\ln m})$. With probability at least $1-O(L)e^{-\Omega(m/L)}$, the first term (right hand side in (50)) is thus bounded by

$$O(\omega L^{3/2})(O(1)+O(\omega L^{5/2}\sqrt{\ln m}))=O(\omega L^{3/2}).\tag{52}$$

In order to bound the second term, which appears in (51), we denote

$$\boldsymbol{d}_k := \boldsymbol{D}'_{l-k-1}(\boldsymbol{g}^{(0)}_{l-k-1}+\boldsymbol{g}'_{l-k-1}), \quad k=0,1,...l-1$$

and

$$\boldsymbol{y}_k := \left(\prod_{j=1}^{k}\boldsymbol{W}^{(0)}_{l-j+1}\boldsymbol{D}^{(0)}_{l-j}\right)\boldsymbol{W}^{(0)}_{l-k}\boldsymbol{d}_k.$$

We show it can be decomposed into $\boldsymbol{y}_k=\boldsymbol{y}_{k,1}+\boldsymbol{y}_{k,2}$, where with probability at least $1-Le^{-\Omega(m/L)}$,

$$\|\boldsymbol{y}_{k,1}\|\le O\left(\frac{(1-\alpha)\omega L^{3/2}}{(1+\alpha^2)^{1/2}\sqrt{m}}\right), \quad \|\boldsymbol{y}_{k,2}\|_\infty \le O\left(\frac{(1-\alpha)\omega L^{3/2}\sqrt{\ln m}}{(1+\alpha^2)^{1/2}\sqrt{m}}\right).$$

Denoting $\boldsymbol{u}_k := \boldsymbol{D}^{(0)}_{l-1}\boldsymbol{W}^{(0)}_{l-1},...\boldsymbol{D}^{(0)}_{l-k}\boldsymbol{W}^{(0)}_{l-k}\boldsymbol{d}_k$ and applying the induction assumption we note that $\|\boldsymbol{d}_k\|_0 < O(m\omega^{2/3}L)$. Next, we apply the third statement of the Lemma B.3 for $\boldsymbol{u}_k$ (instead of $\boldsymbol{v}$) and obtain that with probability at least $1-e^{-\Omega(m/L)}$

$$\|\boldsymbol{u}_k\| \le 4\|\boldsymbol{d}_k\|. \tag{53}$$

We note that $\boldsymbol{y}_k=\boldsymbol{W}^{(0)}_l\boldsymbol{u}_k$ and thus $\boldsymbol{y}_k|\boldsymbol{u}_k \sim N\left(0,\frac{2\|\boldsymbol{u}_k\|^2}{m}\boldsymbol{I}\right)$.

We denote $y_{k,j} := (\boldsymbol{y}_k)_j$ and $\sigma^2 := 2\|\boldsymbol{u}_k\|^2/m$ and we let $b=O(\|\boldsymbol{u}_k\|\sqrt{\ln m/m})$. We investigate the tail probability of the Gaussian random variable $y_{k,j}$ conditioned on $\boldsymbol{u}_k$. It is clear that

$$\mathbb{P}(|y_{k,j}|\ge bt|\boldsymbol{u}_k) \le \frac{1}{\sqrt{2\pi}bt/\sigma}e^{-b^2t^2/2\sigma^2} \quad \forall t\in\mathbb{N}. \tag{54}$$

We denote $R_t := \{j : y_{k,j}\ge bt\}\subset[m]$ and $r_t := \sqrt{m}/((\ln m)^2t^2)$. Using the independence of $\{y_{k,j}\}_{j\in[m]}$ given $\boldsymbol{u}_k$ and applying a union bound for (54) yield

$$\begin{aligned}\mathbb{P}(|R_t|\ge r_t|\boldsymbol{u}_k) &\le \binom{m}{r_t}\times\left(\frac{1}{\sqrt{2\pi}bt/\sigma}e^{-b^2t^2/2\sigma^2}\right)^{r_t}\\ &\le \left(\frac{\|\boldsymbol{u}_k\|}{\sqrt{\pi}bt\sqrt{m(1+\alpha^2)}}\right)^{r_t}\left(\frac{me}{r_t}\right)^{r_t}e^{-\Omega(b^2t^2mr_t)}\\ &\le O(1)\exp\left(-\Omega(b^2t^2mr_t)+\left(\frac{1}{2}\ln m-\ln b-\Omega(1)\right)r_t\right).\end{aligned}$$

Denoting $q := \sqrt{m}/\ln^2 m$, we simplify the above bound as follows

$$\mathbb{P}(|R_t|\ge q/t^2)\le e^{-\Omega(b^2qm)}.$$

We further denote $Q := \{0,1,2,3,..\lfloor\frac{1}{2}\log_2 q\rfloor\}$, $N_Q := \lfloor\frac{1}{2}\log_2 q\rfloor$ and $T := \{2^p : p\in Q\}$. We designate the elements in T by $t_p := 2^p$ for $p\in Q$. Let $t_{N_Q+1} := 2^{\lfloor\frac{1}{2}\log_2 q\rfloor+1}\equiv 2^{N_Q+1}$ and notice that $t^2_{N_Q+1} > q$. Thus, applying the above estimate and a union bound over $t\in T$ and t_{N_Q+1}

$$|R_t| < q/t^2,\ \forall t\in T, \ \text{ and } \ |R_{t_{N_Q+1}}|<1 \ \text{ with probability at least } \ 1-(|T|+1)e^{-\Omega(b^2qm)}.$$

By definition, we note that when $|R_{t_{N_Q+1}}|=0$ and $|y_{k,j}|<t_{N_Q+1}$ for $j\in R_{t_{N_Q}}$. We also note that for $j\in R_{t_p}\setminus R_{t_{p+1}}$,

$|y_{k,j}|<t_{p+1}$. Thus, for $R:=R_1\equiv\{j:|y_{k,j}|\geq b\}$, we bound $\sum_{j\in R}y_{k,j}^2$ with high probability as follows

$$\begin{aligned}\sum_{j\in R}y_{k,j}^2&=\sum_{j\in R/R_{t_{N_Q}}}y_{k,j}^2+\sum_{j\in R_{t_{N_Q}}}y_{k,j}^2\leq\sum_{j\in R/R_{t_{N_Q}}}y_{k,j}^2+|R_{t_{N_Q}}|(bt_{N_Q+1})^2\\&\leq\sum_{j\in R/R_{t_{N_Q}}/R_{t_{N_Q-1}}}y_{k,j}^2+\sum_{j\in R_{t_{N_Q-1}}/R_{t_{N_Q}}}y_{k,j}^2+|R_{t_{N_Q}}|(bt_{N_Q+1})^2\\&\leq\sum_{j\in R/R_{t_{N_Q}}/R_{t_{N_Q-1}}}y_{k,j}^2+|R_{t_{N_Q-1}}|(bt_{N_Q})^2+|R_{t_{N_Q}}|(bt_{N_Q+1})^2\\&\ldots\ldots\\&\leq\sum_{p\in Q}|R_{t_p}|(b2^{p+1})^2\leq\sum_{p\in Q}q/t_p^2(b2^{p+1})^2\\&=\sum_{p\in Q}qb^2 2^2=O(qb^2\ln q)\ \text{ with probability at least } 1-\Omega(|T|)e^{-\Omega(b^2qm)}.\end{aligned}$$

Since $b=O\Big(\|\boldsymbol{u}_k\|\sqrt{\ln m/m}\Big)$ and $q=\sqrt{m}/\ln^2 m$, we express the above bound as

$$\sum_{j\in R}y_{k,j}^2\leq O(\|\boldsymbol{u}_k\|^2/m)\ \text{ with probability at least }\ 1-e^{-\Omega(\frac{m^{1/2}}{\ln m})}. \tag{55}$$

We split vector $\boldsymbol{y}_k$ into $\boldsymbol{y}_k=\boldsymbol{y}_{k,1}+\boldsymbol{y}_{k,2}$ using the indices set R as

$$\boldsymbol{y}_{k,1}=(y_{k,1}1_{1\in R},y_{k,2}1_{2\in R},\ldots,y_{k,m}1_{m\in R})^T, \tag{56}$$
$$\boldsymbol{y}_{k,2}=(y_{k,1}1_{1\notin R},y_{k,2}1_{2\notin R},\ldots,y_{k,m}1_{m\notin R})^T. \tag{57}$$

Using (55) and the definition of R, and then the induction assumption on the bound of $\|\boldsymbol{d}_k\|$ and (53) yield the following estimates with probability at least $1-e^{-\Omega\left(\frac{m^{1/2}}{\ln m}\right)}$:

$$\|\boldsymbol{y}_{k,1}\|\leq O\bigg((\frac{\|\boldsymbol{u}\|}{m^{1/2}}\bigg)\leq O\bigg((\frac{(1-\alpha)\omega L^{3/2}}{(1+\alpha^2)^{1/2}m^{1/2}}\bigg), \tag{58}$$
$$\|\boldsymbol{y}_{k,2}\|_\infty\leq b=O\left(\frac{\|\boldsymbol{u}\|\sqrt{\ln m}}{\sqrt{m}}\right)\leq O\left(\frac{(1-\alpha)\omega L^{3/2}\sqrt{\ln m}}{(1+\alpha^2)^{1/2}\sqrt{m}}\right). \tag{59}$$

Following the later decomposition of $\boldsymbol{y}_k$ (with the components in (56) and (57)), we decompose the term in (51) into $\sum_{k=0}^{l-1}\boldsymbol{y}_{k,1}$ and $\sum_{k=0}^{l-1}\boldsymbol{y}_{k,2}$. We denote $\boldsymbol{g}'_{l,2}:=\sum_{k=0}^{l-1}\boldsymbol{y}_{k,2}$ and $\boldsymbol{g}'_{l,1}:=\boldsymbol{g}'_l-\boldsymbol{g}'_{l,2}$. We note that $\boldsymbol{g}'_{l,1}$ is the sum of the term in (50) and $\sum_{k=0}^{l-1}\boldsymbol{y}_{k,1}$. By using the bound of (50) given in (52) and (58), we bound $\boldsymbol{g}'_{l,1}$ as follows

$$\begin{aligned}\|\boldsymbol{g}'_{l,1}\|&\leq\left\|\sum_{k=0}^{l-1}\left(\prod_{j=1}^{k}\boldsymbol{W}_{l-j+1}^{(0)}\boldsymbol{D}_{l-j}^{(0)}\right)\Big(\boldsymbol{W}'_{l-k}(\boldsymbol{D}_{l-k-1}^{(0)}+\boldsymbol{D}'_{l-k-1})(\boldsymbol{g}_{l-k-1}^{(0)}+\boldsymbol{g}'_{l-k-1})\Big)\right\|\\&\quad+\sum_{k=0}^{l-1}\|\boldsymbol{y}_{k,1}\|\\&\leq O(\omega L^{3/2})+\sum_{k=0}^{l-1}\|\boldsymbol{y}_{k,1}\|\\&\leq O(\omega L^{3/2})+L\max_{k\in\{0,1,\ldots,l-1\}}\|\boldsymbol{y}_{k,1}\|\\&\leq O(\omega L^{3/2})+LO\bigg(\frac{(1-\alpha)\omega L^{3/2}}{(1+\alpha^2)^{1/2}m^{1/2}}\bigg).\end{aligned}$$

Using the fact that $m \geq \Omega(L^2)$, we show the ℓ_2 norm for $\boldsymbol{g}'_{l,1}$ in the second statement of this lemma holds:

$$\|\boldsymbol{g}'_{l,1}\| \leq O(\omega L^{3/2}) + LO\left(\frac{(1-\alpha)\omega L^{3/2}}{(1+\alpha^2)^{1/2} m^{1/2}}\right) \leq O(\omega L^{3/2}).$$

Applying the induction assumption, i.e., $\|\boldsymbol{g}'_{l-k,1}\| \leq O(\omega L^{3/2})$ for $k \in \{0,1,...,l-1\}$, and (59), we conclude the second statement of the lemma for layer l as follows

$$\|\boldsymbol{g}'_{l,2}\|_\infty \leq \sum_{k=0}^{l-1} \|\boldsymbol{y}_{k,2}\|_\infty \leq LO\left(\frac{1-\alpha}{(1+\alpha^2)^{1/2}} \frac{\sqrt{\ln m}\omega L^{3/2}}{\sqrt{m}}\right) = O\left(\frac{\sqrt{\ln m}\omega L^{5/2}}{\sqrt{m}}\right). \tag{60}$$

Finally, we note that $\|\boldsymbol{g}'_l\| \leq \|\boldsymbol{g}'_{l,1}\| + \|\boldsymbol{g}'_{l,2}\|$, and thus the first part $\|\boldsymbol{g}'_{l,1}\|$ is bounded by $O(\omega L^{3/2})$. Furthermore, applying (60), we bound the second part, $\|\boldsymbol{g}'_{l,2}\|$, as follows

$$\|\boldsymbol{g}'_{l,2}\| = \sqrt{\sum_{j \in S_2} g_{l,2,j}^2} \leq \sqrt{m \frac{\ln m \omega^2 L^5}{m}} = \sqrt{\ln m}\omega L^{5/2}.$$

By definition, $\boldsymbol{h}'_l = \boldsymbol{D}\boldsymbol{g}'_l + \boldsymbol{D}'\boldsymbol{g}_l^{(0)} + \boldsymbol{D}'\boldsymbol{g}'_l = \boldsymbol{D}\boldsymbol{g}'_l + \boldsymbol{D}'\boldsymbol{g}_l$. Applying $\|\boldsymbol{D}\| \leq 1$, $\|\boldsymbol{g}'_l\| \leq O(\omega L^{5/2}\sqrt{\ln m})$ and $\|\boldsymbol{D}'\boldsymbol{g}_l\| \leq O(\omega L^{3/2})$, we bound the norm of $\boldsymbol{h}'_l$ in the following way

$$\|\boldsymbol{h}'_l\| \leq O(1)O(\omega L^{5/2}\sqrt{\ln m}) + O(\omega L^{3/2}) = O(\omega L^{5/2}\sqrt{\ln m}).$$

Thus the third statement of this lemma is concluded for layer l. □

Lemma B.5. *For given integer a,b as $1 \leq a < b \leq L$, and if $d < O\big(\frac{m}{L \ln m}\big)$, $\|\boldsymbol{W}'\| \leq \omega < O\Big(\frac{1}{L^{9/2}\ln^{3/2} m}\Big)$. Then we obtain that with probability at least $1 - e^{-\Omega(m/L)}$*

1. $\|\boldsymbol{W}_b^{(0)}(\boldsymbol{D}_{i,b-1}^{(0)} + \boldsymbol{D}'_{i,b-1})\boldsymbol{W}_{b-1}^{(0)} ... (\boldsymbol{D}_{i,a}^{(0)} + \boldsymbol{D}'_{i,a})\boldsymbol{W}_a^{(0)}\| \leq O(\sqrt{L})$.
2. $\|(\boldsymbol{W}_b^{(0)} + \boldsymbol{W}'_b)(\boldsymbol{D}_{i,b-1}^{(0)} + \boldsymbol{D}'_{i,b-1})(\boldsymbol{W}_{b-1}^{(0)} + \boldsymbol{W}'_{b-1}) ... (\boldsymbol{D}_{i,a}^{(0)} + \boldsymbol{D}'_{i,a})(\boldsymbol{W}_a^{(0)} + \boldsymbol{W}'_a)\| \leq O(\sqrt{L})$.
3. $\|\boldsymbol{W}_{b+1}^{(0)}(\boldsymbol{D}_{i,b}^{(0)} + \boldsymbol{D}'_{i,b})(\boldsymbol{W}_b^{(0)} + \boldsymbol{W}'_b) ... (\boldsymbol{D}_{i,a}^{(0)} + \boldsymbol{D}'_{i,a}) - \boldsymbol{W}_{b+1}^{(0)}\boldsymbol{D}_{i,b}^{(0)}\boldsymbol{W}_b^{(0)} ... \boldsymbol{W}_{a+1}^{(0)}\boldsymbol{D}_{i,a}^{(0)}\| \leq O\Big(\frac{1-\alpha}{\sqrt{1+\alpha^2}} L^{3/2}\Big)$.
4. $\|\boldsymbol{B}(\boldsymbol{D}_L^{(0)} + \boldsymbol{D}'_L)(\boldsymbol{W}_L^{(0)} + \boldsymbol{W}'_L) ... (\boldsymbol{D}_{i,a}^{(0)} + \boldsymbol{D}'_{i,a}) - \boldsymbol{B}\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)} ... \boldsymbol{W}_{a+1}^{(0)}\boldsymbol{D}_{i,a}^{(0)}\|$
   $\leq O\Big(\frac{1-\alpha}{\sqrt{1+\alpha^2}} \frac{\omega^{1/3} L^2 \sqrt{m \ln m}}{\sqrt{d}}\Big)$.

The proof of this lemma follows the same arguments of the proofs of Lemmas 8.6 and 8.7 in Allen-Zhu et al. (2019b), but uses instead Lemma B.4 and the fact that $\|\boldsymbol{D}'\| = (1-\alpha)/\sqrt{1+\alpha^2}$.

B.4 Gradient Bounds and Proof of Lemma 4.2

We first introduce two lemmas (Lemmas B.6 and B.7) that provide upper and lower bounds for the Frobenius norm of a certain matrix-valued function $\boldsymbol{G}_{i,l}(\boldsymbol{v}; \boldsymbol{W}^{(0)})$ with randomly initialized parameters $\boldsymbol{W}^{(0)}$. This function, which is defined below in (61) equals the gradient of the loss function when $\boldsymbol{v} = \boldsymbol{e}_i^{(0)} \equiv \boldsymbol{B}\boldsymbol{h}_{L,i}^{(0)} - \boldsymbol{y}_i$. At last, we conclude Lemma 4.2 by applying the perturbation bounds of Lemmas B.4 and B.5 in order to show that the order of the bounds in Lemmas B.6 and B.7 are not affected by a small perturbation $\boldsymbol{W}'$ as long as $\|\boldsymbol{W}'\| \leq \omega < O\Big(\frac{\delta^{3/2}}{n^{3/2} L^{15/2} \ln^{3/2} m}\Big)$.

We remark that the proof of Lemma B.6 is straightforward and follows Allen-Zhu et al. (2019b). The proof of Lemma B.7 follows ideas of Zou and Gu (2019), while adapting it to Leaky ReLUs and improving the lower bound of $\|\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}^{(0)})\|_F^2$ by quantifying lower bounds for layers before L instead of only using $\|\nabla_{\boldsymbol{W}_L} \mathcal{L}(\boldsymbol{W}^{(0)})\|_F^2$ as done in Zou and Gu (2019). This improvement reduces a factor L in the lower bound, which will eventually make the learning rate of the desired theory independent of L. The idea of concluding Lemma 4.2 by examining the effect of a small perturbation on the parameter follows Allen-Zhu et al. (2019b).

We define the matrix-valued function, $\boldsymbol{G}_{i,l}(\boldsymbol{v}; \boldsymbol{W})$, for $l \in [L]$ and $i \in [n]$ and $\boldsymbol{v} \in \mathbb{R}^d$ as follows

$$\boldsymbol{G}_{i,l}(\boldsymbol{v}; \boldsymbol{W}) := \boldsymbol{D}_{i,l}\mathbf{Back}_{i,l}^T \boldsymbol{v}\boldsymbol{h}_{i,l-1}^T = (\mathbf{Back}_{i,l}\boldsymbol{D}_{i,l})^T \boldsymbol{v}\boldsymbol{h}_{i,l-1}^T. \tag{61}$$

We note that $\boldsymbol{G}_{i,l}(\boldsymbol{v};\boldsymbol{W})$ is related to the gradient of the loss function as follows:

$$\boldsymbol{G}_{i,l}(\boldsymbol{e}_i;\boldsymbol{W})\equiv\nabla_{\boldsymbol{W}_l}\mathrm{loss}(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W}).$$

Lemma B.6. *Assume the setup of §2 with randomly initialized $\boldsymbol{W}^{(0)}$. If $d\le O(\frac{m}{L\ln m})$, then with probability at least $1-e^{-\Omega(m/L)}$*

$$\|\boldsymbol{G}_{i,l}(\boldsymbol{v};\boldsymbol{W}^{(0)})\|_F^2\le O\Big(\frac{m}{d}\Big)\|\boldsymbol{v}\|^2. \tag{62}$$

Proof. The second statement in Lemma B.3 implies that $\|\mathbf{Back}_{i,l}\|<O(\sqrt{\frac{m}{d}})$ with probability at least $1-e^{-\Omega(m/L)}$ and therefore

$$\begin{aligned}\|\boldsymbol{G}_{i,l}(\boldsymbol{v};\boldsymbol{W}^{(0)})\|_F^2&\le\|\boldsymbol{D}_{i,l}\mathbf{Back}_{i,l}^{(0)T}\boldsymbol{v}\boldsymbol{h}_{i,l-1}^{(0)T}\|_F^2\\&\le\|\boldsymbol{D}_{i,l}\mathbf{Back}_{i,l}^{(0)T}\boldsymbol{v}\|^2\|\boldsymbol{h}_{i,l-1}^{(0)T}\|^2\\&\le O\Big(\frac{m}{d}\|\boldsymbol{v}_i\|_2^2\Big).\end{aligned}$$

□

Lemma B.7. *Assume the setup of §2 and with randomly initialized $\boldsymbol{W}^{(0)}$. For any set of vector $\{\boldsymbol{v}_i\}_{i=1}^n\subset\mathbb{R}^d$,*

$$\|\sum_{i=1}^n\boldsymbol{G}_{i,l}(\boldsymbol{v}_i;\boldsymbol{W}^{(0)})\|_F^2\ge\Omega\Big(\frac{(1-\alpha)^2}{(1+\alpha^2)}\frac{\delta m}{ndL}\Big)\sum_{i=1}^n\|\boldsymbol{v}_i\|^2\quad\textit{with probability }\ge1-e^{-\Omega(m\delta^2)}.$$

Proof. We separate the proof of this lemma into four parts. In the first part, we define a set in $\mathbb{R}^m$ (see (63) below) and show two important properties of this set (see (64) and (66) below). In the second part, we establish a lower bound for a useful function (as defined in (70) below) with a probability at least 0.5. In the third part, we use this lower bound to establish a lower bound of the loss function with a positive probability. In the fourth part, we conclude the lemma by using all the results proved in the former three parts.

Since we assume randomly initialized parameters without training, we simply denote $\boldsymbol{h}_{i,l}:=\boldsymbol{h}_{i,l}^{(0)}$ and $\boldsymbol{W}:=\boldsymbol{W}^{(0)}$ across this proof.

Part 1. We arbitrarily fix $l\in[L]$ and recall that $\boldsymbol{h}_{i,l}$ is the output of lth layer. We denote

$$\hat{\boldsymbol{h}}_{i,l}:=\boldsymbol{h}_{i,l}/\|\boldsymbol{h}_{i,l}\|.$$

We form an orthogonal matrix $\boldsymbol{Q}_{i,l}\in\mathbb{R}^{m\times m}$ whose first column is $\hat{\boldsymbol{h}}_{i,l}$. We denote the matrix in $\mathbb{R}^{m\times(m-1)}$ which completes this vector by $\tilde{\boldsymbol{Q}}_{i,l}$, that is, $\boldsymbol{Q}_{i,l}:=[\hat{\boldsymbol{h}}_{i,l},\tilde{\boldsymbol{Q}}_{i,l}]$.

For a small constant $c_1>0$ (the choice of c_1 will be determined during the proof), we let $\gamma=c_1\delta/(nL\sqrt{m})$. For $i\in[n]$ and the fixed $l\in[L]$, we define

$$\mathcal{W}_{i,l}:=\{\boldsymbol{w}\in\mathbb{R}^m:|\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w}|<\gamma,|\langle\tilde{\boldsymbol{Q}}_{i,l}\tilde{\boldsymbol{Q}}_{i,l}^T\boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle|>2\gamma,\forall j\in[n],j\ne i\}\subset\mathbb{R}^m. \tag{63}$$

We prove that for any choice of γ the sets $\mathcal{W}_{i,l}$, $i\in[n]$, have no intersection, that is,

$$\mathcal{W}_{i,l}\cap\mathcal{W}_{j,l}=\emptyset,\quad\forall i\ne j\in[n]. \tag{64}$$

For any $\boldsymbol{w}\in\mathcal{W}_{i,l}$, we need to prove that $\boldsymbol{w}\notin\mathcal{W}_{j,l}$, where $j\ne i\in[n]$. We prove this by contradiction. Given $\boldsymbol{w}\in\mathcal{W}_{i,l}$, we assume that there exists $j\ne i\in[n]$ such that $\boldsymbol{w}\in\mathcal{W}_{j,l}$. Since $\tilde{\boldsymbol{Q}}_{j,l}\tilde{\boldsymbol{Q}}_{j,l}^T=\boldsymbol{I}-\hat{\boldsymbol{h}}_{j,l}\hat{\boldsymbol{h}}_{j,l}^T$, we rewrite $\tilde{\boldsymbol{Q}}_{j,l}\tilde{\boldsymbol{Q}}_{j,l}^T\boldsymbol{w}$ as

$$\tilde{\boldsymbol{Q}}_{j,l}\tilde{\boldsymbol{Q}}_{j,l}^T\boldsymbol{w}=(\boldsymbol{I}-\hat{\boldsymbol{h}}_{j,l}\hat{\boldsymbol{h}}_{j,l}^T)\boldsymbol{w}=\boldsymbol{w}-\langle\boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle\hat{\boldsymbol{h}}_{j,l}. \tag{65}$$

Applying (65) and the fact that $\langle \boldsymbol{w},\hat{\boldsymbol{h}}_{i,l}\rangle<\gamma$ and $\langle \boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle<\gamma$ for $\boldsymbol{w}\in\mathcal{W}_{i,l}\cap\mathcal{W}_{j,l}$ results in

$$\begin{aligned}|\langle \tilde{\boldsymbol{Q}}_{j,l}\tilde{\boldsymbol{Q}}_{j,l}^T\boldsymbol{w},\hat{\boldsymbol{h}}_{i,l}\rangle| &= |\langle \boldsymbol{w}-\langle \boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle\hat{\boldsymbol{h}}_{j,l},\hat{\boldsymbol{h}}_{i,l}\rangle| \\ &\le |\langle \boldsymbol{w},\hat{\boldsymbol{h}}_{i,l}\rangle| + |\langle \boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle\langle \hat{\boldsymbol{h}}_{j,l},\hat{\boldsymbol{h}}_{i,l}\rangle| \\ &< \gamma+\gamma|\langle \hat{\boldsymbol{h}}_{j,L},\hat{\boldsymbol{h}}_{i,L}\rangle| \\ &\le 2\gamma.\end{aligned}$$

On the other hand, since $\boldsymbol{w}\in\mathcal{W}_{j,l}$, $|\langle \tilde{\boldsymbol{Q}}_{j,l}\tilde{\boldsymbol{Q}}_{j,l}^T\boldsymbol{w},\hat{\boldsymbol{h}}_{i,l}\rangle|>2\gamma$ for $i\neq j$, which contradicts the above equation. Therefore, we conclude (64).

