Proceedings of Machine Learning Research vol 134:1–101, 2021 34rd Annual Conference on Learning Theory

When Does Gradient Descent with Logistic Loss Interpolate Using Deep Networks with Smoothed ReLU Activations?

Niladri S. Chatterji CHATTERJI@BERKELEY.EDU
University of California, Berkeley

Philip M. Long PLONG@GOOGLE.COM
Google

Peter L. Bartlett PETER@BERKELEY.EDU
University of California, Berkeley & Google

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We establish conditions under which gradient descent applied to fixed-width deep networks drives the logistic loss to zero, and prove bounds on the rate of convergence. Our analysis applies for smoothed approximations to the ReLU, such as Swish and the Huberized ReLU, proposed in previous applied work. We provide two sufficient conditions for convergence. The first is simply a bound on the loss at initialization. The second is a data separation condition used in prior analyses.

Keywords: optimization guarantees, neural networks, interpolating methods, binary classification, deep learning.

1. Introduction

Interest in the properties of interpolating deep learning models trained with first-order optimization methods is surging (Zhang et al., 2017a; Belkin et al., 2019). One important question is to understand how gradient descent with appropriate random initialization routinely finds interpolating (near-zero training loss) solutions to these non-convex optimization problems.

In this paper our focus is to understand when gradient descent drives the logistic loss to zero when applied to fixed-width deep networks using smooth approximations to the ReLU activation function. We derive upper bounds on the rate of convergence under two conditions. The first result only requires that the initial loss is small, but does not require any assumption about the width of the network. It guarantees that if the initial loss is small then gradient descent drives the logistic loss down to zero. The second result is under a separation condition on the data. Under this assumption we demonstrate that the loss decreases adequately in the initial iterations such that the first result applies.

A few ideas that facilitate our analysis are as follows: under the first set of assumptions, when the loss is small, we show that the negative gradient aligns with the weights of the network. This lower bounds the norm of the gradient at the beginning of the gradient step and implies that the loss decreases quickly at the beginning of the step. We then show that the loss is smooth in the neighborhood of the beginning of the step. The smoothness of the loss combined with the lower

bound on the norm of the gradient at the beginning of the step implies that the loss decreases throughout the gradient step when the step-size is small enough.

The second sufficient condition is when the data is separable by a margin using the features obtained by the gradient of the neural network at initialization (see Assumption 3.2). This assumption has previously been studied by Chen et al. (2021). Intuitively, it is weaker than an assumption that the training examples are not too close, as we discuss after its definition. Under this assumption we use a neural tangent kernel (NTK) analysis to show that the loss decreases sufficiently in the first stage of optimization such that we can invoke our first result to guarantee that the loss decreases thereafter in the second stage. To analyze this first stage we borrow ideas from (Allen-Zhu et al., 2019; Zou et al., 2020), because the formulation of their results was most closely aligned with our needs. However we note that their results do not directly apply since they study networks with ReLU activations while we study smooth approximations to the ReLU. In addition to adapting their proofs to our setting, we also worked out some details in the original proofs.

Our first result could be viewed as a tool to establish convergence under a wide variety of conditions. Our second result is one example of how it may be applied. Other separation assumptions on the data like the ones studied by Ji and Telgarsky (2019c); Chen et al. (2021); Zou et al. (2020), could also be used in conjunction with our first result to establish convergence to zero training loss.

Recently Chatterji et al. (2020) showed that gradient descent applied to two-layer neural networks drives the logistic loss to zero when the initial loss is small and the activation functions are Huberized ReLUs. Our work can be viewed as a generalization of their result to the case of deep networks.

Previously, Lyu and Li (2020) studied the margin maximization of ReLU networks for the logistic loss. They also proved that gradient descent applied to deep networks drives the training logistic loss to zero. However, their result requires the neural network to be both positive homogeneous and smooth (see, for example, the proof of Lemma E.7 of their paper), so that a substantially different analysis was needed here. Their assumptions rule out the ReLU and close approximations to it like Swish (Ramachandran et al., 2018) or the Huberized ReLU (Tatro et al., 2020) that are widely used in practice. Their results do apply in case that the ReLU is raised to a power strictly greater than two. As far as we know, the analysis of the alignment between the negative gradient and the weights originated in their paper: in this paper, we establish such alignment under weaker conditions.

Prior work has shown that gradient descent drives the squared loss of fixed-width deep networks to zero (Du et al., 2018, 2019; Allen-Zhu et al., 2019; Oymak and Soltanolkotabi, 2020), using the NTK perspective (Jacot et al., 2018; Chizat et al., 2019). The logistic loss however is qualitatively different. Driving the logistic loss to zero requires the weights to go to infinity, far from their initial values. This means that a Taylor approximation around the initial values cannot be applied. While the NTK framework has also been applied to analyze training with the logistic loss, a typical result (Li and Liang, 2018; Allen-Zhu et al., 2019; Zou et al., 2020) is that after $\mathsf{poly}(1/\varepsilon)$ updates, a network of size or width $\mathsf{poly}(1/\varepsilon)$ achieves ε loss. Thus to guarantee loss very close to zero, these analyses require larger and larger networks. The reason for this appears to be that a key part of these analyses is to show that a wider network can achieve a certain fixed loss by traveling a shorter distance in parameter space. Since, to drive the logistic loss to zero with a fixed-width network, the parameters must travel an unbounded distance, it seems that the NTK approach cannot be applied to obtain the results of this paper.

The remainder of the paper is organized as follows. In Section 2 we introduce notation and definitions. In Section 3 we present our main theorems. We provide a proof of our first result,

Theorem 3.1, in Section 4. We conclude with a discussion in Section 5. Appendix A points to other related work. The proof of our second result, Theorem 3.3, and other technical details, are presented in the remaining appendices.

2. Preliminaries

This section includes notational conventions and a description of the setting.

2.1. Notation

Given a vector v, let $\|v\|$ denote its Euclidean norm, $\|v\|_p$ denote its ℓ_p-norm for any $p \geq 1$, $\|v\|_0$ denote the number of non-zero entries, and $\mathrm{diag}(v)$ denote a diagonal matrix with v along the diagonal. We say a vector v is k-sparse if $\|v\|_0 \leq k$. Given a matrix M, let $\|M\|$ denote its Frobenius norm, $\|M\|_{op}$ denote its operator norm and $\|M\|_0$ denote the number of non-zero entries in the matrix. Given either a matrix or a tensor we let $\mathrm{vec}(\cdot)$ be its vectorization. Given a tensor T, let $\|T\| = \|\mathrm{vec}(T)\|$; we will sometimes call this the Frobenius norm of T. If, for matrices $T_1, \ldots, T_{L+1}$ of different shapes, we refer to them collectively as T, we define $\|T\|$ analogously. Given two tensors A and B let $A \cdot B$ denote the element-wise dot product $\mathrm{vec}(A) \cdot \mathrm{vec}(B)$. For any $k \in \mathbb{N}$, we denote the set $\{1, \ldots, k\}$ by $[k]$. For a number p of inputs, we denote the set of unit-length vectors in $\mathbb{R}^p$ by $\mathbb{S}^{p-1}$. We use the standard "big Oh notation" (see, e.g., Cormen et al., 2009). We will use $c, c', c_1, \ldots$ to denote constants, which may take different values in different contexts.

For a function J of a tensor V, we denote the gradient of J at V by $\nabla_V J(V)$, and define $\mathrm{Lip}(\nabla_V J(V))$ to be the local Lipschitz constant of $\nabla_V J(V)$, as a function of V, with respect to the Euclidean norm. That is

$$\mathrm{Lip}(\nabla_V J(V)) = \limsup_{W \to V} \frac{\|\nabla_V J(V) - \nabla_W J(W)\|}{\|V - W\|}.$$

2.2. The Setting

We will analyze gradient descent applied to minimize the training loss of a multi-layer network.

We assume that the number of inputs is equal to the number of hidden nodes per layer to simplify the presentation of our results. Our techniques can easily extend to the case where there are different numbers of hidden nodes in different layers. Let p denote the number of inputs and the number of hidden nodes per layer, and let L denote the number of hidden layers.

We will denote the activation function by ϕ. Given a vector v let $\phi(v)$ denote a vector with the activation function applied to each coordinate. We study activation functions that are similar to the ReLU activation function but are smooth.

Definition 2.1 *A activation function ϕ is h-smoothly approximately ReLU if,*

- *the function ϕ is differentiable;*
- $\phi(0) = 0$*;*
- ϕ' *is* $\frac{1}{h}$*-Lipschitz and* $|\phi'(z)| \leq 1$*;*
- *for all* $z \in \mathbb{R} : |\phi'(z)z - \phi(z)| \leq h/2$.

It may aid intuition to note that, for small h, the condition that $|\phi'(z)z - \phi(z)| \leq h/2$ can be paraphrased to say that a first-order Taylor approximation of ϕ at z is accurate at the origin. It is easy to verify the activation functions ϕ are contractive with respect to the Euclidean norm. That is, for any $v_1, v_2 \in \mathbb{R}^p$, $\|\phi(v_1) - \phi(v_2)\| \leq \|v_1 - v_2\|$. See Lemma B.9. Here are a couple of examples of activation functions that are h-smoothly approximately ReLU.

1. Huberized ReLU (Tatro et al., 2020):

$$\phi(z) := \begin{cases} 0 & \text{if } z < 0, \\ \frac{z^2}{2h} & \text{if } z \in [0, h], \\ z - \frac{h}{2} & \text{otherwise.} \end{cases} \tag{1}$$

2. Scaled Swish (Ramachandran et al., 2018): $\phi(z) = \frac{z}{1.1(1+\exp(-2z/h))}$. The scaling factor $1/1.1$ ensures that $|\phi'(z)| \leq 1$.

For $i \in \{1, \ldots, L\}$, let $V_i \in \mathbb{R}^{p \times p}$ be the weight matrix of the ith layer and let $V_{L+1} \in \mathbb{R}^{1 \times p}$ be the weight vector corresponding to the outer layer. Let $V = (V_1, \ldots, V_{L+1})$ consist of all of the trainable parameters in the network. Let f_V denote the function computed by the network, which maps x to

$$f_V(x) = V_{L+1}\phi\left(V_L \cdots \phi(V_1 x)\right).$$

Consider a training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{S}^{p-1} \times \{-1, 1\}$. For any sample $s \in [n]$, define $u^V_{0,s} = x^V_{0,s} := x_s$ and for all $\ell \in [L]$, define

$$u^V_{\ell,s} := V_\ell x^V_{\ell-1,s} \quad \text{and,} \quad x^V_{\ell,s} := \phi\left(V_\ell x^V_{\ell-1,s}\right),$$

that is, $u^V_{\ell,s}$ refers to the pre-activation features in layer ℓ, while $x^V_{\ell,s}$ corresponds to the features after applying the activation function in the ℓth layer. Also for any $\ell \in [L]$ and $s \in [n]$ let

$$\Sigma^V_{\ell,s} := \operatorname{diag}\left(\phi'(u_{\ell,s})\right) = \operatorname{diag}\left(\phi'\left(V_\ell x^V_{\ell-1,s}\right)\right).$$

Define the training loss (empirical risk with respect to the logistic loss) J by

$$J(V) := \frac{1}{n}\sum_{s=1}^{n} \log(1 + \exp\left(-y_s f_V(x_s)\right)),$$

and refer to loss on example s by

$$J(V; x_s, y_s) := \log(1 + \exp\left(-y_s f_V(x_s)\right)).$$

The gradient of the loss evaluated at V is

$$\nabla_V J(V) = \frac{1}{n}\sum_{s=1}^{n} \frac{-y_s \nabla_V f_V(x_s)}{1 + \exp\left(y_s f_V(x_s)\right)},$$

and the partial gradient of f_V with respect to V_ℓ has the form (see, e.g., Zou et al., 2020)

$$\frac{\partial f_V(x_s)}{\partial V_\ell} = \left(\Sigma^V_{\ell,s} \prod_{j=\ell+1}^{L} \left(V_j^\top \Sigma^V_{j,s}\right)\right) V_{L+1}^\top x^{V\top}_{\ell-1,s}, \quad \text{when } \ell \in [L], \tag{2a}$$

$$\frac{\partial f_V(x_s)}{\partial V_{L+1}} = x^{V\top}_{L,s}. \tag{2b}$$

We analyze the iterates of gradient descent $V^{(1)}, V^{(2)}, \ldots$ defined by

$$V^{(t+1)} := V^{(t)} - \alpha_t \nabla_V J|_{V=V^{(t)}}$$

in terms of the properties of $V^{(1)}$.

Definition 2.2 *For all iterates t, define $J_{ts} := J(V^{(t)}; x_s, y_s)$ and let $J_t := \frac{1}{n}\sum_{s=1}^{n} J_{ts}$. Additionally for all t, define $\nabla J_t := \nabla_V J|_{V=V^{(t)}}$.*

3. Main Results

In this section we present our theorems and discuss their implications.

3.1. A General Result

Given the initial weight matrix $V^{(1)}$, width p, depth L, and training data $\{x_s, y_s\}_{s\in[n]}$, define $h_{\max}$, $\alpha_{\max}$ and $\widetilde{Q}$ below:

$$h_{\max} := \min\left\{\frac{L^{\frac{L}{2}-3}\log(1/J_1)}{24\sqrt{p}\|V^{(1)}\|^L}, 1\right\}, \tag{3a}$$

$$\alpha_{\max}(h) := \min\left\{\frac{h}{1024\,(L+1)^2\,pJ_1\|V^{(1)}\|^{3L+5}}, \frac{(L+\frac{1}{2})\|V^{(1)}\|^2}{2L(L+\frac{3}{4})^2 J_1 \log^{\frac{2}{L}}(1/J_1)}\right\}, \text{ and} \tag{3b}$$

$$\widetilde{Q}(\alpha) := \frac{L(L+\frac{3}{4})^2\alpha J_1 \log^{\frac{2}{L}}(1/J_1)}{(L+\frac{1}{2})\|V^{(1)}\|^2}. \tag{3c}$$

Theorem 3.1 *For any $L \geq 1$, for all $n \geq 3$, for all $p \geq 1$, for any initial parameters $V^{(1)}$ and dataset $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{S}^{p-1}\times\{-1, 1\}$, for any h-smoothly approximately ReLU activation function with $h < h_{\max}$, any positive $\alpha \leq \alpha_{\max}(h)$ and positive $Q \leq \widetilde{Q}(\alpha)$ the following holds for all $t \geq 1$. If each step-size $\alpha_t = \alpha$, and if $J_1 < 1/n^{1+24L}$ then, for all $t \geq 1$,*

$$J_t \leq \frac{J_1}{Q\cdot(t-1)+1}.$$

We reiterate that this theorem makes no assumption about the width p of the network and makes a very mild assumption on the number of samples required: $n \geq 3$. The only other assumption is that the initial loss is less than $1/n^{1+24L}$. We pick the step-size to be a constant, which leads to a rate that scales with $1/t$.

Next we provide an example where we show that it is possible to arrive at a small loss solution using gradient descent starting from randomly initialized weight matrices.

3.2. Small Loss Guarantees Using NTK Techniques

In this subsection assume that the entries of the initial weight matrices for the layers $\ell \in \{1, \ldots, L\}$ are drawn independently from $\mathcal{N}(0, 2/p)$, and the entries of $V_{L+1}^{(1)}$ are drawn independently from $\mathcal{N}(0, 1)$. In this section we also specialize to the case where the activation function is the Huberized ReLU (see its definition in equation (1)). We make the following assumption on the training data.

Assumption 3.2 *With probability $1-\delta$ over the random initialization, there exists a collection of matrices $W^\star = (W_1^\star, \ldots, W_{L+1}^\star)$ with $\|W^\star\| = 1$, such that for all samples $s \in [n]$*

$$y_s \left(\nabla f_{V^{(1)}}(x_s) \cdot W^\star\right) \geq \sqrt{p}\gamma,$$

for some $\gamma > 0$.

The scaling factor $\sqrt{p}$ on the right hand side is to balance the scale of the norm of the gradient at initialization which will scale with $\sqrt{p}$ as well. This is because the entries of the final layer $V_{L+1}^{(1)}$ are drawn independently from $\mathcal{N}(0,1)$. This assumption is inspired by Assumption 4.1 made by Chen et al. (2021). This assumption can be seen to be implied by stronger conditions that simply require that the training examples are not too close, as employed in (Allen-Zhu et al., 2019; Zou et al., 2020). Here is some rough intuition of why. The components of $\nabla f_{V^{(1)}}(x_s)$ include values computed at the last hidden layer when x_s is processed using $V^{(1)}$ (that is, $\nabla_{V_{L+1}} f_{V^{(1)}}(x_s) = x_{L,s}^{V^{(1)}}$). For wide networks with Huberized ReLU activations, if the values of x_s in the training examples do not have duplicates, their embeddings into the last hidden layer of nodes are in general position with high probability. In fact, the Gaussian Process analysis of infinitely wide deep networks at initialization (Matthews et al., 2018a,b) suggests that, for wide networks, the embeddings will not even be close to failing to be in general position (see Agarwal et al., 2021). If the width $p \gg n$, results from (Cover, 1965) show that they will be linearly separable. The anti-concentration conferred by the Gaussian initialization promotes larger (though not necessarily constant) margins. Assumption 3.2 is more refined than a separation condition, since it captures a sense in which the data is amenable to treatment with neural networks that enables us to provide stronger guarantees in such cases. Furthermore, in Appendix C we show that Assumption 3.2 is satisfied with a constant margin γ by two-layer networks with Huberized ReLUs for data satisfying a clustering condition. Finally, we note that we could also use other assumptions on the data that have been studied in the literature (for example by, Ji and Telgarsky, 2019c) to guarantee that the loss reduces below $1/n^{1+24L}$, as required to invoke Theorem 3.1. However, we provide guarantees only under this assumption in the interest of simplicity.

Define

$$\rho := \frac{c_1}{\sqrt{p}\gamma}\left[\sqrt{\log\left(\frac{n}{\delta}\right)} + \log\left(6n^{(2+24L)}\right)\right], \tag{4}$$

where $c_1 \geq 0$ is a large enough absolute constant. Also set the value of

$$h = h_{\mathsf{NT}} := \frac{(1+24L)\log(n)}{6(6p)^{\frac{L+1}{2}}L^3}. \tag{5}$$

With these choices of ρ and h we are now ready to state our convergence result under Assumption 3.2. The proof of this theorem is presented in Appendix D.

Theorem 3.3 *Consider a network with Huberized ReLU activations. There exists $r(n, L, \delta) = \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ such that for any $L \geq 1$, $n \geq 3$, $\delta > 0$, under Assumption 3.2 with $\gamma \in (0, 1]$ if $h = h_{\mathsf{NT}}$ and $p \geq \frac{r(n,L,\delta)}{\gamma^2}$ then both of the following hold with probability at least $1 - 4\delta$ over the random initialization:*

1. *For all $t \in [T]$, set the step-size $\alpha_t = \alpha_{\mathsf{NT}} = \Theta\left(\frac{1}{pL^5}\right)$, where $T = \left\lceil \frac{3(L+1)\rho^2 n^{2+24L}}{2\alpha_{\mathsf{NT}}} \right\rceil$. Then*

$$\min_{t\in[T]} J_t < \frac{1}{n^{1+24L}}.$$

2. *Set $V^{(T+1)} = V^{(s)}$, where $s \in \arg\min_{s\in[T]} J(V^{(s)})$, and for all $t \geq T+1$, set the step-size $\alpha_t = \alpha_{\max}(h)$. Then for all $t \geq T+1$,*

$$J_t \leq O\left(\frac{L^{\frac{3L+11}{2}}(6p)^{2L+5}}{n^{1+24L}\cdot(t-T-1)}\right).$$

We invite the reader to interpret the result of this theorem in two scenarios. The first is where the depth L is a constant and the margin $\gamma \geq \left(p^{\omega}\mathsf{poly}\left(n, \log\left(\frac{1}{\delta}\right)\right)\right)^{-1}$, for some constant $\omega \in [0, \frac{1}{2})$. In this case the conditions of Theorem 3.3 are satisfied for $p = \mathsf{poly}\left(n, \log\left(\frac{1}{\delta}\right)\right)$, and, for such p, the rate of convergence in the second stage is

$$J_t \leq O\left(\frac{L^{\frac{3L+11}{2}}(6p)^{2L+5}}{n^{1+24L}\cdot(t-T-1)}\right) \leq \frac{\mathsf{poly}\left(n, \log\left(\frac{1}{\delta}\right)\right)}{t}.$$

Another scenario is where the margin γ is at least a constant. Here it suffices for the width $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$. Thus if the number of samples $n \geq \left[L^{\frac{3L+11}{2}}(6p)^{2L+5}\right]^{\frac{1}{1+24L}}$ then the rate of convergence in this second stage is

$$J_t \leq O\left(\frac{L^{\frac{3L+11}{2}}(6p)^{2L+5}}{n^{1+24L}\cdot(t-T-1)}\right) = O\left(\frac{1}{t-T-1}\right).$$

4. Proof of Theorem 3.1

In this section, we prove Theorem 3.1.

4.1. Technical Tools

In this subsection we assemble several technical tools required to prove Theorem 3.1. Their proofs (which in turn depend on additional, more basic, lemmas) can be found in Appendix B.

We start with the following lemma, which is a slight variant of a standard inequality, and provides a bound on the loss after a step of gradient descent when the loss function is locally smooth.

Lemma 4.1 *For $\alpha > 0$, let $V^{(t+1)} = V^{(t)} - \alpha\nabla J_t$. If, for all convex combinations W of $V^{(t)}$ and $V^{(t+1)}$, we have $\mathrm{Lip}(\nabla_W J(W)) \leq M$, then if $\alpha \leq \frac{1}{(L+\frac{1}{2})M}$, we have*

$$J_{t+1} \leq J_t - \frac{\alpha L\|\nabla J_t\|^2}{L+\frac{1}{2}}.$$

To apply Lemma 4.1 we need to show that the loss J is smooth near J_t. The following lemma establishes the smoothness of J, if the weights are large enough. (We will be able to apply it, since the weights must be fairly large to achieve small loss.)

Lemma 4.2 *If $h \leq 1$, for any weights V such that $\|V\| \geq \sqrt{L+1/2}$, we have*

$$\mathrm{Lip}(\nabla_V J(V)) \leq \frac{256(L+1)\sqrt{p}\|V\|^{3L+5}J(V)}{h}.$$

Next, we show that J changes slowly in general, and especially slowly when it is small.

Lemma 4.3 *For any weight matrix V such that $\|V\| \geq \sqrt{L+1/2}$ then*

$$\|\nabla_V J(V)\| \leq \sqrt{(L+1)p}\|V\|^{L+1}\min\{J(V), 1\}.$$

The following lemma applies Lemma 4.1 (along with Lemma 4.2) to show that if the step-size at step t is small enough then the loss decreases by an amount that is proportional to the squared norm of the gradient.

Lemma 4.4 *If $h \leq 1$, $J_t < \frac{1}{n^{1+24L}}$, and*

$$\alpha J_t \leq \frac{h}{1024\,(L+1)^2\sqrt{p}\|V^{(t)}\|^{3L+5}},$$

then

$$J_{t+1} \leq J_t - \frac{\alpha L\|\nabla J_t\|^2}{L+\frac{1}{2}}.$$

The next lemma establishes a lower bound on the norm of the gradient at any iteration in terms of the loss J_t and the norm of the weight matrix $V^{(t)}$.

Lemma 4.5 *For all $L \in \mathbb{N}$ if $h \leq h_{\max}$, $J_t < \frac{1}{n^{1+24L}}$, and $\|V^{(t)}\|^L \leq \log(1/J_t)\frac{\|V^{(1)}\|^L}{\log(1/J_1)}$ then*

$$\|\nabla J_t\| \geq \frac{(L+\frac{3}{4})J_t\log(1/J_t)}{\|V^{(t)}\|}. \tag{6}$$

The lower bound on the gradient is proved by showing that the alignment between the negative gradient $-\nabla J_t$ and $V^{(t)}$ is large when the loss is small. The proof proceeds by showing that when h is sufficiently small and the norm of $V^{(t)}$ is not too large, then the inner product between $-\nabla J_t$ and $V^{(t)}$ can be lower bounded by a function of the loss J_t.

4.2. The Proof

As stated above, the proof goes through for any positive $h \leq h_{\max}$, step-size $\alpha \leq \alpha_{\max}(h)$ and any $Q \leq \widetilde{Q}(\alpha)$ (recall the definitions of $h_{\max}, \alpha_{\max}$ and $\widetilde{Q}$ in equations (3a)-(3c)). We will use the following multi-part inductive hypothesis:

(I1) $J_t \leq \frac{J_1}{Q\cdot(t-1)+1}$;

(I2) $\frac{\log(1/J_t)}{\|V^{(t)}\|^L} \geq \frac{\log(1/J_1)}{\|V^{(1)}\|^L}$;

(I3) $\alpha J_t \leq \frac{h}{1024(L+1)^2p\|V^{(t)}\|^{3L+5}}$.

The first part of the inductive hypothesis will be used to ensure that the loss decreases at the prescribed rate, the second part helps establish a lower bound on the norm of the gradient in light of Lemma 4.5 and the third part will ensure that the step-size is small enough to apply Lemma 4.4 and also allows us to make several useful approximations in our proofs.

The base case is trivially true for the first and second part of the inductive hypothesis. It is true for the third part since the step-size $\alpha \leq \alpha_{\max}(h) \leq \frac{h}{1024(L+1)^2 p J_1 \|V^{(1)}\|^{3L+5}}$. Now let us assume that the inductive hypothesis holds for a step $t \geq 1$ and prove that it holds for the next step $t+1$. We start with Part I1.

Lemma 4.6 *If the inductive hypothesis holds at step t, then*

$$J_{t+1} \leq \frac{J_1}{Qt+1}.$$

Proof Since $\alpha J_t \leq \frac{h}{1024(L+1)^2 p \|V^{(t)}\|^{3L+5}}$ and $J_t \leq J_1 < \frac{1}{n^{1+24L}}$, by invoking Lemma 4.4,

$$J_{t+1} \leq J_t - \frac{L\alpha}{(L+\frac{1}{2})} \|\nabla J_t\|^2.$$

Additionally since $h \leq h_{\max}$ and by Part I2 of the inductive hypothesis $\|V^{(t)}\|^L \leq \frac{\log(1/J_t)\|V^{(1)}\|^L}{\log(1/J_1)}$, we use the lower bound on the norm of the gradient established in Lemma 4.5 to get

$$\begin{aligned} J_{t+1} &\leq J_t - \frac{L(L+\frac{3}{4})^2 \alpha J_t^2 \log^2(1/J_t)}{(L+\frac{1}{2})\|V^{(t)}\|^2} \\ &\overset{(i)}{\leq} J_t \left(1 - \frac{L(L+\frac{3}{4})^2 \alpha J_t \log^{2-\frac{2}{L}}(1/J_t) \log^{\frac{2}{L}}(1/J_1)}{(L+\frac{1}{2})\|V^{(1)}\|^2}\right) \\ &\overset{(ii)}{\leq} J_t \left(1 - \frac{L(L+\frac{3}{4})^2 \alpha J_t \log^{\frac{2}{L}}(1/J_1)}{(L+\frac{1}{2})\|V^{(1)}\|^2}\right), \end{aligned} \tag{7}$$

where (i) follows by Part (I2) of the inductive hypothesis, and (ii) follows since $L \geq 1$ and $J_t \leq J_1 < \frac{1}{n^{1+24L}}$, therefore $\log^{2-\frac{2}{L}}(1/J_t) \geq 1$.

For any $z \geq 0$, the quadratic function

$$z - z^2 \frac{L(L+\frac{3}{4})^2 \alpha \log^{\frac{2}{L}}(1/J_1)}{(L+\frac{1}{2})\|V^{(1)}\|^2}$$

is a monotonically increasing function in the interval

$$\left[0, \frac{(L+\frac{1}{2})\|V^{(1)}\|^2}{2L(L+\frac{3}{4})^2 \alpha \log^{\frac{2}{L}}(1/J_1)}\right].$$

Thus, because $J_t \leq \frac{J_1}{Q(t-1)+1}$, if $\frac{J_1}{Q(t-1)+1} \leq \frac{(L+\frac{1}{2})\|V^{(1)}\|^2}{2L(L+\frac{3}{4})^2 \alpha \log^{\frac{2}{L}}(1/J_1)}$, the RHS of (7) is bounded above by its value when $J_t = \frac{J_1}{Q(t-1)+1}$. But this is easy to check: by our choice of step-size α we

have,

$$\alpha \leq \alpha_{\max} \leq \frac{(L+\frac{1}{2})\|V^{(1)}\|^2}{2L(L+\frac{3}{4})^2 J_1 \log^{\frac{2}{L}}(1/J_1)}$$
$$\Rightarrow J_1 \leq \frac{(L+\frac{1}{2})\|V^{(1)}\|^2}{2L(L+\frac{3}{4})^2 \alpha \log^{\frac{2}{L}}(1/J_1)}$$
$$\Rightarrow \frac{J_1}{Q(t-1)+1} \leq \frac{(L+\frac{1}{2})\|V^{(1)}\|^2}{2L(L+\frac{3}{4})^2 \alpha \log^{\frac{2}{L}}(1/J_1)}.$$

Bounding the RHS of inequality (7) by using the worst case that $J_t = \frac{J_1}{Q(t-1)+1}$, we get that

$$\begin{aligned}
J_{t+1} &\leq \frac{J_1}{Q(t-1)+1}\left(1 - \frac{J_1}{Q(t-1)+1}\frac{L(L+\frac{3}{4})^2\alpha\log^{\frac{2}{L}}(1/J_1)}{(L+\frac{1}{2})\|V^{(1)}\|^2}\right) \\
&= \frac{J_1}{Qt+1}\left(1+\frac{Q}{Q(t-1)+1}\right)\left(1 - \frac{Q}{Q(t-1)+1}\frac{L(L+\frac{3}{4})^2\alpha J_1\log^{\frac{2}{L}}(1/J_1)}{Q(L+\frac{1}{2})\|V^{(1)}\|^2}\right) \\
&\leq \frac{J_1}{Qt+1}\left(1 - \left(\frac{Q}{Q(t-1)+1}\right)^2\right) \qquad \left(\text{since } Q \leq \widetilde{Q}(\alpha) = \frac{L(L+\frac{3}{4})^2\alpha J_1\log^{\frac{2}{L}}(1/J_1)}{(L+\frac{1}{2})\|V^{(1)}\|^2}\right) \\
&\leq \frac{J_1}{Qt+1}.
\end{aligned}$$

This establishes the desired upper bound on the loss at step $t+1$. ■

In the next lemma we shall establish that the second part of the inductive hypothesis holds.

Lemma 4.7 *Under the setting of Theorem 3.1, if the induction hypothesis holds at step t then,*

$$\frac{\log\left(\frac{1}{J_{t+1}}\right)}{\|V^{(t+1)}\|^L} \geq \frac{\log\left(\frac{1}{J_1}\right)}{\|V^{(1)}\|^L}.$$

Proof We know from Lemma 4.4 that

$$J_{t+1} \leq J_t\left(1 - \frac{L\alpha\|\nabla J_t\|^2}{(L+\frac{1}{2})J_t}\right),$$

and by the triangle inequality

$$\|V^{(t+1)}\| \leq \|V^{(t)}\| + \alpha\|\nabla J_t\|,$$

hence

$$\frac{\log\left(\frac{1}{J_{t+1}}\right)}{\|V^{(t+1)}\|^L} \geq \frac{\log\left(\frac{1}{J_t\left(1-\frac{L\alpha}{(L+\frac{1}{2})J_t}\|\nabla J_t\|^2\right)}\right)}{\left(\|V^{(t)}\|+\alpha\|\nabla J_t\|\right)^L} = \frac{\log\left(\frac{1}{J_t}\right)+\log\left(\frac{1}{\left(1-\frac{L\alpha}{(L+\frac{1}{2})J_t}\|\nabla J_t\|^2\right)}\right)}{\left(\|V^{(t)}\|+\alpha\|\nabla J_t\|\right)^L}$$

$$= \frac{\log\left(\frac{1}{J_t}\right)\left(1-\frac{\log\left(1-\frac{L\alpha}{(L+\frac{1}{2})J_t}\|\nabla J_t\|^2\right)}{\log\left(\frac{1}{J_t}\right)}\right)}{\|V^{(t)}\|^L\left(1+\frac{\alpha\|\nabla J_t\|}{\|V^{(t)}\|}\right)^L}$$

$$\overset{(i)}{\geq} \frac{\log\left(\frac{1}{J_t}\right)}{\|V^{(t)}\|^L}\left\{\frac{\left(1+\frac{L\alpha\|\nabla J_t\|^2}{(L+\frac{1}{2})J_t\log\left(\frac{1}{J_t}\right)}\right)}{\left(1+\frac{\alpha\|\nabla J_t\|}{\|V^{(t)}\|}\right)^L}\right\} \tag{8}$$

where (i) follows since $\log(1-z) \leq -z$ for all $z \in (0,1)$ and because

$$\begin{aligned}
&\frac{L\alpha}{(L+\frac{1}{2})J_t}\|\nabla J_t\|^2 \\
&\leq \frac{L\alpha}{(L+\frac{1}{2})}\left[(L+1)pJ_t\|V^{(t)}\|^{2(L+1)}\right] \qquad \text{(by Lemma 4.3)} \\
&= \alpha J_t\left[\frac{L(L+1)p\|V^{(t)}\|^{2(L+1)}}{L+\frac{1}{2}}\right] \\
&< \alpha J_t\left[\frac{1024\,(L+1)^2\,p\|V^{(t)}\|^{3L+5}}{h}\right] \qquad (\|V^{(t)}\|>1 \text{ by Lemma B.5, and } h\leq 1) \\
&\leq 1 \qquad \text{(by Part I3 of the IH).}
\end{aligned}$$

We want to show that the term in curly brackets in inequality (8) is at least 1, that is, we want

$$1+\frac{L\alpha\|\nabla J_t\|^2}{(L+\frac{1}{2})J_t\log\left(\frac{1}{J_t}\right)} \geq \left(1+\frac{\alpha\|\nabla J_t\|}{\|V^{(t)}\|}\right)^L. \tag{9}$$

We will first show that this inequality holds in the case where $L>1$. To show this, note that

$$\begin{aligned}
\frac{\alpha\|\nabla J_t\|}{\|V^{(t)}\|} &\leq \alpha J_t\sqrt{(L+1)p}\|V^{(t)}\|^L \qquad \text{(by Lemma 4.3)} \\
&< \frac{1}{L-1}\cdot\alpha J_t\left[\frac{1024\,(L+1)^2\,p\|V^{(t)}\|^{3L+5}}{h}\right] \qquad (\text{since } \|V^{(t)}\|>1 \text{ by Lemma B.5}) \\
&\leq \frac{1}{L-1} \qquad \text{(by Part I3 of the IH).}
\end{aligned}$$

For any positive $z < \frac{1}{L-1}$ we have the inequality that $(1+z)^L \leq 1 + \frac{Lz}{1-(L-1)z}$, therefore to show that inequality (9) holds it instead suffices to show that

$$
\begin{aligned}
&1 + \frac{L\alpha\|\nabla J_t\|^2}{(L+\frac{1}{2})J_t \log\left(\frac{1}{J_t}\right)} \geq 1 + \frac{L\alpha\|\nabla J_t\|}{\|V^{(t)}\|\left(1 - \frac{(L-1)\alpha\|\nabla J_t\|}{\|V^{(t)}\|}\right)} \\
&\Leftrightarrow \frac{\|\nabla J_t\|}{(L+\frac{1}{2})J_t \log\left(\frac{1}{J_t}\right)} \geq \frac{1}{\|V^{(t)}\|\left(1 - \frac{(L-1)\alpha\|\nabla J_t\|}{\|V^{(t)}\|}\right)} \\
&\Leftrightarrow \|\nabla J_t\| \geq \frac{(L+\frac{1}{2})J_t \log\left(\frac{1}{J_t}\right)}{\|V^{(t)}\|} + \frac{(L-1)\alpha\|\nabla J_t\|^2}{\|V^{(t)}\|} \\
&\Leftarrow \|\nabla J_t\| \geq \frac{(L+\frac{1}{2})J_t \log\left(\frac{1}{J_t}\right)}{\|V^{(t)}\|} + \frac{(L-1)\alpha(L+1)p\|V^{(t)}\|^{2L+2}J_t^2}{\|V^{(t)}\|} \qquad \text{(by Lemma 4.3)} \\
&\Leftarrow \|\nabla J_t\| \geq \frac{\left(L+\frac{1}{2}+\alpha J_t \frac{1024(L^2-1)p\|V^{(t)}\|^{2L+2}}{h} \times \frac{h}{1024\log\left(\frac{1}{J_t}\right)}\right) J_t \log\left(\frac{1}{J_t}\right)}{\|V^{(t)}\|} \\
&\Leftarrow \|\nabla J_t\| \geq \frac{(L+\frac{3}{4})J_t \log\left(\frac{1}{J_t}\right)}{\|V^{(t)}\|},
\end{aligned}
$$

where the last implication follows from Part I3 of the IH, the fact that $h \leq 1$ and because $J_t \leq J_1 \leq 1/n^{1+24L}$ and $n \geq 3$. Now this last inequality holds again because of Lemma 4.5 that guarantees that $\|\nabla J_t\| \geq \frac{(L+\frac{3}{4})J_t \log(1/J_t)}{\|V^{(t)}\|}$. A similar argument can also be used in the case where $L = 1$, without the use of the inequality that was used to upper bound $(1+z)^L$. Thus we have proved that the term in the curly brackets in inequality (8) is at least 1 and hence

$$
\frac{\log\left(\frac{1}{J_{t+1}}\right)}{\|V^{(t+1)}\|^L} \geq \frac{\log\left(\frac{1}{J_t}\right)}{\|V^{(t)}\|^L} \geq \frac{\log\left(\frac{1}{J_1}\right)}{\|V^{(1)}\|^L}.
$$

This proves that the ratio is bounded below at step $t+1$ by its initial value and establishes our claim. ■

Finally we ensure that the third part of the inductive hypothesis holds. This allows us to apply Lemma 4.4 in the next step $t+1$.

Lemma 4.8 *Under the setting of Theorem 3.1, if the induction hypothesis holds at step t, then*

$$
\alpha J_{t+1} \leq \frac{h}{1024\,(L+1)^2\,p\|V^{(t+1)}\|^{3L+5}}.
$$

Proof We know by Lemma 4.7 that $\|V^{(t+1)}\|^L \leq \frac{\log(1/J_{t+1})\|V^{(1)}\|^L}{\log(1/J_1)}$ so it instead suffices to prove that

$$
\alpha J_{t+1} \log^{\frac{3L+5}{L}}\left(\frac{1}{J_{t+1}}\right) \leq \frac{h \log^{\frac{3L+5}{L}}(1/J_1)}{1024\,(L+1)^2\,p\|V^{(1)}\|^{3L+5}}. \tag{10}
$$

Lemma 4.6 establishes that $J_{t+1} \leq J_1 < 1/n^{1+24L}$. The function $z \log^{\frac{3L+5}{L}}(1/z)$ is increasing over the interval $(0, \frac{1}{e^{\frac{3L+5}{L}}})$. Recall that $n \geq 3$ therefore,

$$J_{t+1} \leq J_1 < \frac{1}{3^{1+24L}} < \frac{1}{e^{\frac{3L+5}{L}}}.$$

Thus, the LHS of (10) is maximized at J_1

$$\alpha J_{t+1} \log^{\frac{3L+5}{L}}\left(\frac{1}{J_{t+1}}\right) \leq \alpha J_1 \log^{\frac{3L+5}{L}}\left(\frac{1}{J_1}\right) \leq \frac{h \log^{\frac{3L+5}{L}}(1/J_1)}{1024\,(L+1)^2\,\sqrt{p}\|V^{(1)}\|^{3L+5}}$$

where final inequality holds by choice of the step-size α. This completes the proof. ■

Combining the results of Lemmas 4.6, 4.7 and 4.8 completes the proof of theorem.

5. Discussion

We have shown that deep networks with smoothed ReLU activations trained by gradient descent with logistic loss achieve training loss approaching zero if the loss is initially small enough. We also established conditions under which this happens that formalize the idea that the NTK features are useful. Our analysis applies in the case of networks using the increasingly popular Swish activation function.

While, to simplify our treatment, we concentrated on the case that the number of hidden nodes in each layer is equal to the number of inputs, our analysis should easily be adapted to the case of varying numbers of hidden units.

Analysis of architectures such as Residual Networks and Transformers would be a potentially interesting next step.

Acknowledgments

We thank the anonymous reviewers for alerting us to a mistake in an earlier version of this paper.

We gratefully acknowledge the support of the NSF through grants DMS-2031883 and DMS-2023505 and the Simons Foundation through award 814639.

References

Naman Agarwal, Pranjal Awasthi, and Satyen Kale. A deep conditioning treatment of neural networks. In *Algorithmic Learning Theory*, pages 249–305, 2021.

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

Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International Conference on Machine Learning*, pages 1908–1916, 2014.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7413–7424, 2019a.

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

Peter L Bartlett and Philip M Long. Failures of model-dependent generalization bounds for least-norm interpolation. *arXiv preprint arXiv:2010.08479*, 2020.

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

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems*, pages 2300–2311, 2018.

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

Guy Bresler and Dheeraj Nagaraj. A corrective view of neural networks: Representation, memorization and learning. In *Conference on Learning Theory*, pages 848–901, 2020.

Alon Brutzkus and Amir Globerson. Why do larger models generalize better? A theoretical perspective via the XOR problem. In *International Conference on Machine Learning*, pages 822–830, 2019.

Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *International Conference on Learning Representations*, 2018.

Sébastien Bubeck. Convex optimization: algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

Niladri S Chatterji, Philip M Long, and Peter L Bartlett. When does gradient descent with logistic loss find interpolating two-layer networks? *arXiv preprint arXiv:2012.02409*, 2020.

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

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

Lénaïc Chizat. Analysis of gradient descent on wide two-layer ReLU neural networks, 2020. URL https://www.msri.org/workshops/928/schedules/28397. Talk at MSRI.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, pages 3036–3046, 2018.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, 2020.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2937–2947, 2019.

Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT Press, 2009.

Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14(3):326–334, 1965.

Amit Daniely. Neural networks learning and memorization with (almost) no over-parameterization. In *Advances in Neural Information Processing Systems*, pages 9007–9016, 2020.

Amit Daniely and Eran Malach. Learning parities with neural networks. *Advances in Neural Information Processing Systems*, 33:20356–20365, 2020.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

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

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *International Conference on Learning Representations*, 2018.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841, 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018b.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 91–99, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798, 2019a.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019b.

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

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems*, pages 17176–17186, 2020.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with ReLU activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47, 2018.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and min-ℓ_1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711, 2020.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Alexander Matthews, Jiri Hron, Mark Rowland, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018a.

Alexander Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018b.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2021.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464, 2019.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel J Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*, 2020a.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 2020b.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations (Workshop)*, 2015.

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

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

Rina Panigrahy, Sushant Sachdeva, and Qiuyi Zhang. Convergence results for neural networks via electrodynamics. In *Innovations in Theoretical Computer Science*, 2018.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. In *International Conference on Learning Representations (Workshop)*, 2018.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. In *International Conference on Machine Learning*, pages 4433–4441, 2018.

Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix Chernoff bound. *arXiv preprint arXiv:1906.03593*, 2019.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (1):2822–2878, 2018.

Norman Tatro, Pin-Yu Chen, Payel Das, Igor Melnyk, Prasanna Sattigeri, and Rongjie Lai. Optimizing mode connectivity via neuron alignment. In *Advances in Neural Information Processing Systems*, 2020.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

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

Martin Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pages 9712–9724, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017a.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer ReLU networks via gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 1524–1534, 2019.