Next, we assume $\boldsymbol{w}\sim N(0,\frac{2}{m}\boldsymbol{I})$ and prove that

$$\mathbb{P}(\boldsymbol{w}\in\mathcal{W}_{i,l})\ge\Omega\left(\frac{\delta}{nL}\right). \tag{66}$$

The orthogonality of $\boldsymbol{Q}_{i,l}$ implies that $\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w}$ and $\tilde{\boldsymbol{Q}}_{i,l}^T\boldsymbol{w}$ are independent. We thus express the probability (66) as follows

$$\mathbb{P}(\boldsymbol{w}\in\mathcal{W}_{i,l})=\mathbb{P}(|\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w}|<\gamma)\mathbb{P}(|\langle \tilde{\boldsymbol{Q}}_{i,l}\tilde{\boldsymbol{Q}}_{i,l}^T\boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle|>2\gamma,\ \forall j\in[n],j\neq i). \tag{67}$$

We note that $\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w}\sim N(0,\frac{2}{m})$ and thus express the first multiplicative term in (67) as

$$\mathbb{P}(|\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w}|<\gamma)=\frac{\sqrt{m}}{\sqrt{4\pi}}\int_{-\gamma}^{\gamma}e^{-\frac{mx^2}{4}}dx\ge\Omega\big(\gamma\sqrt{m}\big),\text{ when }\gamma\sqrt{m}<1. \tag{68}$$

To express the second multiplicative term of (67), we first derive the distribution of $\hat{\boldsymbol{h}}_{j,l}^T\tilde{\boldsymbol{Q}}_{i,l}\tilde{\boldsymbol{Q}}_{i,l}^T\boldsymbol{w}$. Since $\tilde{\boldsymbol{Q}}_{i,l}\tilde{\boldsymbol{Q}}_{i,l}^T=\boldsymbol{I}_m-\hat{\boldsymbol{h}}_{i,l}\hat{\boldsymbol{h}}_{i,l}^T$ and $\tilde{\boldsymbol{Q}}_{i,l}^T\tilde{\boldsymbol{Q}}_{i,l}=\boldsymbol{I}_{m-1}$,

$$\begin{aligned}\hat{\boldsymbol{h}}_{j,l}^T\tilde{\boldsymbol{Q}}_{i,l}\tilde{\boldsymbol{Q}}_{i,l}^T\boldsymbol{w} &\sim N\left(0,\frac{2}{m}\hat{\boldsymbol{h}}_{j,l}^T\tilde{\boldsymbol{Q}}_{i,l}\tilde{\boldsymbol{Q}}_{i,l}^T\tilde{\boldsymbol{Q}}_{i,l}\tilde{\boldsymbol{Q}}_{i,l}^T\hat{\boldsymbol{h}}_{j,l}\right) \\ &= N\left(0,\frac{2}{m}\hat{\boldsymbol{h}}_{j,l}^T(\boldsymbol{I}-\hat{\boldsymbol{h}}_{i,l}\hat{\boldsymbol{h}}_{i,l}^T)\hat{\boldsymbol{h}}_{j,l}\right) \\ &= N\left(0,(1-\langle \hat{\boldsymbol{h}}_{j,l},\hat{\boldsymbol{h}}_{i,l}\rangle^2)\frac{2}{m}\right).\end{aligned}$$

By Lemma B.2, we recall that with probability at least $1-e^{-\Omega(\delta^4m/L^4)}$,

$$\langle \hat{\boldsymbol{h}}_{i,L},\hat{\boldsymbol{h}}_{j,L}\rangle^2\le 1-\Omega(\delta^2/L^2),\ \text{ for all } i\neq j\in[n].$$

We thus note that $\hat{\boldsymbol{h}}_{j,l}^T\tilde{\boldsymbol{Q}}_{i,l}\tilde{\boldsymbol{Q}}_{i,l}^T\boldsymbol{w}\sim N(0,\tau^2)$, where τ^2 is greater than $\Omega(\delta^2/mL^2)$. Consequently,

$$\mathbb{P}(|\hat{\boldsymbol{h}}_{j,l}^T\tilde{\boldsymbol{Q}}_{i,l}\tilde{\boldsymbol{Q}}_{i,l}^T\boldsymbol{w}|<2\gamma)=\frac{1}{\sqrt{2\pi\tau^2}}\int_{-2\gamma}^{2\gamma}\exp\left(-\frac{x^2}{2\tau^2}\right)dx\le O\Big(\frac{\gamma}{\tau}\Big)\le O\left(\frac{\gamma L\sqrt{m}}{\delta}\right).$$

Applying a union bound over all $j\in[n]$, $j\neq i$, yields

$$\mathbb{P}\Big(\exists j\in[n],j\neq i\ \text{ such that }|\hat{\boldsymbol{h}}_{j,l}^T\tilde{\boldsymbol{Q}}_{i,l}\tilde{\boldsymbol{Q}}_{i,l}^T\boldsymbol{w}|\le 2\gamma\Big)\le nO\left(\frac{\gamma L\sqrt{m}}{\delta}\right).$$

Consequently,

$$\mathbb{P}\Big(|\hat{\boldsymbol{h}}_{j,l}^T\tilde{\boldsymbol{Q}}_{i,l}\tilde{\boldsymbol{Q}}_{i,l}^T\boldsymbol{w}|>2\gamma\ \forall j\in[n],j\neq i\Big)\ge 1-O\left(\frac{\gamma nL\sqrt{m}}{\delta}\right). \tag{69}$$

Plugging (69) and (68) into (67) yields

$$\begin{aligned}\mathbb{P}(\boldsymbol{w}\in\mathcal{W}_{i,l}) &= \mathbb{P}(|u_{i,1}|<\gamma)\mathbb{P}(|\boldsymbol{v}_{i,j}|>2\gamma,\ \forall j\in[n],j\neq i) \\ &\ge \Omega\big(\gamma\sqrt{m}\big)\left(1-O\left(\frac{\gamma nL\sqrt{m}}{\delta}\right)\right).\end{aligned}$$

Recall that $\gamma = c_1\delta/(nL\sqrt{m})$, we select small c_1 such that both $O(\frac{\gamma nL\sqrt{m}}{\delta}) = O(1)\cdot c_1 < 1$ and $\gamma\sqrt{m} = c_1\delta/(nL) < 1$. We thus conclude this part as follows

$$\mathbb{P}(\boldsymbol{w}\in\mathcal{W}_{i,l}) \geq \Omega\left(\frac{\delta}{nL}\right).$$

Part 2. Given integer $k\in[m]$ and $l\in[L]$, we define the following vector-valued function for $\boldsymbol{a}=(a_1,...a_n)^T\in\mathbb{R}^n$ and $\boldsymbol{w}\in\mathbb{R}^m$:

$$\boldsymbol{b}_{k,l}(\boldsymbol{w},\boldsymbol{a}) := \sum_{i=1}^{n} a_i\tilde{\sigma}'_\alpha(\langle\boldsymbol{w},\boldsymbol{h}_{i,l}\rangle)\boldsymbol{h}_{i,l}. \tag{70}$$

We prove that conditioning on the event $\boldsymbol{w}\in\mathcal{W}_{i,l}$, a certain lower bound of $\|\boldsymbol{b}_{k,l}(\boldsymbol{w},\boldsymbol{a})\|$ is achieved with a probability at least 0.5, that is,

$$P\left(\|\boldsymbol{b}_{k,l}(\boldsymbol{w},\boldsymbol{a})\| \geq \frac{a_i}{2}\frac{(1-\alpha)}{\sqrt{1+\alpha^2}}\|\boldsymbol{h}_{i,l}\| \,\middle|\, \boldsymbol{w}\in\mathcal{W}_{i,l}\right) > \frac{1}{2}.$$

We rewrite $\boldsymbol{w}$ as $\boldsymbol{w} = \boldsymbol{Q}_{i,l}\boldsymbol{Q}_{i,l}^T\boldsymbol{w} = (\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w})\hat{\boldsymbol{h}}_{i,l} + \tilde{\boldsymbol{Q}}_{i,k}\tilde{\boldsymbol{Q}}_{i,k}^T\boldsymbol{w}$,

$$\langle\boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle = (\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w})\langle\hat{\boldsymbol{h}}_{i,l},\hat{\boldsymbol{h}}_{j,l}\rangle + \langle\tilde{\boldsymbol{Q}}_{i,k}\tilde{\boldsymbol{Q}}_{i,k}^T\boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle \quad \text{for} \quad j\neq i$$

Using the following two facts: $\boldsymbol{w}\in\mathcal{W}_{i,l}$ and both $\hat{\boldsymbol{h}}_{i,l}$ and $\hat{\boldsymbol{h}}_{j,l}$ are unit vectors, we bound the absolute value of the first term of the above expression as folows

$$|(\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w})||\langle\hat{\boldsymbol{h}}_{i,l},\hat{\boldsymbol{h}}_{j,l}\rangle| < \gamma.$$

Since $\boldsymbol{w}\in\mathcal{W}_{i,l}$, the magnitude of the second term is greater than 2γ. We note that the sign of $\langle\boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle$ is the same as that of $\langle\tilde{\boldsymbol{Q}}_{i,k}\tilde{\boldsymbol{Q}}_{i,k}^T\boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle$. This and the piecewise linearity of the Leaky ReLU function imply that for $\boldsymbol{w}\in\mathcal{W}_{i,l}$

$$\tilde{\sigma}'_\alpha(\langle\boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle) = \tilde{\sigma}'_\alpha(\langle\tilde{\boldsymbol{Q}}_{i,k}\tilde{\boldsymbol{Q}}_{i,k}^T\boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle\rangle), \quad \text{for} \quad j\neq i. \tag{71}$$

We note (71) implies the following expression for $\boldsymbol{b}_{k,l}(\boldsymbol{w},\boldsymbol{a})$ for $\boldsymbol{w}\in\mathcal{W}_{i,l}$: by ,

$$\begin{aligned}
\boldsymbol{b}_{k,l}(\boldsymbol{w},\boldsymbol{a}) &= a_i\tilde{\sigma}'_\alpha(\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w})\boldsymbol{h}_{i,l} + \sum_{j\neq i} a_j\tilde{\sigma}'_\alpha(\hat{\boldsymbol{h}}_{j,l}^T\boldsymbol{w})\boldsymbol{h}_{j,l} \\
&= a_i\tilde{\sigma}'_\alpha(\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w})\boldsymbol{h}_{i,l} + \sum_{j\neq i} a_j\tilde{\sigma}'_\alpha(\langle\tilde{\boldsymbol{Q}}_{i,k}\tilde{\boldsymbol{Q}}_{i,k}^T\boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle\rangle)\boldsymbol{h}_{j,l} \\
&= a_i\frac{(1-\alpha)}{\sqrt{1+\alpha^2}}1_{\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w}>0}\boldsymbol{h}_{i,l} + a_i\frac{\alpha}{\sqrt{1+\alpha^2}}\boldsymbol{h}_{i,l} + \sum_{j\neq i} a_j\tilde{\sigma}'_\alpha(\langle\tilde{\boldsymbol{Q}}_{i,k}\tilde{\boldsymbol{Q}}_{i,k}^T\boldsymbol{w},\hat{\boldsymbol{h}}_{j,l}\rangle\rangle)\boldsymbol{h}_{j,l}.
\end{aligned}$$

We denote

$$\begin{aligned}
\boldsymbol{b}_1 &:= a_i\frac{(1-\alpha)}{\sqrt{1+\alpha^2}}\boldsymbol{h}_{i,L-1} \\
\boldsymbol{r} &:= a_i\frac{\alpha}{\sqrt{1+\alpha^2}}\boldsymbol{h}_{i,L-1} + \sum_{j\neq i} a_j\phi'_\alpha(\langle\tilde{\boldsymbol{Q}}_i\tilde{\boldsymbol{u}}_i,\hat{\boldsymbol{h}}_{j,L-1}\rangle)\boldsymbol{h}_{j,L-1},
\end{aligned}$$

and thus express $\boldsymbol{b}_{k,l}(\boldsymbol{w},\boldsymbol{a})$ as follows

$$\boldsymbol{b}_{k,l}(\boldsymbol{w},\boldsymbol{a}) = \boldsymbol{b}_1 1_{\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w}>0} + \boldsymbol{r}. \tag{72}$$

By symmetry of normal distribution, we know that $\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w} > 0$ with probability 0.5. We also note that $\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w}$ and $\tilde{\boldsymbol{Q}}_{i,l}^T\boldsymbol{w}$ are independent and thus $1_{\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w}>0}$ is independent with $\boldsymbol{r}$.

We consider two possibility for $\boldsymbol{r}$:

- When $\|\boldsymbol{r}\| \geq \frac{1}{2}\|\boldsymbol{b}_1\|$, we know that with probability 0.5, $\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w} \leq 0$, which implies $\boldsymbol{b}_{k,l}(\boldsymbol{w},\boldsymbol{a}) = \boldsymbol{r}$, and thus $\|\boldsymbol{b}_{k,l}(\boldsymbol{w},\boldsymbol{a})\| \geq \frac{1}{2}\|\boldsymbol{b}_1\|$. We thus note that at least with probability 0.5 that $\|\boldsymbol{b}_{k,l}(\boldsymbol{w},\boldsymbol{a})\| \geq \frac{1}{2}\|\boldsymbol{b}_1\|$.

- When $\|\boldsymbol{r}\| < \frac{1}{2}\|\boldsymbol{b}_1\|$, we note that $\hat{\boldsymbol{h}}_{i,l}^T\boldsymbol{w} > 0$ with probability 0.5, then by triangle inequality, we imply $\|\boldsymbol{b}_{k,l}(\boldsymbol{w},\boldsymbol{a})\| \geq \|\boldsymbol{b}_1\| - \|\boldsymbol{r}\| \geq \frac{1}{2}\|\boldsymbol{b}_1\|$.

We conclude that

$$P\left(\|\boldsymbol{b}_{k,l}(\boldsymbol{w},\boldsymbol{a})\| \geq \frac{a_i}{2}\frac{(1-\alpha)}{\sqrt{1+\alpha^2}}\|\boldsymbol{h}_{i,l}\|\|\,\big|\,\boldsymbol{w}\in\mathcal{W}_{i,l}\right) \geq \frac{1}{2}. \tag{73}$$

Part 3. The proof of this part does not depend on a particular choice of $i\in[n]$. For simplicity, we thus drop the subscript i in this part.

For $\boldsymbol{v}\in\mathbb{R}^d$, $k\in[m]$ and $l\in[L]$, we define $a_{k,l} := \langle(\mathbf{Back}_l)_{\cdot,k},\boldsymbol{v}\rangle$. We want to show that for any integers $k\in[m]$ and $l\in[L]$,

$$\mathbb{P}\left((a_{k,l})^2 \geq O\left(\frac{\|\boldsymbol{v}\|^2}{d}\right)\right) > 1-\exp(-O(1)). \tag{74}$$

To prove the above statement, we also need an auxiliary statement for $l\in\{2,3,...L+1\}$,

$$\|\boldsymbol{D}_{l-1}\mathbf{Back}_l^T\boldsymbol{v}\| \geq (1-\epsilon)\sqrt{\frac{m}{2d}}\|\boldsymbol{v}\| \text{ with probability at least } 1-e^{-\Omega(m\epsilon^2/L^2)}. \tag{75}$$

In order to prove the above two statements (74) and (75), we first prove that $\boldsymbol{W}_l\Big|\boldsymbol{D}_l$ has the same distribution as $\boldsymbol{W}_l$, i.e., $N(0,\frac{2}{m})$. Then we use a similar argument to that in the proof of Lemma B.1 in order to show (75). Finally, by using the distribution of $\boldsymbol{W}_l$ given $\boldsymbol{D}_l$, together with (75), we prove (74) and conclude this part.

We prove a more general statement for conditional distributions: given a normal random vector in $\mathbb{R}^p$ as $\boldsymbol{w}\sim N(0,\sigma^2\boldsymbol{I}_p)$, and a random vector $\boldsymbol{h}\in\mathbb{R}^p$ that satisfies following three properties:

1. $\boldsymbol{h}$ is independent with $\boldsymbol{w}$
2. The norm $\|\boldsymbol{h}\|$ is independent with the direction $\boldsymbol{h}/\|\boldsymbol{h}\|$
3. The direction $\boldsymbol{h}/\|\boldsymbol{h}\|$ is uniform distribution in the unit sphere $\mathcal{S}^{p-1}$

We further define $B := 1_{\boldsymbol{h}^T\boldsymbol{w}>0}$ as a random variable. Then the conditional distribution of $\boldsymbol{w}|B$ is the same as the unconditional distribution of $\boldsymbol{w}$, that is

$$\boldsymbol{w}|B \overset{d}{=} \boldsymbol{w} \sim N(0,\sigma^2\boldsymbol{I}_p). \tag{76}$$

Remark: a normal random vector $N(\boldsymbol{0},\sigma^2\boldsymbol{I})$ satisfies the above three properties and thus $\boldsymbol{w}$ also satisfies above three properties.

We denote the unit vectors $\hat{\boldsymbol{h}} := \boldsymbol{h}/\|\boldsymbol{h}\|$ and $\hat{\boldsymbol{w}} := \boldsymbol{w}/\|\boldsymbol{w}\|$. We first note that $B \equiv 1_{\hat{\boldsymbol{h}}^T\hat{\boldsymbol{w}}>0}$ only depends on the directions of $\boldsymbol{h}$ and $\boldsymbol{w}$. By the former observation and the fact that $\|\boldsymbol{w}\|$ is independent with $\hat{\boldsymbol{w}}$, we thus note $\|\boldsymbol{w}\||B = \|\boldsymbol{w}\|$. We denote the probability density function for a random variable Y by f_Y. We next consider the probability density function $f_{\boldsymbol{w}|B}(\boldsymbol{w})$, by independence of the norm and the direction for $\boldsymbol{w}$, we obtain

$$f_{\boldsymbol{w}|B}(\boldsymbol{w}) = f_{\boldsymbol{w}|B}(\|\boldsymbol{w}\|,\hat{\boldsymbol{w}}) = f_{\|\boldsymbol{w}\||B}(\|\boldsymbol{w}\|)f_{\hat{\boldsymbol{w}}|B}(\hat{\boldsymbol{w}}) = f_{\|\boldsymbol{w}\|}(\|\boldsymbol{w}\|)f_{\hat{\boldsymbol{w}}|B}(\hat{\boldsymbol{w}}). \tag{77}$$

Thus, in order to show (76), it is sufficient suffices to show that $\hat{\boldsymbol{w}}|B \overset{d}{=} \hat{\boldsymbol{w}}$. We prove this by showing that for any set $\mathcal{A}\subset\mathcal{S}^{p-1}$ in unit sphere, $\mathbb{P}(\hat{\boldsymbol{w}}\in\mathcal{A}|B=b) = \mathbb{P}(\hat{\boldsymbol{w}}\in\mathcal{A})$ for any $b=0$ or 1. Given $\hat{\boldsymbol{h}}$ is uniform in unit sphere, we know that for any fixed direction $\hat{\boldsymbol{w}}$, $\mathbb{P}(\hat{\boldsymbol{h}}^T\hat{\boldsymbol{w}}>0) = 0.5$. By Bayes formula, former observation, and $\hat{\boldsymbol{h}}$ is uniform in $\mathcal{S}^{p-1}$

$$\begin{aligned}
\mathbb{P}(\hat{\boldsymbol{w}}\in\mathcal{A}|B=1) &= \frac{\mathbb{P}(\hat{\boldsymbol{w}}\in\mathcal{A},\, B=1)}{\mathbb{P}(B=1)} \\
&= \frac{\mathbb{P}(B=1|\hat{\boldsymbol{w}}\in\mathcal{A})\mathbb{P}(\hat{\boldsymbol{w}}\in\mathcal{A})}{\int_{\hat{\boldsymbol{w}}}\mathbb{P}(\hat{\boldsymbol{h}}^T\hat{\boldsymbol{w}}>0|\hat{\boldsymbol{w}})f_{\hat{\boldsymbol{w}}}(\hat{\boldsymbol{w}})} \\
&= \frac{\int_{\mathcal{A}}\mathbb{P}(B=1|\hat{\boldsymbol{w}}=\hat{\boldsymbol{w}})f_{\hat{\boldsymbol{w}}}(\hat{\boldsymbol{w}})d\hat{\boldsymbol{w}}}{\int_{\mathcal{S}^{p-1}}\mathbb{P}(\hat{\boldsymbol{h}}^T\hat{\boldsymbol{w}}>0|\hat{\boldsymbol{w}})f_{\hat{\boldsymbol{w}}}(\hat{\boldsymbol{w}})d\hat{\boldsymbol{w}}} \\
&= \frac{0.5\int_{\mathcal{A}}f_{\hat{\boldsymbol{w}}}(\hat{\boldsymbol{w}})d\hat{\boldsymbol{w}}}{0.5\int_{\mathcal{S}^{p-1}}f_{\hat{\boldsymbol{w}}}(\hat{\boldsymbol{w}})d\hat{\boldsymbol{w}}} \\
&= \int_{\mathcal{A}}f_{\hat{\boldsymbol{w}}}(\hat{\boldsymbol{w}})d\hat{\boldsymbol{w}} = \mathbb{P}(\hat{\boldsymbol{w}}\in\mathcal{A}).
\end{aligned}$$

A similar argument leads to $\mathbb{P}(\hat{\boldsymbol{w}}\in\mathcal{A}|B=0)=\mathbb{P}(\hat{\boldsymbol{w}}\in\mathcal{A})$. By (77) and above argument, we conclude (76).

Given the symmetry of normal distribution, we conclude that $\boldsymbol{g}_l$ satisfies the three properties we required for $\boldsymbol{h}$ above. Together with the fact that $(\boldsymbol{W}_l)_{k,\cdot}$ is normal $N(0,2/m\boldsymbol{I}_m)$, we thus conclude that $(\boldsymbol{W}_l)_{k,\cdot}|(\boldsymbol{D}_l)_{kk}$ is still normal $N(0,2/m\boldsymbol{I}_m)$.