Yuchen Zhang, Jason D Lee, Martin Wainwright, and Michael Jordan. On the learnability of fully-connected neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 83–91, 2017b.

Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International Conference on Machine Learning*, pages 4140–4149, 2017.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *NeurIPS*, pages 2053–2062, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109(3):467–492, 2020.

Contents

Appendix A. Additional Related Work

Building on the work of Lyu and Li (2020), Ji and Telgarsky (2020) study finite-width deep ReLU neural networks and show that starting from a small loss, gradient flow coupled with logistic loss leads to convergence of the directions of the parameter vectors. They also demonstrate alignment between the parameter vector directions and the negative gradient. However, they do not prove that the training loss converges to zero.

Using mean-field techniques Chizat and Bach (2020), building on (Chizat and Bach, 2018; Mei et al., 2019), show that infinitely wide two-layer squared ReLU networks trained with gradient flow on the logistic loss leads to a max-margin classifier in a particular non-Hilbertian space of functions. See also the videos in a talk about this work (Chizat, 2020). Chen et al. (2020) analyzed regularized training with gradient flow on infinitely wide networks. When training is regularized, the weights also may travel far from their initial values. Previously Brutzkus et al. (2018) studied finite-width two-layer leaky ReLU networks and showed that when the data is linearly separable, these networks can be trained up to zero-loss using stochastic gradient descent with the hinge loss.

Our study is motivated in part by the line of work that has emerged which emphasizes the need to understand the behavior of interpolating (zero training loss/error) classifiers and regressors. A number of recent papers have analyzed the properties of interpolating methods in linear regression (Hastie et al., 2019; Bartlett et al., 2020; Muthukumar et al., 2020b; Tsigler and Bartlett, 2020; Bartlett and Long, 2020), linear classification (Montanari et al., 2019; Chatterji and Long, 2021; Liang and Sur, 2020; Muthukumar et al., 2020a; Hsu et al., 2021), kernel regression (Liang and Rakhlin, 2020; Mei and Montanari, 2021; Liang et al., 2020) and simplicial nearest neighbor methods (Belkin et al., 2018).

There are also many related papers that characterize the implicit bias of the solution obtained by first-order methods (Neyshabur et al., 2015; Soudry et al., 2018; Ji and Telgarsky, 2019a; Gunasekar et al., 2018a,b; Li et al., 2018; Arora et al., 2019a; Ji and Telgarsky, 2019b).

Finally, we note that a number of other recent papers also theoretically study the optimization of neural networks including (Andoni et al., 2014; Li and Yuan, 2017; Zhong et al., 2017; Zhang et al., 2017b; Ge et al., 2018; Panigrahy et al., 2018; Du et al., 2018; Safran and Shamir, 2018; Zhang et al., 2019; Arora et al., 2019b; Brutzkus and Globerson, 2019; Wei et al., 2019; Ji and Telgarsky, 2019c; Nitanda et al., 2019; Song and Yang, 2019; Zou and Gu, 2019; Bresler and Nagaraj, 2020; Daniely, 2020; Daniely and Malach, 2020).

Appendix B. Omitted Proofs from Section 4.1

In this section we present the proofs of Lemmas 4.1-4.5.

B.1. Additional Definitions

Definition B.1 *For any weight matrix V, define $g_s(V) := \frac{1}{1+\exp(y_s f_V(x_s))}$. We will often use g_s as shorthand for $g_s(V)$ when V can be determined from context. Further, for all $t \in \{0, 1, \ldots\}$, define $g_{ts} := g_s(V^{(t)})$.*

Informally, $g_s(V)$ is the size of the contribution of example s to the gradient.

Definition B.2 *For all iterates t, all $\ell \in [L+1]$ and all $s \in [n]$, define $x_{\ell,s}^{(t)} := x_{\ell,s}^{V^{(t)}}$, $u_{\ell,s}^{(t)} := u_{\ell,s}^{V^{(t)}}$ and $\Sigma_{\ell,s}^{(t)} := \Sigma_{\ell,s}^{V^{(t)}}$.*

B.2. Basic Lemmas

To prove Lemmas 4.1-4.5, we will need some more basic lemmas, which we first prove.

Lemma B.3 *For any $x \in \mathbb{R}^p$ and $y \in \{-1, 1\}$ and any weight matrix V we have the following:*

1.
$$\frac{1}{1+\exp\left(yf_V(x)\right)} \leq \log(1+\exp(-yf_V(x))) = J(V; x, y).$$

2.
$$\frac{\exp\left(yf_V(x)\right)}{\left(1+\exp\left(yf_V(x)\right)\right)^2} \leq \frac{1}{1+\exp\left(yf_V(x)\right)} \leq J(V; x, y).$$

Proof Part 1 follows since for any $z \in \mathbb{R}$, we have the inequality $(1+\exp(z))^{-1} \leq \log(1+\exp(-z))$.

Part 2 follows since for any $z \in \mathbb{R}^d$, we have the inequality

$$\exp(z)/\left(1+\exp(z)\right)^2 \leq \left(1+\exp(z)\right)^{-1}.$$

■

The following lemma is useful for establishing a relatively simple lower bound on a sum of applications of a concave function.

Lemma B.4 *If $\psi : [0, M] \to \mathbb{R}$ is a concave function with $\psi(0) = 0$. Then the minimum of $\sum_{i=1}^n \psi(z_i)$ subject to $z_1, \ldots, z_n \geq 0$ and $\sum_{i=1}^n z_i = M$ is $\psi(M)$.*

Proof Let $z_1, \ldots, z_n$ be any solution, and let i be the least index such that $z_i > 0$. Then, since ψ is concave and non-negative, we have that

$$\psi(z_1 + z_i) + \psi(0) = \psi(z_1 + z_i) \leq \psi(z_1) + \psi(z_i).$$

Thus, replacing z_1 with $z_1 + z_i$, and replacing z_i with 0, produces a solution with one fewer nonzero entries that it at least as good. Repeating this for each $i > 1$ implies that the solution with $z_1 = M$ and $z_2 = \ldots = z_n = 0$ is optimal. ■

The next lemma shows that large weights are needed to achieve small loss.

Lemma B.5 *For any $L \in \mathbb{N}$ and any weight matrix V if $J(V) \leq \frac{2}{n^{1+24L}}$ then, $\|V\| > \sqrt{L+1} \geq \sqrt{2}$.*

Proof Since ϕ is 1-Lipschitz and $\phi(0) = 0$, for all z, $|\phi(z)| \leq |z|$, and thus, given any sample s,

$$\begin{aligned}
J(V; x_s, y_s) &= \log\left(1 + \exp\left(-y_s V_{L+1}\phi(V_L \cdots \phi(V_1 x_s))\right)\right) \\
&\geq \log\left(1 + \exp\left(-\prod_{j=1}^{L+1}\|V_j\|_{op}\|x_s\|\right)\right) \\
&\geq \log\left(1 + \exp\left(-\prod_{j=1}^{L+1}\|V_j\|_{op}\right)\right) \qquad \text{(since } \|x_s\| = 1\text{)} \\
&\geq \log\left(1 + \exp\left(-\prod_{j=1}^{L+1}\|V_j\|\right)\right).
\end{aligned}$$

By the AM-GM inequality

$$\left(\prod_{j=1}^{L+1}\|V_j\|^2\right)^{\frac{1}{L+1}} \leq \frac{\sum_{j=1}^{L+1}\|V_j\|^2}{L+1} = \frac{\|V\|^2}{L+1}.$$

Therefore

$$J(V; x_s, y_s) \geq \log\left(1 + \exp\left(-\left(\frac{\|V\|}{\sqrt{L+1}}\right)^{L+1}\right)\right).$$

Now we know that

$$\frac{2}{n^{1+24L}} > J(V) = \frac{1}{n}\sum_{s\in[n]} J(V; x_s, y_s) \geq \log\left(1 + \exp\left(-\left(\frac{\|V\|}{\sqrt{L+1}}\right)^{L+1}\right)\right).$$

Solving for $\|V\|$ leads to the implication

$$\sqrt{L+1}\log^{\frac{1}{L+1}}\left(\frac{1}{\exp\left(\frac{2}{n^{1+24L}}\right) - 1}\right) < \|V\|.$$

Since for any $z \in [0,1]$, $\exp(z) \leq 1 + 2z$ and $n \geq 3$, hence

$$\begin{aligned}
\|V\| > \sqrt{L+1}\log^{\frac{1}{L+1}}\left(\frac{n^{1+24L}}{4}\right) \geq \sqrt{L+1}\log^{\frac{1}{L+1}}\left(\frac{3^{1+24L}}{4}\right) &> \sqrt{L+1}\log^{\frac{1}{L+1}}\left(3^{23L}\right) \\
&= \sqrt{L+1}(23L)^{\frac{1}{L+1}}\log^{\frac{1}{L+1}}(3) \\
&> \sqrt{L+1} \geq \sqrt{2}.
\end{aligned}$$

■

Lemma B.6 *For any $L \in \mathbb{N}$:*

1. *if $\|V\| > \sqrt{L+1}$, then $\max_{k\in[L]} \prod_{j=k+1}^{L+1} \|V_j\| \le \left(\frac{\|V\|}{\sqrt{L}}\right)^L \le \|V\|^L$;*
2. *if $\|V\| > 1$, then $\max_{k\in[L]} \prod_{j=k+1}^{L+1} \|V_j\| \le \|V\|^L$.*

Proof Let $\eta^2 = \|V\|^2$. Then for any $k \in [L]$,

$$\prod_{j=k+1}^{L+1} \|V_j\|$$

is maximized subject to $\sum_{j=k+1}^{L+1} \|V_j\|^2 \le \eta^2$ when every $\|V_j\|^2 = \eta^2/(L-k+1)$; this follows by the AM-GM inequality.

Therefore we have

$$\max_{k\in[L]} \prod_{j=k+1}^{L+1} \|V_j\|_{op} \le \max_{k\in[L]} \left(\frac{\eta}{\sqrt{L-k+1}}\right)^{L-k+1}.$$

If $\|V\| \ge \sqrt{L+1}$,

$$\max_{k\in[L]} \left(\frac{\eta}{\sqrt{L-k+1}}\right)^{L-k+1} \le \left(\frac{\eta}{\sqrt{L}}\right)^L \le \eta^L.$$

and if $\|V\| > 1$ then

$$\max_{k\in[L]} \left(\frac{\eta}{\sqrt{L-k+1}}\right)^{L-k+1} \le \eta^L.$$

■

The next lemma bounds the product of the operator norms of matrices in terms of a "collective Frobenius norm".

Lemma B.7 *For matrices $A_1, \ldots, A_{L+1}$ and $M_1, \ldots, M_{L+1}$, let $A = (A_1, \ldots, A_{L+1})$. For all $i \in [L+1]$, $\|M_i\|_{op} \le 1$. Then, for any nonempty $\mathcal{I} \subseteq [L+1]$*

$$\prod_{i\in\mathcal{I}} \|A_i\|_{op}\|M_i\|_{op} \le \max\left\{\frac{\|A\|^{L+1}}{(L+1)^{\frac{L+1}{2}}}, \|A\|\right\}.$$

Proof We know that for all $i \in [L+1]$, $\|A_i\|_{op} \le \|A_i\|$, therefore, by the AM-GM inequality

$$\begin{aligned}\prod_{i\in\mathcal{I}} \left(\|A_i\|_{op}^2\|M_i\|_{op}^2\right) \le \prod_{i\in\mathcal{I}} \|A_i\|_{op}^2 \le \prod_{i\in\mathcal{I}} \|A_i\|^2 &\le \left(\frac{\sum_{i\in\mathcal{I}}\|A_i\|^2}{|\mathcal{I}|}\right)^{|\mathcal{I}|} \\ &\le \left(\frac{\|A\|^2}{|\mathcal{I}|}\right)^{|\mathcal{I}|} \\ &\le \max\left\{\frac{\|A\|^{2(L+1)}}{(L+1)^{L+1}}, \|A\|^2\right\}.\end{aligned}$$

Taking square roots completes the proof. ■

The next lemma bounds bounds how much perturbing the factors changes a product of matrices.

Lemma B.8 *Let $A_1, \ldots, A_{L+1}$, $B_1, \ldots, B_{L+1}$, $M_1, \ldots, M_{L+1}$ and $N_1, \ldots, N_{L+1}$ be matrices, and let $A = (A_1, \ldots, A_{L+1})$ and $B = (B_1, \ldots, B_{L+1})$. Assume*

- $\|A\| \geq \sqrt{L + 1/2}$,
- *for all $i \in [L+1]$, $\|M_i\|_{op} \leq 1$ and $\|N_i\|_{op} \leq 1$ and*
- *for all $i \in [L+1]$, $\|M_i - N_i\|_{op} \leq \kappa$,*

then

$$\left\| \prod_{i=1}^{L+1} (A_i M_i) - \prod_{i=1}^{L+1} (B_i N_i) \right\|_{op} \leq \frac{3}{2} (\|A\| + \|A - B\|)^{L+1} \left(\kappa \|A\| + \|A - B\| \right).$$

Proof By the triangle inequality

$$\begin{aligned}
&\left\| \prod_{i=1}^{L+1} (A_i M_i) - \prod_{i=1}^{L+1} (B_i N_i) \right\|_{op} \\
&= \left\| \sum_{j=1}^{L+1} \left(\left(\prod_{i=1}^{j} A_i M_i \right) \left(\prod_{i=j+1}^{L+1} B_i N_i \right) - \left(\prod_{i=1}^{j-1} A_i M_i \right) \left(\prod_{i=j}^{L+1} B_i N_i \right) \right) \right\|_{op} \\
&\leq \sum_{j=1}^{L+1} \left\| \left(\prod_{i=1}^{j} A_i M_i \right) \left(\prod_{i=j+1}^{L+1} B_i N_i \right) - \left(\prod_{i=1}^{j-1} A_i M_i \right) \left(\prod_{i=j}^{L+1} B_i N_i \right) \right\|_{op} \\
&= \sum_{j=1}^{L+1} \left\| (A_j M_j - B_j N_j) \left(\prod_{i=1}^{j-1} A_i M_i \right) \left(\prod_{i=j+1}^{L+1} B_i N_i \right) \right\|_{op} \\
&\leq \sum_{j=1}^{L+1} \|A_j M_j - B_j N_j\|_{op} \left\| \left(\prod_{i=1}^{j-1} A_i M_i \right) \left(\prod_{i=j+1}^{L+1} B_i N_i \right) \right\|_{op}. \qquad (11)
\end{aligned}$$

For some j, consider $T(j) := (A_1, \ldots, A_{j-1}, B_{j+1}, B_{L+1})$. By the triangle inequality,

$$\|T(j)\| \leq \|A\| + \|A - B\|.$$

Thus, Lemma B.7 implies

$$\begin{aligned}
\left\| \left(\prod_{i=1}^{j-1} A_i M_i \right) \left(\prod_{i=j+1}^{L+1} B_i N_i \right) \right\|_{op} &\leq \left(\prod_{i=1}^{j-1} \|A_i M_i\|_{op} \right) \left(\prod_{i=j+1}^{L+1} \|B_i N_i\|_{op} \right) \\
&\leq \left(\prod_{i=1}^{j-1} \|A_i\|_{op} \right) \left(\prod_{i=j+1}^{L+1} \|B_i\|_{op} \right) \\
&\leq \max \left\{ \frac{(\|A\| + \|A - B\|)^{L+1}}{(L+1)^{\frac{L+1}{2}}}, \|A\| + \|A - B\| \right\}.
\end{aligned}$$

Returning to (11),

$$
\begin{aligned}
&\left\|\prod_{i=1}^{L+1} A_i M_i - \prod_{i=1}^{L+1} B_i N_i\right\|_{op} \\
&\leq \max\left\{\frac{(\|A\| + \|A-B\|)^{L+1}}{(L+1)^{\frac{L+1}{2}}}, \|A\| + \|A-B\|\right\} \sum_{j=1}^{L+1} \|A_j M_j - B_j N_j\|_{op} \\
&= \max\left\{\frac{(\|A\| + \|A-B\|)^{L+1}}{(L+1)^{\frac{L+1}{2}}}, \|A\| + \|A-B\|\right\} \sum_{j=1}^{L+1} \|A_j M_j - A_j N_j + A_j N_j - B_j N_j\|_{op} \\
&\leq \max\left\{\frac{(\|A\| + \|A-B\|)^{L+1}}{(L+1)^{\frac{L+1}{2}}}, \|A\| + \|A-B\|\right\} \sum_{j=1}^{L+1} \left(\|A_j(M_j - N_j)\|_{op} + \|A_j - B_j\|_{op}\right) \\
&\leq \max\left\{\frac{(\|A\| + \|A-B\|)^{L+1}}{(L+1)^{\frac{L+1}{2}}}, \|A\| + \|A-B\|\right\} \left[\sum_{j=1}^{L+1} \|A_j\|_{op} \|M_j - N_j\|_{op} + \sum_{j=1}^{L+1} \|A_j - B_j\|_{op}\right] \\
&\leq \max\left\{\frac{(\|A\| + \|A-B\|)^{L+1}}{(L+1)^{\frac{L+1}{2}}}, \|A\| + \|A-B\|\right\} \left[\kappa \sum_{j=1}^{L+1} \|A_j\| + \sum_{j=1}^{L+1} \|A_j - B_j\|\right] \\
&\leq \max\left\{\frac{(\|A\| + \|A-B\|)^{L+1}}{(L+1)^{\frac{L+1}{2}}}, \|A\| + \|A-B\|\right\} \left[\sqrt{L+1}\,(\kappa\|A\| + \|A-B\|)\right] \\
&= \max\left\{\frac{(\|A\| + \|A-B\|)^{L+1}}{(L+1)^{\frac{L}{2}}}, \sqrt{L+1}\,(\|A\| + \|A-B\|)\right\} (\kappa\|A\| + \|A-B\|) \\
&\leq \frac{3}{2}(\|A\| + \|A-B\|)^{L+1}\,(\kappa\|A\| + \|A-B\|)\,,
\end{aligned}
$$

where the last inequality holds since $\|A\| \geq \sqrt{L+1/2}$. This completes the proof. ■

The next lemma shows that h-smoothly approximately ReLU activations are contractive maps.

Lemma B.9 *Given an h-smoothly approximately ReLU activation ϕ, for any $v_1, v_2 \in \mathbb{R}^p$ we have $\|\phi(v_1) - \phi(v_2)\| \leq \|v_1 - v_2\|$. That is, ϕ is a contractive map with respect to the Euclidean norm.*

Proof Let $(v)_j$ denote the jth coordinate of a vector v. For each $j \in [p]$, by the mean value theorem for some $\tilde{v}_j \in [(v_2)_j, (v_1)_j]$

$$(\phi(v_1) - \phi(v_2))_j = \phi'(\tilde{v}_j)(v_1 - v_2)_j.$$

Thus,

$$
\begin{aligned}
\|\phi(v_1) - \phi(v_2)\|^2 = \sum_{j\in[p]} \left(\phi'(\tilde{v}_j)(v_1 - v_2)_j\right)^2 = \sum_{j\in[p]} \left(\phi'(\tilde{v}_j)\right)^2 (v_1 - v_2)_j^2 &\overset{(i)}{\leq} \sum_{j\in[p]} (v_1 - v_2)_j^2 \\
&= \|v_1 - v_2\|^2,
\end{aligned}
$$

where (i) follows because $|\phi'(z)| \leq 1$ for all $z \in \mathbb{R}$ for h-smoothly approximately ReLU activations. Taking square roots completes the proof. ■

B.3. Proof of Lemma 4.1

Lemma 4.1 *For $\alpha > 0$, let $V^{(t+1)} = V^{(t)} - \alpha \nabla J_t$. If, for all convex combinations W of $V^{(t)}$ and $V^{(t+1)}$, we have $\mathrm{Lip}(\nabla_W J(W)) \leq M$, then if $\alpha \leq \frac{1}{(L+\frac{1}{2})M}$, we have*

$$J_{t+1} \leq J_t - \frac{\alpha L \|\nabla J_t\|^2}{L + \frac{1}{2}}.$$

Proof Along the line segment joining $V^{(t)}$ to $V^{(t+1)}$, the function $J(\cdot)$ is M-smooth, therefore by using a standard argument (see, e.g., Bubeck, 2015, Lemma 3.4) we get that

$$\begin{aligned}
J_{t+1} &\leq J_t + \nabla J_t \cdot (V^{(t+1)} - V^{(t)}) + \frac{M}{2}\|V^{(t+1)} - V^{(t)}\|^2 \\
&= J_t - \alpha\|\nabla J_t\|^2 + \frac{\alpha^2 M}{2}\|\nabla J_t\|^2 \\
&= J_t - \alpha\left(1 - \frac{\alpha M}{2}\right)\|\nabla J_t\|^2 \\
&\leq J_t - \frac{L}{L+\frac{1}{2}}\alpha\|\nabla J_t\|^2.
\end{aligned}$$

This completes the proof. ■

B.4. Proof of Lemma 4.2

The proof of Lemma 4.2 is built up in stages, through a series of lemmas.

The first lemma bounds the norm of the difference between the pre-activation ($u^V_{j,s}$) and post-activation features ($x^V_{j,s}$) at any layer j, when the weight matrix of a single layer is swapped. It also provides a bound on the norm of the pre-activation and post-activation features at any layer in terms of the norm of the weight matrix.

Lemma B.10 *Consider $V = (V_1, \ldots, V_{L+1})$ and $W = (W_1, \ldots, W_{L+1})$, and $\ell \in [L+1]$. Suppose that $V_j = W_j$ for all $j \neq \ell$, and $\|V\|, \|W\| > \sqrt{L + 1/2}$. Then, for all examples s and all layers j,*

1. $\|u^V_{j,s}\| \leq \|V\|^{L+1}$;
2. $\|x^V_{j,s}\| \leq \|V\|^{L+1}$;
3. $\|u^V_{j,s} - u^W_{j,s}\| \leq \|V_\ell - W_\ell\|_{op}\|V\|^{L+1}$; *and*
4. $\|x^V_{j,s} - x^W_{j,s}\| \leq \|V_\ell - W_\ell\|_{op}\|V\|^{L+1}$.