Next, we estimate the norm of $\boldsymbol{D}_{l-1}^T\mathbf{Back}_l^T\boldsymbol{v}$. We define vector $\boldsymbol{z}_l:=\boldsymbol{D}_l\mathbf{Back}_{l+1}^T\boldsymbol{v}$ for $l\in[L]$ and $\boldsymbol{z}_{L+1}:=\boldsymbol{v}$. We first note $\boldsymbol{z}_L=\boldsymbol{D}_L\boldsymbol{B}^T\boldsymbol{v}$ and $(\boldsymbol{B}^T\boldsymbol{v})_j\sim N(0,\|\boldsymbol{v}\|^2/d)$ for $j\in[m]$. By denoting Bournulli random variables $B_{L,j}:=1_{(\boldsymbol{g}_L)_j>0}$, each index of $\boldsymbol{z}_L$ can be expressed as

$$(\boldsymbol{z}_L)_j^2=\frac{B_{L,j}+\alpha^2(1-B_{L,j})}{1+\alpha^2}(\boldsymbol{B}^T\boldsymbol{v})_j^2\ \text{ for } j\in[m].$$

We denote $Q_L:=\{j:B_{L,j}=1\}$. Conditioning on Q_L, denote two independent random variables $H_{L,1}\sim\chi^2(|Q_L|)$ and $H_{L,2}\sim\chi^2(m-|Q_L|)$, we note

$$\|\boldsymbol{z}_L\|^2\Big|Q_L\overset{d}{=}\frac{\|\boldsymbol{v}\|^2}{d(1+\alpha^2)}H_{L,1}+\frac{\alpha^2\|\boldsymbol{v}\|^2}{d(1+\alpha^2)}H_{L,2}.$$

By symmetry of random variables before L layer, we know $B_{L,j}\sim$ Bournulli(0.5) and then by Chernoff bound on binomial distribution, we note that with probability at least $1-e^{-\Omega(m\epsilon^2)}$, $|Q_L|\in[(0.5-\epsilon/2)m,(0.5+\epsilon/2)m]$. Given this even happen, by using tail probability for chi-squared distribution, we note that

$$\mathbb{P}(H_{L,1}<0.5m(1-\epsilon))<e^{-\Omega(m\epsilon^2)}.$$

Similarly,

$$\mathbb{P}(H_{L,2}<0.5m(1-\epsilon))<e^{-\Omega(m\epsilon^2)}.$$

By taking event $|Q_L|\in[(0.5-\epsilon/2)m,(0.5+\epsilon/2)m]$ and using above probabilities, we conclude the lower bound for $\|\boldsymbol{z}_L\|$

$$\|\boldsymbol{z}_L\|^2\geq\frac{m\|\boldsymbol{v}\|^2}{2d}(1-\epsilon)\ \text{ with probability at least } 1-\Omega(e^{-\Omega(m\epsilon^2)}).\tag{78}$$

We note that $\boldsymbol{z}_{l-1}=\boldsymbol{D}_{l-1}^T\boldsymbol{W}_l^T\boldsymbol{z}_l$. Conditioning on $\boldsymbol{z}_l$, we note that $\boldsymbol{W}_l|\boldsymbol{z}_l\equiv\boldsymbol{W}_l|\boldsymbol{D}_l$ is a random matrix whose entries are i.i.d $N(0,2/m)$. We denote a random variable $B_{l,j}:=1_{(\boldsymbol{g}_l)_j>0}$, then

$$\|\boldsymbol{z}_{l-1}\|^2|\boldsymbol{z}_l=\sum_{j=1}^m\left(\frac{B_{l,j}+\alpha^2(1-B_{l,j})}{1+\alpha^2}\left(\sum_i(\boldsymbol{W}_l)_{i,j}(\boldsymbol{z}_l)_i\right)^2\right)\Big|\boldsymbol{z}_l.$$

We note that $(\sum_i\boldsymbol{W}_l)_{i,j}(\boldsymbol{z}_l)_i|\boldsymbol{z}_l\sim N(0,2\|\boldsymbol{z}_l\|^2/m)$. We denote the indices set where $B_{l,j}=1$ by $Q_l:=\{j:B_{l,j}=1\}$ and conditioning on Q_l, we further denote two independent random variables $H_{l,1}\sim\chi^2(|Q_l|)$ and $H_{l,2}\sim\chi^2(m-|Q_l|)$. We note that conditioning on Q_l, by similar argument we used above in proof of Lemma B.1, we know that

$$\|\boldsymbol{z}_{l-1}\|^2\Big|\boldsymbol{z}_l,Q_l\overset{d}{=}\frac{2\|\boldsymbol{z}_l\|^2}{m(1+\alpha^2)}H_{l,1}+\frac{2\alpha^2\|\boldsymbol{z}_l\|^2}{m(1+\alpha^2)}H_{l,2}.\tag{79}$$

By the same argument to derive (78), we know that by Chernoff bound for binomial distribution, with probability at least $1-e^{-\Omega(m\epsilon^2)}$, $|Q_l|\in[(0.5-\epsilon/2)m,(0.5+\epsilon/2)m]$, thus we note that

$$\mathbb{P}(H_{l,1}<0.5m(1-\epsilon))<e^{-\Omega(m\epsilon^2)},\ \mathbb{P}(H_{l,2}<0.5m(1-\epsilon))<e^{-\Omega(m\epsilon^2)}.$$

Consequently,

$$\|\boldsymbol{z}_{l-1}\|^2\geq\|\boldsymbol{z}_l\|^2(1-\epsilon)\ \text{ with probability at least } 1-\Omega(e^{-\Omega(m\epsilon^2)}).\tag{80}$$

For any positive number ϵ_0, when we choose $\epsilon=\epsilon_0/L$ in (78) and (80), and then by $(1-\epsilon_0/L)^L>1-\epsilon_0$, we conclude that

$$\|\boldsymbol{z}_l\|^2\geq\frac{m}{2d}\|\boldsymbol{v}\|^2(1-\epsilon_0)\text{ for all } l\in[L],\ \text{ with probability at least } 1-\Omega(L)e^{-\Omega(m\epsilon_0^2/L^2)}.\tag{81}$$

Finally, recall that $a_{k,l} = \langle (\mathbf{Back}_l)_{\cdot,k}, \boldsymbol{v} \rangle$ and by definition of $\boldsymbol{z}_l$ in above proof, we note that $a_{k,l} \equiv \langle (\boldsymbol{W}_l)_{\cdot,k}, \boldsymbol{z}_l \rangle$. We note that $(\boldsymbol{W}_l)_{\cdot,k}|\boldsymbol{z}_l = (\boldsymbol{W}_l)_{\cdot,k}|\boldsymbol{D}_l$, by first statement we proved in this part, we further can derive that $(\boldsymbol{W}_l)_{\cdot,k}|\boldsymbol{z}_l \sim N(0, 2/m\boldsymbol{I})$. Thus, we know that conditioning on $\boldsymbol{z}_l$,

$$a_{k,l}|\boldsymbol{z}_l \sim N(0, \frac{2\|\boldsymbol{z}_l\|^2}{m}),$$

By the tail probability of normal, we note that the with a constant probability that $a_{k,l}$ is lower bounded as

$$\mathbb{P}\bigg((a_{k,l})^2 \geq O\bigg(\frac{2\|\boldsymbol{z}_l\|^2}{m} \bigg) \bigg) > 1 - \exp(-\Omega(1)).$$

Combining with (81), which holds with an overwhelming probability, with a small constant choice of ϵ_0, we conclude

$$\mathbb{P}\bigg((a_{k,l})^2 \geq O\bigg(\frac{\|\boldsymbol{v}\|^2}{d} \bigg) \bigg) > 1 - \exp(-\Omega(1)) \text{ for} l \in [L].$$

Lastly, we also show this is also true for $l = L+1$. Recall that $\mathbf{Back}_{L+1} \equiv \boldsymbol{B}$ and that $a_{k,L+1} \equiv \langle \boldsymbol{B}_{\cdot,k}, \boldsymbol{v} \rangle \sim N\Big(0, \frac{\|\boldsymbol{v}\|^2}{d}\Big)$. By using normal distribution property,

$$\mathbb{P}\bigg((a_{k,L+1})^2 \geq O\bigg(\frac{\|\boldsymbol{v}\|^2}{d} \bigg) \bigg) > 1 - \exp(-\Omega(1)).$$

We conclude this part by the final statement that

$$\mathbb{P}\bigg((a_{k,l})^2 \geq O\bigg(\frac{\|\boldsymbol{v}\|^2}{d} \bigg) \bigg) > 1 - \exp(-\Omega(1)) \quad \text{for } l \in [L+1]. \tag{82}$$

Part 4. We denote a vector $\boldsymbol{a}_{k,l} \in \mathbb{R}^n$ by denoting its entries as $(\boldsymbol{a}_{k,l})_i := \langle (\mathbf{Back}_{i,l})_{\cdot,k}, \boldsymbol{v}_i \rangle$ for $i \in [n]$. By definition (70), we note that $\boldsymbol{b}_{k,l-1}((\boldsymbol{W}_l)_{k,\cdot}, \boldsymbol{a}_{k,l+1}) \equiv (\sum_{i=1}^n \boldsymbol{G}_{i,l}(\boldsymbol{v}_i; \boldsymbol{W}))_{k,\cdot}$, by the definition of Frobenius norm of a vector of matrices,

$$\bigg\| \sum_{i=1}^n \boldsymbol{G}_{i,l}(\boldsymbol{v}_i; \boldsymbol{W}) \bigg\|_F^2 = \sum_{k=1}^m \|\boldsymbol{b}_{k,l-1}((\boldsymbol{W}_l)_{k,\cdot}, \boldsymbol{a}_{k,l+1})\|^2. \tag{83}$$

Due to (64), for any vector $\boldsymbol{w} \in \mathbb{R}^m$ and any integer $l \in [L]$, we note

$$1 \geq \sum_{i=1}^n 1_{\boldsymbol{w} \in \mathcal{W}_{i,l-1}}. \tag{84}$$

It follows from (83) and (84),

$$\begin{aligned} \bigg\| \sum_{i=1}^n \boldsymbol{G}_{i,l}(\boldsymbol{v}_i; \boldsymbol{W}) \bigg\|_F^2 &\geq \sum_{k=1}^m \|\boldsymbol{b}_{k,l-1}((\boldsymbol{W}_l)_{k,\cdot}, \boldsymbol{a}_{k,l+1})\|^2 \sum_{i=1}^n 1_{(\boldsymbol{W}_l)_{k,\cdot} \in \mathcal{W}_{i,l-1}} \\ &= \sum_{k=1}^m \sum_{i=1}^n \|\boldsymbol{b}_{k,l-1}((\boldsymbol{W}_l)_{k,\cdot}, \boldsymbol{a}_{k,l+1})\|^2 1_{(\boldsymbol{W}_l)_{k,\cdot} \in \mathcal{W}_{i,l-1}} \end{aligned}$$

By (73), we know that with probability at least 0.5, conditioning on $(\boldsymbol{W}_l)_{k,\cdot} \in \mathcal{W}_{i,l-1}$,

$$\|\boldsymbol{b}_{k,l-1}((\boldsymbol{W}_l)_{k,\cdot}, \boldsymbol{a}_{k,l+1})\|^2 \geq \frac{(\boldsymbol{a}_{k,l+1})_i^2}{4} \frac{(1-\alpha)^2}{1+\alpha^2} \|\boldsymbol{h}_{i,l-1}\|^2$$

We introduce the following new event $\mathcal{V}_{i,l}$ as follows

$$\mathcal{V}_{i,l} := \Big\{ (\boldsymbol{W}_l)_{k,\cdot} \in \mathcal{W}_{i,l-1},\ (\boldsymbol{a}_{k,l+1})_i^2 \geq \frac{\|\boldsymbol{v}_i\|^2}{2d},\ \|\boldsymbol{h}_{i,l-1}\| \geq \frac{1}{2} \Big\}.$$

Using this event, the observation $\mathcal{V}_{i,l} \subset \{(\boldsymbol{W}_l)_{k,\cdot} \in \mathcal{W}_{i,l-1}\}$, the definition of $\mathcal{V}_{i,l}$ and (73), we obtain the following lower bound on the squared norm in (83):

$$\begin{aligned}\left\|\sum_{i=1}^{n} \boldsymbol{G}_{i,l}(\boldsymbol{v}_i;\boldsymbol{W})\right\|_F^2 &\geq \sum_{k=1}^{m}\sum_{i=1}^{n} \|\boldsymbol{b}_{k,l-1}((\boldsymbol{W}_l)_{k,\cdot},\boldsymbol{a}_{k,l+1})\|^2 1_{(\boldsymbol{W}_l)_{k,\cdot}\in\mathcal{W}_{i,l-1}} \\ &\geq \sum_{k=1}^{m}\sum_{i=1}^{n} \|\boldsymbol{b}_{k,l-1}((\boldsymbol{W}_l)_{k,\cdot},\boldsymbol{a}_{k,l+1})\|^2 1_{\mathcal{V}_{i,l}} \\ &\geq \sum_{k=1}^{m}\sum_{i=1}^{n} \frac{\|\boldsymbol{v}_i\|^2}{32d}\frac{(1-\alpha)^2}{1+\alpha^2}\mathbb{P}(\mathcal{V}_{i,l}).\end{aligned}$$

For simplicity, we denote

$$Z_k := \sum_{i=1}^{n} \frac{\|\boldsymbol{v}_i\|^2}{32d}\frac{(1-\alpha)^2}{1+\alpha^2} 1_{\mathcal{V}_{i,l}}.$$

To lower bound the probability $\mathbb{P}(\mathcal{V}_{i,l})$, we note that $\boldsymbol{W}_l$, $\boldsymbol{a}_{k,l+1}$ and $\boldsymbol{h}_{i,l-1}$ are independent because they depend on $\boldsymbol{W}_{l+k}$ for $k\in[L-l+1]$, $\boldsymbol{W}_l$ and $\boldsymbol{W}_{l-k}$ for $k\in[l]$. We note that $(\boldsymbol{a}_{k,l+1})_i$ is corresponding to $a_{k,l+1}$ with selecting $\boldsymbol{v}=\boldsymbol{v}_i$ in the statement proven in the previous part (74). Then by using (67), (74) and applying Lemma B.1

$$\begin{aligned}\mathbb{P}(\mathcal{V}_{i,l}) &= \mathbb{P}((\boldsymbol{W}_l)_{k,\cdot}\in\mathcal{W}_{i,l-1})\mathbb{P}\left((\boldsymbol{a}_{k,l+1})_i^2 \geq \frac{\|\boldsymbol{v}_i\|^2}{2d}\right)\mathbb{P}\left(\|\boldsymbol{h}_{i,l-1}\|\geq\frac{1}{2}\right) \\ &\geq \Omega\left(\frac{\delta}{nL}\right)\times(1-\exp(-\Omega(1)))\times\left(1-e^{-\Omega(m/L)}\right) = \Omega\left(\frac{\delta}{nL}\right).\end{aligned}$$

By property of indicator function, we note that

$$\mathbb{E}Z_k = \sum_{i=1}^{n} \frac{\|\boldsymbol{e}_i\|^2}{32d}\frac{(1-\alpha)^2}{1+\alpha^2}\mathbb{P}(\mathcal{V}_{i,l})$$

and

$$\mathrm{Var}Z_k = \frac{\|\boldsymbol{e}_i\|^2}{32d}\frac{(1-\alpha)^2}{1+\alpha^2} 1_{\mathcal{V}_{i,l}}\mathbb{P}(\mathcal{V}_{i,l})(1-\mathbb{P}(\mathcal{V}_{i,l}).$$

Then, by using Hoeffding inequality, with probability at least $1-e^{-\Omega(m\delta^2/L^2)}$ that

$$\begin{aligned}\sum_{k=1}^{m} Z_k &\geq \frac{m}{2}\sum_{i=1}^{n}\frac{\|\boldsymbol{e}_i\|^2}{32d}\frac{(1-\alpha)^2}{1+\alpha^2}\mathbb{P}(\mathcal{V}_i) \\ &\geq \frac{Cm}{2}\sum_{i=1}^{n}\frac{\|\boldsymbol{v}_i\|^2}{32d}\frac{(1-\alpha)^2}{1+\alpha^2}\Omega\left(\frac{\delta}{nL}\right).\end{aligned}$$

Thus we conclude the Lemma, for all $l\in[L]$, as follows:

$$\left\|\sum_{i=1}^{n} \boldsymbol{G}_{i,l}(\boldsymbol{v}_i;\boldsymbol{W}^{(0)})\right\|_F^2 \geq \sum_{k=1}^{m} Z_k \geq \Omega\left(\frac{(1-\alpha)^2}{(1+\alpha^2)}\frac{\delta m}{ndL}\right)\sum_{i=1}^{n}\|\boldsymbol{v}_i\|^2.$$

□

At last, we conclude the proof of Lemma 4.2.

Proof of Lemma 4.2. In order to prove the lower and upper bounds for the gradient for parameters $\boldsymbol{W}$ close to $\boldsymbol{W}^{(0)}$, we need leverage Lemma B.5 to show that after perturbation from $\boldsymbol{W}^{(0)}$, the change in gradient has a smaller order than the upper bound in Lemma B.6 and the lower bound in Lemma B.7. Then the same upper and lower bounds hold for $\boldsymbol{W}$ such that $\|\boldsymbol{W}^{(0)}-\boldsymbol{W}\|<\omega$ and thus conclude Lemma 4.2.

We denote a perturbation of the function $\boldsymbol{G}_{i,l}(\boldsymbol{v};\boldsymbol{W}^{(0)})$ with respect to $\boldsymbol{W}^{(0)}$,

$$\begin{aligned}
&\boldsymbol{G}_{i,l}(\boldsymbol{v};\boldsymbol{W})-\boldsymbol{G}_{i,l}(\boldsymbol{v};\boldsymbol{W}^{(0)})\\
&=(\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{D}_{i,L}\boldsymbol{W}_L...\boldsymbol{D}_{i,l+1}\boldsymbol{W}_{l+1}\boldsymbol{D}_{i,l})^T\boldsymbol{h}_{i,l-1}^T-(\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{D}_{i,L}^{(0)}\boldsymbol{W}_L^{(0)}...\boldsymbol{D}_{i,l+1}^{(0)}\boldsymbol{W}_{l+1}^{(0)}\boldsymbol{D}_{i,l}^{(0)})^T\boldsymbol{h}_{i,l-1}^{(0)T}\\
&=(\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{D}_{i,L}\boldsymbol{W}_L...\boldsymbol{D}_{i,l+1}\boldsymbol{W}_{l+1}\boldsymbol{D}_{i,l})^T\boldsymbol{h}_{li,-1}^T-(\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{D}_{i,L}^{(0)}\boldsymbol{W}_L^{(0)}...\boldsymbol{D}_{i,l+1}^{(0)}\boldsymbol{W}_{l+1}^{(0)}\boldsymbol{D}_{i,l}^{(0)})^T\boldsymbol{h}_{i,l-1}^T\\
&+(\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{D}_{i,L}^{(0)}\boldsymbol{W}_L^{(0)}...\boldsymbol{D}_{i,l+1}^{(0)}\boldsymbol{W}_{l+1}^{(0)}\boldsymbol{D}_{i,l}^{(0)})^T\boldsymbol{h}_{i,l-1}^T-(\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{D}_{i,L}^{(0)}\boldsymbol{W}_L^{(0)}...\boldsymbol{D}_{i,l+1}^{(0)}\boldsymbol{W}_{l+1}^{(0)}\boldsymbol{D}_{i,l}^{(0)})^T\boldsymbol{h}_{i,l-1}^{(0)T}
\end{aligned}$$

Using $\|\boldsymbol{u}\boldsymbol{v}^T\|_F\leq\|\boldsymbol{u}\|\|\boldsymbol{v}\|$, and denoting vectors $\boldsymbol{v}_i\in\mathbb{R}^d$, we derive the bound for the change of the gradient by

$$\begin{aligned}
&\|\sum_{i=1}^n\boldsymbol{G}_{i,l}(\boldsymbol{v}_i;\boldsymbol{W})-\boldsymbol{G}_{i,l}(\boldsymbol{v}_i;\boldsymbol{W}^{(0)})\|_F\\
&\leq\sum_{i=1}^n\|\boldsymbol{v}_i^T(\boldsymbol{B}\boldsymbol{D}_L\boldsymbol{W}_L...\boldsymbol{D}_{l+1}\boldsymbol{W}_{l+1}\boldsymbol{D}_l-\boldsymbol{B}\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}...\boldsymbol{D}_{l+1}^{(0)}\boldsymbol{W}_{l+1}^{(0)})\boldsymbol{D}_l^{(0)}\|\|\boldsymbol{h}_{l-1}\|\\
&\quad+\|\boldsymbol{v}_i^T\boldsymbol{B}\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}...\boldsymbol{D}_{l+1}^{(0)}\boldsymbol{W}_{l+1}^{(0)}\boldsymbol{D}_l^{(0)}\|\|\boldsymbol{h}_{l-1}-\boldsymbol{h}_{l-1}^{(0)}\|.
\end{aligned}\tag{85}$$

By Lemma B.5,

$$\begin{aligned}
&\|(\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{D}_L\boldsymbol{W}_L\cdots\boldsymbol{D}_{l+1}\boldsymbol{W}_{l+1})-(\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}\cdots\boldsymbol{D}_{l+1}^{(0)}\boldsymbol{W}_{l+1}^{(0)})\|\\
&\leq O\left(\omega^{1/3}L^2\frac{\sqrt{m\ln m}}{\sqrt{d}}\right)\|\boldsymbol{v}\|\ \text{ with probability at least } 1-e^{-\Omega(m/L)}.
\end{aligned}\tag{86}$$

By Lemma B.1,

$$\|\boldsymbol{h}^{(0)}\|\leq 1.1\ \text{ with probability at least } 1-e^{-\Omega(m/L)}.\tag{87}$$

By Lemma B.4,

$$\|\boldsymbol{h}_{l-1}-\boldsymbol{h}_{l-1}^{(0)}\|\leq O(\omega L^{5/2}\sqrt{\ln m})\ \text{ with probability at least } 1-e^{-\Omega(m/L)}.\tag{88}$$

We note that the combination of (87), (88) and the bound $\omega<O\Big(\frac{1}{L^{5/2}\sqrt{\ln m}}\Big)$ (which is a weaker bound than the one stated in the lemma) imply

$$\|\boldsymbol{h}\|\leq O(1)\ \text{ with probability at least } 1-e^{-\Omega(m/L)}.\tag{89}$$

By applying (86), (87), (88) and (89) to (85), we conclude that with probability at least $1-e^{-\Omega(m/L)}$

$$\left\|\sum_{i=1}^n\boldsymbol{G}_{i,l}(\boldsymbol{v}_i;\boldsymbol{W})-\boldsymbol{G}_{i,l}(\boldsymbol{v}_i;\boldsymbol{W}^{(0)})\right\|_F^2\leq O\left(\omega^{2/3}L^4\frac{m\ln m}{d}\right)\sum_{i=1}^n\|\boldsymbol{v}_i\|^2.\tag{90}$$

We note that for $l\in[L]$, $i\in[n]$,

$$\nabla_{\boldsymbol{W}_l}\text{loss}(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W})=\boldsymbol{G}_{i,l}(\boldsymbol{e}_i;\boldsymbol{W}).$$

and thus

$$\nabla_{\boldsymbol{W}_l}\mathcal{L}(\boldsymbol{W})=\sum_{i=1}^n\boldsymbol{G}_{i,l}(\boldsymbol{e}_i;\boldsymbol{W}).\tag{91}$$

Therefore, substituting $\boldsymbol{v}_i=\boldsymbol{e}_i$ in (90), the left-hand side of (90) becomes the perturbation of the gradient of the loss function. Since $\omega<O\Big(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m}\Big)$ and $\delta<c_0$,

$$\omega^{2/3}L^4\frac{m\ln m}{d}<O\left(\frac{\delta m}{ndL}\right)<O\Big(\frac{mn}{d}\Big).\tag{92}$$

For the upper bound, by Lemma B.6, (91), (90) and then by (92), with probability at least $1-e^{-\Omega(m/L)}$,

$$\begin{aligned}\|\nabla_{\boldsymbol{W}_l}\mathcal{L}(\boldsymbol{W})\|_F^2&=\left\|\sum_{i=1}^n \boldsymbol{G}_{i,l}(\boldsymbol{e}_i;\boldsymbol{W})\right\|_F^2\\&\le 2\left\|\sum_{i=1}^n \boldsymbol{G}_{i,l}(\boldsymbol{e}_i;\boldsymbol{W}^{(0)})\right\|_F^2+2\left\|\sum_{i=1}^n \boldsymbol{G}_{i,l}(\boldsymbol{e}_i;\boldsymbol{W}^{(0)})-\boldsymbol{G}_{i,l}(\boldsymbol{e}_i;\boldsymbol{W})\right\|_F^2\\&\le\left(O\Big(\frac{mn}{d}\Big)+O\bigg(\omega^{2/3}L^4\frac{m\ln m}{d}\bigg)\right)\sum_{i=1}^n\|\boldsymbol{e}_i\|^2\\&\le O\Big(\frac{mn}{d}\Big)\sum_{i=1}^n\|\boldsymbol{e}_i\|^2\\&=O\Big(\frac{mn}{d}\Big)\mathcal{L}(\boldsymbol{W}).\end{aligned}$$

By definition, we further conclude that

$$\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W})\|^2\le\max_{l\in[L]}\|\nabla_{\boldsymbol{W}_l}\mathcal{L}(\boldsymbol{W})\|_F^2\le O\Big(\frac{mn}{d}\Big)\mathcal{L}(\boldsymbol{W}).$$

For the lower bound, by Lemma B.7, (91), (90) and then by (92), with probability at least $1-e^{-\Omega(m\delta^2)}$,

$$\begin{aligned}\|\nabla_{\boldsymbol{W}_l}\mathcal{L}(\boldsymbol{W})\|_F^2&=\left\|\sum_{i=1}^n \boldsymbol{G}_{i,l}(\boldsymbol{e}_i;\boldsymbol{W})\right\|_F^2\\&\ge\frac{1}{2}\left\|\sum_{i=1}^n \boldsymbol{G}_{i,l}(\boldsymbol{e}_i;\boldsymbol{W}^{(0)})\right\|_F^2-\left\|\sum_{i=1}^n \boldsymbol{G}_{i,l}(\boldsymbol{e}_i;\boldsymbol{W}^{(0)})-\boldsymbol{G}_{i,l}(\boldsymbol{e}_i;\boldsymbol{W})\right\|_F^2\\&\ge\Omega\bigg(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{\delta m}{ndL}-O\bigg(\omega^{2/3}L^4\frac{m\ln m}{d}\bigg)\bigg)\sum_{i=1}^n\|\boldsymbol{e}_i\|^2\\&\ge\Omega\bigg(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{\delta m}{ndL}\bigg)\mathcal{L}(\boldsymbol{W}).\end{aligned}$$

By definition, we conclude that

$$\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W})\|_F^2=\sum_{l\in[L]}\|\nabla_{\boldsymbol{W}_l}\mathcal{L}(\boldsymbol{W})\|_F^2\ge\Omega\bigg(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{\delta m}{nd}\bigg)\mathcal{L}(\boldsymbol{W}).\tag{93}$$

□

B.5 Proof of Lemma 4.1

We prove Lemma 4.1 by adapting the arguments of the proof of Theorem 4 in Allen-Zhu et al. (2019b) to Leaky ReLUs.

Let us first introduce some notation. We let $\boldsymbol{W}^*$ be a vector of matrices satisfying $\|\boldsymbol{W}^*-\boldsymbol{W}^{(0)}\|<\omega$, where we think of $\boldsymbol{W}^*$ as a vector of matrices at an arbitrary training step (we will apply the lemma in this way). We denote a perturbation of $\boldsymbol{W}^*$ by $\boldsymbol{W}'$ and the perturbed matrix by $\boldsymbol{W}:=\boldsymbol{W}^*+\boldsymbol{W}'$. Additional notation corresponding to the original, perturbation and perturbed settings (of $\boldsymbol{W}^*$, $\boldsymbol{W}$ and $\boldsymbol{W}'$, respectively) is summarized as follows:

$$\begin{array}{lll}\boldsymbol{g}^*_{i,l}=\boldsymbol{W}^*_l\boldsymbol{h}^*_{i,l-1}, & \boldsymbol{g}_{i,l}=\boldsymbol{W}_l\boldsymbol{h}_{i,l-1} & \boldsymbol{g}'_{i,l}=\boldsymbol{g}_{i,l}-\boldsymbol{g}^*_{i,l}\\(\boldsymbol{D}_{i,l})^*_{jj}=\dfrac{1_{(\boldsymbol{g}^*_{i,l})_j\ge0}+\alpha1_{(\boldsymbol{g}^*_{i,l})_j<0}}{\sqrt{1+\alpha^2}}, & (\boldsymbol{D}_{i,l})_{jj}=\dfrac{1_{(\boldsymbol{g}_{i,l})_j\ge0}+\alpha1_{(\boldsymbol{g}_{i,l})_j<0}}{\sqrt{1+\alpha^2}}, & \boldsymbol{D}'_{i,l}=\boldsymbol{D}_{i,l}-\boldsymbol{D}^*_{i,l}\\\boldsymbol{h}^*_{i,l}=\tilde\sigma_\alpha(\boldsymbol{W}^*_l\boldsymbol{h}^*_{l-1})\equiv\tilde\sigma_\alpha(\boldsymbol{g}^*_{i,l}), & \boldsymbol{h}_{i,l}=\tilde\sigma_\alpha(\boldsymbol{W}_l\boldsymbol{h}_{l-1})\equiv\tilde\sigma_\alpha(\boldsymbol{g}_{i,l}), & \boldsymbol{h}'_{i,l}=\boldsymbol{h}_{i,l}-\boldsymbol{h}^*_{i,l}\\\boldsymbol{e}^*_{i,l}=\boldsymbol{y}_i-\boldsymbol{B}\boldsymbol{h}^*_{i,L}, & \boldsymbol{e}_{i,l}=\boldsymbol{y}_i-\boldsymbol{B}\boldsymbol{h}^*_{i,L} & \boldsymbol{e}'_{i,l}=\boldsymbol{e}_{i,l}-\boldsymbol{e}^*_{i,l}.\end{array}$$

The loss functions at $\boldsymbol{W}^*$ and $\boldsymbol{W}$ are expressed as

$$\mathcal{L}(\boldsymbol{W}^*)=\frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{e}_i^*\|^2,\ \mathcal{L}(\boldsymbol{W})=\frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{e}_i\|^2. \tag{94}$$

We introduce an auxiliary lemma before proving Lemma 4.1.

Lemma B.8. *There exists a set of diagonal matrices $\boldsymbol{D}''_{i,l}\in[-\sqrt{2},\sqrt{2}]^{m\times m}$ so that*

$$\boldsymbol{h}'_{i,l}=\boldsymbol{h}_{i,l}-\boldsymbol{h}^*_{i,l}=\sum_{a=1}^{l}(\boldsymbol{D}^*_{i,l}+\boldsymbol{D}''_{i,l})\boldsymbol{W}^*_l(\boldsymbol{D}^*_{i,l-1}+\boldsymbol{D}''_{i,l-1})...\boldsymbol{W}^*_{a+1}(\boldsymbol{D}^*_{i,a}+\boldsymbol{D}''_{i,a})\boldsymbol{W}'_a\boldsymbol{h}_{i,a-1}.$$

Furthermore, the following bounds hold

$$\|\boldsymbol{h}'_{i,l}\|\le O(L^{3/2})\|\boldsymbol{W}'\|,\quad \|\boldsymbol{B}\boldsymbol{h}'_{i,L}\|\le O(L\sqrt{m/d})\|\boldsymbol{W}'\|\ \ \textit{and}\ \ \|\boldsymbol{D}''_{i,l}\|_0\le O(m\omega^{2/3}L).$$

The proof of this lemma is identical to the proof of Claim 11.2 in Allen-Zhu et al. (2019b). It is obtained by replacing $|D''_{k,k}|\le 1$ in the second statement of Proposition 11.3 in Allen-Zhu et al. (2019b) with $|D''_{k,k}|\le\sqrt{2}$ in order to fit the setting of Leaky ReLUs.

The rest of this section provides a detailed proof of Lemma 4.1.

Proof of Lemma 4.1. We first express the loss function at $\boldsymbol{W}$ as follows

$$\begin{aligned}
&\text{loss}(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W})\\
&=\frac{1}{2}\|\boldsymbol{B}\boldsymbol{h}_{i,L}-\boldsymbol{y}_i\|^2=\frac{1}{2}\|\boldsymbol{B}(\boldsymbol{h}_{i,L}-\boldsymbol{h}^*_{i,L})+\boldsymbol{B}\boldsymbol{h}^*_{i,L}-\boldsymbol{y}_i\|^2\\
&=\frac{1}{2}\|\boldsymbol{e}_i^*+\boldsymbol{B}(\boldsymbol{h}_{i,L}-\boldsymbol{h}^*_{i,L})\|^2=\frac{1}{2}\|\boldsymbol{e}_i^*\|^2+\frac{1}{2}\|\boldsymbol{B}(\boldsymbol{h}_{i,L}-\boldsymbol{h}^*_{i,L})\|^2+\langle\boldsymbol{e}_i^*,\boldsymbol{B}(\boldsymbol{h}_{i,L}-\boldsymbol{h}^*_{i,L})\rangle\\
&=\text{loss}_i^*+\frac{1}{2}\|\boldsymbol{B}(\boldsymbol{h}_{i,L}-\boldsymbol{h}^*_{i,L})\|^2+\boldsymbol{e}_i^{*T}\boldsymbol{B}(\boldsymbol{h}_{i,L}-\boldsymbol{h}^*_{i,L})=\text{loss}_i^*+\frac{1}{2}\|\boldsymbol{B}(\boldsymbol{h}_{i,L}-\boldsymbol{h}^*_{i,L})\|^2+\boldsymbol{e}_i^{*T}\boldsymbol{B}\boldsymbol{h}'_{i,L}.
\end{aligned}\tag{95}$$

Then we expand $\langle\nabla\mathcal{L}(\boldsymbol{W}^*),\boldsymbol{W}'\rangle$ as

$$\begin{aligned}
&\langle\nabla\mathcal{L}(\boldsymbol{W}^*),\boldsymbol{W}'\rangle\\
&=\sum_{l=1}^{L}\langle\nabla_{\boldsymbol{W}_l}\mathcal{L}(\boldsymbol{W}^*),\boldsymbol{W}'_l\rangle=\sum_{l=1}^{L}\sum_{i=1}^{n}\langle\boldsymbol{D}^*_{i,l}\boldsymbol{W}^{*T}_{l+1}\boldsymbol{D}^*_{i,l+1}...\boldsymbol{D}^*_{i,L}\boldsymbol{B}^T\boldsymbol{e}_i^*\boldsymbol{h}^{*T}_{l-1}(\boldsymbol{x}_i),\boldsymbol{W}'_l\rangle\\
&=\sum_{l=1}^{L}\sum_{i=1}^{n}\langle\boldsymbol{D}^*_{i,l}\boldsymbol{W}^{*T}_{l+1}\boldsymbol{D}^*_{i,l+1}...\boldsymbol{D}^*_{i,L}\boldsymbol{B}^T\boldsymbol{e}_i^*\boldsymbol{h}^{*T}_{l-1}(\boldsymbol{x}_i),\boldsymbol{W}'_l\rangle\\
&=\sum_{l=1}^{L}\sum_{i=1}^{n}\boldsymbol{e}_i^{*T}\boldsymbol{B}\boldsymbol{D}^*_{i,L}\boldsymbol{W}^*_L...\boldsymbol{D}^*_{i,l+1}\boldsymbol{W}^*_{l+1}\boldsymbol{D}^*_{i,l}\boldsymbol{W}'_l\boldsymbol{h}^*_{l-1}(\boldsymbol{x}_i).
\end{aligned}$$

The above two equations imply the following estimate

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{W}^*+\boldsymbol{W}')-\mathcal{L}(\boldsymbol{W}^*)-\langle\nabla\mathcal{L}(\boldsymbol{W}^*),\boldsymbol{W}'\rangle\\
&=-\langle\nabla\mathcal{L}(\boldsymbol{W}^*),\boldsymbol{W}'\rangle+\sum_{i=1}^{n}(\mathrm{loss}_i-\mathrm{loss}_i^*)\\
&=-\sum_{l=1}^{L}\sum_{i=1}^{n}\boldsymbol{e}_i^{*T}\boldsymbol{B}\boldsymbol{D}_{i,L}^*\boldsymbol{W}_L^*...\boldsymbol{D}_{i,l+1}^*\boldsymbol{W}_{l+1}^*\boldsymbol{D}_{i,l}^*\boldsymbol{W}_l'\boldsymbol{h}_{l-1}^*(\boldsymbol{x}_i)\\
&\quad+\sum_{i=1}^{n}\big(\frac{1}{2}\|\boldsymbol{B}(\boldsymbol{h}_{i,L}-\boldsymbol{h}_{i,L}^*)\|^2+\boldsymbol{e}_i^{*T}\boldsymbol{B}(\boldsymbol{h}_{i,L}-\boldsymbol{h}_{i,L}^*)\big)\\
&=\sum_{i=1}^{n}e_i^{*T}\boldsymbol{B}\Big((\boldsymbol{h}_{i,L}-\boldsymbol{h}_{i,L}^*)-\sum_{l=1}^{L}\boldsymbol{D}_{i,L}^*\boldsymbol{W}_L^*...\boldsymbol{D}_{i,l+1}^*\boldsymbol{W}_{l+1}^*\boldsymbol{D}_{i,l}^*\boldsymbol{W}_l'\boldsymbol{h}_{l-1}^*(\boldsymbol{x}_i)\Big) \qquad (96)\\
&\quad+\frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{B}(\boldsymbol{h}_{i,L}-\boldsymbol{h}_{i,L}^*)\|^2. \qquad (97)
\end{aligned}$$

Lemma B.8 provides the following upper bound for (97)

$$\frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{B}(\boldsymbol{h}_{i,L}-\boldsymbol{h}_{i,L}^*)\|^2\le O(nL^2m/d)\|\boldsymbol{W}'\|^2. \qquad (98)$$

We note that (96) can be differently expressed by using Lemma B.8 to replace $\boldsymbol{h}-\boldsymbol{h}^*$ with some diagonal matrices, $\boldsymbol{D}_{i,l}''$, and by adding and subtracting the term $\sum_{l=1}^{L}\boldsymbol{D}_{i,L}^*\boldsymbol{W}_L^*...\boldsymbol{D}_{i,l+1}^*\boldsymbol{W}_{l+1}^*\boldsymbol{D}_{i,l}^*\boldsymbol{W}_l'\boldsymbol{h}_{i,l}$ as follows

$$\begin{aligned}
&\boldsymbol{e}_i^{*T}\boldsymbol{B}\Big((\boldsymbol{h}_{i,L}-\boldsymbol{h}_{i,L}^*)-\sum_{l=1}^{L}\boldsymbol{D}_{i,L}^*\boldsymbol{W}_L^*...\boldsymbol{D}_{i,l+1}^*\boldsymbol{W}_{l+1}^*\boldsymbol{D}_{i,l}^*\boldsymbol{W}_l'\boldsymbol{h}_{l-1}^*(\boldsymbol{x}_i)\Big)\\
&=\boldsymbol{e}_i^{*T}\boldsymbol{B}\Big(\sum_{l=1}^{L}(\boldsymbol{D}_{i,L}^*+\boldsymbol{D}_{i,L}'')\boldsymbol{W}_L^*...\boldsymbol{W}_{l+1}^*(\boldsymbol{D}_{i,l}^*+\boldsymbol{D}_{i,l}'')\boldsymbol{W}_l'\boldsymbol{h}_{i,l-1}\\
&\quad-\sum_{l=1}^{L}\boldsymbol{D}_{i,L}^*\boldsymbol{W}_L^*...\boldsymbol{D}_{i,l+1}^*\boldsymbol{W}_{l+1}^*\boldsymbol{D}_{i,l}^*\boldsymbol{W}_l'\boldsymbol{h}_{l-1}^*(\boldsymbol{x}_i)\Big)\\
&=\boldsymbol{e}_i^{*T}\boldsymbol{B}\Big(\sum_{l=1}^{L}\big((\boldsymbol{D}_{i,L}^*+\boldsymbol{D}_{i,L}'')\boldsymbol{W}_L^*...\boldsymbol{W}_{l+1}^*(\boldsymbol{D}_{i,l}^*+\boldsymbol{D}_{i,l}'')\boldsymbol{W}_l'-\boldsymbol{D}_{i,L}^*\boldsymbol{W}_L^*...\boldsymbol{W}_{l+1}^*\boldsymbol{D}_{i,l}^*\boldsymbol{W}_l'\big)\boldsymbol{h}_{i,l-1} \qquad (99)\\
&\quad-\sum_{l=1}^{L}\boldsymbol{D}_{i,L}^*\boldsymbol{W}_L^*...\boldsymbol{D}_{i,l+1}^*\boldsymbol{W}_{l+1}^*\boldsymbol{D}_{i,l}^*\boldsymbol{W}_l'\big(\boldsymbol{h}_{i,l-1}-\boldsymbol{h}_{l-1}^*(\boldsymbol{x}_i)\big)\Big). \qquad (100)
\end{aligned}$$

Next, we upper bound (99) and (100). In order to bound (99), we first use Lemma B.5 to obtain the following bound

$$\begin{aligned}
&\|\boldsymbol{B}(\boldsymbol{D}_{i,L}^*+\boldsymbol{D}_{i,L}'')\boldsymbol{W}_L^*...\boldsymbol{W}_{l+1}^*(\boldsymbol{D}_{i,l}^*+\boldsymbol{D}_{i,l}'')\boldsymbol{W}_l'-\boldsymbol{B}\boldsymbol{D}_{i,L}^*\boldsymbol{W}_L^*...\boldsymbol{W}_{l+1}^*\boldsymbol{D}_{i,l}^*\boldsymbol{W}_l'\|\\
&\le O\left(\frac{1-\alpha}{\sqrt{1+\alpha^2}}\frac{\omega^{1/3}L^2\sqrt{m\ln m}}{\sqrt{d}}\right)\|\boldsymbol{W}_l'\|. \qquad (101)
\end{aligned}$$

Using (94), we note that $(\sum_{i=1}^{n}\|\boldsymbol{e}_i^*\|)^2\le n\sum_{i=1}^{n}\|\boldsymbol{e}_i^*\|^2=n\mathcal{L}(\boldsymbol{W}^*)$. Combining this fact and (101) yields the following bound for (99):

$$\begin{aligned}
&\sum_{i=1}^{n}\boldsymbol{e}_i^{*T}\boldsymbol{B}\Big(\sum_{l=1}^{L}\big((\boldsymbol{D}_{i,L}^*+\boldsymbol{D}_{i,L}'')\boldsymbol{W}_L^*...\boldsymbol{W}_{l+1}^*(\boldsymbol{D}_{i,l}^*+\boldsymbol{D}_{i,l}'')\boldsymbol{W}_l'-\boldsymbol{D}_{i,L}^*\boldsymbol{W}_L^*...\boldsymbol{W}_{l+1}^*\boldsymbol{D}_{i,l}^*\boldsymbol{W}_l'\big)\boldsymbol{h}_{i,l-1}\Big)\\
&\le\sqrt{n\mathcal{L}(\boldsymbol{W}^*)}O\left(\frac{1-\alpha}{\sqrt{1+\alpha^2}}\frac{\omega^{1/3}L^2\sqrt{m\ln m}}{\sqrt{d}}\right)\|\boldsymbol{W}'\|. \qquad (102)
\end{aligned}$$

In order to bound (100), we apply Lemma B.3 and Lemma B.5 to obtain

$$\begin{aligned}
&\|\boldsymbol{B}\boldsymbol{D}^*{}_{i,L}\boldsymbol{W}^*{}_L...\boldsymbol{D}^*{}_{i,l+1}\boldsymbol{W}^*{}_{l+1}\boldsymbol{D}^*{}_{i,l}\| \\
&\leq \|\boldsymbol{B}\boldsymbol{D}^0{}_{i,L}\boldsymbol{W}^0{}_L...\boldsymbol{D}^0{}_{i,l+1}\boldsymbol{W}^0{}_{l+1}\boldsymbol{D}^0{}_{i,l}\| \\
&\quad + \|\boldsymbol{B}\boldsymbol{D}^0{}_{i,L}\boldsymbol{W}^0{}_L...\boldsymbol{D}^0{}_{i,l+1}\boldsymbol{W}^0{}_{l+1}\boldsymbol{D}^0{}_{i,l} - \boldsymbol{B}\boldsymbol{D}^*{}_{i,L}\boldsymbol{W}^*{}_L...\boldsymbol{D}^*{}_{i,l+1}\boldsymbol{W}^*{}_{l+1}\boldsymbol{D}^*{}_{i,l}\| \\
&\leq O(\sqrt{Lm/d}) + O\left(\frac{1-\alpha}{\sqrt{1+\alpha^2}}\frac{\omega^{1/3}L^2\sqrt{m\ln m}}{\sqrt{d}}\right). \qquad (103)
\end{aligned}$$

Lemma B.8 implies that $\|\boldsymbol{h}-\boldsymbol{h}^*\| = \|\boldsymbol{h}'\| \leq O(L^{3/2}\|\boldsymbol{W}'\|)$. Combining this observation and (103) results in

$$\left\|\sum_{i=1}^n \boldsymbol{e}_i^{*T}\sum_{l=1}^L \boldsymbol{B}\boldsymbol{D}^*_{i,L}\boldsymbol{W}^*_L...\boldsymbol{D}^*_{i,l+1}\boldsymbol{W}^*_{l+1}\boldsymbol{D}^*_{i,l}\boldsymbol{W}'_l\boldsymbol{h}'_{i,l-1}\right\| \leq \sum_{i=1}^n \|\boldsymbol{e}_i^*\| O(L^2\sqrt{m/d})\|\boldsymbol{W}'\|^2. \qquad (104)$$

In order to bound $\|\boldsymbol{e}_i^*\|$, we first note that at initialization

$$\|\boldsymbol{e}_i^{(0)}\| = \|\boldsymbol{B}\boldsymbol{h}_{L,i}^{(0)} - \boldsymbol{y}_i\| \leq \|\boldsymbol{y}_i\| + \|\boldsymbol{B}\boldsymbol{h}_{L,i}^{(0)}\|,$$

where

$$\boldsymbol{B}\boldsymbol{h}_{L,i}^{(0)} \sim N\left(0, \frac{\|\boldsymbol{h}\|^2}{d}\boldsymbol{I}_d\right).$$

For this Gaussian distribution and $d < O(1)$,

$$\mathbb{P}\left(\|(\boldsymbol{B}\boldsymbol{h}_{L,i}^{(0)})\|^2 > \frac{\sqrt{m}}{\sqrt{d}}\right) \leq e^{-\Omega(\frac{m}{d})} = e^{-\Omega(m)}.$$

Therefore, with probability at least $1-e^{-\Omega(m)}$,

$$\|\boldsymbol{e}_i^{(0)}\| \leq O\left(\frac{\sqrt{m}}{\sqrt{d}}\right).$$

For general $\boldsymbol{e}_i^*$, $\|\boldsymbol{e}_i^*\| = \left\|\boldsymbol{B}\left(\boldsymbol{h}_{i,L}^{(0)} + (\boldsymbol{h}_{i,L}^* - \boldsymbol{h}_{i,L}^{(0)})\right) - \boldsymbol{y}_i\right\|$. Lemma B.8 implies that if $\omega \leq O(1/L)$

$$\|\boldsymbol{e}_i^*\| \leq \|\boldsymbol{e}_i^{(0)}\| + \|\boldsymbol{B}(\boldsymbol{h}_{i,L}^* - \boldsymbol{h}_{i,L}^{(0)})\| \leq O\left(\frac{\sqrt{m}}{\sqrt{d}}\right). \qquad (105)$$

The combination of (105) and (104) results in the following bound on the term specified in (100)

$$\left\|\sum_{i=1}^n \boldsymbol{e}_i^{*T}\sum_{l=1}^L \boldsymbol{B}\boldsymbol{D}^*_{i,L}\boldsymbol{W}^*_L...\boldsymbol{D}^*_{i,l+1}\boldsymbol{W}^*_{l+1}\boldsymbol{D}^*_{i,l}\boldsymbol{W}'_l\boldsymbol{h}'_{i,l-1}\right\| \leq O\left(\frac{nL^2m}{d}\right)\|\boldsymbol{W}'\|^2 \qquad (106)$$

Combining the bounds in (98), (102) and (106) we bound the terms in (96) and (97) with the above specified probability. We thus conclude the desired result, that is, if $\boldsymbol{W}^*$ is such that $\|\boldsymbol{W}^* - \boldsymbol{W}^{(0)}\| < \omega$, then with probability at least $1-e^{-\Omega(m)}$

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{W}^*+\boldsymbol{W}') - \mathcal{L}(\boldsymbol{W}^*) - \langle\nabla\mathcal{L}(\boldsymbol{W}^*), \boldsymbol{W}'\rangle \\
&\leq \sqrt{n\mathcal{L}(\boldsymbol{W}^*)}O\left(\frac{1-\alpha}{\sqrt{1+\alpha^2}}\frac{\omega^{1/3}L^2\sqrt{m\ln m}}{\sqrt{d}}\right)\|\boldsymbol{W}'\| + O(nL^2m/d)\|\boldsymbol{W}'\|^2.
\end{aligned}$$

□

B.6 Conclusion of the Proof of Theorem 3.1

Most of the proof of Theorem 3.1 was given in §4. The only part that remains unverified is to show that during training

$$\|\boldsymbol{W}^t-\boldsymbol{W}^{(0)}\|<\omega<O\bigg(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m}\bigg).$$

For this purpose, we establish Lemma B.9 below.

Lemma B.9. *Assume the setup of §2, where the learning rate satisfies* $\eta<\frac{\delta^{3/2}d^{1/2}}{n^3L^{15/2}m^{1/2}\ln^2 m}$ *and the width* m *of the neural network satisfies* $\frac{m}{\ln^4 m}>\Omega\Big(\frac{1+\alpha^2}{(1-\alpha)^2}\frac{n^5L^{15}d}{\delta^4}\Big)$. *Then in the training stage described by Algorithm 2*

$$\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\|<O\bigg(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m}\bigg)\quad \textit{with probability at least}\quad 1-e^{-\Omega(\ln m)}. \tag{107}$$

Proof. We first establish the bound

$$\mathcal{L}(\boldsymbol{W}^{(0)})<O(n\ln^{1/2}m)\ \text{ with probability at least } 1-e^{-\Omega(\ln m)}. \tag{108}$$

We note that $\boldsymbol{B}\boldsymbol{h}_{i,L}\sim N\Big(0,\frac{\|\boldsymbol{h}_{i,L}\|^2}{d}\Big)$ and thus $\frac{d}{\|\boldsymbol{h}_{i,L}\|^2}\|\boldsymbol{B}\boldsymbol{h}_{i,L}\|^2|\boldsymbol{h}_{i,L}\sim\chi^2(d)$. Applying this observation and Lemma B.1 (i.e., $\|\boldsymbol{h}_i\|\in[0.5,1.5]$, with probability at least $1-e^{-\Omega(m/L)}$) yields

$$\mathbb{P}\bigg(\frac{d}{\|\boldsymbol{h}_{i,L}\|^2}\|\boldsymbol{B}\boldsymbol{h}_{i,L}\|^2>(1+\epsilon)d\bigg)<e^{-\Omega(d\epsilon^2)}.$$

Choosing $\epsilon=\sqrt{\ln m}$ and applying a union bound over $i\in[n]$ (but noting that since $m>\Omega(n)$ the probability $1-ne^{-\Omega(d\ln m)}$ is of the same order as $1-e^{-\Omega(d\ln m)}$), we obtain the bound

$$\|\boldsymbol{B}\boldsymbol{h}_{i,L}\|^2\le O(\sqrt{\ln m})\ \text{ with probability at least } 1-e^{-\Omega(d\ln m)}. \tag{109}$$

Therefore, we conclude (108) as follows:

$$\mathcal{L}(\boldsymbol{W}^{(0)})=\sum_{i=1}^n\|\boldsymbol{y}_i-\boldsymbol{B}\boldsymbol{h}_{i,L}\|^2\le n(O(1)+O(\sqrt{\ln m}))=O(n\sqrt{\ln m}).$$

Next, we prove (107) by induction on $t=1,\cdots$. It is trivial that the statement holds for $t=0$.

To prove the induction step we follow ideas that were introduced in the proof of Lemma 4.1 in Zou and Gu (2019). Using the induction assumption, we can apply (16) and then (14) (indeed, the conditions for these bounds are guaranteed by the induction assumption) and consequently obtain

$$\begin{aligned}\sqrt{\mathcal{L}(\boldsymbol{W}^{(s)})}-\sqrt{\mathcal{L}(\boldsymbol{W}^{(s+1)})}&=\frac{\mathcal{L}(\boldsymbol{W}^{(s)})-\mathcal{L}(\boldsymbol{W}^{(s+1)})}{\sqrt{\mathcal{L}(\boldsymbol{W}^{(s)})}+\sqrt{\mathcal{L}(\boldsymbol{W}^{(s+1)})}}\ge\Omega(1)\frac{\eta\|\nabla\mathcal{L}(\boldsymbol{W}^{(s)})\|_F^2}{\sqrt{\mathcal{L}(\boldsymbol{W}^{(s)})}}\\&\ge\frac{(1-\alpha)}{\sqrt{1+\alpha^2}}\Omega\bigg(\sqrt{\frac{\delta m}{nd}}\bigg)\eta\|\nabla\mathcal{L}(\boldsymbol{W}^{(s)})\|_F,\end{aligned}$$

or equivalently,

$$\eta\|\nabla\mathcal{L}(\boldsymbol{W}^{(s)})\|_F\le\frac{\sqrt{1+\alpha^2}}{(1-\alpha)}\Omega\bigg(\sqrt{\frac{nd}{\delta m}}\bigg)\Big(\sqrt{\mathcal{L}(\boldsymbol{W}^{(s)})}-\sqrt{\mathcal{L}(\boldsymbol{W}^{(s+1)})}\Big). \tag{110}$$

Combining the training procedure with (110) yields

$$\begin{aligned}\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\|&\le\eta\sum_{s=0}^{t-1}\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{(s)})\|\\&\le\frac{\sqrt{1+\alpha^2}}{(1-\alpha)}\Omega\bigg(\sqrt{\frac{nd}{\delta m}}\bigg)\Big(\sqrt{\mathcal{L}(\boldsymbol{W}^{(0)})}-\sqrt{\mathcal{L}(\boldsymbol{W}^{(t)})}\Big)\\&\le\frac{\sqrt{1+\alpha^2}}{(1-\alpha)}\Omega\bigg(\sqrt{\frac{nd}{\delta m}}\bigg)\sqrt{\mathcal{L}(\boldsymbol{W}^{(0)})}.\end{aligned} \tag{111}$$

Applying (108) to the bound above we conclude that when $\frac{m}{\ln^4 m} > \frac{1+\alpha^2}{(1-\alpha)^2}\Omega(n^5L^{15}d/\delta^4)$,

$$\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\| \leq O\left(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m}\right) \text{ with probability at least } 1-e^{-\Omega(\ln m)}.$$

□

B.7 Proof of Theorem 3.2

Throughout this proof we assume that $\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\| \leq O(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m})$ during training, which is a sufficient condition for some of the propositions used, such as for Lemma 4.1. After finalizing the proof under this assumption, we establish Lemma B.10 that guarantees this assumption.