Proof *Proof of Parts 1 and 2:* For any sample s and layer j we have

$$\begin{aligned}\|u^V_{j,s}\| = \|V_j\phi(u^V_{j-1,s})\| \le \|V_j\|_{op}\|\phi(u^V_{j-1,s})\| &\overset{(i)}{\le} \|V_j\|_{op}\|u^V_{j-1,s}\| \le \prod_{k=1}^{j}\|V_k\|_{op}\|x_s\| \\ &\overset{(ii)}{\le} \prod_{k=1}^{j}\|V_k\|_{op} \\ &\overset{(iii)}{\le} \|V\|^{L+1},\end{aligned}$$

where (i) follows since ϕ is contractive (Lemma B.9), (ii) is because $\|x_s\| = 1$ and (iii) is by Lemma B.7. This completes the proof of Part 1 of this lemma. Again since ϕ is contractive, $\|x^V_{j,s}\| = \|\phi(u^V_{j,s})\| \le \|V\|^{L+1}$, which establishes the second part of the lemma.

Proof of Parts 3 and 4: For any $j < \ell$, $u^V_{j,s} = u^W_{j,s}$ and $x^V_{j,s} = x^W_{j,s}$, since $V_j = W_j$ for all $j \neq \ell$. For $j = \ell$ we have

$$\begin{aligned}\|u^V_{\ell,s} - u^W_{\ell,s}\| = \|V_\ell x^V_{\ell-1,s} - W_\ell x^W_{\ell-1,s}\| = \|(V_\ell - W_\ell)x^V_{\ell-1,s}\| &\le \|V_\ell - W_\ell\|_{op}\|x^V_{\ell-1,s}\| \\ &= \|V_\ell - W_\ell\|_{op}\|\phi\left(V_{\ell-1}x^V_{\ell-2,s}\right)\| \\ &\overset{(i)}{\le} \|V_\ell - W_\ell\|_{op}\prod_{k<\ell}\|V_k\|_{op} \\ &\le \|V_\ell - W_\ell\|_{op}\prod_{k<\ell}\|V_k\| \\ &\overset{(ii)}{\le} \|V_\ell - W_\ell\|_{op}\|V\|^{L+1}\end{aligned}$$

where (i) follows since ϕ is a contractive map (Lemma B.9) and because $\|x_s\| = 1$, and (ii) follows by applying Lemma B.7. Since ϕ is contractive we also have that

$$\|x^V_{\ell,s} - x^W_{\ell,s}\| \le \|V_\ell - W_\ell\|_{op}\|V\|^{L+1}.$$

When $j > \ell$, it is possible to establish our claim by mirroring the argument in the $j = \ell$ case which completes the proof of the last two parts.

■

The next lemma upper bounds difference between the $\Sigma^V_{j,s}$ and $\Sigma^W_{j,s}$, when the weight matrices differ in a single layer.

Lemma B.11 *Consider $V = (V_1, \ldots, V_L)$ and $W = (W_1, \ldots, W_L)$, and $\ell \in [L]$. Suppose that $V_j = W_j$ for all $j \neq \ell$, and $\|V\|, \|W\| > \sqrt{L + 1/2}$. Then, for all examples s and all layers j,*

$$\|\Sigma^V_{j,s} - \Sigma^W_{j,s}\|_{op} \le \frac{\|V_\ell - W_\ell\|_{op}\|V\|^{L+1}}{h}.$$

Proof For any $j \in [L]$ and any $s \in [n]$, $\Sigma^V_{j,s}$ and $\Sigma^W_{j,s}$ are both diagonal matrices, and hence

$$\begin{aligned}\|\Sigma^V_{j,s} - \Sigma^W_{j,s}\|_{op} &= \|\phi'(u^V_{j,s}) - \phi'(u^W_{j,s})\|_\infty \\ &\le \frac{\|u^V_j - u^W_j\|_\infty}{h} && \text{(since } \phi' \text{ is } (1/h)\text{-Lipschitz)} \\ &\le \frac{\|V_\ell - W_\ell\|_{op}\|V\|^{L+1}}{h},\end{aligned}$$

by Lemma B.10. ■

The following lemma bounds the difference between $g_s(V)$ and $g_s(W)$ for any sample s when the weight matrices V and W differ in a single layer.

Lemma B.12 *Consider $V = (V_1, \dots, V_{L+1})$ and $W = (W_1, \dots, W_{L+1})$, and $\ell \in [L+1]$. Suppose that $V_j = W_j$ for all $j \neq \ell$, with $\|V_\ell - W_\ell\|_{op}\|V\|^{L+1} \leq 1$, and $\|V\|, \|W\| > \sqrt{L+1/2}$. Also suppose that, for all examples s, for all convex combinations $\widetilde{W}$ of V and W, we have $J_s(\widetilde{W}) \leq 2J_s(V)$. Then*

$$|g_s(V) - g_s(W)| \leq 2J_s(V)\|V_\ell - W_\ell\|_{op}\|V\|^{L+1}.$$

Proof By Taylor's theorem applied to the function $1/(1+\exp(z))$ we can bound

$$\begin{aligned}
&|g_s(V) - g_s(W)| \\
&= \left| \frac{1}{1+\exp\left(y_s f_V(x_s)\right)} - \frac{1}{1+\exp\left(y_s f_W(x_s)\right)} \right| \\
&\leq \underbrace{\frac{\exp\left(y_s f_V(x_s)\right)}{\left(1+\exp\left(y_s f_V(x_s)\right)\right)^2} |y_s f_W(x_s) - y_s f_V(x_s)|}_{=:\Xi_1} \\
&+ \underbrace{\frac{\left(y_s f_W(x_s) - y_s f_V(x_s)\right)^2}{2} \max_{\widetilde{W} \in [V,W]} \left| \frac{2\exp(2y_s f_{\widetilde{W}}(x_s))}{(\exp(y_s f_{\widetilde{W}}(x_s))+1)^3} - \frac{\exp(y_s f_{\widetilde{W}}(x_s))}{(\exp(y_s f_{\widetilde{W}}(x_s))+1)^2} \right|}_{=:\Xi_2}. \qquad (12)
\end{aligned}$$

The first term Ξ_1 can be bounded as

$$\begin{aligned}
\Xi_1 &= \frac{1}{1+\exp\left(y_s f_V(x_s)\right)} \frac{\exp\left(y_s f_V(x_s)\right)}{1+\exp\left(y_s f_V(x_s)\right)} |y_s f_W(x_s) - y_s f_V(x_s)| \\
&\leq \frac{1}{1+\exp\left(y_s f_V(x_s)\right)} |f_W(x_s) - f_V(x_s)| \\
&= g_s(V) \left\| u^W_{L+1,s} - u^V_{L+1,s} \right\| \overset{(i)}{\leq} g_s(V)\|V_\ell - W_\ell\|_{op}\|V\|^{L+1} \overset{(ii)}{\leq} J_s(V)\|V_\ell - W_\ell\|_{op}\|V\|^{L+1},
\end{aligned}$$

where (i) follows by applying Lemma B.10 and (ii) follows since $g_s(V) \leq J_s(V)$ by Lemma B.3. The second term Ξ_2

$$\begin{aligned}
\Xi_2 &= \frac{\left(y_s f_W(x_s) - y_s f_V(x_s)\right)^2}{2} \max_{\widetilde{W} \in [V,W]} \left| \frac{2\exp(2y_s f_{\widetilde{W}}(x_s))}{(\exp(y_s f_{\widetilde{W}}(x_s))+1)^3} - \frac{\exp(y_s f_{\widetilde{W}}(x_s))}{(\exp(y_s f_{\widetilde{W}}(x_s))+1)^2} \right| \\
&\overset{(i)}{\leq} \frac{\left(f_W(x_s) - f_V(x_s)\right)^2}{2} \max_{\widetilde{W} \in [V,W]} \log(1+\exp(-y_s f_{\widetilde{W}}(x_s))) \\
&= \frac{\left(f_W(x_s) - f_V(x_s)\right)^2}{2} \max_{\widetilde{W} \in [V,W]} J_s(\widetilde{W}) \\
&\overset{(ii)}{\leq} J_s(V)\left(f_W(x_s) - f_V(x_s)\right)^2 \\
&= J_s(V)\left(u^V_{L+1,s} - u^W_{L+1,s}\right)^2 \overset{(iii)}{\leq} J_s(V)\|V_\ell - W_\ell\|^2_{op}\|V\|^{2(L+1)} \overset{(iv)}{\leq} J_s(V)\|V_\ell - W_\ell\|_{op}\|V\|^{L+1},
\end{aligned}$$

where (i) follows since for every $z \in \mathbb{R}$

$$\left| \frac{2\exp(2z)}{(\exp(z)+1)^3} - \frac{\exp(z)}{(\exp(z)+1)^2} \right| \leq \log(1+\exp(-z)),$$

(ii) is by our assumption that for any $\widetilde{W} \in [V, W]$, $J_s(\widetilde{W}) \leq 2J_s(V)$, (iii) follows by invoking Lemma B.10 and finally (iv) is by the assumption that $\|V_\ell - W_\ell\|_{op}\|V\|^{L+1} \leq 1$. By using our bounds on Ξ_1 and Ξ_2 in conjunction with inequality (12) we obtain the bound

$$|g_s(V) - g_s(W)| \leq 2J_s(V)\|V_\ell - W_\ell\|_{op}\|V\|^{L+1}$$

completing the proof. ■

By using Lemmas B.10, B.11 and B.12 we will now bound the norm of the difference of the gradients of the loss at V and W, when these weight matrices differ in a single layer.

Lemma B.13 *Let $h \leq 1$, and consider $V = (V_1, \ldots, V_{L+1})$ and $W = (W_1, \ldots, W_{L+1})$, and $\ell \in [L+1]$. Suppose that $V_j = W_j$ for all $j \neq \ell$, and*

- $\|V_\ell - W_\ell\|_{op} \leq 1$;
- $\|V - W\| \leq \frac{\|V\|}{2(L+1)}$;
- $\|V\| > \sqrt{L+1/2}$ *and* $\|W\| > \sqrt{L+1/2}$;
- *for all s and all convex combinations $\widetilde{W}$ of V and W, $J_s(\widetilde{W}) \leq 2J_s(V)$.*

Then,

$$\|\nabla_V J_s(V) - \nabla_W J_s(W)\| \leq \frac{64\sqrt{(L+1)p}J_s(V)\|V\|^{3L+5}\|V_\ell - W_\ell\|}{h}.$$

Proof We can decompose $\|\nabla_V J_s(W) - \nabla_W J_s(V)\|^2$ into contributions from different layers as follows:

$$\|\nabla_V J_s(W) - \nabla_W J_s(V)\|^2 = \sum_{k=1}^{L+1} \|\nabla_{V_k} J_s(W) - \nabla_{W_k} J_s(V)\|^2 \tag{13}$$

First we seek a bound on $\|\nabla_{V_k} J_s(V) - \nabla_{W_k} J_s(W)\|_{op}$ when $k \in [L]$. (We will handle the output layer separately.) We have

$$
\begin{aligned}
&\|\nabla_{V_k} J_s(V) - \nabla_{W_k} J_s(W)\|_{op} \\
&= \left\| g_s(W) \left(\Sigma_{k,s}^W \prod_{j=k+1}^{L} \left(W_j^\top \Sigma_{j,s}^W \right) \right) W_{L+1}^\top x_{k-1,s}^{W\top} - g_s(V) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) V_{L+1}^\top x_{k-1,s}^{V\top} \right\|_{op} \\
&= \left\| g_s(W) \left(\Sigma_{k,s}^W \prod_{j=k+1}^{L} \left(W_j^\top \Sigma_{j,s}^W \right) - \Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) + \Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) W_{L+1}^\top x_{k-1,s}^{W\top} \right. \\
&\qquad \left. - g_s(V) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) V_{L+1}^\top x_{k-1,s}^{V\top} \right\|_{op} \\
&= \left\| g_s(W) \left(\Sigma_{k,s}^W \prod_{j=k+1}^{L} \left(W_j^\top \Sigma_{j,s}^W \right) - \Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) W_{L+1}^\top x_{k-1,s}^{W\top} \right. \\
&\qquad + g_s(W) \Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) W_{L+1}^\top x_{k-1,s}^{W\top} \\
&\qquad \left. - g_s(V) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) V_{L+1}^\top x_{k-1,s}^{V\top} \right\|_{op} \\
&= \left\| g_s(W) \left(\Sigma_{k,s}^W \prod_{j=k+1}^{L} \left(W_j^\top \Sigma_{j,s}^W \right) - \Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) W_{L+1}^\top x_{k-1,s}^{W\top} \right. \\
&\qquad + g_s(W) \Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) W_{L+1}^\top (x_{k-1,s}^{W\top} - x_{k-1,s}^{V\top} + x_{k-1,s}^{V\top}) \\
&\qquad \left. - g_s(V) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) V_{L+1}^\top x_{k-1,s}^{V\top} \right\|_{op} \\
&= \left\| g_s(W) \left(\Sigma_{k,s}^W \prod_{j=k+1}^{L} \left(W_j^\top \Sigma_{j,s}^W \right) - \Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) W_{L+1}^\top x_{k-1,s}^{W\top} \right. \\
&\qquad + g_s(W) \Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) W_{L+1}^\top (x_{k-1,s}^{W\top} - x_{k-1,s}^{V\top}) \\
&\qquad + g_s(W) \Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) W_{L+1}^\top x_{k-1,s}^{V\top} \\
&\qquad \left. - g_s(V) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) V_{L+1}^\top x_{k-1,s}^{V\top} \right\|_{op}.
\end{aligned}
$$

Applying the triangle inequality

$$
\begin{aligned}
&\|\nabla_{V_k} J_s(V) - \nabla_{W_k} J_s(W)\|_{op} \\
&\leq \underbrace{\left\| g_s(W) \left(\Sigma^V_{k,s} \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) W_{L+1}^\top (x^{W\top}_{k-1,s} - x^{V\top}_{k-1,s}) \right\|_{op}}_{=:\Xi_1} \\
&\quad + \underbrace{\left\| g_s(W) \left(\Sigma^V_{k,s} \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) W_{L+1}^\top x^{V\top}_{k-1,s} - g_s(V) \left(\Sigma^V_{k,s} \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) V_{L+1}^\top x^{V\top}_{k-1,s} \right\|_{op}}_{=:\Xi_2} \\
&\quad + \underbrace{\left\| g_s(W) \left(\Sigma^W_{k,s} \prod_{j=k+1}^{L} \left(W_j^\top \Sigma^W_{j,s} \right) - \Sigma^V_{k,s} \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) W_{L+1}^\top x^{W\top}_{k-1,s} \right\|_{op}}_{=:\Xi_3}. \qquad (14)
\end{aligned}
$$

We will control each of these three terms separately in lemmas below. First in Lemma B.14 we establish that

$$\Xi_1 \leq 4J_s(V)\|V_\ell - W_\ell\|_{op}\|V\|^{2(L+1)},$$

then in Lemma B.15 we prove that

$$\Xi_2 \leq 4J_s(V)\|V\|^{3(L+1)}\|V_\ell - W_\ell\|_{op},$$

and in Lemma B.16 we establish that

$$\Xi_3 \leq \frac{56J_s(V)\|V\|^{3L+5}\|V_\ell - W_\ell\|}{h}.$$

These three bound combined with the decomposition in (14) tells us that for any $k \in [L]$

$$
\begin{aligned}
\|\nabla_{V_k} J_s(V) - \nabla_{W_k} J_s(W)\|_{op} &\leq 4J_s(W)\|V_\ell - W_\ell\|_{op}\|V\|^{2(L+1)} + 4J_s(V)\|V\|^{3(L+1)}\|V_\ell - W_\ell\|_{op} \\
&\quad + \frac{56J_s(V)\|V\|^{3L+5}\|V_\ell - W_\ell\|}{h} \\
&\leq \frac{64J_s(V)\|V\|^{3L+5}\|V_\ell - W_\ell\|}{h},
\end{aligned}
$$

where the previous inequality follows since $h < 1$ and $\|V\| > 1$. Since V_k and W_k are a $p \times p$-dimensional matrices, we find that

$$
\begin{aligned}
\|\nabla_{V_k} J_s(V) - \nabla_{W_k} J_s(W)\| &\leq \sqrt{p}\|\nabla_{V_k} J_s(V) - \nabla_{W_k} J_s(W)\|_{op} \\
&\leq \frac{64\sqrt{p}J_s(V)\|V\|^{3L+5}\|V_\ell - W_\ell\|}{h}. \qquad (15)
\end{aligned}
$$

For the final layer we know that

$$\begin{aligned}
\|\nabla_{V_{L+1}} J_s(V) - \nabla_{W_{L+1}} J_s(W)\| &= \|g_s(V)x_{L,s}^V - g_s(W)x_{L,s}^W\| \\
&= \|(g_s(V) - g_s(W) + g_s(W))x_{L,s}^V - g_s(W)x_{L,s}^W\| \\
&\leq |g_s(V) - g_s(W)|\|x_{L,s}^V\| + g_s(W)\|x_{L,s}^V - x_{L,s}^W\| \\
&\overset{(i)}{\leq} 2J_s(V)\|V_\ell - W_\ell\|_{op}\|V\|^{2(L+1)} + g_s(W)\|V_\ell - W_\ell\|_{op}\|V\|^{L+1} \\
&\overset{(ii)}{\leq} 2J_s(V)\|V_\ell - W_\ell\|_{op}\|V\|^{2(L+1)} + 2J_s(V)\|V_\ell - W_\ell\|_{op}\|V\|^{L+1} \\
&\overset{(iii)}{\leq} 4J_s(V)\|V_\ell - W_\ell\|_{op}\|V\|^{2(L+1)}, \qquad (16)
\end{aligned}$$

where (i) follows by invoking Lemma B.12 and Lemma B.10, (ii) follows since $g_s(W) \leq J_s(W)$ by Lemma B.3 and because by assumption $J_s(W) \leq 2J_s(V)$, and (iii) follows since $\|V\| > 1$. This previous inequality along with (15) and (13) yield

$$\begin{aligned}
&\|\nabla_V J_s(W) - \nabla_W J_s(V)\|^2 \\
&\qquad \leq L\left(\frac{64\sqrt{p}J_s(V)\|V\|^{3L+5}\|V_\ell - W_\ell\|}{h}\right)^2 + \left(4J_s(V)\|V_\ell - W_\ell\|_{op}\|V\|^{2(L+1)}\right)^2 \\
&\qquad \leq (L+1)\left(\frac{64\sqrt{p}J_s(V)\|V\|^{3L+5}\|V_\ell - W_\ell\|}{h}\right)^2.
\end{aligned}$$

Taking square roots completes the proof. ■

As promised in the proof of Lemma B.13 we now bound Ξ_1.

Lemma B.14 *Borrowing the setting and notation of Lemma B.13, if Ξ_1 is as defined in (14), we have*

$$\Xi_1 \leq 4J_s(V)\|V_\ell - W_\ell\|_{op}\|V\|^{2(L+1)}.$$

Proof Unpacking using the definition of Ξ_1

$$
\begin{aligned}
\Xi_1 &= \left\| g_s(W) \left(\Sigma^V_{k,s} \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) W_{L+1}^\top (x^{W\top}_{k-1,s} - x^{V\top}_{k-1,s}) \right\|_{op} \\
&= \left\| g_s(W) \left(\Sigma^V_{k,s} \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) (W_{L+1}^\top - V_{L+1}^\top + V_{L+1}^\top)(x^{W\top}_{k-1,s} - x^{V\top}_{k-1,s}) \right\|_{op} \\
&\leq g_s(W) \left\| \Sigma^V_{k,s} \right\|_{op} \left(\prod_{j=k+1}^{L} \left\| V_j^\top \right\|_{op} \left\| \Sigma^V_{j,s} \right\|_{op} \right) \left\| W_{L+1}^\top - V_{L+1}^\top + V_{L+1}^\top \right\|_{op} \left\| x^{W\top}_{k-1,s} - x^{V\top}_{k-1,s} \right\| \\
&\overset{(i)}{\leq} g_s(W) \left(\prod_{j=k+1}^{L} \left\| V_j^\top \right\|_{op} \right) \left\| W_{L+1}^\top - V_{L+1}^\top + V_{L+1}^\top \right\|_{op} \left\| x^{W\top}_{k-1,s} - x^{V\top}_{k-1,s} \right\| \\
&\leq g_s(W) \left(\prod_{j=k+1}^{L} \left\| V_j^\top \right\|_{op} \right) \left(\left\| W_{L+1}^\top - V_{L+1}^\top \right\|_{op} + \left\| V_{L+1}^\top \right\|_{op} \right) \left\| x^{W\top}_{k-1,s} - x^{V\top}_{k-1,s} \right\| \\
&\overset{(ii)}{\leq} g_s(W) \|V_\ell - W_\ell\|_{op} \|V\|^{L+1} \left(\prod_{j=k+1}^{L} \left\| V_j^\top \right\|_{op} \right) \left(\left\| W_{L+1}^\top - V_{L+1}^\top \right\|_{op} + \left\| V_{L+1}^\top \right\|_{op} \right) \\
&= g_s(W) \|V_\ell - W_\ell\|_{op} \|V\|^{L+1} \left(\|W_\ell - V_\ell\|_{op} \prod_{j=k+1}^{L} \left\| V_j^\top \right\|_{op} + \prod_{j=k+1}^{L+1} \left\| V_j^\top \right\|_{op} \right) \\
&\overset{(iii)}{\leq} g_s(W) \|V_\ell - W_\ell\|_{op} \|V\|^{L+1} \left(\|W_\ell - V_\ell\|_{op} \|V\|^{L+1} + \|V\|^{L+1} \right) \\
&\overset{(iv)}{\leq} 2 J_s(V) \|V_\ell - W_\ell\|_{op} \|V\|^{2(L+1)} \left(\|W_\ell - V_\ell\|_{op} + 1 \right) \\
&\overset{(v)}{\leq} 4 J_s(V) \|V_\ell - W_\ell\|_{op} \|V\|^{2(L+1)},
\end{aligned}
$$

where (i) follows since $\|\Sigma^V_{k,s}\|_{op} \leq 1$, (ii) follows from invoking Lemma B.10, (iii) is by Lemma B.7, (iv) follows since $g_s(W) \leq J_s(W)$ by Lemma B.3 and because by assumption $J_s(W) \leq 2J_s(V)$. Finally (v) follows since we have assumed that $\|V_\ell - W_\ell\|_{op} \leq 1$. ■

We continue and now bound Ξ_2 which as defined in the proof of Lemma B.13.