Applying Lemma 4.1 and taking expectations yield

$$\begin{aligned}\mathbb{E}\mathcal{L}(\boldsymbol{W}^{(t+1)}) &= \mathbb{E}\mathcal{L}(\boldsymbol{W}^{(t)}-\eta\nabla_{\boldsymbol{W}}\mathcal{L}_B(\boldsymbol{W}^{(t)})) \\ &\leq \mathbb{E}\mathcal{L}(\boldsymbol{W}^{(t)})-\mathbb{E}\eta\langle\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{(t)}),\nabla_{\boldsymbol{W}}\mathcal{L}_B(\boldsymbol{W}^{(t)})\rangle \\ &\quad+\frac{\eta(1-\alpha)\omega^{\frac{1}{3}}L^2\sqrt{mn\mathcal{L}(\boldsymbol{W}^{(t)})\ln m}}{\sqrt{d(1+\alpha^2)}}\mathbb{E}O\Big(\|\nabla_{\boldsymbol{W}}\mathcal{L}_B(\boldsymbol{W}^{(t)})\|\Big) \\ &\quad+\frac{\eta^2nL^2m}{d}\mathbb{E}O\Big(\|\nabla_{\boldsymbol{W}}\mathcal{L}_B(\boldsymbol{W}^{(t)})\|^2\Big).\end{aligned} \tag{112}$$

Applying the following basic observations:

$$\mathbb{E}\langle\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{(t)}),\nabla_{\boldsymbol{W}}\mathcal{L}_B(\boldsymbol{W}^{(t)})\rangle = \frac{b}{n}\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{(t)})\|_F^2,$$

$$\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{(t)})\| = \max_{l\in[L]}\|\nabla_{\boldsymbol{W}_l}\mathcal{L}(\boldsymbol{W}^{(t)})\| \leq \max_{l\in[L]}\|\nabla_{\boldsymbol{W}_l}\mathcal{L}(\boldsymbol{W}^{(t)})\|_F \leq \|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{(t)})\|_F,$$

while selecting $\omega < \frac{\delta^{3/2}}{n^3L^6\ln^{3/2}m}$ and $\eta < \frac{d}{bL^2m}$, to (112) results in

$$\mathbb{E}\mathcal{L}(\boldsymbol{W}^{(t+1)}) \leq \mathcal{L}(\boldsymbol{W}^{(t)})-\frac{\eta b}{n}\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{(t)})\|_F^2 \leq \left(1-\Omega\left(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{\eta\delta mb}{n^2d}\right)\right)\mathcal{L}(\boldsymbol{W}^{(t)}). \tag{113}$$

For simplicity, we define

$$\gamma := \left(1-\Omega\left(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{\eta\delta mb}{n^2d}\right)\right),$$

and (113) becomes

$$\mathbb{E}\mathcal{L}(\boldsymbol{W}^{(t+1)}) \leq \gamma\mathcal{L}(\boldsymbol{W}^{(t)}). \tag{114}$$

Next, we establish a bound for $\mathcal{L}(\boldsymbol{W}^{(t+1)})$ without expectation. We note that (9) implies

$$\|\nabla_{\boldsymbol{W}_l}\mathcal{L}_B(\boldsymbol{W}^{(t)})\|_F^2 \leq (bm/d)\mathcal{L}(\boldsymbol{W}^{(t)}),$$

and consequently

$$\|\nabla_{\boldsymbol{W}}\mathcal{L}_B(\boldsymbol{W}^{(t)})\|_F^2 \leq \frac{bmL}{d}\mathcal{L}(\boldsymbol{W}^{(t)}) \text{ and } \|\nabla_{\boldsymbol{W}}\mathcal{L}_B(\boldsymbol{W}^{(t)})\|^2 \leq \frac{bm}{d}\mathcal{L}(\boldsymbol{W}^{(t)}). \tag{115}$$

The application of Lemma 4.1, (115) and our choice of η results in

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}^{(t+1)}) &\leq \mathcal{L}(\boldsymbol{W}^{(t)})+\eta\|\nabla_{\boldsymbol{W}}\mathcal{L}_B(\boldsymbol{W}^{(t)})\|_F\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{(t)})\|_F+\eta\frac{b^2mn}{d}\mathcal{L}(\boldsymbol{W}^{(t)}) \\ &\leq \left(1+O\left(\frac{\eta mL\sqrt{nb}}{d}\right)\right)\mathcal{L}(\boldsymbol{W}^{(t)}).\end{aligned} \tag{116}$$

For simplicity, we define $\beta := 1+O(\eta mL\sqrt{nb}/d)$, and (116) becomes

$$\mathcal{L}(\boldsymbol{W}^{(t+1)}) \leq \beta\mathcal{L}(\boldsymbol{W}^{(t)}). \tag{117}$$

We denote

$$\mathcal{L}^t := \mathcal{L}(\boldsymbol{W}^{(t)})$$

and define the filtration

$$\mathcal{F}_t := \sigma(\boldsymbol{W}^{(0)},..\boldsymbol{W}^{(t)}).$$

We further define

$$Y_t := \ln\mathcal{L}^t - \ln\mathcal{L}^{t-1} - \mathbb{E}(\ln\mathcal{L}^t - \ln\mathcal{L}^{t-1}|\mathcal{F}_{t-1})$$

and

$$X_t := \sum_{s=1}^{t} Y_s.$$

We note that $\{X_t\}$ is a martingale.

We will use Azuma's inequality to bound X_t. We thus need to show that $\{X_t\}$ is $\boldsymbol{c}-$Lipschitz (i.e., $|Y_t| \le c_t$,). We verify the $\boldsymbol{c}-$Lipschitz property by applying the definition of Y_t, (117) and (114) as follows:

$$\begin{aligned} |Y_{t+1}| &= |\ln\mathcal{L}^{(t+1)} - \ln\mathcal{L}^{(t)} - \mathbb{E}\ln\mathcal{L}^{(t+1)} - \ln\mathcal{L}^{(t)}|\mathcal{F}_t| \\ &\le \ln\beta - \ln\gamma = \ln\frac{\beta}{\gamma}. \end{aligned}$$

Then by Azuma's inequality,

$$\mathbb{P}(|X_t - EX_t| \ge \lambda) \le 2\exp\left(-\frac{\lambda^2}{2t\ln^2\beta/\gamma}\right). \tag{118}$$

Choosing $\lambda = \sqrt{t}\ln(\beta/\gamma)\ln m$ in (118) yields

$$|X_t| \le \sqrt{t}\ln(\beta/\gamma)\ln m \;\text{ with probability at least } 1 - e^{-\Omega(\ln^2 m)}.. \tag{119}$$

Applying the definition of Y_t and (114) results in

$$\ln\mathcal{L}^t = X_t + \ln\mathcal{L}^{(0)} + \sum_{s=1}^{t}\mathbb{E}(Y_s - Y_{s-1}|\mathcal{F}_{s-1}) \le X_t + \ln\mathcal{L}^{(0)} + t\ln\gamma.$$

We further apply the above observation and (119) to conclude that with probability at least $1 - e^{-\Omega(\ln^2 m)}$

$$\begin{aligned} \ln\mathcal{L}^{(t)} &\le \ln\mathcal{L}^{(0)} + t\ln\gamma + \sqrt{t}\ln\left(\frac{\beta}{\gamma}\right)\ln m \\ &\le \ln\mathcal{L}^{(0)} + \frac{\ln^2\left(\frac{\beta}{\gamma}\right)\ln^2 m}{4|\ln\gamma|} - \left(\sqrt{|\ln\gamma|}\sqrt{t} - \frac{\ln\frac{\beta}{\gamma}\ln m}{2\sqrt{|\ln\gamma|}}\right)^2. \end{aligned}$$

We note that for $f(x) = (ax+b)^2$ and $x > 4b/a$, $f(x) \ge \frac{1}{2}a^2x^2$. Using this fact, we conclude that when $\sqrt{t} > \frac{2\ln\frac{\beta}{\gamma}\ln m}{|\ln\gamma|}$, or equivalently, when $t > \frac{4\ln^2\frac{\beta}{\gamma}\ln^2 m}{\ln^2\gamma}$,

$$\ln\mathcal{L}^{(t)} \le \ln\mathcal{L}^{(0)} + \frac{\ln^2\frac{\beta}{\gamma}\ln^2 m}{4|\ln\gamma|} + t\times 1_{\left\{t > \frac{4\ln^2\frac{\beta}{\gamma}\ln^2 m}{\ln^2\gamma}\right\}}\ln\gamma \;\text{ with probability } \ge 1 - e^{-\Omega(\ln^2 m)}. \tag{120}$$

This implies that when $t > \frac{4\ln^2\frac{\beta}{\gamma}\ln^2 m}{\ln^2\gamma}$ we achieve linear convergence with a convergence rate of γ. By our choice of η, the additional term in (120) is bounded as follows

$$: \frac{\ln^2(\beta/\gamma)\ln^2 m}{|\ln\gamma|} \le O\left(\frac{(\beta-1)^2}{1-\gamma}\ln^2 m\right) = O\left(\eta\frac{mn^3L^2\ln^2 m}{d\delta}\right) < O(1).$$

The above lower bound of t that guarantees linear convergence can be further simplified. Since $\frac{x}{1+x}\leq\ln(1+x)\leq x$, we note $t\geq\frac{((\beta-1)^2+\frac{(\gamma-1)^2}{\gamma^2})\gamma^2\ln^2 m}{(\gamma-1)^2}$ and thus

$$t\geq\ln^2 m\times\left(1+\frac{(\beta-1)^2}{(\gamma-1)^2}\right).$$

Recalling the expressions for β and γ, we conclude that linear convergence is achieved when

$$t>\Omega\left(\frac{(1+\alpha^2)^2}{(1-\alpha)^4}\frac{n^5L^2}{\delta^2 b}\ln^2 m\right). \tag{121}$$

The above argument holds for one training step with probability at least $1-e^{-\Omega(m)}$. It extends to T steps with probability at least $1-Te^{-\Omega(m)}$. We note that the number of epochs T can be bounded using the bound ϵ on the training error, the convergence rate in (6), and (108), as follows:

$$T=\ln(\epsilon/C_0\mathcal{L}(\boldsymbol{W}^{(0)}))/\ln\gamma<\Theta(\ln(\epsilon/C_0 n\sqrt{\ln m})/\ln\gamma)\leq O\left(\frac{\eta b\delta m}{n^2 d}(\ln\epsilon^{-1}+\ln(C_0 n\sqrt{\ln m}))\right).$$

Therefore, the probability that ensures T-steps training with training error lower than ϵ is at least

$$1-O\left(\frac{\eta b\delta m}{n^2 d}(\ln\epsilon^{-1}+\ln(C_0 n\sqrt{\ln m}))\right)e^{-\Omega(m)}.$$

Because $m>\Omega(\text{poly}(n,L,d,\delta^{-1},b))$ and also $m>\Omega(\ln\ln\epsilon^{-1})$, this probability is of order $1-e^{-\Omega(m)}$.

Lemma B.10. *Assume the setup of §2 with learning rate $\eta<\frac{\delta^{3/2}d^{1/2}}{b^{1/2}n^3L^{15/2}m^{1/2}\ln^2 m}$ and neural network width m satisfying $\frac{m}{\ln^4 m}>\frac{(1+\alpha^2)^4}{(1-\alpha)^8}\Omega\left(\frac{n^8L^{15}d}{\delta^5 b}\right)$. Then during training according to Algorithm 3,*

$$\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\|<O\left(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m}\right)\quad\textit{with probability at least } 1-e^{-\Omega(\ln m)}.$$

Proof. The proof is similar to Lemma B.9, but with different bounds in this SGD setting. We first show bound the perturbation at initialization as follows:

$$\|\boldsymbol{W}^{(1)}-\boldsymbol{W}^{(0)}\|=\eta\|\nabla_{\boldsymbol{W}}\mathcal{L}_B(\boldsymbol{W}^{(0)})\|\leq\eta O\left(\sqrt{\frac{mnb}{d}}\right)\mathcal{L}(\boldsymbol{W}^{(0)})<O\left(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m}\right).$$

We denote $T_0:=\Omega\left(\frac{(1+\alpha^2)^2}{(1-\alpha)^4}\frac{n^5L^2}{\delta^2 b}\ln^2 m\right)$. Combining the SGD update step, (115), (120), (121) and our choice of η yields

$$\begin{aligned}
\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\|&\leq\eta\sum_{s=0}^{t-1}\|\nabla_{\boldsymbol{W}}\mathcal{L}_B(\boldsymbol{W}^{(s)})\|\leq\eta\sum_{s=0}^{t-1}\sqrt{\frac{mb}{d}}\sqrt{\mathcal{L}(\boldsymbol{W}^{(s)})}\\
&\leq\sqrt{\frac{mb}{d}}\eta\left(\frac{1}{1-\sqrt{\gamma}}+T_0\right)\sqrt{\mathcal{L}(\boldsymbol{W}^{(0)})}\\
&\leq O(1)\sqrt{\frac{mb}{d}}\eta\left(\frac{1+\alpha^2}{(1-\alpha)^2}\frac{n^2L\eta}{\delta mb}+\Omega\left(\frac{(1+\alpha^2)^2}{(1-\alpha)^4}\frac{n^5L^2}{\delta^2 b}\ln^2 m\right)\right)\sqrt{\mathcal{L}(\boldsymbol{W}^{(0)})}\\
&\leq O(1)\sqrt{\frac{mb}{d}}\eta\Omega\left(\frac{(1+\alpha^2)^2}{(1-\alpha)^4}\frac{n^5L^2}{\delta^2 b}\ln^2 m\right)\sqrt{\mathcal{L}(\boldsymbol{W}^{(0)})}\\
&\leq O(1)\sqrt{\frac{mb}{d}}\frac{d\delta}{mn^3L^2\ln^2 m}\Omega\left(\frac{(1+\alpha^2)^2}{(1-\alpha)^4}\frac{n^5L^2}{\delta^2 b}\ln^2 m\right)\sqrt{\mathcal{L}(\boldsymbol{W}^{(0)})}\\
&=O(1)\frac{(1+\alpha^2)^2}{(1-\alpha)^4}\frac{\sqrt{d}n^2}{\sqrt{mb}\delta}\sqrt{\mathcal{L}(\boldsymbol{W}^{(0)})}.
\end{aligned}$$

It is thus clear that when $\frac{m}{\ln^4 m}>\left(\frac{1+\alpha^2}{(1-\alpha)^2}\right)^4\Omega\left(\frac{n^8L^{15}d}{\delta^5 b}\right)$, $\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\|\leq O\left(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m}\right)$. □

B.8 Proof of Lemma 4.3

To simplify the proof, we study the generalization error of each output coordinate separately. We denote the $k-$th row of the matrix $\boldsymbol{B}$ by $\boldsymbol{B}_{k,\cdot}$ and treat it as a column vector. For $k\in[d]$, we define function

$$f_k(\boldsymbol{x};\boldsymbol{W}):=\boldsymbol{B}_{k,\cdot}^T\boldsymbol{h}_L(\boldsymbol{x}),$$

that is, $f_k(\boldsymbol{x};\boldsymbol{W})$ is the $k-$th coordinate of the NN output vector. The loss function can be written as $\text{loss}_k(\boldsymbol{x},\boldsymbol{y};\boldsymbol{W}):=(f_k(\boldsymbol{x};\boldsymbol{W})-\boldsymbol{y}_k)^2$. Recall that the underlying measurable function $F(\boldsymbol{x})$ (i.e., $\boldsymbol{y}_i=F(\boldsymbol{x}_i)$) is a $d-$dimensional vector-valued function and we denote by $F_k(\boldsymbol{x})$ the $k-$th coordinate of $F(\boldsymbol{x})$. The generalization error is similarly defined as $R_k(\boldsymbol{W}):=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\boldsymbol{X}}}(f_k(\boldsymbol{x};\boldsymbol{W})-F_k(\boldsymbol{x}))^2$. We also denote

$$B_\omega(\boldsymbol{W}^{(0)}):=\{\boldsymbol{W}:\|\boldsymbol{W}-\boldsymbol{W}^{(0)}\|<\omega\}.$$

From Lemma B.9 and Lemma B.10, with high probability, $\boldsymbol{W}^{(t)}$ is close to $\boldsymbol{W}^{(0)}$ during the training. Therefore, we just need to only consider NN functions whose parameters $\boldsymbol{W}$ fall in a small ball around $\boldsymbol{W}^{(0)}$, i.e. $\|\boldsymbol{W}-\boldsymbol{W}^{(0)}\|<\omega$, where $\omega<O\Big(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m}\Big)$. For a given $k\in[d]$, we denote the corresponding function class as

$$\mathcal{G}_{k,\omega}:=\{g:(x,y)\mapsto f_k(\boldsymbol{x};\boldsymbol{W}):\|\boldsymbol{W}-\boldsymbol{W}^{(0)}\|<\omega\}$$

We introduce the empirical Rademacher complexity on the dataset $\{x_i,y_i\}_{i=1}^n$ as follows:

$$\hat{\mathcal{R}}(\mathcal{G}_{k,\omega}):=\mathbb{E}_\sigma\sup_{g\in\mathcal{G}_{k,\omega}}\sum_{i=1}^n\sigma_i g(x_i,y_i)$$

For $k\in[d]$, we first bound the generalization error on the $k-$th coordinate of the output vector.

We first note that by (109) and Lemma B.4, with high probability that $f_k(\boldsymbol{x}_i;\boldsymbol{W})<O(\ln^{1/4}m)$ for all $i\in[n]$ and $\|\boldsymbol{W}-\boldsymbol{W}^{(0)}\|<\omega$. We apply Theorem 11.3 in Mohri et al. (2018) with function class $\mathcal{G}_{k,\omega}$, and thus bound $R_k(\boldsymbol{W})$ with probability at least $1-\Omega(1/m)$ by

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\boldsymbol{X}}}\text{loss}_k(\boldsymbol{x},F(\boldsymbol{x});\boldsymbol{W})\le\frac{1}{n}\sum_{i=1}^n\text{loss}_k(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W})+2O\Big(\ln^{1/4}m\Big)\hat{\mathcal{R}}(\mathcal{G}_{k,\omega})+O\Big(\ln^{1/4}m\Big)O\left(\sqrt{\frac{\ln2m}{2n}}\right).\tag{122}$$

We note that the first term in (122) is bounded by the previously discussed training error, and the third term in (122) is very small when we collect a sufficiently large dataset since m is polynomially dependent on n. Next, we estimate the bound for the second term, the empirical Rademacher complexity.

$$\begin{aligned}\hat{\mathcal{R}}(\mathcal{G}_{k,\omega})=&\mathbb{E}_\sigma\sup_{\boldsymbol{W}\in B_\omega(\boldsymbol{W}^{(0)})}\frac{1}{n}\sum_{i=1}^n\sigma_i\Big(f_k(x_i,y_i;\boldsymbol{W})-f_k(x_i,y_i;\boldsymbol{W}^{(0)})-\langle\nabla_{\boldsymbol{W}}f_k(x_i,y_i;\boldsymbol{W}^{(0)}),\boldsymbol{W}-\boldsymbol{W}^{(0)}\rangle\\&+f_k(x_i,y_i;\boldsymbol{W}^{(0)})+\langle\nabla_{\boldsymbol{W}}f_k(x_i,y_i;\boldsymbol{W}^{(0)}),\boldsymbol{W}-\boldsymbol{W}^{(0)}\rangle\Big)\\\le&\sup_{\boldsymbol{W}\in B_\omega(\boldsymbol{W}^{(0)})}\sup_i\Big|f_k(x_i,y_i;\boldsymbol{W})-f_k(x_i,y_i;\boldsymbol{W}^{(0)})-\langle\nabla_{\boldsymbol{W}}f_k(x_i,y_i;\boldsymbol{W}^{(0)}),\boldsymbol{W}-\boldsymbol{W}^{(0)}\rangle\Big|\qquad(123)\\&+\mathbb{E}_\sigma\sup_{\boldsymbol{W}\in B_\omega(\boldsymbol{W}^{(0)})}\frac{1}{n}\sum_{i=1}^n\sigma_i f_k(x_i,y_i;\boldsymbol{W}^{(0)})\qquad(124)\\&+\mathbb{E}_\sigma\sup_{\boldsymbol{W}\in B_\omega(\boldsymbol{W}^{(0)})}\frac{1}{n}\sum_{i=1}^n\sigma_i\langle\nabla_{\boldsymbol{W}}f_k(x_i,y_i;\boldsymbol{W}^{(0)}),\boldsymbol{W}-\boldsymbol{W}^{(0)}\rangle.\qquad(125)\end{aligned}$$

We first consider (124), since there is no dependence on $\boldsymbol{W}$, it is clear that

$$\mathbb{E}_\sigma\sup_{\boldsymbol{W}\in B_\omega(\boldsymbol{W}^{(0)})}\frac{1}{n}\sum_{i=1}^n\sigma_i f_k(x_i,y_i;\boldsymbol{W}^{(0)})=\frac{1}{n}\sum_{i=1}^n f_k(x_i,y_i;\boldsymbol{W}^{(0)})\mathbb{E}_\sigma\sigma_i=0.\tag{126}$$

In order to help bound (123), which is the first term in our bound of the empirical Rademacher complexity for the NN function class, we introduce the following lemma.

Lemma B.11. *If* $\omega < O\Big(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m}\Big)$ *and* $\|\boldsymbol{W}-\boldsymbol{W}^{(0)}\|<\omega$, *then for any* $\boldsymbol{x}\in\boldsymbol{R}^p$, *with probability at least* $1-\exp(-\Omega(\sqrt{m}/\ln m)$,

$$\Big|f_k(\boldsymbol{x};\boldsymbol{W})-f_k(\boldsymbol{x};\boldsymbol{W}^{(0)})-\langle\nabla_{\boldsymbol{W}}f_k(\boldsymbol{x};\boldsymbol{W}^{(0)}),\boldsymbol{W}-\boldsymbol{W}^{(0)}\rangle\Big|<\frac{1-\alpha}{\sqrt{1+\alpha^2}}O(\omega^{4/3}L^2\sqrt{m\ln m}). \tag{127}$$

We prove Lemma B.11 at the end of this section after we finalize the proof of Lemma 4.3 (while applying Lemma B.11). Using Lemma B.11, the term (123) can be bounded as

$$\begin{aligned}&\sup_i\Big|f_k(x_i,y_i;\boldsymbol{W})-f_k(x_i,y_i;\boldsymbol{W}^{(0)})-\langle\nabla_{\boldsymbol{W}}f_k(x_i,y_i;\boldsymbol{W}^{(0)}),\boldsymbol{W}-\boldsymbol{W}^{(0)}\rangle\Big|\\&\leq\frac{1-\alpha}{\sqrt{1+\alpha^2}}\omega^{4/3}L^2\sqrt{m\ln m}.\end{aligned}$$

Applying Cauchy-Schwarz inequality and Jessen's inequality to (125) and using Lemma 4.2, we conclude

$$\begin{aligned}&\Big|\mathbb{E}_\sigma\sup_{\boldsymbol{W}\in B_\omega(\boldsymbol{W}^{(0)})}\frac{1}{n}\sum_{i=1}^n\sigma_i\langle\nabla_{\boldsymbol{W}}f_k(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W}^{(0)}),\boldsymbol{W}-\boldsymbol{W}^{(0)}\rangle\Big|\\&\leq\omega\mathbb{E}_\sigma\sup_{\boldsymbol{W}\in B_\omega(\boldsymbol{W}^{(0)})}\frac{1}{n}\sum_{i=1}^n\sum_{l=1}^L\|\nabla_{\boldsymbol{W}_l}f_k(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W}^{(0)})\|_F\\&\leq\frac{\omega}{n}\sum_{l=1}^L\sqrt{\sum_{i=1}^n\|\nabla_{\boldsymbol{W}_l}f_k(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W}^{(0)})\|_F^2}\\&\leq\frac{\omega}{n}L\sqrt{mn}\leq\frac{\omega L\sqrt{m}}{\sqrt{n}}\end{aligned}$$

Using the bound for (123), (124) and (125) in (122), it follows that with probability at least $1-e^{-\Omega(\ln m)}$

$$R_k(\boldsymbol{W})\leq\frac{1}{n}\mathrm{loss}_k(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W})+\frac{1-\alpha}{\sqrt{1+\alpha^2}}O\Big(\ln m\sqrt{m}L^2\omega^{4/3}\Big)+\omega O\Big(L\sqrt{m\ln m/n}\Big)+O\left(\sqrt{\frac{\ln m}{n}}\right).$$

Summing over $k\in[d]$, we conclude Lemma 4.3 as follows

$$R(\boldsymbol{W})\leq\frac{1}{n}\mathrm{loss}(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W})+\frac{1-\alpha}{\sqrt{1+\alpha^2}}O\Big(d\ln m\sqrt{m}L^2\omega^{4/3}\Big)+O\Big(d\sqrt{m\ln m/n}L\omega\Big)+O\left(d\sqrt{\frac{\ln m}{n}}\right).$$

Finally, we complete this section by presenting the proof of Lemma B.11.