Lemma B.15 *Borrowing the setting and notation of Lemma B.13, if Ξ_2 is as defined in (14) then*

$$\Xi_2 \leq 4 J_s(V) \|V\|^{3(L+1)} \|V_\ell - W_\ell\|_{op}.$$

Proof Unpacking the term Ξ_2

$$
\begin{aligned}
&\Xi_2 \\
&= \left\| g_s(W) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) W_{L+1}^\top x_{k-1,s}^{V\top} - g_s(V) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) V_{L+1}^\top x_{k-1,s}^{V\top} \right\|_{op} \\
&= \Bigg\| g_s(W) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) (W_{L+1}^\top - V_{L+1}^\top + V_{L+1}^\top) x_{k-1,s}^{V\top} \\
&\qquad - g_s(V) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) V_{L+1}^\top x_{k-1,s}^{V\top} \Bigg\|_{op} \\
&\leq \left\| g_s(W) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) (W_{L+1}^\top - V_{L+1}^\top) x_{k-1,s}^{V\top} \right\|_{op} \\
&\qquad + \left\| g_s(W) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) V_{L+1}^\top x_{k-1,s}^{V\top} - g_s(V) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) V_{L+1}^\top x_{k-1,s}^{V\top} \right\|_{op} \\
&= \underbrace{\left\| g_s(W) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) (W_{L+1}^\top - V_{L+1}^\top) x_{k-1,s}^{V\top} \right\|_{op}}_{\spadesuit_2} \\
&\qquad + \underbrace{\left\| (g_s(W) - g_s(V)) \left(\Sigma_{k,s}^V \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}^V \right) \right) V_{L+1}^\top x_{k-1,s}^{V\top} \right\|_{op}}_{=:\blacktriangledown_2} . \qquad (17)
\end{aligned}
$$

The first term

$$
\begin{aligned}
\spadesuit_2 &= \left\| g_s(W) \left(\Sigma^V_{k,s} \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) (W_{L+1}^\top - V_{L+1}^\top) x_{k-1,s}^{V\top} \right\|_{op} \\
&\le g_s(W) \left\| \Sigma^V_{k,s} \right\|_{op} \left(\prod_{j=k+1}^{L} \|V_j^\top\|_{op} \|\Sigma^V_{j,s}\|_{op} \right) \|W_{L+1}^\top - V_{L+1}^\top\|_{op} \|x_{k-1,s}^{V\top}\| \\
&\overset{(i)}{\le} g_s(W) \left(\prod_{j=k+1}^{L} \|V_j^\top\|_{op} \right) \|W_{L+1}^\top - V_{L+1}^\top\|_{op} \|x_{k-1,s}^{V\top}\| \\
&\overset{(ii)}{\le} g_s(W) \|V\|^{L+1} \|W_{L+1}^\top - V_{L+1}^\top\|_{op} \|x_{k-1,s}^{V\top}\| \\
&\le g_s(W) \|V\|^{L+1} \|V_\ell - W_\ell\|_{op} \|x_{k-1,s}^{V\top}\| \\
&\overset{(iii)}{\le} g_s(W) \|V\|^{L+1} \|V_\ell - W_\ell\|_{op} \|V\|^{L+1} \\
&= g_s(W) \|V\|^{2(L+1)} \|V_\ell - W_\ell\|_{op} \\
&\overset{(iv)}{\le} 2J_s(V) \|V\|^{2(L+1)} \|V_\ell - W_\ell\|_{op}
\end{aligned}
$$

where (i) follows since $\|\Sigma^V_{k,s}\|_{op} \le 1$, (ii) is by invoking Lemma B.7, (iii) follows due to Lemma B.10, and (iv) is because $g_s(W) \le J_s(W)$ by Lemma B.3 and by the assumption $J_s(W) \le 2J_s(V)$.

Moving on to $\heartsuit_2$,

$$
\begin{aligned}
\heartsuit_2 &= \left\| (g_s(W) - g_s(V)) \left(\Sigma^V_{k,s} \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) V_{L+1}^\top x_{k-1,s}^{V\top} \right\|_{op} \\
&\le |g_s(W) - g_s(V)| \|\Sigma^V_{k,s}\|_{op} \left(\prod_{j=k+1}^{L} \|V_j^\top\|_{op} \|\Sigma^V_{j,s}\|_{op} \right) \|V_{L+1}^\top\|_{op} \|x_{k-1,s}^{V\top}\| \\
&\le |g_s(W) - g_s(V)| \left(\prod_{j=k+1}^{L+1} \|V_j^\top\|_{op} \right) \|x_{k-1,s}^{V\top}\| \\
&\overset{(i)}{\le} |g_s(W) - g_s(V)| \|V\|^{L+1} \|x_{k-1,s}^{V\top}\| \\
&\overset{(ii)}{\le} |g_s(W) - g_s(V)| \|V\|^{2(L+1)} \\
&\overset{(iii)}{\le} 2J_s(V) \|V_\ell - W_\ell\|_{op} \|V\|^{3(L+1)},
\end{aligned}
$$

where (i) follows by Lemma B.7, (ii) is by Lemma B.10 and (iii) is by invoking Lemma B.12. Combining the bounds on $\spadesuit_2$ and $\heartsuit_2$ along with (17) we find that

$$
\begin{aligned}
\Xi_2 &\le 2J_s(V) \|V\|^{2(L+1)} \|V_\ell - W_\ell\|_{op} + 2J_s(V) \|V\|^{3(L+1)} \|V_\ell - W_\ell\|_{op} \\
&\le 4J_s(V) \|V\|^{3(L+1)} \|V_\ell - W_\ell\|_{op},
\end{aligned}
$$

where the previous inequality follows since $\|V\| > 1$. ■

Finally we bound Ξ_3 which as defined in the proof of Lemma B.13.

Lemma B.16 *Borrowing the setting and notation of Lemma B.13, if Ξ_3 is as defined in (14) then*

$$\Xi_3 \leq \frac{56 J_s(V)\|V\|^{3L+5}\|V_\ell - W_\ell\|}{h}.$$

Proof Since $g_s(W) \leq J_s(W)$ and $J_s(W) \leq 2J_s(V)$ (by assumption) we have that

$$\begin{aligned}
\Xi_3 &= \left\| g_s(W) \left(\Sigma^W_{k,s} \prod_{j=k+1}^{L} \left(W_j^\top \Sigma^W_{j,s} \right) - \Sigma^V_{k,s} \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) W^\top_{L+1} x^{W\top}_{k-1,s} \right\|_{op} \\
&\leq J_s(W) \left\| \left(\Sigma^W_{k,s} \prod_{j=k+1}^{L} \left(W_j^\top \Sigma^W_{j,s} \right) - \Sigma^V_{k,s} \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) W^\top_{L+1} x^{W\top}_{k-1,s} \right\|_{op} \\
&\leq 2J_s(V)\|W^\top_{L+1}\|_{op}\|x^{W\top}_{k-1,s}\| \left\| \Sigma^W_{k,s} \prod_{j=k+1}^{L} \left(W_j^\top \Sigma^W_{j,s} \right) - \Sigma^V_{k,s} \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right\|_{op} \\
&= 2J_s(V)\|W^\top_{L+1}\|_{op}\|x^{W\top}_{k-1,s}\| \left\| \Sigma^W_{k,s} \prod_{j=k+1}^{L} \left(W_j^\top \Sigma^W_{j,s} \right) - (\Sigma^V_{k,s} - \Sigma^W_{k,s} + \Sigma^W_{k,s}) \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right\|_{op} \\
&\leq \underbrace{2J_s(V)\|W^\top_{L+1}\|_{op}\|x^{W\top}_{k-1,s}\| \left\| \Sigma^W_{k,s} \left(\prod_{j=k+1}^{L} \left(W_j^\top \Sigma^W_{j,s} \right) - \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) \right\|_{op}}_{=:\spadesuit_3} \\
&\quad + \underbrace{2J_s(V)\|W^\top_{L+1}\|_{op}\|x^{W\top}_{k-1,s}\| \left\| (\Sigma^V_{k,s} - \Sigma^W_{k,s}) \prod_{j=k+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right\|_{op}}_{=:\clubsuit_3}. \qquad (18)
\end{aligned}$$

Before we bound $\spadesuit_3$ and $\clubsuit_3$, let us establish a few useful bounds. First note that for any layer j by Lemma B.11

$$\|\Sigma^V_{j,s} - \Sigma^W_{j,s}\|_{op} \leq \frac{\|V_\ell - W_\ell\|_{op}\|V\|^{L+1}}{h}. \qquad (19)$$

Also we know that

$$\begin{aligned}
\|x^{W\top}_{k-1,s}\| \leq \|x^{V\top}_{k-1,s}\| + \|x^{W\top}_{k-1,s} - x^{V\top}_{k-1,s}\| &\leq \|V\|^{L+1} + \|V\|^{L+1}\|V_\ell - W_\ell\|_{op} \\
&\leq \|V\|^{L+1}\left(1 + \|V_\ell - W_\ell\|_{op}\right) \\
&\leq 2\|V\|^{L+1}. \qquad (20)
\end{aligned}$$

Finally,

$$\begin{aligned}\|W_{L+1}\|_{op} \le \|V_{L+1}\|_{op} + \|V_{L+1} - W_{L+1}\|_{op} &\le \|V\| + \|V_\ell - W_\ell\|_{op} \\ &\le 2\|V\|, \end{aligned} \tag{21}$$

where the last inequality follows by our assumptions that $\|V_\ell - W_\ell\|_{op} \le 1$ and $\|V\| > 1$.

With these bounds in place we are ready to bound $\spadesuit_3$:

$$\begin{aligned}
\spadesuit_3 &= 2J_s(V)\|W_{L+1}^\top\|_{op}\|x_{k-1,s}^{W\top}\| \left\|\Sigma_{k,s}^W \left(\prod_{j=k+1}^{L}\left(W_j^\top\Sigma_{j,s}^W\right) - \prod_{j=k+1}^{L}\left(V_j^\top\Sigma_{j,s}^V\right)\right)\right\|_{op} \\
&\le 2J_s(V)\|W_{L+1}^\top\|_{op}\|x_{k-1,s}^{W\top}\| \left\|\Sigma_{k,s}^W\right\| \left\|\prod_{j=k+1}^{L}\left(W_j^\top\Sigma_{j,s}^W\right) - \prod_{j=k+1}^{L}\left(V_j^\top\Sigma_{j,s}^V\right)\right\|_{op} \\
&\overset{(i)}{\le} 8J_s(V)\|V\|^{L+2} \left\|\prod_{j=k+1}^{L}\left(W_j^\top\Sigma_{j,s}^W\right) - \prod_{j=k+1}^{L}\left(V_j^\top\Sigma_{j,s}^V\right)\right\|_{op} \\
&\overset{(ii)}{\le} 12J_s(V)\|V\|^{L+2}\left(\|V\| + \|V-W\|\right)^{L+1}\left(\left(\frac{\|V_\ell - W_\ell\|_{op}\|V\|^{L+1}}{h}\right)\|V\| + \|V-W\|\right) \\
&= \frac{12J_s(V)\|V\|^{3L+5}\|V-W\|}{h}\left(1+\frac{\|V-W\|}{\|V\|}\right)^{L+1}\left(\frac{\|V_\ell - W_\ell\|_{op}}{\|V-W\|} + \frac{h}{\|V\|^{L+2}}\right) \\
&\overset{(iii)}{\le} \frac{24J_s(V)\|V\|^{3L+5}\|V-W\|}{h}\left(1+\frac{\|V-W\|}{\|V\|}\right)^{L+1} \\
&\overset{(iv)}{\le} \frac{24J_s(V)\|V\|^{3L+5}\|V-W\|}{h}\left(1+\frac{2(L+1)\|V-W\|}{\|V\|}\right) \\
&\overset{(v)}{\le} \frac{48J_s(V)\|V\|^{3L+5}\|V-W\|}{h}
\end{aligned} \tag{22}$$

where (i) follows by using the bounds in (20) and (21), (ii) follows by invoking Lemma B.8 and using (19), (iii) follows since $h \le 1$ and $\|V\| \ge 1$ by assumption, and therefore

$$\frac{\|V_\ell - W_\ell\|_{op}}{\|V-W\|} + \frac{h}{\|V\|^{L+2}} \le 2,$$

inequality (iv) follows since for any $0 < z < \frac{1}{L}$, $(1+z)^{L+1} \le 1 + 2(L+1)z$ and because by assumption $\|V-W\| \le \|V\|/(2(L+1))$, and finally (v) is again because $\|V-W\| \le \|V\|/(2(L+1))$.

Let's turn our attention to $\clubsuit_3$.

$$\begin{aligned}
\clubsuit_3 &= 2J_s(V)\|W_{L+1}^\top\|_{op}\|x_{k-1,s}^{W\top}\| \left\|(\Sigma_{k,s}^W - \Sigma_{k,s}^V)\prod_{j=k+1}^{L}\left(V_j^\top \Sigma_{j,s}^V\right)\right\|_{op} \\
&\overset{(i)}{\leq} 8J_s(V)\|V\|^{L+2}\left\|(\Sigma_{k,s}^W - \Sigma_{k,s}^V)\prod_{j=k+1}^{L}\left(V_j^\top \Sigma_{j,s}^V\right)\right\|_{op} \\
&\leq 8J_s(V)\|V\|^{L+2}\left\|\Sigma_{k,s}^W - \Sigma_{k,s}^V\right\|_{op}\prod_{j=k+1}^{L}\|V_j^\top\|_{op}\|\Sigma_{j,s}^V\|_{op} \\
&\overset{(ii)}{\leq} 8J_s(V)\|V\|^{2L+3}\left\|\Sigma_{k,s}^W - \Sigma_{k,s}^V\right\|_{op} \\
&\overset{(iii)}{\leq} \frac{8J_s(V)\|V\|^{3L+4}\left\|V_\ell - W_\ell\right\|_{op}}{h},
\end{aligned} \tag{23}$$

where (i) follows from the bounds in (20) and (21), (ii) follows by invoking Lemma B.7 and (iii) is by inequality (19).

By combining the bounds in (22) and (23) we have a bound on Ξ_3.

$$\begin{aligned}
\Xi_3 &\leq \frac{48J_s(V)\|V\|^{3L+5}}{h} + \frac{8J_s(V)\|V\|^{3L+4}\left\|V_\ell - W_\ell\right\|_{op}}{h} \\
&\leq \frac{56J_s(V)\|V\|^{3L+5}\|V - W\|}{h} \\
&= \frac{56J_s(V)\|V\|^{3L+5}\|V_\ell - W_\ell\|}{h},
\end{aligned}$$

which completes the proof. ■

Lemma B.13 provides a bound on the norm of the difference between $\nabla_V J_s(V)$ and $\nabla_V J_s(W)$, when the weight matrices V and W differ only at a single layer. We next invoke Lemma B.13 $(L+1)$ times to bound the norm of the difference between the gradients of the loss at V and W when they potentially differ in all of the layers.

Lemma B.17 *Let $h \leq 1$, and consider $V = (V_1, \ldots, V_{L+1})$ and $W = (W_1, \ldots, W_{L+1})$, such that the following are satisfied for all $j \in [L+1]$:*

- $\|V - W\| \leq \frac{\|V\|}{6L+10}$;
- $\|V\| > \sqrt{L + 1/2}$ *and* $\|W\| > \sqrt{L + 1/2}$.

For every $j \in \{0, \ldots, L+1\}$ define $T(j) := (W_1, W_2, \ldots, W_j, V_{j+1}, \ldots, V_{L+1})$. Suppose that for all $j \in [L+1]$, for all examples s, and for all convex combinations $\widetilde{W}$ of $T(j)$ and $T(j+1)$, $J_s(\widetilde{W}) \leq 2J_s(T(j)) \leq 4J_s(V)$. Then

$$\|\nabla_V J(V) - \nabla_W J(W)\| \leq \frac{256(L+1)\sqrt{p}J(V)\|V\|^{3L+5}\|V - W\|}{h}.$$

Proof We may transform V into W by swapping one layer at a time. For any $s \in [n]$ Lemma B.13 bounds the norm of difference in each swap, thus,

$$\begin{aligned}
\|\nabla_V J_s(V) - \nabla_W J_s(W)\| &= \left\| \sum_{k=0}^{L} \left(\nabla_{T(k)} J_s(T(k)) - \nabla_{T(k+1)} J_s(T(k+1)) \right) \right\| \\
&\leq \sum_{k=0}^{L} \left\| \nabla_{T(k)} J_s(T(k)) - \nabla_{T(k+1)} J_s(T(k+1)) \right\| \\
&\leq \sum_{k=1}^{L+1} \frac{64\sqrt{(L+1)p} J_s(T(k)) \|T(k)\|^{3L+5} \|V_k - W_k\|}{h} \\
&= \frac{64\sqrt{(L+1)p}}{h} \sum_{k=1}^{L+1} J_s(T(k)) \|T(k)\|^{3L+5} \|V_k - W_k\| \\
&\leq \frac{128\sqrt{(L+1)p} J_s(V)}{h} \sum_{k=1}^{L+1} \|T(k)\|^{3L+5} \|V_k - W_k\|, \qquad (24)
\end{aligned}$$

where the final inequality follows from the assumption that $J_s(T(k)) \leq 2J_s(V)$. For any $k \in [L+1]$

$$\begin{aligned}
\|T(k)\|^{3L+5} = \|V\|^{3L+5} \left(\frac{\|T(k)\|}{\|V\|} \right)^{3L+5} &= \|V\|^{3L+5} \left(\frac{\|T(k) - V + V\|}{\|V\|} \right)^{3L+5} \\
&\leq \|V\|^{3L+5} \left(1 + \frac{\|T(k) - V\|}{\|V\|} \right)^{3L+5} \\
&\leq \|V\|^{3L+5} \left(1 + \frac{\|W - V\|}{\|V\|} \right)^{3L+5} \\
&\overset{(i)}{\leq} \|V\|^{3L+5} \left(1 + \frac{(3L+5)\|W - V\|}{\|V\|} \right) \\
&\overset{(ii)}{\leq} 2\|V\|^{3L+5},
\end{aligned}$$

where (i) follows since for any non-negative $z < \frac{1}{3L+5}$, $(1+z)^{3L+5} \leq 1 + (6L+10)z$ and because by assumption $\|V - W\|/\|V\| \leq \frac{1}{6L+10}$, and (ii) again follows by our assumption that $\|V - W\|/\|V\| \leq \frac{1}{6L+10}$. Using this bound in inequality (24)

$$\begin{aligned}
\|\nabla_V J_s(V) - \nabla_W J_s(W)\| &\leq \frac{256\sqrt{(L+1)p} J_s(V) \|V\|^{3L+5}}{h} \sum_{k=1}^{L+1} \|V_k - W_k\| \\
&\leq \frac{256\sqrt{(L+1)p} J_s(V) \|V\|^{3L+5}}{h} \left(\sqrt{L+1} \|V - W\| \right) \\
&= \frac{256(L+1)\sqrt{p} J_s(V) \|V\|^{3L+5} \|V - W\|}{h}.
\end{aligned}$$

Thus,

$$\begin{aligned}\|\nabla_V J(V) - \nabla_W J(W)\| &= \left\| \frac{1}{n} \sum_{s\in[n]} \nabla_V J_s(V) - \nabla_W J_s(W) \right\| \\ &\le \frac{1}{n} \sum_{s\in[n]} \|\nabla_V J_s(V) - \nabla_W J_s(W)\| \\ &\le \frac{1}{n} \sum_{s\in[n]} \frac{256(L+1)\sqrt{p} J_s(V)\|V\|^{3L+5}\|V-W\|}{h} \\ &= \frac{256(L+1)\sqrt{p} J(V)\|V\|^{3L+5}\|V-W\|}{h}\end{aligned}$$

completing the proof. ■

Lemma 4.2 *If $h \le 1$, for any weights V such that $\|V\| \ge \sqrt{L+1/2}$, we have*

$$\mathrm{Lip}(\nabla_V J(V)) \le \frac{256(L+1)\sqrt{p}\|V\|^{3L+5} J(V)}{h}.$$

Proof Since the function $J(\cdot)$ is continuous, for all close enough W the assumptions of Lemma B.17 are satisfied. ■