Proof of Lemma B.11. Using the notation of §B.3 (in particular, $\boldsymbol{h}_l$ and $\boldsymbol{h}_l^{(0)}$) and the definitions of $f_k(\boldsymbol{x};\boldsymbol{W}^{(0)})$ and $f_k(\boldsymbol{x};\boldsymbol{W})$ and recalling that $\boldsymbol{h}_0=\boldsymbol{h}_0^{(0)}$ and $\boldsymbol{h}_l=\boldsymbol{D}_l\boldsymbol{W}_l\boldsymbol{h}_{l-1}$ we derive the following expression:

$$\begin{aligned}&f_k(\boldsymbol{x};\boldsymbol{W})-f_k(\boldsymbol{x};\boldsymbol{W}^{(0)})=\boldsymbol{B}_{k,\cdot}^T\Big(\boldsymbol{D}_L\boldsymbol{W}_L\cdots\boldsymbol{D}_1\boldsymbol{W}_1-\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}\cdots\boldsymbol{D}_1^{(0)}\boldsymbol{W}_1^{(0)}\Big)\boldsymbol{A}\boldsymbol{x}\\&=\boldsymbol{B}_{k,\cdot}^T\Big(\boldsymbol{D}_L\boldsymbol{W}_L\boldsymbol{h}_{L-1}-\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}\boldsymbol{h}_{L-1}\\&\quad+\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}\boldsymbol{h}_{L-1}-\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}\boldsymbol{D}_{L-1}^{(0)}\boldsymbol{W}_{L-1}^{(0)}\boldsymbol{h}_{L-2}\\&\quad+\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}\boldsymbol{D}_{L-1}^{(0)}\boldsymbol{W}_{L-1}^{(0)}\boldsymbol{h}_{L-2}-\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}\boldsymbol{D}_{L-1}^{(0)}\boldsymbol{W}_{L-1}^{(0)}\boldsymbol{D}_{L-2}^{(0)}\boldsymbol{W}_{L-2}^{(0)}\boldsymbol{h}_{L-3}\\&\cdots\ \cdots\\&+\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}\cdots\boldsymbol{D}_2^{(0)}\boldsymbol{W}_2^{(0)}\boldsymbol{h}_1-\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}\cdots\boldsymbol{D}_2^{(0)}\boldsymbol{W}_2^{(0)}\boldsymbol{D}_1^{(0)}\boldsymbol{W}_1^{(0)}\boldsymbol{h}_0\Big).\end{aligned} \tag{128}$$

Let a and b be two integers in $[1,L]$. If $b\leq a$, we denote

$$(\boldsymbol{D}^{(0)}\boldsymbol{W}^{(0)})_{a\mapsto b}:=\boldsymbol{D}_a^{(0)}\boldsymbol{W}_a^{(0)}\boldsymbol{D}_{a-1}^{(0)}\boldsymbol{W}_{a-1}^{(0)}\cdots\boldsymbol{D}_b^{(0)}\boldsymbol{W}_b^{(0)}.$$

If $a<b$, we denote

$$(\boldsymbol{D}^{(0)}\boldsymbol{W}^{(0)})_{a\mapsto b}:=\boldsymbol{I}.$$

Applying $\boldsymbol{h}_l=\boldsymbol{D}_l\boldsymbol{W}_l\boldsymbol{h}_{l-1}$ for the first term in each line of (128), (128) can be written as

$$\begin{aligned}
&f_k(\boldsymbol{x};\boldsymbol{W})-f_k(\boldsymbol{x};\boldsymbol{W}^{(0)})\\
&=\boldsymbol{B}_{k,\cdot}^T\Big((\boldsymbol{D}_L\boldsymbol{W}_L-\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)})\boldsymbol{h}_{L-1}\\
&+(\boldsymbol{D}^{(0)}\boldsymbol{W}^{(0)})_{L\mapsto L}(\boldsymbol{D}_{L-1}\boldsymbol{W}_{L-1}-\boldsymbol{D}_{L-1}^{(0)}\boldsymbol{W}_{L-1}^{(0)})\boldsymbol{h}_{L-2}\\
&+(\boldsymbol{D}^{(0)}\boldsymbol{W}^{(0)})_{L\mapsto L-1}(\boldsymbol{D}_{L-2}\boldsymbol{W}_{L-2}-\boldsymbol{D}_{L-2}^{(0)}\boldsymbol{W}_{L-2}^{(0)})\boldsymbol{h}_{L-3}\\
&\cdots\ \cdots\\
&+(\boldsymbol{D}^{(0)}\boldsymbol{W}^{(0)})_{L\mapsto 2}(\boldsymbol{D}_1\boldsymbol{W}_1-\boldsymbol{D}_1^{(0)}\boldsymbol{W}_1^{(0)})\boldsymbol{h}_0\Big)\\
&=\boldsymbol{B}_{k,\cdot}^T\sum_{l=1}^{L}(\boldsymbol{D}^{(0)}\boldsymbol{W}^{(0)})_{L\mapsto l+1}(\boldsymbol{D}_l\boldsymbol{W}_l-\boldsymbol{D}_l^{(0)}\boldsymbol{W}_l^{(0)})\boldsymbol{h}_{l-1}\\
&=\boldsymbol{B}_{k,\cdot}^T\sum_{l=1}^{L}(\boldsymbol{D}^{(0)}\boldsymbol{W}^{(0)})_{L\mapsto l+1}(\boldsymbol{D}_l-\boldsymbol{D}_l^{(0)})\boldsymbol{W}_l\boldsymbol{h}_{l-1} \qquad (129)\\
&\quad+\boldsymbol{B}_{k,\cdot}^T\sum_{l=1}^{L}(\boldsymbol{D}^{(0)}\boldsymbol{W}^{(0)})_{L\mapsto l+1}\boldsymbol{D}_l^{(0)}(\boldsymbol{W}_l-\boldsymbol{W}_l^{(0)})(\boldsymbol{h}_{l-1}-\boldsymbol{h}_{l-1}^{(0)}) \qquad (130)\\
&\quad+\boldsymbol{B}_{k,\cdot}^T\sum_{l=1}^{L}(\boldsymbol{D}^{(0)}\boldsymbol{W}^{(0)})_{L\mapsto l+1}\boldsymbol{D}_l^{(0)}(\boldsymbol{W}_l-\boldsymbol{W}_l^{(0)})\boldsymbol{h}_{l-1}^{(0)}. \qquad (131)
\end{aligned}$$

According statement 1 in Lemma B.4, with probability at least $1-e^{-\Omega(\sqrt{m}/\ln m)}$, $\|\boldsymbol{D}_l'\boldsymbol{g}_l\|<(1-\alpha)/\sqrt{1+\alpha^2}O(L^{3/2}\omega)$ and $\|\boldsymbol{D}_l'\|_0\leq O(m\omega^{2/3}L)$. Combining this with statement 4 in Lemma B.3 with $\boldsymbol{v}=(1,1,\cdots,1)^T\in\mathbb{R}^d$, we bound the norm $\|(\boldsymbol{B}_k^T\boldsymbol{D}^{(0)}\boldsymbol{W}^{(0)})_{L\mapsto l+1}\|$ by $O(\omega^{1/3}\sqrt{mL\ln m})$ with probability at least $1-\exp(-\Omega(m\omega^{3/2}L\ln m)$. Then (129) can be bounded (with the same probability) by

$$\frac{1-\alpha}{\sqrt{1+\alpha^2}}O(\omega^{4/3}L^2\sqrt{m\ln m}). \qquad (132)$$

By using statement 3 in Lemma B.4, i.e., $\|\boldsymbol{h}_l-\boldsymbol{h}_l^{(0)}\|<O(\omega L^{5/2}\ln m)$ with probability at least $1-e^{-\Omega(\sqrt{m}/\ln m)}$, we note the norm of the summation in (130) is bounded by $O(\omega^2L^{5/2}\ln m)$ (with the latter probability), which is much smaller than (132) when ω is small as given.

By noting that the gradient of $f_k(\boldsymbol{x};\boldsymbol{W})$ with respect to $\boldsymbol{W}_l$ can be written as

$$\nabla_{\boldsymbol{W}_l}f_k(\boldsymbol{x};\boldsymbol{W}^{(0)})=(\boldsymbol{B}_{k,\cdot}^T\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}\cdots\boldsymbol{W}_{l+1}^{(0)}\boldsymbol{D}_l^{(0)})^T\boldsymbol{h}_{l-1}^{(0)T},$$

we express the summands in (131) as follows

$$\langle\nabla_{\boldsymbol{W}_l}f_k(\boldsymbol{x};\boldsymbol{W}^{(0)}),\boldsymbol{W}_l-\boldsymbol{W}_l^{(0)}\rangle\equiv\boldsymbol{B}_{k,\cdot}^T\boldsymbol{D}_L^{(0)}\boldsymbol{W}_L^{(0)}\cdots\boldsymbol{W}_{l+1}^{(0)}\boldsymbol{D}_l^{(0)}(\boldsymbol{W}_l-\boldsymbol{W}_l^{(0)})\boldsymbol{h}_{l-1}^{(0)}. \qquad (133)$$

Using (133) and bounding (129) and (130) by (132), we conclude that with probability at least $1-e^{-\Omega(\sqrt{m}\ln m)}$

$$\Big|f_k(\boldsymbol{x};\boldsymbol{W})-f_k(\boldsymbol{x};\boldsymbol{W}^{(0)})-\langle\nabla_{\boldsymbol{W}}f_k(\boldsymbol{x};\boldsymbol{W}^{(0)}),\boldsymbol{W}-\boldsymbol{W}^{(0)}\rangle\Big|<\frac{1-\alpha}{\sqrt{1+\alpha^2}}O(\omega^{4/3}L^2\sqrt{m\ln m}).$$

□

B.9 Proof of Theorem 3.4

The key idea of the proof of this theorem is to establish a bound for ω, such that $\|\boldsymbol{W}^{(t)}-\boldsymbol{W}\|<\omega$ during training. Considering the learning rate η and training steps t, we first establish a simple bound for ω as

$$\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\|\leq\sum_{t=0}^{t-1}\eta\|\nabla_{\boldsymbol{W}}\mathcal{L}^{(t)}\|\leq\eta\sqrt{\frac{mn}{d}}\sum_{t=0}^{t}\sqrt{\mathcal{L}^{(t)}}\leq\eta t\sqrt{\frac{nm\ln m}{d}}.$$

Furthermore, in the proof of Lemma B.9, the following universal bound of ω was introduced:

$$\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\|\leq O\left(\sqrt{\frac{nd}{\delta m}}\right)\sqrt{\mathcal{L}^{(0)}}.$$

By combining these two bounds with the universal bound $\omega<O(\frac{\delta^{3/2}}{n^{3/2}L^{15/2}\ln^{3/2}m})$ and using Lemma 4.3, we conclude the theorem.

B.10 Generalization Error Bound for SGD

We present a theorem similar to Theorem 3.4 that establishes the upper bound of the generalization error for SGD.

Theorem B.12. *Assume the setup of §2 with SGD, where* $m=\Theta\left(\frac{n^{13+2\epsilon}L^{15+2\epsilon}d^{1+2\epsilon}}{b\delta^{5-2\epsilon}}\right)$ *for* $\epsilon>0$ *and* $\eta=\Theta(\frac{d\delta}{n^3L^3m\ln^2m})$. *Assume further that* m *is larger than its lower bound and* η *is smaller than its upper bound in Theorem 3.2 (by an appropriate choice of the hidden constants in* Θ *and in comparison to the constants hidden in the lower bound of* m *and the upper bound of* η *in Theorem 3.2). Then at a given training epoch* t*, with probability at least* $1-e^{-\Omega(\ln m)}$*, the generalization error is bounded as follows*

$$\begin{aligned}R(\boldsymbol{W}^{(t)})\leq\gamma^tO(\ln m)+\min\left\{\left(\frac{1-\alpha}{\sqrt{1+\alpha^2}}\right)O\left(\frac{d^{1/3}t^{4/3}}{m^{1/6}n^{10/3}L^2\ln^{8/3}m}\right),O\left(\frac{d^{3/2+\epsilon}n^{2+\epsilon}}{b^{1/2}\delta^{1/2-\epsilon}L^{1/2-\epsilon}\ln m}\right)\right\}+\\ \min\left\{O\left(\frac{\sqrt{d}\,t}{n^3L^2\ln^{3/2}m}\right),O\left(\frac{n^{2+\epsilon}L^{2+\epsilon}d^{1/2+\epsilon}}{b^{1/2}\delta^{1-\epsilon}\ln m}\right)\right\}+O\left(d\sqrt{\frac{\ln m}{n}}\right).\end{aligned}\tag{134}$$

The proof is similar to the proof of Theorem 3.4. We estimate the bound of ω when t is small as

$$\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\|<O\left(\frac{mn\eta t}{d}\right)\sqrt{\mathcal{L}^{(0)}}.$$

Also, the bound of ω in the entire training for SGD can be obtained in the proof of Lemma B.10 as

$$\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\|<O\left(\frac{d\sqrt{n}}{\delta\sqrt{mb}}\sqrt{\mathcal{L}^{(0)}}\right).$$

Then combining these two bounds of ω and using Lemma 4.3, we could conclude the theorem.

B.11 Special dataset

In this section, we consider a special class of datasets and improve our theory for datasets from this class. We first introduce the special dataset and establish the assumption, then present theorems to bound the convergence rate and generalization error under this assumption. The proof will be given in Appendix B.13.

First, with the parameters $\boldsymbol{W}^{(0)}$ before the $l-$th layer, we denote the output at the $l-$th layer as

$$\mathcal{N}_l(\boldsymbol{x};\boldsymbol{A},\boldsymbol{W}_1^{(0)},\boldsymbol{W}_2^{(0)},\cdots\boldsymbol{W}_{l-1}^{(0)},\boldsymbol{u})=\tilde{\sigma}_\alpha(\boldsymbol{u}^T\tilde{\sigma}_\alpha(\boldsymbol{W}_{l-1}^{(0)}\tilde{\sigma}_\alpha(\boldsymbol{W}_{l-2}^{(0)}\cdots\tilde{\sigma}_\alpha(\boldsymbol{W}_1^{(0)}\boldsymbol{A}\boldsymbol{x})))),$$

and define the following class of functions:

$$\begin{aligned}\mathcal{F}_l:=\Big\{&\boldsymbol{f}(\boldsymbol{x})=(f_1(\boldsymbol{x}),\cdots,f_d(\boldsymbol{x}))^T:\mathbb{R}^p\mapsto\mathbb{R}^d,\text{ where}\\ &f_j(\boldsymbol{x})=\mathbb{E}_{\boldsymbol{u}}c_j(\boldsymbol{u})\mathcal{N}_l(\boldsymbol{x};\boldsymbol{A},\boldsymbol{W}_1^{(0)}\cdots\boldsymbol{W}_{l-1}^{(0)},\boldsymbol{u})\text{ for }\boldsymbol{u}\in\mathbb{R}^m\sim N\left(0,\frac{2}{m}\right)\text{ and}\\ &c_j:\mathbb{R}^m\mapsto\mathbb{R}\text{ such that }|c_j(\cdot)|\leq1\text{ for }j\in[d]\Big\}.\end{aligned}\tag{135}$$

We note that this function class $\mathcal{F}_l$ includes functions defined by an $l-$layer leaky ReLU neural network, where the first $l-1$ layers use the initialized parameters $\boldsymbol{W}^{(0)}$ and only the parameters of the $l-$th layer are tuned with a certain regularization condition ($\|c_j\|_\infty<1$). Given this function class, we restrict our discussion to datasets satisfying the assumption below. For such datasets, we can improve the upper bound for the convergence rate, the lower bound for the width m, and the upper bound for the generalization error.

Assumption B.13. For any small constant $0<\lambda<\frac{1}{2\sqrt{nd}}$, there exists $f\in\mathcal{F}_{L-1}$ such that

$$\|\boldsymbol{f}(\boldsymbol{x}_i)-(\boldsymbol{y}_i-\hat{\boldsymbol{y}}_i)\|<\lambda, \text{ for all } \ i\in[n].$$

Theorem B.14. *Assume the setup in §2 with a dataset satisfying Assumption B.13, where both $m/ln^4m>\Omega(d^5nL^{12})$ and $m>\Omega(\ln\ln\epsilon^{-1})$, and the NN is trained according to Algorithm 2, with learning rate $\eta\le O(\frac{d}{nL^2m})$. Then with probability at least $1-e^{-\Omega(\ln m)}$*

$$\mathcal{L}(\boldsymbol{W}^{(T)})<\epsilon \text{ and } \ \mathcal{L}(\boldsymbol{W}^{(t)})\le\gamma^t\mathcal{L}(\boldsymbol{W}^{(0)}),\ \forall t\le T,$$

where

$$\gamma=1-\Omega\left(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{\eta m}{d^2}\right) \text{ and } T=\frac{\ln\big(\epsilon/\mathcal{L}(\boldsymbol{W}^{(0)})\big)}{\ln\gamma}.$$

Theorem B.15. *Assume the setup of §2 with GD, a dataset satisfying Assumption B.13, $m=\Theta(n^{1+2\tau}L^{12+2\tau}d^{5+2\tau})$ for $\tau>0$ and $\eta=\Theta(\frac{d}{nL^2m})$. Assume further that m is larger than its lower bound and η is smaller than its upper bound in Theorem B.14 (by an appropriate choice of the hidden constants in Θ and compared to the constants hidden in the lower bound of m and in the upper bound of η in Theorem B.14). Then at a given training epoch $t\le T$ (see (4) for T), with probability at least $1-e^{-\Omega(\ln m)}$, the generalization error is bounded as follows*

$$R(\boldsymbol{W}^{(t)})\le\gamma^t\mathcal{L}(\boldsymbol{W}^{(0)})+\min\left\{O\left(\frac{d^{3/2+\tau}n^{1/2+\tau}L^\tau}{\ln m}\right),O\left(\frac{1-\alpha}{\sqrt{1+\alpha^2}}\frac{d^{1/3}t^{4/3}}{m^{1/6}n^{2/3}L^{2/3}}\right)\right\}$$
$$+\min\left\{O\left(\frac{\sqrt{d\ln m}\ t}{nL}\right),O\left(\frac{n^\tau L^{1+\tau}d^{2+\tau}}{\ln m}\right)\right\}+O\left(d\sqrt{\frac{\ln m}{n}}\right).$$

We notice that for datasets satisfying Assumption B.13 several significant improvements from the previous estimates are obtained. Firstly, the lower bound for m is improved to linear dependence on n, whereas in the general scenario the lower bound grows as n^5. Secondly, the bound of $1-\gamma$ is improved in Theorem B.14 by a factor of $\frac{n}{\delta d}$. Thirdly, several terms in the generalization error bound in Theorem B.15 are improved from Theorem 3.4, including the first term in the first minimum is improved by a factor of $\frac{1}{L}$ and the second term in the second minimum is improved by a factor of $\frac{\sqrt{d^3\delta}}{L\sqrt{n}}$. On the other hand, the optimal choice of α remains the same as the dependence on α is the same as that in Theorems 3.1 and 3.4.

B.12 Convergence Theorem for General Convex Loss Functions

We extend our convergence theory, in particular Theorem 3.1, to convex loss functions, i.e., loss functions of the form

$$\mathcal{L}_{\text{convex}}(\boldsymbol{W})=\sum_i^n l(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i), \text{ where } l(\boldsymbol{y}_i,\cdot) \text{ is convex.} \tag{136}$$

These include common loss functions for classification, such as the binary cross entropy and categorical cross entropy. Furthermore, it also includes the following loss function suggested in (Kumar et al., 2023):

$$\mathcal{L}_{\exp}(\boldsymbol{W}):=\frac{1}{2}\sum_i^n e^{\lambda\|\boldsymbol{y}_i-\hat{\boldsymbol{y}}_i\|^2}., \tag{137}$$

Kumar et al. (2023) obtained a special bound for the generalization error when using this loss function. We later use the following theorem and the proposition of Kumar et al. (2023) to infer that $\alpha=-1$ is also optimal for generalization when using re-weighted gradient descent and overparameterized neural networks.

We next formulate the main theorem using the following definition. Let $\boldsymbol{W}^*$ denote the matrix of parameters minimizing the loss function and define

$$\mathcal{E}^{(t)}:=\mathcal{L}_{\text{convex}}(\boldsymbol{W}^{(t)})-\mathcal{L}_{\text{convex}}(\boldsymbol{W}^*).$$

Theorem B.16. *Assume the setup of §2 with the convex loss function defined in (136), where the width m satisfies both $m/\ln^4 m \geq \frac{1+\alpha^2}{(1-\alpha)^2}\Omega(\frac{n^6L^{16}d}{\delta^4})$ and $m > \Omega(\ln(\epsilon^{-1}\ln\epsilon^{-1}))$ and the training is according to Algorithm 2 with learning rate $\eta \leq O(\frac{d}{nL^2m})$. Then with probability at least $1-\exp^{-\Omega(\ln m)}$,*

$$\mathcal{E}^{(T)} < \epsilon \text{ and } \mathcal{E}^{(t+1)} \leq \gamma^{(t)}\mathcal{E}^{(t)}, \ \forall t \leq T, \ \text{where} \tag{138}$$

where

$$\gamma^{(t)} = 1 - \mathcal{E}^{(t)}\Omega\Big(\frac{(1-\alpha)^2}{(1+\alpha^2)}\frac{\delta\eta m}{n^2dL}\Big) \ \text{ and } \ T \leq O\Big(\frac{1+\alpha^2}{(1-\alpha)^2}\frac{n^2dL}{\eta\delta m\epsilon}(\ln\epsilon^{-1} + \ln(n\sqrt{\ln m}))\Big). \tag{139}$$

Combining (138) and the expression for $\gamma^{(t)}$ in (139), we note that the rate of convergence is slower than the one in (3) and that $\alpha = -1$ corresponds to the smallest upper bound for the number of epochs needed for the training error to be smaller than ϵ.

Proposition 3.1 in Kumar et al. (2023) implies that minimizing the generalization error bound is equivalent to minimizing the training error when using the modified loss function given in (137). Applying (139) of Theorem B.16, we conclude that when using the modified loss function in (137) for training, the choice of $\alpha = -1$ yields the smallest bound for the required number of training epochs to achieve a bound ϵ on the training error. Using this observation with arbitrarily small ϵ and the above discussed theory of Kumar et al. (2023) we can conclude that $\alpha = -1$ minimizes the generalization error bound with the smallest upper bound on the number of training epochs for which the training error is guaranteed to be less than ϵ. Nevertheless, this discussion involves an upper bound we obtained for the number of epochs and does not apply to the actual number of epochs. Consequently, the above stated prediction may not be precise, that is, it is possible that at a smaller number of epochs than the bound, one may obtain an error less than ϵ by $\alpha \neq -1$. For a synthetic dataset, we empirically verified the predicted optimal choice of $\alpha = -1$ (see Figure 6).

Proof of Theorem B.16. We consider the modified loss function defined in (136), and, for simplicity, we let $\gamma = 1$ in this section, while the proof can be easily extended for any $\gamma > 0$. By introducing the convex function $l(\boldsymbol{y}, \boldsymbol{z})$, the loss function for each data point $i \in [n]$ can be written as

$$\text{loss}_{\text{convex}}(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{W}) := l(\boldsymbol{y}_i, \boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i; \boldsymbol{W}))$$

We denote the following notation in this subsection,

$$\boldsymbol{e}_i := \nabla_{\boldsymbol{z}} l(\boldsymbol{y}_i, \boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i; \boldsymbol{W})). \tag{140}$$

We will establish similar gradient bounds as in Lemma 4.2 and a similar semi-smoothness inequality as in Lemma 4.1 for the loss function defined in (136), and then we will prove the convergence theory with this loss function using some of the above established results.

For the first part, in order to achieve the gradient bound, we follow the proof of Lemma 4.2 with the modified loss function defined in (136). The proof is mostly the same as the one in Appendix B.4. The only difference is the use of the previous definition in (61). We remark that it is straightforward to verify that $\boldsymbol{G}_{i,l}(\boldsymbol{e}_i; \boldsymbol{W}) \equiv \nabla_{\boldsymbol{W}_l} \text{loss}_{\text{convex}}(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{W})$ by using the definition of $\boldsymbol{e}_i$ in (140). By Lemma B.6 and Lemma B.7

$$\|\nabla_{\boldsymbol{W}_l} \mathcal{L}_{\text{convex}}(\boldsymbol{W})\|_F^2 \leq \sum_{i=1}^n \|\boldsymbol{e}_i\|^2 O\Big(\frac{mn}{d}\Big), \quad \text{for } \ l \in [L] \tag{141}$$

$$\|\nabla_{\boldsymbol{W}} \mathcal{L}_{\text{convex}}(\boldsymbol{W})\|_F^2 \geq \sum_{i=1}^n \|\boldsymbol{e}_i\|^2 \Omega\Big(\frac{(1-\alpha)^2}{(1+\alpha^2)}\frac{\delta m}{nd}\Big). \tag{142}$$

Remark: In the above bounds, when we use the MSE loss function, $\sum_{i=1}^n \|\boldsymbol{e}_i\|^2 \equiv \mathcal{L}$, which yields the original bounds provided in Lemma 4.2.

In order to show the semi-smoothness, we follow the proof of Lemma 4.1 in Appendix B.5, where most parts are exactly the same. By using (142) and (141), and plugging the notation of $\boldsymbol{e}_i$ (defined in (140)) into (95), the semi-smoothness

inequality becomes

$$\begin{aligned}\mathcal{L}_{\text{convex}}(\boldsymbol{W}+\boldsymbol{W}')\leq\mathcal{L}_{\text{convex}}(\boldsymbol{W})+\langle\nabla_{\boldsymbol{W}}\mathcal{L}_{\text{convex}}(\boldsymbol{W}),\boldsymbol{W}'\rangle+\frac{nL^2m}{d}O(\|\boldsymbol{W}'\|_2^2)\\+\frac{(1-\alpha)\omega^{1/3}L^2\sqrt{mn\sum_{i=1}^n\|\boldsymbol{e}_i^{(t)}\|^2\ln m}}{\sqrt{d(1+\alpha^2)}}O(\|\boldsymbol{W}'\|_2).\end{aligned}\tag{143}$$

Lastly, we prove the convergence for this loss function. By using (143) and the same argument discussed in §4.1, the inequality (16) still holds. Then using the lower bound of the gradient in (142), (16) becomes

$$\mathcal{L}_{\text{convex}}(\boldsymbol{W}^{(t+1)})\leq\mathcal{L}_{\text{convex}}(\boldsymbol{W}^{(t)})-\Omega\bigg(\frac{(1-\alpha)^2}{(1+\alpha^2)}\frac{\delta\eta m}{nd}\bigg)\sum_{i=1}^n\|\boldsymbol{e}_i^{(t)}\|^2.\tag{144}$$

By convexity of $l(\boldsymbol{y},\boldsymbol{z})$, we first establish that for any $i\in[n]$ and any $\boldsymbol{y},\boldsymbol{z}\in\mathbb{R}^d$,

$$l(\boldsymbol{y}_i,\boldsymbol{y})-\boldsymbol{y}_{\boldsymbol{y}_i,\boldsymbol{z}}\leq\langle\nabla_{\boldsymbol{y}}l(\boldsymbol{y}_i,\boldsymbol{y}),\boldsymbol{y}-\boldsymbol{z}\rangle\leq\|\nabla_{\boldsymbol{y}}l(\boldsymbol{y}_i,\boldsymbol{y})\|\|\boldsymbol{y}-\boldsymbol{z}\|.$$

We denote by $\boldsymbol{W}^*$ the optimal parameter that minimizes $\mathcal{L}_{\exp}(\boldsymbol{W})$. Letting $\boldsymbol{y}:=\boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i;\boldsymbol{W}^{(t)})$ and $\boldsymbol{z}:=\boldsymbol{g}_{L+1}(\boldsymbol{x}_i;\boldsymbol{W}^*)$ and using the above inequality result in

$$\|\boldsymbol{e}_i^{(t)}\|\geq\frac{l(\boldsymbol{y}_i,\boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i;\boldsymbol{W}^{(t)}))-l(\boldsymbol{y}_i,\boldsymbol{g}_{L+1}(\boldsymbol{x}_i;\boldsymbol{W}^*))}{\|\boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i;\boldsymbol{W}^{(t)})-\boldsymbol{g}_{L+1}(\boldsymbol{x}_i;\boldsymbol{W}^*)\|}.\tag{145}$$

Using Lemma B.5, yields that, with probability at least $1-e^{-\Omega(\ln m)}$,

$$\begin{aligned}&\|\boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i;\boldsymbol{W}^*)-\boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i;\boldsymbol{W}^{(t)})\|\leq\|\boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i;\boldsymbol{W}^*)-\boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i;\boldsymbol{W}^{(0)})\|\\&\quad+\|\boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i;\boldsymbol{W}^{(t)})-\boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i;\boldsymbol{W}^{(0)})\|\\&\leq(\|\boldsymbol{W}^{(0)}-\boldsymbol{W}^*\|+\omega)\|\nabla_{\boldsymbol{W}}\boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i;\boldsymbol{W}^{(0)})\|\leq O(\sqrt{L}).\end{aligned}\tag{146}$$

Applying (146) to (145) results in

$$\|\boldsymbol{e}_i^{(t)}\|\geq\Big(l(\boldsymbol{y}_i,\boldsymbol{g}_{i,L+1}(\boldsymbol{x}_i;\boldsymbol{W}^{(t)}))-l(\boldsymbol{y}_i,\boldsymbol{g}_{L+1}(\boldsymbol{x}_i;\boldsymbol{W}^*))\Big)/O(\sqrt{L}).\tag{147}$$

Applying (147) to (144), we derive that

$$\begin{aligned}&\mathcal{L}_{\text{convex}}(\boldsymbol{W}^{(t+1)})-\mathcal{L}_{\text{convex}}(\boldsymbol{W}^*)\\&\leq\mathcal{L}_{\text{convex}}(\boldsymbol{W}^{(t)})-\mathcal{L}_{\text{convex}}(\boldsymbol{W}^*)-\Omega\bigg(\frac{(1-\alpha)^2}{(1+\alpha^2)}\frac{\delta\eta m}{n^2dL}\bigg)\Big(\mathcal{L}_{\text{convex}}(\boldsymbol{W}^{(t)})-\mathcal{L}_{\text{convex}}(\boldsymbol{W}^*)\Big)^2.\end{aligned}\tag{148}$$

In order to derive a bound for the number of training epoch T that is required for $\mathcal{L}^{(T)}-\mathcal{L}^*<\epsilon$, for $t<T$, by assuming $\mathcal{L}^{(t)}-\mathcal{L}^*>\epsilon$, the above equation is bounded by

$$\mathcal{L}_{\text{convex}}(\boldsymbol{W}^{(t+1)})-\mathcal{L}_{\text{convex}}(\boldsymbol{W}^*)\leq\bigg(1-\epsilon\Omega\bigg(\frac{(1-\alpha)^2}{(1+\alpha^2)}\frac{\delta\eta m}{n^2dL}\bigg)\bigg)\Big(\mathcal{L}_{\text{convex}}(\boldsymbol{W}^{(t)})-\mathcal{L}_{\text{convex}}(\boldsymbol{W}^*)\Big).$$

The lower bound for m becomes $m\ln^4m>\frac{1+\alpha^2}{(1-\alpha)^2}\Omega(n^6L^{16}d/\delta^4)$ to ensure the same perturbation bound in Lemma B.9. Denoting $\gamma:=\Big(1-\epsilon\Omega\Big(\frac{(1-\alpha)^2}{(1+\alpha^2)}\frac{\delta\eta m}{n^2dL^2}\Big)\Big)$, it follows that

$$T=\frac{\ln\epsilon^{-1}+\ln(\mathcal{L}_{\text{convex}}(\boldsymbol{W}^{(0)})-\mathcal{L}_{\text{convex}}(\boldsymbol{W}^*))}{\ln\gamma^{-1}}\leq O\bigg(\frac{n^2dL}{\eta\delta m\epsilon}(\ln\epsilon^{-1}+\ln(n\sqrt{\ln m}))\bigg).$$

We follow the exact same steps in the proof of Lemma B.9 in Appendix B.6 and verify that when $m>\ln(\epsilon^{-1}\ln\epsilon^{-1})$, the probability that (148) holds for $T-$steps is at least $1-e^{-\Omega(\ln m)}$. □

B.13 Proofs for a special class of datasets

This section includes the proof of Theorems B.14 and B.15 in Appendix B.11. We first present several lemmas and their proofs, then use these lemmas to prove those theorems.

Lemma B.17. *Consider a dataset $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i\in[n]}$ satisfying Assumption B.13, where $m\geq\Omega(nd)$, then there exists a vector $\boldsymbol{u}_{l,j}\in\mathcal{B}_1^m\subset\mathbb{R}^m$, such that*

$$|\langle\boldsymbol{u}_{l,j},\boldsymbol{h}_{i,L-1}^{(0)}\rangle-(y_{i,j}-\hat{y}_{i,j})|\leq O(\lambda),\ \ \textit{for all}\ \ i\in[n],j\in[d],\ \ \textit{with probability at least}\ \ 1-e^{-\Omega(\lambda^2 m)}.$$

Proof. We complete the proof by constructing the following unit vector $\boldsymbol{u}_j$:

$$\boldsymbol{u}_j=\frac{1}{\sqrt{2m}}\Big(c_j(\sqrt{m/2}(\boldsymbol{W}_{L-1}^{(0)})_{1,\cdot}),c_j(\sqrt{m/2}(\boldsymbol{W}_{L-1}^{(0)})_{2,\cdot}),\cdots,c_j(\sqrt{m/2}(\boldsymbol{W}_{L-1}^{(0)})_{m,\cdot})\Big)^T.$$

One can easily verify that $\boldsymbol{u}_j\in\mathcal{B}_1^m$ by using the fact that $|c_j|<1$, which is guaranteed by the definition of the function class in (135).

The inner product of $\boldsymbol{u}_j$ and $\boldsymbol{h}_{i,L-1}^{(0)}$ is given by

$$\begin{aligned}\langle\boldsymbol{u}_j,\boldsymbol{h}_{i,L-1}^{(0)}\rangle&=\frac{1}{\sqrt{2m}}\sum_{k=1}^m c_j(\sqrt{m/2}(\boldsymbol{W}_{L-1}^{(0)})_{k,\cdot})\mathcal{N}_{L-1}(\boldsymbol{x};\boldsymbol{A},\boldsymbol{W}_1^{(0)},\cdots\boldsymbol{W}_{L-2}^{(0)},\boldsymbol{W}_{L-1}^{(0)})\\&=\frac{1}{\sqrt{2m}}\sum_{k=1}^m c_j(\sqrt{m/2}(\boldsymbol{W}_{L-1}^{(0)})_{k,\cdot})\sqrt{\frac{2}{m}}\mathcal{N}_{L-1}(\boldsymbol{x};\boldsymbol{A},\boldsymbol{W}_1^{(0)},\cdots\boldsymbol{W}_{L-2}^{(0)},\sqrt{m/2}\boldsymbol{W}_{L-1}^{(0)})\\&=\frac{1}{m}\sum_{k=1}^m c_j(\sqrt{m/2}(\boldsymbol{W}_{L-1}^{(0)})_{k,\cdot})\mathcal{N}_{L-1}(\boldsymbol{x};\boldsymbol{A},\boldsymbol{W}_1^{(0)},\cdots\boldsymbol{W}_{L-2}^{(0)},\sqrt{m/2}\boldsymbol{W}_{L-1}^{(0)})\end{aligned}$$

For simplicity, we denote that $Z_{k,i,j}:=c_j(\sqrt{m/2}(\boldsymbol{W}_{L-1}^{(0)})_{k,\cdot})\mathcal{N}_{L-1}(\boldsymbol{x};\boldsymbol{A},\boldsymbol{W}_1^{(0)},\cdots\boldsymbol{W}_{L-2}^{(0)},\sqrt{m/2}\boldsymbol{W}_{L-1}^{(0)})$, and above equation becomes

$$\langle\boldsymbol{u}_j,\boldsymbol{h}_{i,L-1}^{(0)}\rangle=\frac{1}{m}\sum_{k=1}^m Z_{k,i,j}.\tag{149}$$

Noting that $\sqrt{m/2}(\boldsymbol{W}_{L-1}^{(0)})_{k,\cdot}\sim N(0,1)$, by using (135), it implies that $\mathbb{E}_{\boldsymbol{W}^{(0)}}Z_{k,i,j}((\boldsymbol{W}_{L-1})_{k,\cdot}^{(0)})=f_j(\boldsymbol{x}_i)$.

Since $|c_j(\cdot)|<1$ is bounded, $Z_{k,i,j}$ is a sub-Gaussian random variable, therefore we can apply Hoeffding's inequality and conclude that

$$\left|\frac{1}{m}\sum_{k=1}^m Z_{k,i,j}-f_j(\boldsymbol{x}_i)\right|\leq\lambda,\ \ \text{with probability at least}\ \ 1-e^{-\Omega(\lambda^2 m)}.\tag{150}$$

Using (149), (150) and Assumption B.13, it follows that with probability at least $1-(nd)e^{-\Omega(\lambda^2 m)}$

$$|\langle\boldsymbol{u}_{l,j},\boldsymbol{h}_{i,L-1}^{(0)}\rangle-(y_{i,j}-\hat{y}_{i,j})|\leq|\langle\boldsymbol{u}_{l,j},\boldsymbol{h}_{i,L-1}^{(0)}\rangle-f_j(\boldsymbol{x}_i)|+|f_j(\boldsymbol{x}_i)-(y_{i,j}-\hat{y}_{i,j})|\leq 2\lambda.\ \ \text{for all}\ \ i\in[n],\ j\in[d].$$

We conclude the Lemma by noting that the probability is at least $1-e^{-\Omega(\lambda^2 m)}$ when $m\geq\Omega(nd)$. □

Lemma B.18. *Under Assumption B.13, when $m\geq\Omega(nd)$, the lower bound for the gradient of the loss function becomes*

$$\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{(0)})\|_F^2\geq\Omega\left(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{m}{d^2}\right)\mathcal{L}(\boldsymbol{W}^{(0)}),\ \ \textit{with probability at least}\ \ 1-e^{-\Omega(m)}.$$

Proof. We first note that by definition, $\|\nabla_{\boldsymbol{W}}\mathcal{L}\|_F\geq\|\nabla_{\boldsymbol{W}_L}\mathcal{L}\|$. Then by definition of matrix $F-$norm and (61), it follows that $\|\nabla_{\boldsymbol{W}_L}\mathcal{L}(\boldsymbol{W}^{(0)})\|_F^2=\sum_{k=1}^m\big\|\sum_{i=1}^n(\boldsymbol{G}_{i,L}(\boldsymbol{e}_i;\boldsymbol{W}^{(0)}))_{k,\cdot}\big\|_2^2$. We write that the $k-$th row of the matrix $\sum_{i=1}^n\boldsymbol{G}_{i,L}(\boldsymbol{e}_i,\boldsymbol{W}^{(0)})$ by

$$\left\|\sum_{i=1}^n(G_{i,L}(\boldsymbol{e}_i;\boldsymbol{W}^{(0)}))_{k,\cdot}\right\|_2^2=\left\|\sum_{i=1}^n\boldsymbol{B}_{k,\cdot}^T\boldsymbol{e}_i D_{i,L,kk}\boldsymbol{h}_{i,L-1}\right\|_2^2.\tag{151}$$

Using the vector $\boldsymbol{u}_j \in \mathcal{B}_1^m$ chosen in Lemma B.17, and denoting $\boldsymbol{u} := \frac{1}{d}\sum_{j=1}^d \boldsymbol{u}_j$, we note that $\|\boldsymbol{u}\| \le 1$. Thus we conclude that

$$\begin{aligned}
&\left\|\sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i D_{i,L,kk} \boldsymbol{h}_{i,L-1}\right\|_2^2 \ge \left|\left\langle \sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i D_{i,L,kk} \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\right\rangle\right|^2 \\
&= \left|\sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i D_{i,L,kk} \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle\right|^2 \\
&= \frac{1}{1+\alpha^2}\left|\sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i (\alpha + (1-\alpha) 1_{\boldsymbol{h}_{i,L,k}>0}) \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle\right|^2 \\
&= \frac{1}{1+\alpha^2}\left|(1-\alpha)\sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i 1_{\boldsymbol{h}_{i,L,k}>0} \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle + \alpha \sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle\right|^2
\end{aligned} \tag{152}$$

By Jensen's inequality, we note that the expectation of (152) becomes

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{h}_{L-1}}\left|(1-\alpha)\sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i 1_{\boldsymbol{h}_{i,L,k}>0} \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle + \alpha \sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle\right|^2 \\
&\ge \left|\mathbb{E}_{\boldsymbol{h}_{L-1}}\left((1-\alpha)\sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i 1_{\boldsymbol{h}_{i,L,k}>0} \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle + \alpha \sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle\right)\right|^2.
\end{aligned} \tag{153}$$

Since $1_{\boldsymbol{h}_{i,L,k}>0}$ is independent with $\boldsymbol{h}_{i,L-1}$, for any integer N, the conditional expectation can be given as

$$\mathbb{E}_{\boldsymbol{h}_{L-1}}\left(\sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i 1_{\boldsymbol{h}_{i,L,k}>0} \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle \middle| \sum 1_{\boldsymbol{h}_{i,L,k}>0} = N\right) = \frac{N}{n}\mathbb{E}_{\boldsymbol{h}_{L-1}}\left(\sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle\right) \tag{154}$$

Moreover, since $1_{\boldsymbol{h}_{i,L,k}>0}$ a Bernoulli random variable $B(0.5)$, when $n > 100$, using an approximation of the probability by the central limit theorem, we know that

$$\sum_i 1_{\boldsymbol{h}_{i,L,k}>0} > n/2 + \sqrt{n}, \text{ with probability at least } 0.1, \tag{155}$$

$$\sum_i 1_{\boldsymbol{h}_{i,L,k}>0} < n/2 - \sqrt{n}, \text{ with probability at least } 0.1. \tag{156}$$

To find a lower bound for (153), we consider two cases for the second term, when $|\mathbb{E}_{\boldsymbol{h}_{L-1}} \alpha \sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i \langle \boldsymbol{h}_{i,L-1,\boldsymbol{u}}\rangle| > \frac{n}{2}|\mathbb{E}_{\boldsymbol{h}_{L-1}}(\boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle)|$, then by (156) and (154), we know that with probability at least 0.1 that

$$\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{h}_{L-1}}\left((1-\alpha)\sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i 1_{\boldsymbol{h}_{i,L,k}>0} \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle + \alpha \sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle\right)\right| \\
&\ge \frac{1}{\sqrt{n}}\left|\mathbb{E}_{\boldsymbol{h}_{L-1}}\left(\sum_{i=1}^n (1-\alpha)\boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle\right)\right|.
\end{aligned} \tag{157}$$

Using similar argument, when $|\mathbb{E}_{\boldsymbol{h}_{L-1}} \alpha \sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i \langle \boldsymbol{h}_{i,L-1,\boldsymbol{u}}\rangle| \le \frac{n}{2}|\mathbb{E}_{\boldsymbol{h}_{L-1}}(\boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle)|$, by using (155), it also follows that with probability at least 0.1 that (157) holds. Thus we conclude that

$$\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{h}_{L-1}}\left((1-\alpha)\sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i 1_{\boldsymbol{h}_{i,L,k}>0} \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle + \alpha \sum_{i=1}^n \boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle\right)\right| \\
&\ge \frac{1}{\sqrt{n}}\left|\mathbb{E}_{\boldsymbol{h}_{L-1}}\left(\sum_{i=1}^n (1-\alpha)\boldsymbol{B}_{k,\cdot}^T \boldsymbol{e}_i \langle \boldsymbol{h}_{i,L-1}, \boldsymbol{u}\rangle\right)\right|, \text{ with probability at least } 0.1.
\end{aligned} \tag{158}$$

Combining (152), (153), and (158), it follows that

$$\mathbb{E}_{\boldsymbol{h}_{L-1}}\left\|\sum_{i=1}^{n}\boldsymbol{B}_{k,\cdot}^{T}\boldsymbol{e}_i D_{i,L,kk}\boldsymbol{h}_{i,L-1}\right\|\geq\frac{(1-\alpha)^2}{n(1+\alpha^2)}\left|\mathbb{E}_{\boldsymbol{h}_{L-1}}\sum_{i=1}^{n}\boldsymbol{B}_{k,\cdot}^{T}\boldsymbol{e}_i\langle\boldsymbol{h}_{i,L-1},\boldsymbol{u}\rangle\right|^2,\ \text{with probability at least}\ \ 0.1. \tag{159}$$

By using the lower bound of $|\boldsymbol{B}_{\cdot,k}^{T}\boldsymbol{a}|$ derived in (74), we obtain that with at least a constant probability $p_0:=1-\exp(-\Omega(1))$

$$\left|\mathbb{E}_{\boldsymbol{h}_{L-1}}\sum_{i=1}^{n}\boldsymbol{B}_{k,\cdot}^{T}\boldsymbol{e}_i\langle\boldsymbol{h}_{i,L-1},\boldsymbol{u}\rangle\right|^2=\left|\boldsymbol{B}_{k,\cdot}^{T}\left(\mathbb{E}_{\boldsymbol{h}_{L-1}}\sum_{i=1}^{n}\boldsymbol{e}_i\langle\boldsymbol{h}_{i,L-1},\boldsymbol{u}\rangle\right)\right|^2\geq\sum_{j=1}^{d}\left|\mathbb{E}_{\boldsymbol{h}_{L-1}}\sum_{i=1}^{n}\boldsymbol{e}_{i,j}\langle\boldsymbol{h}_{i,L-1},\boldsymbol{u}\rangle\right|^2. \tag{160}$$

We thus conclude that by (151), Hoeffding inequality, (159) and (160), with probability at least $1-e^{-\Omega(m)}$

$$\sum_{k=1}^{m}\left\|\sum_{i=1}^{n}(\boldsymbol{G}_{i,L}(\boldsymbol{e}_i;\boldsymbol{W}^{(0)}))_{k,\cdot}\right\|=\sum_{k=1}^{m}\mathbb{E}_{\boldsymbol{h}_{L-1}}\left\|\sum_{i=1}^{n}\boldsymbol{B}_{k,\cdot}^{T}\boldsymbol{e}_i D_{i,L,kk}\boldsymbol{h}_{i,L-1}\right\|\geq\frac{0.1p_0m}{2}\frac{(1-\alpha)^2}{n(1+\alpha^2)}\sum_{j=1}^{d}\left|\mathbb{E}_{\boldsymbol{h}_{L-1}}\sum_{i=1}^{n}\boldsymbol{e}_{i,j}\langle\boldsymbol{h}_{i,L-1},\boldsymbol{u}\rangle\right|^2. \tag{161}$$

Using (161), Assumption B.13 and the fact that $\mathbb{E}\boldsymbol{e}_{i,j}\boldsymbol{e}_{i,j'}=0$ if $j\neq j'$, when $\mathcal{L}>1$ imply

$$\begin{aligned}
&\sum_{j=1}^{d}\left|\mathbb{E}_{\boldsymbol{h}_{L-1}}\sum_{i=1}^{n}\boldsymbol{e}_{i,j}\langle\boldsymbol{h}_{i,L-1},\boldsymbol{u}\rangle\right|=\frac{1}{d}\left|\mathbb{E}\sum_{i=1}^{n}\boldsymbol{e}_{i,j}(\langle\boldsymbol{h}_{i,L-1},\boldsymbol{u}_j\rangle-y_{i,j}+\hat{y}_{i,j}+y_{i,j}-\hat{y}_{i,j})\right|\\
&\geq\frac{1}{d}\sum_{j=1}^{d}\left|\sum_{i=1}^{n}\boldsymbol{e}_{i,j}(y_{i,j}-\hat{y}_{i,j})\right|-\frac{1}{d}\sum_{j=1}^{d}\left|\sum_{i=1}^{n}\boldsymbol{e}_{i,j}(\langle\boldsymbol{h}_{i,L-1},\boldsymbol{u}_j\rangle-(y_{i,j}-\hat{y}_{i,j}))\right|\\
&\geq\frac{1}{d}\sum_{j=1}^{d}\left|\sum_{i=1}^{n}\boldsymbol{e}_{i,j}^2\right|-\frac{\lambda}{d}\sum_{j=1}^{d}\left|\sum_{i=1}^{n}\boldsymbol{e}_{i,j}\right|\geq\frac{1}{d}\mathcal{L}-\lambda\frac{\sqrt{n}}{\sqrt{d}}\mathcal{L}\geq\frac{1}{2d}\mathcal{L}.
\end{aligned}$$

Applying (161) and (162), we conclude that

$$\sum_{k=1}^{m}\left\|\sum_{i=1}^{n}\boldsymbol{B}_{k,\cdot}^{T}\boldsymbol{e}_i D_{i,L,kk}\boldsymbol{h}_{i,L-1}\right\|_2^2\geq\frac{m}{nd^2}\frac{(1-\alpha)^2}{1+\alpha^2}\mathcal{L}^2,\ \text{with probability at least}\ \ 1-e^{-\Omega(m)}. \tag{162}$$

Considering at initial parameter $\boldsymbol{W}^{(0)}$,

$$\mathcal{L}(\boldsymbol{W}^{(0)})=\sum_{i=1}^{n}\|\boldsymbol{y}_i-\boldsymbol{B}\boldsymbol{h}_{i,L}\|^2=\sum_{i=1}^{n}\|\boldsymbol{y}_i\|^2+\|\boldsymbol{B}\boldsymbol{h}_{i,L}\|^2-2\langle\boldsymbol{y}_i,\boldsymbol{B}\boldsymbol{h}_{i,L}\rangle.$$

Using the fact that $\mathbb{E}\langle\boldsymbol{y}_i,\boldsymbol{B}\boldsymbol{h}_{i,L}\rangle=0$ and Lemma B.1, we establish a lower bound for $\mathcal{L}(\boldsymbol{W}^{(0)})$,

$$\mathcal{L}(\boldsymbol{W}^{(0)})\geq\Omega(n). \tag{163}$$

Applying the definition of the gradient norms, and (162) with the lower bound of $\mathcal{L}(\boldsymbol{W}^{(0)})$ in (163), we conclude that

$$\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{(0)})\|_F^2\geq\sum_{k=1}^{m}\left\|\sum_{i=1}^{n}\boldsymbol{B}_{k,\cdot}^{T}\boldsymbol{e}_i D_{i,L,kk}\boldsymbol{h}_{i,L-1}\right\|_2^2\geq\Omega\left(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{m}{d^2}\right)\mathcal{L}(\boldsymbol{W}^{(0)}),\ \text{with probability at least}\ \ 1-e^{-\Omega(m)}.$$

□

Lemma B.19. *Assume the setup of §2 and the dataset satisfy Assumption B.13, when* $\|\boldsymbol{W}-\boldsymbol{W}^{(0)}\|<\omega<O(\frac{1}{d^{3/2}L^6\ln^{3/2}m})$*, with probability at least* $1-e^{-\Omega(m)}$*,*

$$\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W})\|\geq\Omega\left(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{m}{d^2}\right)\mathcal{L}(\boldsymbol{W}).$$

Proof. For $\boldsymbol{W}$ such that $\|\boldsymbol{W}^{(0)}-\boldsymbol{W}\|<\omega$, we use the same argument in the proof of Lemma 4.2. It is straightforward to verify that $\omega^{2/3}L^4 m\ln m/d$ (this is the right hand side of (90)) is smaller than $O(m/d^2)$ by plugging in $\omega<O(\frac{1}{d^{3/2}L^6\ln^{3/2}m})$. We conclude that for $\boldsymbol{W}$ such that $\|\boldsymbol{W}-\boldsymbol{W}^{(0)}\|<\omega$, with probability at least $1-e^{-\Omega(m)}$

$$\|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W})\|\geq\Omega\bigg(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{m}{d^2}\bigg)\mathcal{L}(\boldsymbol{W}). \tag{164}$$

□

Compared to the conclusion in Lemma 4.2 with Lemma B.19, the lower bound is improved by a factor of $\frac{n}{\delta d}$ when the dataset satisfies Assumption B.13.