B.5. Proof of Lemma 4.3

Lemma 4.3 *For any weight matrix V such that $\|V\| \ge \sqrt{L+1/2}$ then*

$$\|\nabla_V J(V)\| \le \sqrt{(L+1)p}\|V\|^{L+1} \min\{J(V), 1\}.$$

Proof For any $\ell \in [L]$ the formula for the gradient of the loss with respect to V_ℓ is given by (see equation (2a))

$$\frac{\partial J(V; x_s, y_s)}{\partial V_\ell} = g_s(V) \left(\Sigma^V_{\ell,s} \prod_{j=\ell+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) V_{L+1}^\top x^{V\top}_{\ell-1,s},$$

therefore its operator norm

$$\begin{aligned}\left\| \frac{\partial J(V; x_s, y_s)}{\partial V_\ell} \right\|_{op} &= g_s(V) \left\| \left(\Sigma^V_{\ell,s} \prod_{j=\ell+1}^{L} \left(V_j^\top \Sigma^V_{j,s} \right) \right) V_{L+1}^\top x^{V\top}_{\ell-1,s} \right\|_{op} \\ &\le g_s(V) \|\Sigma^V_{\ell,s}\|_{op} \left(\prod_{j=\ell+1}^{L} \|V_j^\top\|_{op} \|\Sigma^V_{j,s}\|_{op} \right) \|V_{L+1}^\top\|_{op} \|x^{V\top}_{\ell-1,s}\| \\ &\le g_s(V) \left(\prod_{j=\ell+1}^{L+1} \|V_j\|_{op} \right) \|x^V_{\ell-1,s}\| \qquad (25)\end{aligned}$$

where the last step follows since $\|\Sigma_{j,s}^V\|_{op} \leq \max_z |\phi'(z)| < 1$. By its definition

$$\begin{aligned}\|x_{\ell-1,s}^V\| = \|\phi\left(V_{\ell-1}\phi\left(\cdots\phi(V_1x_s)\right)\right)\| &\overset{(i)}{\leq} \|V_{\ell-1}\phi\left(\cdots\phi(V_1x_s)\right)\| \\ &\leq \|V_{\ell-1}\|_{op}\|\phi\left(\cdots\phi(V_1x)\right)\| \\ &\leq \left(\prod_{j=1}^{\ell-1}\|V_j\|_{op}\right)\|x_s\| \overset{(ii)}{\leq} \left(\prod_{j=1}^{\ell-1}\|V_j\|_{op}\right),\end{aligned}$$

where (i) follows since ϕ is contractive (Lemma B.9) and (ii) is because $\|x_s\| = 1$. Along with inequality (25) this implies

$$\left\|\frac{\partial J(V;x_s,y_s)}{\partial V_\ell}\right\|_{op} \leq g_s(V)\prod_{j\neq\ell}\|V_j\|_{op} \leq g_s(V)\prod_{j\neq\ell}\|V_j\| \leq g_s(V)\|V\|^{L+1},$$

where the last inequality follows from Lemma B.7. Therefore we have

$$\left\|\frac{\partial J(V)}{\partial V_\ell}\right\|_{op} = \left\|\frac{1}{n}\sum_{s\in[n]}\frac{\partial J(V;x_s,y_s)}{\partial V_\ell}\right\|_{op} \leq \frac{1}{n}\sum_{s\in[n]}\left\|\frac{\partial J(V;x_s,y_s)}{\partial V_\ell}\right\|_{op} \leq \frac{\|V\|^{L+1}}{n}\sum_{s\in[n]}g_s(V).$$

We know that $g_s(V) \leq J_s(V)$ by Lemma B.3 and also that $g_s(V) < 1$. Therefore,

$$\left\|\frac{\partial J(V)}{\partial V_\ell}\right\|_{op} \leq \frac{\|V\|^{L+1}}{n}\min\left\{\sum_s J_s(V), n\right\} \leq \|V\|^{L+1}\min\left\{J(V), 1\right\}.$$

Given that V_ℓ is a $p\times p$ matrix we infer

$$\left\|\frac{\partial J(V)}{\partial V_\ell}\right\| \leq \sqrt{p}\left\|\frac{\partial J(V)}{\partial V_\ell}\right\|_{op} \leq \sqrt{p}\|V\|^{L+1}\min\left\{J(V), 1\right\}. \tag{26}$$

When $\ell = L+1$

$$\frac{\partial J(V;x_s,y_s)}{\partial V_{L+1}} = g_s(V)x_{L,s}^{V\top},$$

by using the same chain of logic as in the case of $\ell < L+1$ we can obtain the bound

$$\left\|\frac{\partial J(V)}{\partial V_\ell}\right\| \leq \sqrt{p}\|V\|^{L+1}\min\left\{J(V), 1\right\}.$$

Summing up over all layers

$$\|\nabla J(V)\|^2 = \sum_{\ell=1}^{L+1}\left\|\frac{\partial J(V)}{\partial V_\ell}\right\|^2 \leq (L+1)p\|V\|^{2(L+1)}\left(\min\left\{J(V), 1\right\}\right)^2,$$

hence, taking squaring roots completes the proof. ■

B.6. Proof of Lemma 4.4

Lemma 4.4 *If* $h \leq 1$, $J_t < \frac{1}{n^{1+24L}}$, *and*

$$\alpha J_t \leq \frac{h}{1024\,(L+1)^2\sqrt{p}\|V^{(t)}\|^{3L+5}},$$

then

$$J_{t+1} \leq J_t - \frac{\alpha L\|\nabla J_t\|^2}{L+\frac{1}{2}}.$$

Proof Since, by assumption, $J_t < \frac{1}{n^{1+24L}}$, Lemma B.5 implies $\|V^{(t)}\| > \sqrt{L+1}$. We would like to apply Lemmas 4.2 and 4.3. To apply these lemmas, we first bound the norm of all convex combinations of $V^{(t)}$ and $V^{(t+1)}$ from above and below. Consider $W = \eta V^{(t)} + (1-\eta)V^{(t+1)} = V^{(t)} - (1-\eta)\alpha\nabla J_t$ for any $\eta \in [0,1]$. An upper bound on the norm raised to the $3L+5$th power is

$$\begin{aligned}
\|W\|^{3L+5} = \|V^{(t)} - (1-\eta)\alpha\nabla J_t\|^{3L+5} &= \|V^{(t)}\|^{3L+5}\left(\frac{\|V^{(t)} - (1-\eta)\alpha\nabla J_t\|}{\|V^{(t)}\|}\right)^{3L+5} \\
&\leq \|V^{(t)}\|^{3L+5}\left(\frac{\|V^{(t)}\| + \alpha\|\nabla J_t\|}{\|V^{(t)}\|}\right)^{3L+5} \\
&= \|V^{(t)}\|^{3L+5}\left(1 + \frac{\alpha\|\nabla J_t\|}{\|V^{(t)}\|}\right)^{3L+5} \\
&\overset{(i)}{\leq} \|V^{(t)}\|^{3L+5}\left(1 + \frac{\alpha(\sqrt{(L+1)p}J_t\|V^{(t)}\|^{L+1})}{\|V^{(t)}\|}\right)^{3L+5} \\
&= \|V^{(t)}\|^{3L+5}\left(1 + \alpha(\sqrt{(L+1)p}J_t\|V^{(t)}\|^{L})\right)^{3L+5} \\
&\overset{(ii)}{\leq} \|V^{(t)}\|^{3L+5}\left(1 + (6L+10)(\sqrt{(L+1)p}\alpha J_t\|V^{(t)}\|^{L}\right) \\
&\leq 2\|V^{(t)}\|^{3L+5} \qquad (27)
\end{aligned}$$

where (i) follows by invoking Lemma 4.3 and (ii) follows since for any $0 < z < 1/(3L+5)$, $(1+z)^{3L+5} \leq 1 + (6L+10)z$ and because the step-size α is chosen such that

$$\begin{aligned}
\alpha(\sqrt{(L+1)p}J_t\|V^{(t)}\|^{L}) &\leq \frac{h}{1024(L+1)^2\sqrt{p}J_t\|V^{(t)}\|^{3L+5}} \cdot \sqrt{(L+1)p}J_t\|V^{(t)}\|^{L} \\
&= \frac{h}{1024(L+1)^{3/2}\|V^{(t)}\|^{2L+5}} \leq \frac{1}{6L+10}.
\end{aligned}$$

Thus, we have shown that the norm of $W \in [V^{(t)}, V^{(t+1)}]$ raised to the $3L+5$th power is bounded by $2\|V^{(t)}\|^{3L+5}$. Next we lower bound the norm of W,

$$\begin{aligned}
\|W\| = \|V^{(t)} - (1-\eta)\alpha\nabla J_t\| &\geq \|V^{(t)}\|\left(1 - \frac{\alpha\|\nabla J_t\|}{\|V^{(t)}\|}\right) \\
&\overset{(i)}{\geq} \|V^{(t)}\|\left(1 - \alpha\sqrt{(L+1)p}J_t\|V^{(t)}\|^L\right) \\
&\overset{(ii)}{>} \sqrt{L+1}\left(1 - \alpha\sqrt{(L+1)p}J_t\|V^{(t)}\|^L\right) \\
&\overset{(iii)}{>} \sqrt{L+1}\left(1 - \frac{1}{6L+10}\right) \\
&> \sqrt{L+1/2},
\end{aligned}$$

where (i) follows by again invoking Lemma 4.3, (ii) is by Lemma B.5 that guarantees that $\|V^{(t)}\| > \sqrt{L+1}$ since $J_t < \frac{1}{n^{1+24L}}$ and (iii) is by the logic above that guarantees that

$$\alpha\left(\sqrt{(L+1)p}J_t\|V^{(t)}\|^L\right) \leq \frac{1}{6L+10}.$$

Thus we have also shown that $\|W\| > \sqrt{L+1/2}$ for any $W \in [V^{(t)}, V^{(t+1)}]$.

In order to apply Lemma 4.1 (that shows that the loss decreases along a gradient step when the loss is smooth along the path), we would like to bound $\mathrm{Lip}(\nabla_W J(W))$ for all convex combinations W of $V^{(t)}$ and $V^{(t+1)}$. For $N = \left\lceil \frac{2\sqrt{(L+1)p}\|V^{(t)}\|^{L+1}\|V^{(t+1)}-V^{(t)}\|}{J_t} \right\rceil$, (similarly to the proof of Lemma E.8 of (Lyu and Li, 2020)) we will prove the following by induction

> For all $s \in \{0, \ldots, N\}$, for all $\eta \in [0, s/N]$, for $W = \eta V^{(t+1)} + (1-\eta)V^{(t)}$, $\mathrm{Lip}(\nabla_W J(W)) \leq \frac{1024(L+1)\sqrt{p}J_t\|V^{(t)}\|^{3L+5}}{h}$.

The base case, where $s = 0$, follows directly from Lemma 4.2. Now, assume that the inductive hypothesis holds from some s, and, for $\eta \in (s/N, (s+1)/N]$, consider $W = \eta V^{(t+1)} + (1-\eta)V^{(t)}$. Let $\widetilde{W} = (s/N)V^{(t+1)} + (1-s/N)V^{(t)}$. Since the step-size α is small enough, applying Lemma 4.1 along with the inductive hypothesis yields $J(\widetilde{W}) \leq J_t$. Applying Lemma 4.3 (which provides a bound on the Lipschitz constant of J)

$$\begin{aligned}
J(W) &\leq J(\widetilde{W}) + (\sqrt{(L+1)p}\max_{\bar{W}\in[W,\widetilde{W}]}\|\bar{W}\|^{L+1})\|W - \widetilde{W}\| \\
&\overset{(i)}{\leq} J(\widetilde{W}) + (2\sqrt{(L+1)p})\|V^{(t)}\|^{L+1}\|W - \widetilde{W}\| \\
&\leq J(\widetilde{W}) + \frac{(2\sqrt{(L+1)p})\|V^{(t)}\|^{L+1}\|V^{(t+1)} - V^{(t)}\|}{N} \\
&= J(\widetilde{W}) + J_t \\
&\leq 2J_t,
\end{aligned}$$

where (i) follows since $\max_{\bar{W}\in[W,\widetilde{W}]}\|\bar{W}\|^{L+1} \leq 2\|V^{(t)}\|^{L+1}$ by using the same logic used to arrive at inequality (27). Applying Lemmas B.5 and 4.2, this implies that for any $W \in [V^{(t)}, V^{(t+1)}]$

$$\mathrm{Lip}(\nabla_W J(W)) \leq \frac{256(L+1)\sqrt{p}J(W)\|W\|^{3L+5}}{h} \leq \frac{1024(L+1)\sqrt{p}J_t\|V^{(t)}\|^{3L+5}}{h},$$

completing the proof of the inductive step.

So, now we know that, for all convex combinations W of $V^{(t)}$ and $V^{(t+1)}$, $\mathrm{Lip}(\nabla_W J(W)) \leq \frac{1024(L+1)\sqrt{p}J_t\|V^{(t)}\|^{3L+5}}{h}$. By our choice of step size $\alpha < \frac{1}{L+\frac{1}{2}} \cdot \frac{h}{1024(L+1)\sqrt{p}J_t\|V^{(t)}\|^{3L+5}}$, so by applying Lemma 4.1, we have that

$$J_{t+1} \leq J_t - \frac{L}{L+\frac{1}{2}}\alpha\|\nabla J_t\|^2$$

which is the desired result. ■

B.7. Proof of Lemma 4.5

Lemma 4.5 *For all $L \in \mathbb{N}$ if $h \leq h_{\max}$, $J_t < \frac{1}{n^{1+24L}}$, and $\|V^{(t)}\|^L \leq \log(1/J_t)\frac{\|V^{(1)}\|^L}{\log(1/J_1)}$ then*

$$\|\nabla J_t\| \geq \frac{(L+\frac{3}{4})J_t\log(1/J_t)}{\|V^{(t)}\|}. \tag{6}$$

Proof We have

$$\begin{aligned}\|\nabla J_t\| = \sup_{a:\|a\|=1}(\nabla J_t \cdot a) &\geq (\nabla J_t)\cdot\left(\frac{-V^{(t)}}{\|V^{(t)}\|}\right)\\ &= \frac{1}{\|V^{(t)}\|}\sum_{\ell\in[L+1]}\nabla_{V_\ell}J_t\cdot\left(-V_\ell^{(t)}\right).\end{aligned} \tag{28}$$

Note that by definition,

$$\nabla_{V_\ell}J_t\cdot\left(-V_\ell^{(t)}\right) = \frac{1}{n}\sum_{s\in[n]}\nabla_{V_\ell}J_{ts}\cdot\left(-V_\ell^{(t)}\right). \tag{29}$$

Consider two cases.

Case 1: (When $\ell = L+1$) In this case, for any $s \in [n]$ by the formula for the gradient in (2b) we have

$$\nabla_{V_{L+1}}J_{ts}\cdot\left(-V_{L+1}^{(t)}\right) = g_{ts}y_sV_{L+1}^{(t)}x_{L,s}^{(t)} = g_{ts}y_sf_{V^{(t)}}(x_s)$$

and therefore

$$\nabla_{V_{L+1}}J_t\cdot\left(-V_{L+1}^{(t)}\right) = \frac{1}{n}\sum_{s\in[n]}g_{ts}y_sf_{V^{(t)}}(x_s). \tag{30}$$

Case 2: (When $\ell \in [L]$) Below we will prove the claim (in Lemma B.18) that for any $\ell \in [L]$

$$\nabla_{V_\ell^{(t)}}J_t\cdot\left(-V_\ell^{(t)}\right) \geq \frac{1}{n}\sum_{s\in[n]}g_{ts}\left[y_sf_{V^{(t)}}(x_s) - \frac{\sqrt{p}h\|V^{(t)}\|^L}{2L^{\frac{L}{2}-1}}\right]. \tag{31}$$

By combining this with the results of inequalities (28) and (30)

$$\begin{aligned}\|\nabla J_t\| &\geq \frac{L+1}{n\|V^{(t)}\|}\sum_{s\in[n]} g_{ts}y_s f_{V^{(t)}}(x_s) - \frac{L\sqrt{p}h\|V^{(t)}\|^L}{2L^{\frac{L}{2}-1}\|V^{(t)}\|}\left[\frac{1}{n}\sum_{s\in[n]} g_{ts}\right] \\ &\overset{(i)}{\geq} \frac{L+1}{n\|V^{(t)}\|}\sum_{s\in[n]} g_{ts}y_s f_{V^{(t)}}(x_s) - \frac{L\sqrt{p}h\|V^{(t)}\|^L}{2L^{\frac{L}{2}-1}\|V^{(t)}\|}J_t \\ &\overset{(ii)}{\geq} \frac{L+1}{n\|V^{(t)}\|}\sum_{s\in[n]} g_{ts}y_s f_{V^{(t)}}(x_s) - \frac{L\sqrt{p}h\|V^{(1)}\|^L\log(1/J_t)}{2\log(1/J_1)L^{\frac{L}{2}-1}\|V^{(t)}\|}J_t \end{aligned} \tag{32}$$

where (i) follows because $g_{ts} \leq J_{ts}$ by Lemma B.3 and (ii) follows by our assumption on $\|V^{(t)}\|$.

For every sample s, $J_{ts} = \log\left(1+\exp\left(-y_s f_{V^{(t)}}(x_s)\right)\right)$ which implies

$$y_s f_{V^{(t)}}(x_s) = \log\left(\frac{1}{\exp(J_{ts})-1}\right) \quad \text{and} \quad g_{ts} = \frac{1}{1+\exp(y_s f_{V^{(t)}}(x_s))} = 1-\exp\left(-J_{ts}\right).$$

Plugging this into inequality (32) we derive,

$$\|\nabla J_t\| \geq \frac{L+1}{n\|V\|}\sum_{s=1}^{n}\left(1-\exp\left(-J_{ts}\right)\right)\log\left(\frac{1}{\exp(J_{ts})-1}\right) - \frac{L\sqrt{p}h\|V^{(1)}\|^L}{2\log(1/J_1)L^{\frac{L}{2}-1}\|V^{(t)}\|}J_t\log(1/J_t).$$

Observe that the function $(1-\exp(-z))\log\left(\frac{1}{\exp(z)-1}\right)$ is continuous and concave (when the inputs lie between 0 and 1) with

$$\lim_{z\to 0^+}(1-\exp(-z))\log\left(\frac{1}{\exp(z)-1}\right) \;=\; 0.$$

Also recall that $\sum_s J_{ts} = J_t n$ and that $J_{ts} \leq J_t n \leq 1/n^{24L} \leq 1$. Therefore applying Lemma B.4 to the function ψ with $\psi(0)=0$ and $\psi(z) = (1-\exp(-z))\log\left(\frac{1}{\exp(z)-1}\right)$ for $z>0$, we get that

$$\begin{aligned}\|\nabla J_t\| \geq \frac{L+1}{\|V^{(t)}\|}&\left[\frac{1-\exp(-J_t n)}{n}\log\left(\frac{1}{\exp(J_t n)-1}\right)\right] \\ &- \frac{L\sqrt{p}h\|V^{(1)}\|^L}{2\log(1/J_1)L^{\frac{L}{2}-1}\|V^{(t)}\|}J_t\log(1/J_t).\end{aligned} \tag{33}$$

We know that for any $z \in [0,1]$

$$\exp(z) \leq 1+2z \quad \text{and} \quad \exp(-z) \leq 1-z+z^2.$$

Since $J_t < \frac{1}{n^{1+24L}}$ and $n \geq 3$, these bounds on the exponential function combined with inequality (33) yields

$$\begin{aligned}
&\|\nabla J_t\| \\
&\geq \frac{L+1}{\|V^{(t)}\|}\left[(J_t - nJ_t^2)\log\left(\frac{1}{2J_t n}\right)\right] - \frac{L\sqrt{p}h\|V^{(1)}\|^L}{2\log(1/J_1)L^{\frac{L}{2}-1}\|V^{(t)}\|}J_t\log(1/J_t) \\
&= \frac{(L+1)J_t\log(1/J_t)}{\|V^{(t)}\|}\left[1 - nJ_t - \frac{\log(2n)}{\log(1/J_t)} - \frac{\sqrt{p}h\|V^{(1)}\|^L}{2\log(1/J_1)L^{\frac{L}{2}-2}}\right] \\
&= \frac{(L+3/4)J_t\log(1/J_t)}{\|V^{(t)}\|}\left[1 + \frac{1}{4(L+\frac{3}{4})}\right]\left[1 - nJ_t - \frac{\log(2n)}{\log(1/J_t)} - \frac{\sqrt{p}h\|V^{(1)}\|^L}{2\log(1/J_1)L^{\frac{L}{2}-2}}\right]. \quad (34)
\end{aligned}$$

By the choice of $h \leq h_{\max}$ we have

$$\frac{\sqrt{p}h\|V^{(1)}\|^L}{2\log(1/J_1)L^{\frac{L}{2}-2}} \leq \frac{1}{48L}.$$

Next, since $J_t < \frac{1}{n^{1+24L}}$ and $n \geq 3$

$$\frac{\log(2n)}{\log(1/J_t)} \leq \frac{\log(2)+\log(n)}{(1+24L)\log(n)} \leq \frac{1}{12L}$$

and

$$nJ_t < \frac{1}{3^{24L}} \leq \frac{1}{48L}.$$

Therefore, using these three bounds in conjunction with inequality (34) yields the bound

$$\|\nabla J_t\| \geq \frac{(L+3/4)J_t\log(1/J_t)}{\|V^{(t)}\|}\left[1 + \frac{1}{4(L+\frac{3}{4})}\right]\left[1 - \frac{1}{8L}\right] \geq \frac{(L+\frac{3}{4})J_t\log(1/J_t)}{\|V^{(t)}\|},$$

which establishes the desired bound. ■

As promised above we now lower bound the inner product between the gradient of the loss with respect to $V_\ell^{(t)}$ and the weight matrix for any $\ell \in [L]$.

Lemma B.18 *Under the conditions of Lemma 4.5 and borrowing all notation from its proof, for all* $\ell \in [L]$

$$\nabla_{V_\ell} J_t \cdot \left(-V_\ell^{(t)}\right) \geq \frac{1}{n}\sum_{s\in[n]} g_{ts}\left[y_s f_{V^{(t)}}(x_s) - \frac{\sqrt{p}h\|V^{(t)}\|^L}{2L^{\frac{L}{2}-1}}\right].$$

Proof To ease notation, let us drop the (t) in the superscript and refer to $V^{(t)}$ as V. Recall that for any matrices A and B, $A \cdot B = \mathrm{vec}(A) \cdot \mathrm{vec}(B) = \mathrm{Tr}(A^\top B)$. Also recall the formula for the

gradient of the loss in (2a), therefore, for any $s \in [n]$

$$
\begin{aligned}
&\nabla_{V_\ell} J_{ts} \cdot (-V_\ell) \\
&= -\mathrm{Tr}\left(V_\ell^\top \nabla_{V_\ell} J_{ts}\right) \\
&= g_{ts} y_s \mathrm{Tr}\left(V_\ell^\top \left(\Sigma_{\ell,s} \prod_{j=\ell+1}^{L} V_j^\top \Sigma_{j,s}\right) V_{L+1}^\top x_{\ell-1,s}^\top\right) \\
&= g_{ts} y_s \mathrm{Tr}\left(\prod_{j=\ell}^{L} \left(V_j^\top \Sigma_{j,s}\right) V_{L+1}^\top x_{\ell-1,s}^\top\right) \\
&\overset{(i)}{=} g_{ts} y_s \mathrm{Tr}\left(x_{\ell-1,s}^\top \prod_{j=\ell}^{L} \left(V_j^\top \Sigma_{j,s}\right) V_{L+1}^\top\right) \\
&= g_{ts} y_s x_{\ell-1,s}^\top \prod_{j=\ell}^{L} \left(V_j^\top \Sigma_{j,s}\right) V_{L+1}^\top \\
&\overset{(ii)}{=} g_{ts} y_s x_{L,s}^\top V_{L+1}^\top + g_{ts} y_s \sum_{k=\ell}^{L-1} \left(\left(x_{k-1,s}^\top V_k^\top \Sigma_{k,s} - x_{k,s}^\top\right) \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}\right) V_{L+1}^\top\right) \\
&\qquad + g_{ts} y_s \left(x_{L-1,s}^\top V_L^\top \Sigma_{L,s} - x_{L,s}^\top\right) V_{L+1}^\top \\
&\overset{(iii)}{=} g_{ts} y_s f_{V^{(t)}}(x_s) + g_{ts} y_s \sum_{k=\ell}^{L-1} \left(\left(x_{k-1,s}^\top V_k^\top \Sigma_{k,s} - x_{k,s}^\top\right) \prod_{j=k+1}^{L} \left(V_j^\top \Sigma_{j,s}\right) V_{L+1}^\top\right) \\
&\qquad + g_{ts} y_s \left(x_{L-1,s}^\top V_L^\top \Sigma_{L,s} - x_{L,s}^\top\right) V_{L+1}^\top,
\end{aligned}
$$

where (i) follows by the cyclic property of the trace, and (ii) follows since the second term and third term in the equation form a telescoping sum, and (iii) is because $f_{V^{(t)}}(x_s) = V_{L+1} x_{L,s}$ by definition. By the property of h-smoothly approximately ReLU activations, for any $z \in \mathbb{R}$ we know that $|\phi'(z)z - \phi(z)| \le \frac{h}{2}$. Therefore for any $k \in [L]$, $\|x_{k-1,s}^\top V_k^\top \Sigma_{k,s} - x_{k,s}^\top\|_\infty \le \frac{h}{2}$ and hence $\|x_{k-1,s}^\top V_k^\top \Sigma_{k,s} - x_{k,s}^\top\| \le \frac{\sqrt{p}h}{2}$. Continuing from the previous displayed equation, by applying the

Cauchy-Schwarz inequality we find

$$
\begin{aligned}
&\nabla_{V_\ell} J_{ts} \cdot (-V_\ell) \\
&\geq g_{ts} y_s f_{V^{(t)}}(x_s) - g_{ts} \sum_{k=\ell}^{L-1} \|x_{k-1,s}^\top V_k^\top \Sigma_{k,s} - x_{k,s}^\top\| \prod_{j=k+1}^{L} \left\|V_j^\top\right\|_{op} \|\Sigma_{j,s}\|_{op} \left\|V_{L+1}^\top\right\| \\
&\qquad\qquad - g_{ts} \|x_{L-1,s}^\top V_L^\top \Sigma_{L,s} - x_{L,s}^\top\| \left\|V_{L+1}^\top\right\| \\
&\geq g_{ts} y_s f_{V^{(t)}}(x_s) - \frac{\sqrt{p} h g_{ts}}{2} \sum_{k=\ell}^{L} \prod_{j=k+1}^{L+1} \|V_j\|_{op} \|\Sigma_{j,s}\|_{op} \\
&\overset{(i)}{\geq} g_{ts} y_s f_{V^{(t)}}(x_s) - \frac{\sqrt{p} h g_{ts}}{2} \sum_{k=\ell}^{L} \prod_{j=k+1}^{L+1} \|V_j\|_{op} \\
&\geq g_{ts} y_s f_{V^{(t)}}(x_s) - \frac{\sqrt{p} L h g_{ts}}{2} \max_{k\in[L]} \prod_{j=k+1}^{L+1} \|V_j\|_{op} \\
&\overset{(ii)}{\geq} g_{ts} y_s f_{V^{(t)}}(x_s) - \frac{\sqrt{p} L h g_{ts}}{2} \max_{k\in[L]} \prod_{j=k+1}^{L+1} \|V_j\| \\
&\overset{(iii)}{\geq} g_{ts} y_s f_{V^{(t)}}(x_s) - \frac{\sqrt{p} h \|V\|^L g_{ts}}{2L^{\frac{L}{2}-1}}
\end{aligned} \tag{35}
$$

where (i) follows since $\phi' \leq 1$ and therefore $\|\Sigma_{j,s}\|_{op} \leq 1$, (ii) follows since for any matrix M, $\|M\|_{op} \leq \|M\|$ and inequality (iii) follows by invoking Lemma B.6 since we know that $\|V\| > \sqrt{L+1}$ by Lemma B.5. The previous display along with the decomposition in equation (29) yields

$$
\nabla_{V_\ell} J_t \cdot (-V_\ell) \geq \frac{1}{n} \sum_{s\in[n]} g_{ts} \left[y_s f_{V^{(t)}}(x_s) - \frac{\sqrt{p} h \|V\|^L}{2L^{\frac{L}{2}-1}} \right]
$$

which completes our proof of this claim. ■

Now that we have proved all the lemmas stated in Section 4.1, the reader can next jump to Section 4.2.

Appendix C. An Example Where the Margin in Assumption 3.2 is Constant

In this section we provide an example where the margin γ in Assumption 3.2 is constant. Consider a two-layer Huberized ReLU network. In this section we always let ϕ denote the Huberized ReLU activation (see its definition in equation (1)). Since here we are only concerned with the properties of the network at initialization, let $V^{(1)}$ be denoted simply by V. The first layer $V_1 \in \mathbb{R}^{p\times p}$ has its entries drawn independently from $\mathcal{N}\left(0, \frac{2}{p}\right)$ and $V_2 \in \mathbb{R}^{1\times p}$ has its entries drawn independently from $\mathcal{N}(0,1)$.

Let $V_{1,i}$ denote the ith row of V_1 and let $V_{2,i}$ denote the ith coordinate of V_2. The network computed by these weights is $f_V(x) = V_2\phi(V_1 x)$.

Consider data in which examples of each class are clustered. There is a unit vector $\mu \in \mathbb{S}^{p-1}$ such that, for all s with $y_s = 1$, $\|x_s - \mu\| \leq r$, and, for all s with $y_s = -1$, $\|x_s - (-\mu)\| \leq r$. Let us say that such data is r-*clustered*. (Recall that $\|x_s\| = 1$ for all s.)

Proposition C.1 *For any $\delta > 0$, suppose that $h \leq \frac{\sqrt{\pi}}{2p}$, $r \leq \min\left\{\frac{1}{16}, \frac{\sqrt{p}h}{c'\sqrt{\log\left(\frac{3pn}{\delta}\right)}}\right\}$, and $p \geq \log^{c'}(n/\delta)$ for a large enough constant $c' > 0$. If the data r-clustered then, with probability $1 - \delta$ there exists $W^\star = (W_1^\star, W_2^\star)$ with $\|W^\star\| = 1$ such that*

$$\text{for all } \; s \in [n], \qquad y_s \left(\nabla_V f_V(x_s) \cdot W^\star\right) \geq c\sqrt{p}$$

where c is a positive absolute constant.

Proof Define a set

$$\mathcal{S} := \left\{ i \in [p] : \frac{1}{2} \leq |V_{2,i}| \leq 2 \right\},$$

and also define

$$\mathcal{S}_+ := \{i \in \mathcal{S} : V_{1,i} \cdot \mu \geq 4h\} \quad \text{ and } \quad \mathcal{S}_- := \{i \in \mathcal{S} : -V_{1,i} \cdot \mu \geq 4h\}.$$

Consider an event $\mathcal{E}_{\mathsf{margin}}$ such that all of the following simultaneously occur:

(a) $p\left(\frac{1}{4} - o_p(1)\right) \leq |\mathcal{S}_+| \leq p\left(\frac{1}{2} + o_p(1)\right)$;

(b) $p\left(\frac{1}{4} - o_p(1)\right) \leq |\mathcal{S}_-| \leq p\left(\frac{1}{2} + o_p(1)\right)$;

(c) for all $s \in [n]$ and $i \in [p]$, $|V_{1,i} \cdot (x_s - y_s\mu)| \leq 2h$.

Using simple concentration arguments in Lemma C.2 below we will show that $\mathbb{P}\left[\mathcal{E}_{\mathsf{margin}}\right] \geq 1 - \delta$. Let us assume that the event $\mathcal{E}_{\mathsf{margin}}$ holds for the remainder of the proof.

The gradient of f with respect to $V_{1,i}$ is

$$\nabla_{V_{1,i}} f_V(x) = x\left(V_{2,i}\phi'(V_{1,i} \cdot x)\right).$$

Consider a sample with index s with $y_s = 1$. For any $i \in \mathcal{S}_+$

$$\begin{aligned}
\operatorname{sign}(V_{2,i})\left(\mu \cdot \nabla_{V_{1,i}} f_V(x_s)\right) &= \mu \cdot x_s \left(|V_{2,i}|\phi'(V_{1,i} \cdot x_s)\right) \\
&= \left(|V_{2,i}|\phi'(V_{1,i} \cdot x_s)\right) + \mu \cdot (x_s - \mu)\left(|V_{2,i}|\phi'(V_{1,i} \cdot x_s)\right) \\
&\overset{(i)}{\geq} \frac{1}{2}\phi'(V_{1,i} \cdot \mu + V_{1,i} \cdot (x_s - \mu)) - 2\mu \cdot (x_s - \mu)\phi'(V_{1,i} \cdot x_s) \\
&\overset{(ii)}{\geq} \frac{1}{2}\phi'(V_{1,i} \cdot \mu + V_{1,i} \cdot (x_s - \mu)) - \frac{1}{8} \\
&\overset{(iii)}{\geq} \frac{\phi'(2h)}{2} - \frac{1}{8} \overset{(iv)}{=} \frac{1}{2} - \frac{1}{8} = \frac{3}{8}
\end{aligned} \tag{36}$$

where (i) follows since $\frac{1}{2} \leq |V_{2,i}| \leq 2$ when $i \in \mathcal{S}_+$. Inequality (ii) follows since ϕ' is bounded by 1 and because $\|x_s - y_s\mu\| \leq r \leq 1/16$. Inequality (iii) follows since $i \in \mathcal{S}_+$ and therefore

$(V_{1,i}) \cdot \mu \geq 4h$, under event $\mathcal{E}_{\mathsf{margin}}$, $(V_{1,i}) \cdot (x_s - \mu) \geq -2h$, and since ϕ' is a monotonically increasing function. Equation (iv) follows since $\phi'(2h) = 1$. On the other hand, for any $i \in \mathcal{S}_-$:

$$\begin{aligned}\operatorname{sign}(V_{2,i})\mu \cdot \nabla_{V_{1,i}} f_V(x_s) &= \mu \cdot x_s \left(|V_{2,i}|\phi'(V_{1,i} \cdot x_s)\right) \\ &= |V_{2,i}|\phi'(V_{1,i} \cdot x_s) + \mu \cdot (x_s - \mu) \left(|V_{2,i}|\phi'(V_{1,i} \cdot x_s)\right) \\ &\overset{(i)}{\geq} -2\mu \cdot (x_s - \mu)\phi'(V_{1,i} \cdot x_s) \overset{(ii)}{\geq} -\frac{1}{8} \end{aligned} \tag{37}$$

where (i) follows since $|V_{2,i}| \leq 2$ when $i \in \mathcal{S}_-$ and ϕ' is always non-negative. Inequality (ii) again follows since ϕ' is bounded by 1 and because $\|x_s - y_s\mu\| \leq r \leq 1/16$.

Similarly we can also show that for a sample s with $y_s = -1$, for any $i \in \mathcal{S}_-$

$$\operatorname{sign}(V_{2,i})\mu \cdot \nabla_{V_{1,i}} f_V(x_s) \leq \frac{-3}{8} \tag{38}$$

and for any $i \in \mathcal{S}_+$

$$\operatorname{sign}(V_{2,i})\mu \cdot \nabla_{V_{1,i}} f_V(x_s) \leq \frac{1}{8}. \tag{39}$$

With these calculations in place let us construct $W^\star = (W_1^\star, W_2^\star)$ where, $W_1^\star \in \mathbb{R}^{p \times p}$, $W_2^\star \in \mathbb{R}^{1 \times p}$ and $\|W^\star\| = 1$. Set $W_2^\star = 0$. For all $i \in \mathcal{S}_+ \cup \mathcal{S}_-$ set

$$W_{1,i}^\star = \operatorname{sign}(V_{2,i})\mu \frac{1}{\sqrt{|\mathcal{S}_+| + |\mathcal{S}_-|}}$$

and for all $i \notin \mathcal{S}_+ \cup \mathcal{S}_-$, set $W_{1,i}^\star = 0$. We can easily check that $\|W^\star\| = 1$ (since $\|\mu\| = 1$). Thus, for any sample s with $y_s = 1$

$$\begin{aligned} & y_s \left(\nabla_V f_V(x_s) \cdot W^\star\right) \\ &= \nabla_{V_1} f_V(x_s) \cdot W_1^\star \\ &= \sum_{i \in \mathcal{S}_+} \nabla_{V_{1,i}} f_V(x_s) \cdot W_{1,i}^\star + \sum_{i \in \mathcal{S}_-} \nabla_{V_{1,i}} f_V(x_s) \cdot W_{1,i}^\star \\ &= \frac{1}{\sqrt{|\mathcal{S}_+| + |\mathcal{S}_-|}} \left[\sum_{i \in \mathcal{S}_+} \operatorname{sign}(V_{2,i})\mu \cdot \nabla_{V_{1,i}} f_V(x_s) + \sum_{i \in \mathcal{S}_-} \operatorname{sign}(V_{2,i})\mu \cdot \nabla_{V_{1,i}} f_V(x_s)\right] \\ &\overset{(i)}{\geq} \frac{1}{\sqrt{|\mathcal{S}_+| + |\mathcal{S}_-|}} \left[\frac{3|\mathcal{S}_+|}{8} - \frac{|\mathcal{S}_-|}{8}\right] \\ &= \frac{1}{8\sqrt{|\mathcal{S}_+| + |\mathcal{S}_-|}} \left[3|\mathcal{S}_+| - |\mathcal{S}_-|\right] \\ &\overset{(ii)}{\geq} \frac{1}{8\sqrt{p\left(1 + o_p(1)\right)}} \left[3p\left(\frac{1}{4} - o_p(1)\right) - p\left(\frac{1}{2} + o_p(1)\right)\right] \geq c\sqrt{p} \end{aligned}$$

where (i) follows by using inequalities (36) and (37) and (ii) follows by Parts (a) and (b) of the event $\mathcal{E}_{\mathsf{margin}}$. The final inequality follows since we assume that p is greater than a constant. This shows that it is possible to achieve a margin of $c\sqrt{p}$ on the positive examples. By mirroring the logic

above and using inequalities (38) and (39) we can show that a margin of $c\sqrt{p}$ can also be attained on the negative examples. This completes our proof. ■

As promised we now show that the event $\mathcal{E}_{\mathsf{margin}}$ defined above occurs with probability at least $1-\delta$.

Lemma C.2 *For the event $\mathcal{E}_{\mathsf{margin}}$ be defined in the proof of Proposition C.1 above,*

$$\mathbb{P}\left[\mathcal{E}_{\mathsf{margin}}\right] \geq 1-\delta.$$

Proof We shall show that each of the three sub-events in the definition of the event $\mathcal{E}_{\mathsf{margin}}$ occur with probability at least $1-\delta/3$. Then a union bound establishes the statement of the lemma.

Proof of Part (a): Recall the definition of the set $\mathcal{S}$

$$\mathcal{S} := \left\{ i \in [p] : \frac{1}{2} \leq |V_{2,i}| \leq 2 \right\},$$

and also the definition of the set $\mathcal{S}_+$

$$\mathcal{S}_+ := \left\{ i \in \mathcal{S} : V_{1,i} \cdot \mu \geq 4h \right\}.$$

We will first derive a high probability bound the size of the set $\mathcal{S}$, and then use this bound to control the size of $\mathcal{S}_+$. A trivial upper bound is $|\mathcal{S}| \leq p$. Let us derive a lower bound on its size. Define the random variable $\zeta_i = \mathbb{I}\left[\frac{1}{2} \leq |V_{2,i}| \leq 2\right]$. It is easy to check that $|\mathcal{S}| = \sum_{i\in[p]} \zeta_i$. The expected value of this random variable

$$\begin{aligned}\mathbb{E}\left[\zeta_i\right] = 1 - \mathbb{P}\left[|V_{2,i}| \leq \frac{1}{2}\right] - \mathbb{P}\left[|V_{2,i}| \geq 2\right] &\overset{(i)}{\geq} 1 - \frac{1/2-(-1/2)}{\sqrt{2\pi}} - \mathbb{P}\left[|V_{2,i}| \geq 2\right] \\ &\overset{(ii)}{\geq} 1 - \frac{1}{\sqrt{2\pi}} - \frac{\exp\left(-2\right)}{\sqrt{2\pi}} > \frac{1}{2}\end{aligned}$$

where (i) follows since $V_{2,i} \sim \mathcal{N}(0,1)$ so its density is upper bounded bounded by $1/\sqrt{2\pi}$, and (ii) follows by a Mill's ratio bound to upper bound $\mathbb{P}\left[|V_{2,i}| \geq z\right] \leq 2 \times \frac{\exp(-z^2/2)}{\sqrt{2\pi}z}$. A Hoeffding bound (see Theorem F.5) implies that for any $\eta \geq 0$

$$\mathbb{P}\left[|\mathcal{S}| \geq p\mathbb{E}\left[\zeta_i\right] - \frac{\eta p}{2}\right] \geq 1 - \exp\left(-c_1\eta^2 p\right).$$

Setting $\eta = 1/p^{1/4}$ we get

$$\mathbb{P}\left[|\mathcal{S}| \geq p\left(\frac{1}{2} - \frac{1}{p^{1/4}}\right)\right] \geq 1 - \exp\left(-c_1\sqrt{p}\right). \tag{40}$$

We now will bound $|\mathcal{S}_+|$ conditioned on the event in the previous display: $p\left(\frac{1}{2} - \frac{1}{p^{1/4}}\right) \leq |\mathcal{S}| \leq p$.

For each $i \in \mathcal{S}$, the random variable $V_{1,i} \cdot \mu \sim \mathcal{N}\left(0, \frac{2}{p}\right)$ since each entry of $V_{1,i}$ is drawn independently from $\mathcal{N}\left(0, \frac{2}{p}\right)$ and because $\|\mu\| = 1$. Define a random variable $\xi_i := \mathbb{I}\left[V_{1,i} \cdot \mu \geq 4h\right]$. It is easy to check that $|\mathcal{S}_+| = \sum_{i\in\mathcal{S}} \xi_i$. The expected value of ξ_i

$$\begin{aligned}\left|\mathbb{E}\left[\xi_i\right] - \frac{1}{2}\right| = \left|\mathbb{P}\left[V_{1,i}\cdot\mu \geq 4h\right] - \frac{1}{2}\right| &= \left|\mathbb{P}\left[V_{1,i}\cdot\mu \geq 0\right] - \mathbb{P}\left[V_{1,i}\cdot\mu \in [0,4h)\right] - \frac{1}{2}\right| \\ &= \mathbb{P}\left[V_{1,i}\cdot\mu \in [0,4h)\right] \overset{(i)}{\leq} \frac{4h\sqrt{p}}{\sqrt{2\pi}\sqrt{2}} \overset{(ii)}{\leq} \frac{1}{\sqrt{p}}\end{aligned}$$

where (i) follows since the density of this Gaussian is upper bounded by $\frac{1}{\sqrt{2\pi}\times\left(\frac{\sqrt{2}}{\sqrt{p}}\right)}$ and (ii) is by the assumption that $h \leq \frac{\sqrt{\pi}}{2p}$. Thus we have shown that $\frac{1}{2} - \frac{1}{\sqrt{p}} \leq \mathbb{E}[\xi_i] \leq \frac{1}{2} + \frac{1}{\sqrt{p}}$. Again a Hoeffding bound (see Theorem F.5) implies that for any $\eta \geq 0$

$$\mathbb{P}\left[\left|\sum_{i\in\mathcal{S}} \xi_i - |\mathcal{S}|\mathbb{E}[\xi_i]\right| \leq \eta p \;\middle|\; p\left(\frac{1}{2} - \frac{1}{p^{1/4}}\right) \leq |\mathcal{S}| \leq p\right] \geq 1 - 2\exp\left(-c_2\eta^2 p\right).$$

By setting $\eta = 1/p^{1/4}$ we get that

$$\mathbb{P}\left[\left|\,|\mathcal{S}_+| - |\mathcal{S}|\mathbb{E}[\xi_i]\right| \leq p^{3/4} \;\middle|\; p\left(\frac{1}{2} - \frac{1}{p^{1/4}}\right) \leq |\mathcal{S}| \leq p\right] \geq 1 - 2\exp\left(-c_2\sqrt{p}\right). \tag{41}$$

By a union bound over the events in (40) and (41) we get that

$$\mathbb{P}\left[p\left(\frac{1}{4} - o_p(1)\right) \leq |\mathcal{S}_+| \leq p\left(\frac{1}{2} + o_p(1)\right)\right] \geq 1 - \exp\left(-c_1\sqrt{p}\right) - 2\exp\left(-c_2\sqrt{p}\right).$$

By assumption $p \geq \log^{c'}(n/\delta)$ for a large enough constant c', thus

$$\mathbb{P}\left[p\left(\frac{1}{4} - o_p(1)\right) \leq |\mathcal{S}_+| \leq p\left(\frac{1}{2} + o_p(1)\right)\right] \geq 1 - \delta/3$$

which completes our proof of the first part.