Proof of Theorem B.14. We follow the exact same proof of Theorem 3.1, with the lower bound of the gradient given in Lemma B.19. It is straight-forward to derive the same inequality (16), and by the lower bound in (164), we obtain that with probability $1-e^{-\Omega(m)}$

$$\mathcal{L}^{(t+1)}\leq\bigg(1-\Omega(\frac{(1-\alpha)^2}{1+\alpha^2}\frac{\eta m}{d^2})\bigg)\mathcal{L}^{(t)}.$$

We conclude the theorem by verifying that during the training process, we always have $|\boldsymbol{W}^{(0)}-\boldsymbol{W}^{(0)}|<\omega<O(\frac{1}{d^{3/2}L^6\ln^{3/2}m})$, which satisfies the condition for both Lemmas B.19 and 4.1. Following the same argument that derives (111) and using the lower bound in (164), we achieve that

$$\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\|\leq\frac{\sqrt{1+\alpha^2}}{1-\alpha}\Omega\bigg(\frac{d}{\sqrt{m}}\bigg)\sqrt{\mathcal{L}^{\boldsymbol{W}^{(0)}}}. \tag{165}$$

By further applying (108), we verify that when $m/\ln^4 m>\frac{1+\alpha^2}{(1-\alpha)^2}\Omega(d^5nL^{12})$, we can derive that

$$\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\|\leq O\bigg(\frac{1}{d^{3/2}L^6\ln^{3/2}m}\bigg). \tag{166}$$

□

Proof of Theorem B.15. The universal bound of $\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\|$ is improved as shown in (165) and (166). Then, we can conclude the theorem by following exactly the same as the proof of Theorem 3.4 which is shown in Appendix B.9. □

C Supplemental numerical experiments and details for the previous experiments

Section C.1 provides the full details of implementation for both the previous and the new experiments. Section C.2 describes new numerical experiments.

C.1 Details of Implementation

We provide some general implementation details and also details specific to the different datasets. Two datasets are new to this section. For completeness, we repeat some information that was provided in Section 5.1.

General implementation details: Throughout the numerical experiments, we applied Algorithm 1 to initialize the parameters of the neural networks. In order to implement the rescaled leaky ReLU as given in (2), we introduce a MULTIPLIER(c) layer, which simply does element-wise multiplication with a given constant c. By combining Leaky ReLU(α) and MULTIPLIER$\big(1/\sqrt{1+\alpha^2}\big)$, we replicate the rescaled Leaky ReLU with parameter α.

In the experiments, we train the NN on the training set and report the error on a reserved testing set (we view it as an approximation for the generalization error). For the synthetic dataset, we generated additional 500 synthetic data points for the testing set. For the real dataset, we performed a standard training-testing split for each dataset.

Synthetic dataset: The architecture of the NNs that we used for the synthetic dataset is shown in Table 2. We generate 1,000 data points as the training dataset and 500 data points as the testing dataset (following the model and sampling procedure described in the main text). We train the NN with GD and a learning rate of 10^{-4}.

California housing: We use an updated version of the California housing dataset, which can be downloaded from Kaggle (`https://www.kaggle.com/datasets/camnugent/california-housing-prices`) and is licensed by CC0.

This dataset was drawn from the 1990 U.S. Census and contains 20,640 observations with 10 different characteristics. Nine of them are numerical ones (e.g., the median income for households and the median value of the houses within a block) and are given in the original dataset Pace and Barry (1997). An additional categorical characteristic is the ocean proximity. Borisov et al. (2022) used this dataset as a benchmark for regression, where one needs to predict the value of the house given the other numerical characteristics. The last characteristic, which is the median house value for households within a block, provides labels for the dependent variable. We follow a similar setting of regression, but we also use the categorical feature of the updated dataset. We standardize the 9 numerical characteristics using the respective means and standard deviations of the training data. We generated a one-hot coding vector for the feature "ocean proximity", including 5 categories, <1H OCEAN, INLAND, NEAR OCEAN, NEAR BAY, and ISLAND. In total, the input $\boldsymbol{x}$ is a 13−dimensional vector. The training data contains 15,480 data points and the testing data contains 5,160 data points.

We built NNs to predict the median housing value in the dataset. The architecture of the NNs is given in Table 3. We applied Algorithm 3 with a batch size of 512 and a learning rate of 10^{-5} to train the NNs.

Table 2: Architecture of the NNs with Leaky ReLU parameter α used for the synthetic dataset.

Layer		Parameter
	Linear	(5, 5000)
Repeat 5 times	Linear	(5000, 5000)
	Leaky Relu	α
	Multipler	$1/(1+\alpha^2)^{1/2}$
	Linear	(5000, 1)

MNIST: We used the MNIST dataset of 28 × 28 images of handwritten digits in order to classify handwritten digits. This dataset is licensed by CC BY-SA 3.0. We flattened each image to a vector of length 784. We randomly sample 2,100 data points from the training set of MNIST as our training set, and use the rest as our testing set. We normalized the training data with 0.5 mean and 0.5 standard deviation. We applied SGD with batch size 64 and a learning rate of 10^{-3}. The architecture of the NN is presented in Table 4 and we use leaky ReLUs with $\alpha \in \{-2,-1,0,0.01,0.05\}$. MNIST was also used to test the Transformer networks. In this case, we normalized the training set with 0.1307 mean and 0.3081 standard deviation. Furthermore, we used the Vision Transformer (ViT) (Dosovitskiy et al., 2020) architecture and applied SGD with batch size 100 and learning rate 10^{-4}. The details of this architecture are shown in Table 5.

F-MNIST: We used the F-MNIST dataset of 28 × 28 images of clothing or accessory items for classification. This dataset is licensed by MIT. We flattened each image to vectors of length 784. We randomly sample 3000 data points from the training set of F-MNIST as the training set, and use the rest for testing. We normalized the training data with 0.5 mean and 0.5 standard deviation. We applied SGD with batch size 64 and a learning rate of 10^{-5}. The architecture of the NN is presented in Table 6 and we use leaky ReLUs with $\alpha \in \{-2,-1,0,0.01,0.05\}$.

CIFAR-10: We used the CIFAR-10 dataset of 32×32 RGB images with 10 categories for classification. This dataset is licensed by MIT. We randomly sample 2560 data points from the training set of CIFAR-10 as our training set, and randomly sample 2560 data points from the testing set of CIFAR-10 as our testing set. We normalized the training data with 0.5 mean and 0.5 standard deviation. Then we applied SGD with batch size 64 and a learning rate of 10^{-6}. The architecture of the NN is presented in Table 7.

IMDB movie reviews dataset: This is a dataset of highly popular movie review paragraphs and it is used for positive or negative sentiment classification (Maas et al., 2011). We downloaded it from the following URL: `http://ai.stanford.edu/~amaas/data/sentiment/`. We randomly sample 5,000 data points from the IMDB movie reviews as the training dataset and use the rest as the testing dataset. We processed the data as follows. We first recorded the words that appeared at least once in the training dataset. For each word, a unique integer was assigned to index it. Then we mapped each paragraph to a vector, whose i-th entry is the assigned index of the i-th word of the paragraph. Finally, we padded each vector with zeros, so that each vector was of the same length. We used the zero-padded vectors as the input of our neural network. After preprocessing, we applied SGD with batch size 50 and learning rate 10^{-5} with an LSTM network, whose architecture is presented in Table 8.

C.2 Additional numerical results

We describe here two additional experiments.

Table 3: Architecture of the NNs with Leaky ReLU parameter α used for California housing.

	Layer	Parameter
	Linear	(13,5000)
Repeat 7 times	Linear	(5000,5000)
	Leaky Relu	α
	Multipler	$1/(1+\alpha^2)^{1/2}$
	Linear	(5000,1)

Table 4: Architecture of the neural networks with α Leaky ReLU parameter used for MNIST.

Layer	Parameter
Linear	(784, 2000)
Linear	(2000, 2000)
Leaky ReLU	α
Multiplier	$\frac{1}{\sqrt{1+\alpha^2}}$
Linear	(2000, 2000)
Leaky ReLU	α
Multiplier	$\frac{1}{\sqrt{1+\alpha^2}}$
Linear	(2000, 10)

Table 5: Architecture of the Transformer neural networks with α Leaky ReLU parameter used for IMDB movie reviews.

Layer	Parameter
Positional Embedding	(49, 64)
Transformer Encoder	head dim = 64, output dim = 64, number of heads = 8 number of layers = 6 mlp dim = 8192
Linear	(64,8192)
Leaky ReLU	α
Multiplier	$\frac{1}{\sqrt{1+\alpha^2}}$
Linear	(8192, 10)

Table 6: Architecture of the neural networks with α Leaky ReLU parameter with width of m and depth of L used for Fashion MNIST.

	Layer	Parameter
	Linear	(784,m)
Repeat L times	Linear	(m,m)
	Leaky Relu	α
	Multipler	$1/(1+\alpha^2)^{1/2}$
	Linear	(m,10)

Additional experiments using MNIST and California housing: We ran the same experiments done in Section 5.1, but with MNIST and California housing. Figure 3 demonstrates the training errors (top) and testing errors (bottom) for the two datasets. For MNIST (left) we used the cross entropy loss and for California housing (right) the MSE loss. For the training errors, we note that the convergence rate is the fastest at $\alpha=-1$ for both datasets and this observation aligns with our theoretical prediction and the previous experiments. The testing errors are rather similar for different choices of α. For the California housing dataset, we note that $\alpha=-1$ achieves the smallest testing error at an early epoch (about $t=20$), but the advantage is marginal compared to other α's. For MNIST, we note that $\alpha=-1$ gets to the

Table 7: Architecture of the neural networks with α Leaky ReLU parameter used for CIFAR-10.

Layer	Parameter
CNN	conv3-512
Leaky ReLU	α
Multiplier	$\frac{1}{\sqrt{1+\alpha^2}}$
CNN	conv3-512
Leaky ReLU	α
Multiplier	$\frac{1}{\sqrt{1+\alpha^2}}$
Max Pooling	2×2
CNN	conv3-512
Leaky ReLU	α
Multiplier	$\frac{1}{\sqrt{1+\alpha^2}}$
CNN	conv3-512
Leaky ReLU	α
Multiplier	$\frac{1}{\sqrt{1+\alpha^2}}$
Max Pooling	2×2
Flatten	
Linear	(32,768, 512)
Leaky ReLU	α
Multiplier	$\frac{1}{\sqrt{1+\alpha^2}}$
Linear	(512, 10)

Table 8: Architecture of the neural networks with α Leaky ReLU parameter used for IMDB movie reviews dataset

Layer	Parameter
Embedding	(1000, 64)
LSTM	input dim = 64, hidden dim = 256, number of layers = 2
Dropout	0.3
Linear	(256,4096)
Leaky ReLU	α
Multiplier	$\frac{1}{\sqrt{1+\alpha^2}}$
Linear	(4096, 1)

same level of testing error as the other αs from a much larger initial testing error. We note that $\alpha=-1$ is not optimal for the testing error. This might happen because the number of samples and the depth are not sufficiently large enough.

Experiments with Long Short-Term Memory (LSTM) and Transformer networks: We ran the same experiment done in Section 5 on MNIST and IMDB with Transformer and LSTM networks, respectively. These architectures are described in Section C.1. Figure 4 demonstrates the training errors (top) and testing errors (bottom). For MNIST (right) we used the negative log likelihood loss and for IMDB (left) we used the binary cross entropy loss. For both algorithms, the training errors converge fastest with $\alpha=-1$. This observation agrees with our theoretical findings and previous experiments. The testing error for IMDB decreases during the first 100 epochs and then increases for the rest of the training. This is because of severe overfitting that is due to the following property of IMDB: the training dataset is small (we used randomly sampled 5,000 data points to be able to deal with sufficiently large widths for overparameterization) compared to the input data dimension (we have 1,000 unique words). The testing error on MNIST, on the other hand, is also the lowest for $\alpha=-1$, but there's no overfitting phenomenon since MNIST is a simple dataset.

Dependence on m and L: We demonstrate the dependence of the training error on m and the testing error on L. We thus ran additional experiments on F-MNIST with different choices of L and m and $\alpha\in\{-2,-1,0,0.01,0.05\}$. First, we fixed $L=2$ and tested $m\in\{1{,}000,2{,}000,5{,}000,10{,}000\}$. Next, we fixed $m=5{,}000$ and tested $L\in\{2,3,4\}$. The architectures of thes NNs were given in Table 6, but with the latter choices of m and L.

Figure 5 shows the dependence of training errors on m and the dependence of testing errors on L. We note that the

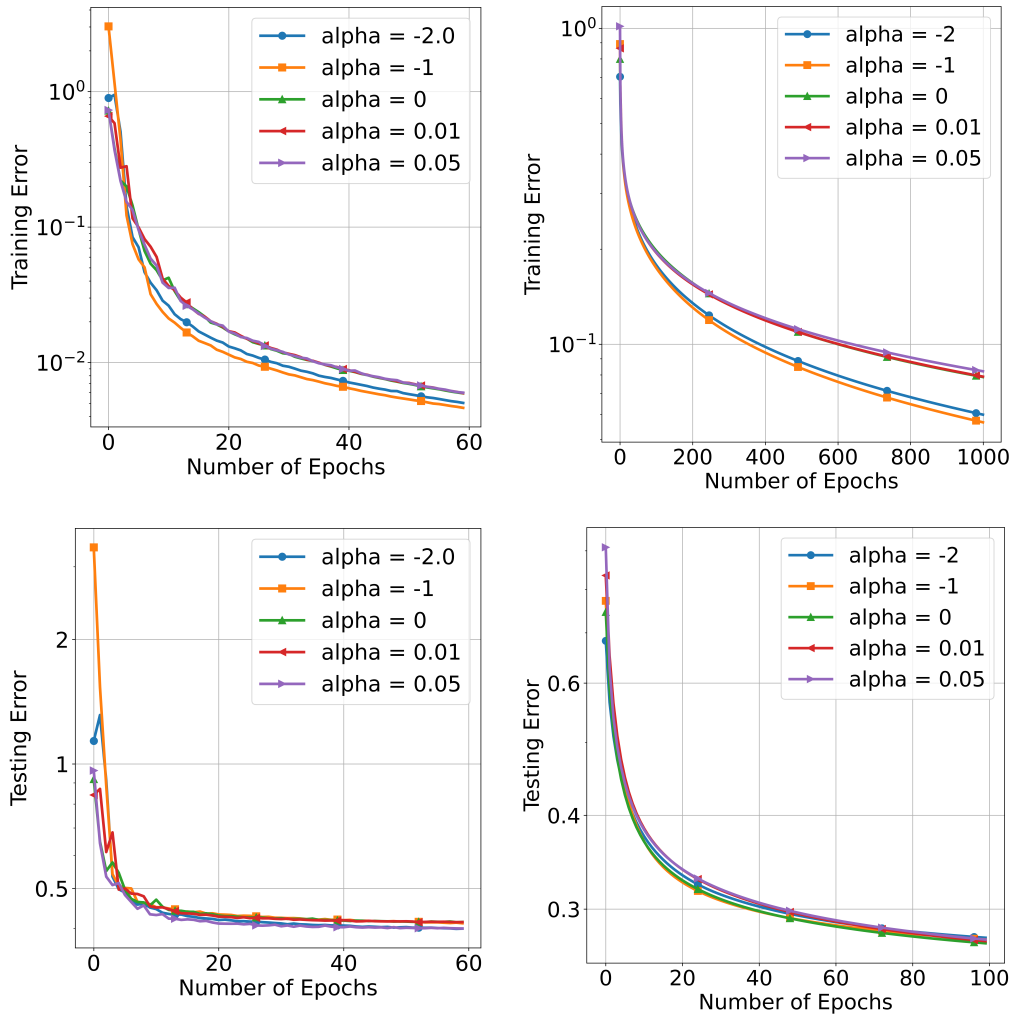

Figure 3: Log-scale training and testing errors using different datasets and different α's. Left: cross entropy errors for MNIST; Right: MSE for California housing. Top row: training errors. Bottom row: testing errors.

training error is monotonically decreasing w.r.t. the width m. We also note that the minimal training error is always achieved at $\alpha=-1$ for different choices of m. This matches our theoretical predictions. We note that the testing error is decreasing for $L\leq 4$. This aligns with equation (134) when t is small. Moreover, when $L=4$, we observe that the minimal testing error is achieved at $\alpha=-1$.

Results using the loss function given in (137): We perform numerical experiments using the exponential loss function in the two datasets for the regression task, the synthetic dataset and the California housing dataset. Figure 6 reports the results. We observe that $\alpha=-1$ achieves both the optimal training error and the optimal generalization error in the synthetic dataset. Furthermore, $\alpha=-1$ achieves optimal training error and second optimal generalization error for the California housing dataset, though the difference in the testing error is small.

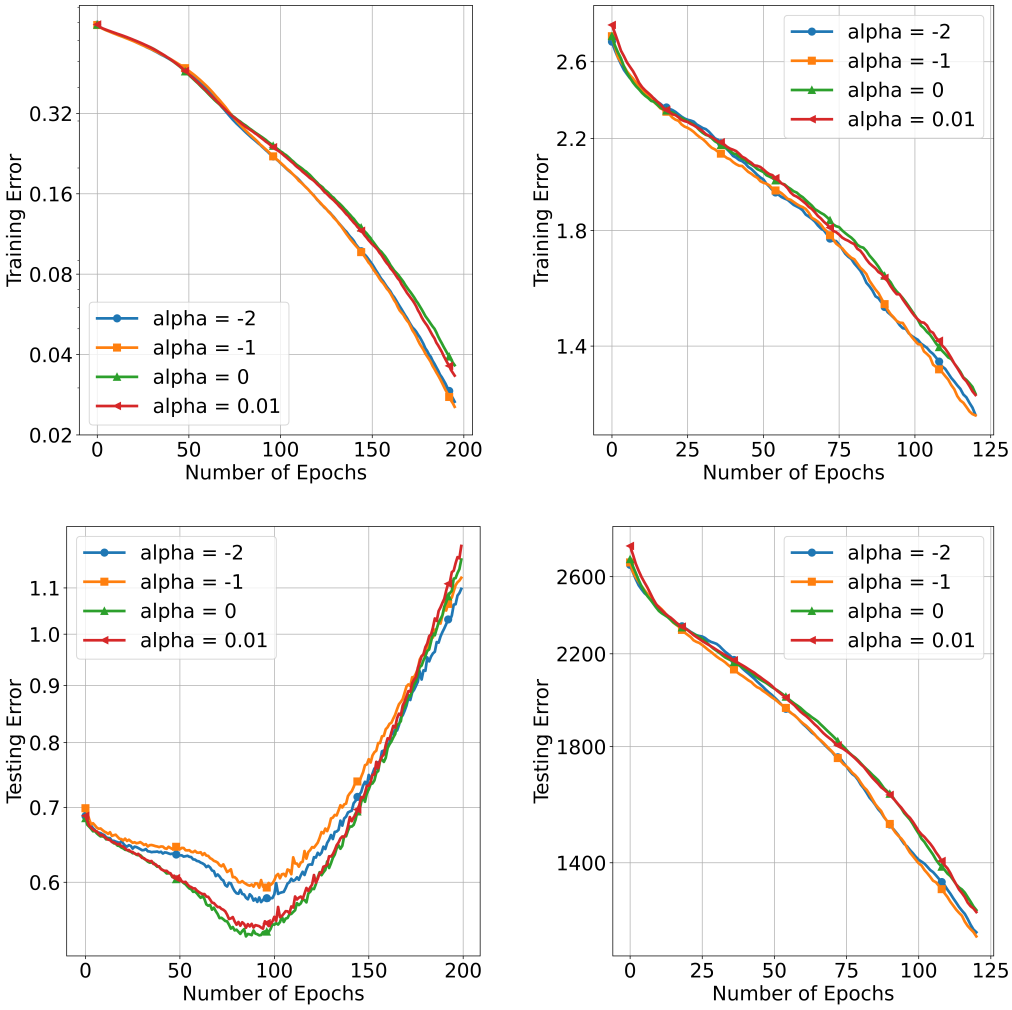

Figure 4: Log-scale training and testing errors using different datasets and different α's. Left: binary entropy errors for IMDB; Right: negative log likelihood errors for Transformer on MNIST. Top row: training errors. Bottom row: testing errors.

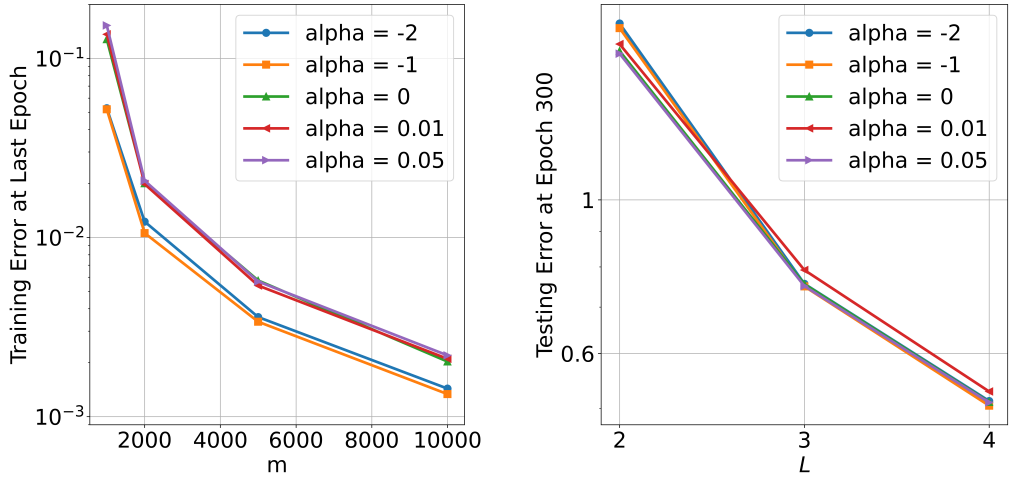

Figure 5: Log-scale errors on F-MNIST with different α's. Left: training errors at the last epoch with $L=2$ and different widths (ms); Right: testing errors at the epoch $t=300$ with $m=5000$ and different depths (Ls).

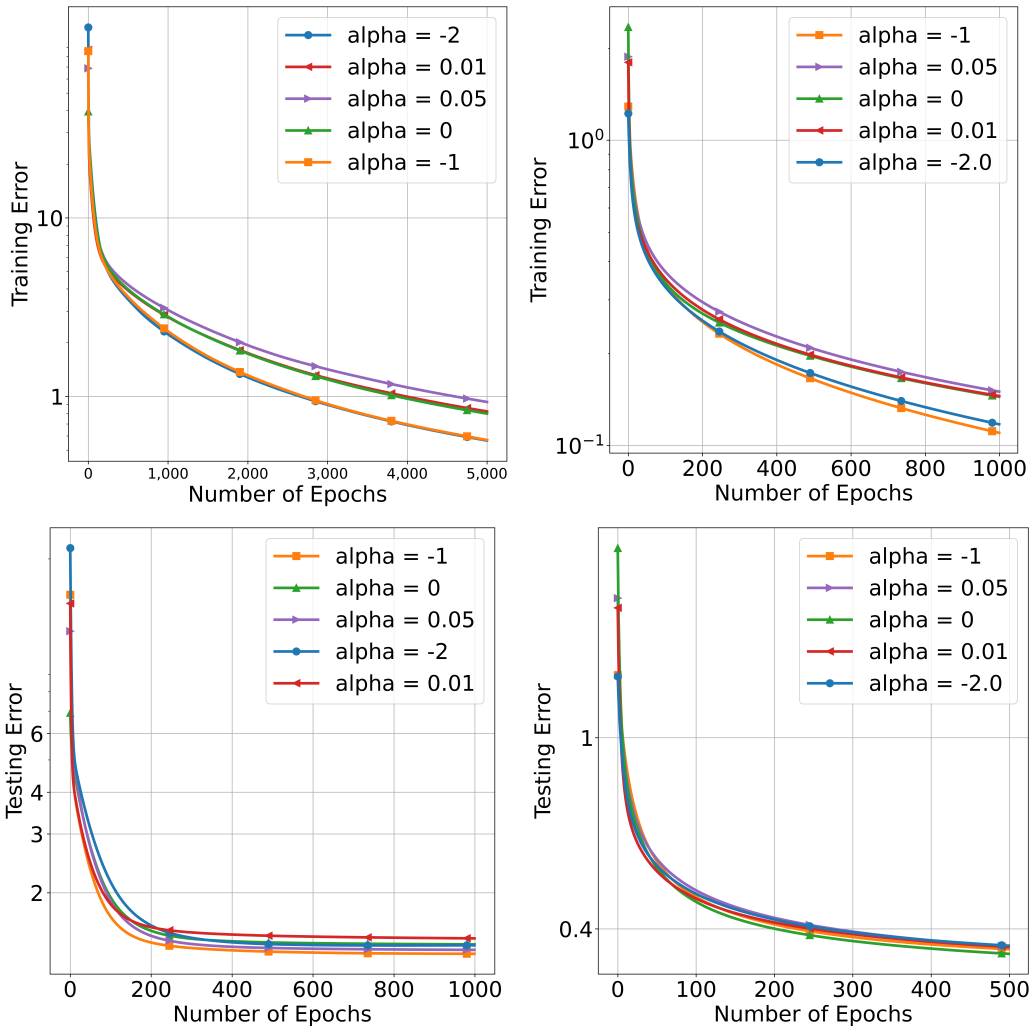

Figure 6: Log-scaled mean square errors for the synthetic dataset with different α's using the loss function of (137). Left: MSE for the synthetic dataset. Right: MSE for California housing dataset. Top row: training errors. Bottom row: testing errors.