Proof of Part (b): The proof of this second part follows by exactly the same logic as Part (a).

Proof of Part (c): Fix any $i \in [p]$ and $s \in [n]$. Recall that $V_{1,i} \sim \mathcal{N}\left(0, \frac{2}{p}I\right)$ and by assumption $\|x_s - y_s\mu\| \leq r$. Thus the random variable $V_{1,i} \cdot (x_s - y_s\mu)$ is a zero-mean Gaussian random variable with variance at most $\frac{2r^2}{p}$. A standard Gaussian concentration bound implies that

$$\mathbb{P}\left[|V_{1,i} \cdot (x_s - y_s\mu)| \leq 2h\right] \geq 1 - 2\exp\left(-\frac{c_2 p h^2}{r^2}\right). \tag{42}$$

By a union bound over all $i \in [p]$ and all $s \in [n]$ we get

$$\mathbb{P}\left[\exists\, i \in [p], s \in [n] \;:\; |(V_{1,i}) \cdot (x_s - y_s\mu)| \leq 2h\right] \geq 1 - 2np\exp\left(-\frac{c_2 p h^2}{r^2}\right) \geq 1 - \frac{\delta}{3}$$

where the last inequality follows since $r^2 \leq \frac{ph^2}{(c')^2\log\left(\frac{3pn}{\delta}\right)}$ and because $p \geq \log^{c'}(n/\delta)$ for a large enough constant $c' > 0$. This completes our proof. ■

Appendix D. Omitted Proofs from Section 3.2

In this section we prove Theorem 3.3. We largely follow the high-level analysis strategy presented in (Chen et al., 2021) to prove that, with high probability, if the the width of the network is large enough then gradient descent drives down the loss to at most $\frac{1}{n^{1+24L}}$ under Assumption 3.2. After that we use our general result, Theorem 3.1, to prove that gradient descent continues to reduce the loss beyond this point. We begin by introducing some definitions that are useful in our proofs in this section. All the results in this section are specialized to the case of the Huberized ReLU activation function (see its definition in equation (1)).

D.1. Additional Definitions and Notation

Following Chen et al. (2021), we define the Neural Tangent random features (henceforth NT) function class. These definitions depend on the initial weights $V^{(1)}$ and radii $\tau, \rho > 0$. We shall choose the value of these radii in terms of problem parameters in the sequel. Define a ball around the initial parameters.

Definition D.1 *For any $V^{(1)}$ and $\rho > 0$ define a ball around this weight matrix as*

$$\mathcal{B}(V^{(1)}, \rho) := \left\{ V : \max_{\ell \in [L+1]} \|V_\ell - V_\ell^{(1)}\| \le \rho \right\}.$$

We then define the neural tangent kernel function class.

Definition D.2 *Given initial weights $V^{(1)}$, define the function*

$$F_{V^{(1)},V}(x) := f_{V^{(1)}}(x) + (\nabla f_{V^{(1)}}(x)) \cdot (V - V^{(1)}),$$

then the NT function class with radius $\rho > 0$ is as follows

$$\mathcal{F}(V^{(1)}, \rho) := \left\{ F_{V^{(1)},V}(x) : V \in \mathcal{B}(V^{(1)}, \rho) \right\}.$$

We continue to define the minimal error achievable by any function in this NT function class.

Definition D.3 *For any $V^{(1)}$ and any $\rho > 0$ define*

$$\varepsilon_{\mathsf{NT}}(V^{(1)}, \rho) := \min_{V \in \mathcal{B}(V^{(1)}, \rho)} \frac{1}{n} \sum_{s=1}^{n} \log(1 + \exp(-y_s F_{V^{(1)},V}(x_s))),$$

that is, it is the minimal training loss achievable by functions in the NT function class centered at $V^{(1)}$. Also let $V^\star(V^{(0)}, \rho) \in \mathcal{B}(V^{(0)}, \rho)$ be an arbitrary minimizer:

$$V^\star \in \operatorname*{arg\,min}_{V \in \mathcal{B}(V^{(1)}, \rho)} \frac{1}{n} \sum_{s=1}^{n} \log(1 + \exp(-y_s F_{V^{(1)},V}(x_s))).$$

We will be concerned with the maximum approximation error of this tangent kernel around a ball of the initial weight matrix.

Definition D.4 *For any $V^{(1)}$ and any $\tau > 0$ define*

$$\varepsilon_{\mathsf{app}}(V^{(1)}, \tau) := \sup_{s \in [n]} \sup_{\hat{V}, \widetilde{V} \in \mathcal{B}(V^{(1)}, \tau)} \left| f_{\hat{V}}(x_s) - f_{\widetilde{V}}(x_s) - \nabla f_{\widetilde{V}}(x_s) \cdot \left(\hat{V} - \widetilde{V} \right) \right|.$$

Finally we define the maximum norm of the gradient with respect to the weights of any layer.

Definition D.5 *For any initial weights $V^{(1)}$ and any $\tau > 0$ define*

$$\Gamma(V^{(1)}, \tau) := \sup_{s \in [n]} \sup_{\ell \in [L+1]} \sup_{V \in \mathcal{B}(V^{(1)}, \tau)} \|\nabla_{V_\ell} f_V(x_s)\|.$$

D.2. Technical Tools Required for the Neural Tangent Kernel Proofs

We borrow (Chen et al., 2021, Lemma 5.1) that bounds the average empirical risk in the first T iterations when the iterates remain in a ball around the initial weight matrix. We have translated the lemma into our notation.

Lemma D.6 *Set the step-size $\alpha_t = \alpha = O\left(\frac{1}{L\Gamma(V^{(1)},\tau)^2}\right)$ for all $t \in [T]$. Suppose that given an initialization $V^{(1)}$ and radius $\rho > 0$ we pick $\tau > 0$ such that $V^\star \in \mathcal{B}(V^{(1)},\tau)$ and $V^{(t)} \in \mathcal{B}(V^{(1)},\tau)$ for all $t \in [T]$, and that $\varepsilon_{\mathsf{app}}(V^{(1)},\tau) < 3/8$. Then*

$$\frac{1}{T}\sum_{t=1}^{T} J(V^{(t)}) \leq \frac{\|V^{(1)} - V^\star\|^2 - \|V^{(T+1)} - V^\star\|^2 + 2T\alpha\varepsilon_{\mathsf{NT}}(V^{(1)},\rho)}{T\alpha\left(\frac{3}{2} - 4\varepsilon_{\mathsf{app}}(V^{(1)},\tau)\right)}.$$

Technically the setting studied by Chen et al. (2021) differs from the setting that we study in our paper. They deal with neural networks with ReLU activations instead of Huberized ReLU activations that we consider here. However, it is easy to scan through the proof of their lemma to verify that it does not rely on any specific properties of ReLUs.

The next lemma bounds the approximation error of the neural tangent kernel in a neighbourhood around the initial weight matrix and provides a bound on the maximum norm of the gradient. The proof of this lemma below relies on several different lemmas that are collected and proved in Appendix E.

Lemma D.7 *For any $\delta > 0$, suppose that $\tau = \Omega\left(\frac{\log^2\left(\frac{nL}{\delta}\right)}{p^{\frac{3}{2}}L^3}\right)$ and, for a sufficiently small positive constant c, we have $\tau \leq \frac{c}{L^{12}\log^{\frac{3}{2}}(p)}$, $h \leq \frac{\tau}{\sqrt{p}}$ and $p = \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for some sufficiently large polynomial. Then, with probability at least $1-\delta$ over the random initialization $V^{(1)}$, we have*

(a) $\varepsilon_{\mathsf{app}}(V^{(1)},\tau) \leq O(\sqrt{p\log(p)}L^5\tau^{4/3})$, *and*

(b) $\Gamma(V^{(1)},\tau) \leq O(\sqrt{p}L^2)$.

Having provided a bound on the approximation error, let us continue and show that gradient descent reaches a weight matrix whose error is comparable to $\varepsilon_{\mathsf{NT}}$.

Lemma D.8 *For any $L \in \mathbb{N}$, $\delta > 0$,*

$$\tau = \Omega\left(\frac{\log^2\left(\frac{nL}{\delta}\right)}{p^{\frac{3}{2}}L^3}\right) \quad \text{and} \quad \tau \leq \frac{c}{(p\log(p))^{\frac{3}{8}}L^{\frac{15}{4}}},$$

where c is a small enough positive constant, $\rho = \frac{\tau}{3L}$, $h \leq \frac{\tau}{\sqrt{p}}$, and $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial, if we run gradient descent with a constant step-size $\alpha_t = \alpha = \Theta\left(\frac{1}{pL^5}\right)$, for $T = \left\lceil \frac{(L+1)\rho^2}{4\alpha\cdot\varepsilon_{\mathsf{NT}}(V^{(1)},\rho)} \right\rceil$ iterations, with probability $1-\delta$ over the random initialization

$$\min_{t\in[T]} J(V^{(t)}) \leq 6\varepsilon_{\mathsf{NT}}(V^{(1)},\rho).$$

Our proof closely follows the proof of (Chen et al., 2021, Theorem 3.3).
Proof Recall the definition of

$$V^\star \in \underset{V\in\mathcal{B}(V^{(1)},\rho)}{\arg\min}\ \frac{1}{n}\sum_{s=1}^{n}\log(1+\exp(-y_s F_{V^{(1)},V}(x_s))).$$

We would like to apply Lemma D.6 to show that the average loss of the iterates of gradient descent decreases. To do so we must first ensure that all iterates $V^{(t)}$ and $V^\star$ remain in a ball of radius τ around initialization.

We have assumed that $\tau \le \frac{c}{(p\log(p))^{\frac{3}{8}}L^{\frac{15}{4}}}$ and that $p \ge \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial. Therefore if this polynomial is large enough we have that $\tau \le \frac{c_1}{L^{12}\log^{\frac{3}{2}}(p)}$, for an arbitrarily small positive constant c_1. This means we can invoke Lemma D.7 which guarantees that with probability at least $1-\delta$, the approximation error $\varepsilon_{\mathsf{app}}(V^{(1)},\tau) \le O(\sqrt{p\log(p)}L^5\tau^{4/3})$ and the maximum norm of the gradient $\Gamma(V^{(1)},\tau) \le O(\sqrt{p}L^2)$. Again recall that $\tau \le \frac{c}{(p\log(p))^{\frac{3}{8}}L^{\frac{15}{4}}}$, where c is a small enough positive constant. Thus for a small enough value of c the approximation error $\varepsilon_{\mathsf{app}}(V^{(1)},\tau) \le \frac{1}{8}$. Let us assume that this is the case going forward.

Since $\rho = \frac{\tau}{3L} \le \tau$, $V^\star$ is clearly in $\mathcal{B}(V^{(1)},\tau)$. We will now show that the iterates $\{V^{(t)}\}_{t\in[T]}$ also lie in this ball by induction. The base case when $t=1$ is trivially true. So now assume that $V^{(1)},\dots,V^{(t-1)}$ lie in this ball and we will proceed to show that $V^{(t)}$ also lies in this ball. Since $\varepsilon_{\mathsf{app}}(V^{(1)},\tau) \le 1/8$, by Lemma D.6 we infer that

$$\frac{1}{t-1}\sum_{t'=1}^{t-1}J(V^{(t')}) \le \frac{\|V^{(1)}-V^\star\|^2 - \|V^{(t)}-V^\star\|^2 + 2(t-1)\alpha\varepsilon_{\mathsf{NT}}(V^{(1)},\rho)}{(t-1)\alpha},$$

which in turn implies that

$$\begin{aligned}\sum_{\ell\in[L+1]}\|V_\ell^{(t)}-V_\ell^\star\|^2 = \|V^{(t)}-V^\star\|^2 &\le \|V^{(1)}-V^\star\|^2 + 2\alpha(t-1)\varepsilon_{\mathsf{NT}}(V^{(1)},\rho) - \alpha\sum_{t'=1}^{t-1}J(V^{(t')}) \\ &\le \|V^{(1)}-V^\star\|^2 + 2\alpha(t-1)\varepsilon_{\mathsf{NT}}(V^{(1)},\rho) \\ &\overset{(i)}{\le} (L+1)\rho^2 + \frac{(L+1)\rho^2}{2} \le \frac{3(L+1)\rho^2}{2} \le 3L\rho^2\end{aligned}$$

where (i) follows since $V^* \in \mathcal{B}(V^{(1)},\rho)$ and $t \le T = \left\lceil \frac{(L+1)\rho^2}{4\alpha\varepsilon_{\mathsf{NT}}(V^{(1)},\rho)} \right\rceil$. Taking square roots implies that for each $\ell \in [L+1]$, $\|V_\ell^{(t)} - V_\ell^\star\| \le \sqrt{3L}\rho$. By the triangle inequality for any $\ell \in [L+1]$

$$\|V_\ell^{(t)} - V_\ell^{(1)}\| \le \|V_\ell^{(t)} - V_\ell^\star\| + \|V_\ell^\star - V_\ell^{(1)}\| \le \sqrt{3L}\rho + \rho < 3L\rho = \tau.$$

This shows that $V_\ell^{(t)} \in \mathcal{B}(V^{(1)},\tau)$ and completes the induction.

Now that we have established that $V^\star$ and $V^{(t)}$ are all in a ball of radius τ around $V^{(1)}$ we can again invoke Lemma D.6 (recall from above that $\varepsilon_{\mathsf{app}}(V^{(1)},\tau) \le \frac{1}{8}$ and $\Gamma(V^{(1)},\tau) \le O(\sqrt{p}L^2)$)

to infer that

$$\min_{t\in[T]} J(V^{(t)}) \le \frac{1}{T}\sum_{t=1}^{T} J(V^{(t)}) \le \frac{\|V^{(1)}-V^\star\|^2 - \|V^{(T+1)}-V^\star\|^2 + 2T\alpha\varepsilon_{\mathsf{NT}}(V^{(1)},\rho)}{T\alpha}$$
$$\le \frac{\|V^{(1)}-V^\star\|^2 + 2T\alpha\varepsilon_{\mathsf{NT}}(V^{(1)},\rho)}{T\alpha}$$
$$= 2\varepsilon_{\mathsf{NT}}(V^{(1)},\rho) + \frac{\|V^{(1)}-V^\star\|^2}{T\alpha} \le 6\varepsilon_{\mathsf{NT}}(V^{(1)},\rho),$$

where the last inequality follows since $V^\star \in \mathcal{B}(V^{(1)},\rho)$, therefore $\|V^{(1)}-V^\star\|^2 \le (L+1)\rho^2$ and because $T = \left\lceil \frac{(L+1)\rho^2}{4\alpha\varepsilon_{\mathsf{NT}}(V^{(1)},\rho)} \right\rceil$. This completes our proof. ■

Finally we shall show that under Assumption 3.2 the error $\varepsilon_{\mathsf{NT}}(V^{(1)},\rho)$ is bounded with high probability. Recall the assumption on the data.

Assumption 3.2 *With probability $1-\delta$ over the random initialization, there exists a collection of matrices $W^\star = (W_1^\star,\dots,W_{L+1}^\star)$ with $\|W^\star\| = 1$, such that for all samples $s \in [n]$*

$$y_s\left(\nabla f_{V^{(1)}}(x_s)\cdot W^\star\right) \ge \sqrt{p}\gamma,$$

for some $\gamma > 0$.

Lemma D.9 *Under the Assumption 3.2, for any $\varepsilon,\delta > 0$, if the radius*

$$\rho \ge \frac{c\left[\sqrt{\log(n/\delta)} + \log\left(\frac{1}{\exp(\varepsilon)-1}\right)\right]}{\sqrt{p}\gamma}$$

for some large enough positive absolute constant c then, with probability $1-2\delta$ over the randomness in the initialization

$$\varepsilon_{\mathsf{NT}}(V^{(1)},\rho) = \min_{V\in\mathcal{B}(V^{(1)},\rho)} \frac{1}{n}\sum_{s=1}^{n} \log(1+\exp(-y_s F_{V^{(1)},V}(x_s))) \le \varepsilon.$$

Proof Recall that, by definition,

$$F_{V^{(1)},V}(x) = f_{V^{(1)}}(x) + (\nabla f_{V^{(1)}}(x))\cdot(V-V^{(1)}).$$

By Assumption 3.2 we know that, with probability $1-\delta$, there exists $W^\star$ with $\|W^\star\| = 1$, such that for all $s \in [n]$

$$y_i\left(\nabla f_{V^{(1)}}(x_s)\cdot W^\star\right) \ge \sqrt{p}\gamma. \tag{43}$$

By Lemma E.11 proved below with know that

$$\mathbb{P}\left[|f_{V^{(1)}}(x_s)| \le c_1\sqrt{\log(n/\delta)}\right] \ge 1-\delta. \tag{44}$$

For the remainder of the proof let's assume that both events in (43) and (44) occur. This happens with probability at least $1-2\delta$. Thus, for any positive λ

$$y_i\left[f_{V^{(1)}}(x_s)+\lambda\nabla_V f_{V^{(1)}}(x_s)\cdot W^\star\right]\geq\lambda\sqrt{p}\gamma-c_1\sqrt{\log(n/\delta)}.$$

Setting $\lambda=\frac{c_1\sqrt{\log(n/\delta)}+\log\left(\frac{1}{\exp(\varepsilon)-1}\right)}{\sqrt{p}\gamma}$ we infer that

$$y_i\left[f_{V^{(1)}}(x_s)+\lambda\left(\nabla_V f_{V^{(1)}}(x_s)\cdot W^\star\right)\right]\geq\lambda\sqrt{p}\gamma-c_1\sqrt{\log(n/\delta)}=\log\left(\frac{1}{\exp(\varepsilon)-1}\right). \quad (45)$$

Set $V=V^{(1)}+\lambda W^\star$. The neural tangent kernel function at this weight vector is

$$F_{V^{(1)},V}(x)=f_V^{(1)}(x)+\nabla_V f_{V^{(1)}}(x)\cdot(V-V^{(1)})=f_V^{(1)}(x)+\lambda\nabla_V f_{V^{(1)}}(x)\cdot W^\star.$$

Thus by using (45)

$$\begin{aligned}\frac{1}{n}\sum_{s=1}^n\log\left(1+\exp\left(-y_iF_{V^{(1)},V}(x_s)\right)\right)&\leq\frac{1}{n}\sum_{s=1}^n\log\left(1+\exp\left(-\log\left(\frac{1}{\exp(\varepsilon)-1}\right)\right)\right)\\&\leq\varepsilon.\end{aligned}$$

We can conclude that if we choose the radius $\rho\geq\lambda\|W^\star\|=\lambda=\frac{c_1\sqrt{\log(n/\delta)}+\log\left(\frac{1}{\exp(\varepsilon)-1}\right)}{\sqrt{p}\gamma}$ (since $\|W^\star\|=1$ by assumption) then there exists a function in the NT function class with training error at most ε. This completes our proof. ■

D.3. Proof of Theorem 3.3

Theorem 3.3 *Consider a network with Huberized ReLU activations. There exists* $r(n,L,\delta)=\mathsf{poly}\left(L,\log\left(\frac{n}{\delta}\right)\right)$ *such that for any* $L\geq1$, $n\geq3$, $\delta>0$, *under Assumption 3.2 with* $\gamma\in(0,1]$ *if* $h=h_{\mathsf{NT}}$ *and* $p\geq\frac{r(n,L,\delta)}{\gamma^2}$ *then both of the following hold with probability at least* $1-4\delta$ *over the random initialization:*

1. *For all* $t\in[T]$, *set the step-size* $\alpha_t=\alpha_{\mathsf{NT}}=\Theta\left(\frac{1}{pL^5}\right)$, *where* $T=\left\lceil\frac{3(L+1)\rho^2n^{2+24L}}{2\alpha_{\mathsf{NT}}}\right\rceil$. *Then*

$$\min_{t\in[T]}J_t<\frac{1}{n^{1+24L}}.$$

2. *Set* $V^{(T+1)}=V^{(s)}$, *where* $s\in\arg\min_{s\in[T]}J(V^{(s)})$, *and for all* $t\geq T+1$, *set the step-size* $\alpha_t=\alpha_{\max}(h)$. *Then for all* $t\geq T+1$,

$$J_t\leq O\left(\frac{L^{\frac{3L+11}{2}}(6p)^{2L+5}}{n^{1+24L}\cdot(t-T-1)}\right).$$

Proof *Proof of Part 1:* Define two events

$$\mathcal{E}_a := \left\{ \varepsilon_{\mathsf{NT}}(V^{(1)}, \rho) \leq \frac{1}{6n^{2+24L}} \right\} \quad \text{and} \quad \mathcal{E}_b := \left\{ \min_{t \in [T]} J(V^{(t)}) \leq 6\varepsilon_{\mathsf{NT}}(V^{(1}, \rho) \right\}.$$

We will show that the $\mathcal{E}_1 := \mathcal{E}_a \cap \mathcal{E}_b$ occurs with probability at least $1 - 3\delta$. That is,

$$\mathbb{P}\left[\mathcal{E}_1 = \left\{ \min_{t \in [T]} J(V^{(t)}) \leq \frac{1}{n^{2+24L}} \right\} \right] \geq 1 - 3\delta. \tag{46}$$

The value of ρ is set to be (this was done in equation (4))

$$\begin{aligned} \rho &= \frac{c_1}{\sqrt{p}\gamma} \left[\sqrt{\log\left(\frac{n}{\delta}\right)} + \log\left(6n^{(2+24L)}\right) \right] \\ &> \frac{c_1}{\sqrt{p}\gamma} \left[\sqrt{\log\left(\frac{n}{\delta}\right)} + \log\left(\frac{1}{\exp\left(\frac{1}{6n^{(2+24L)}}\right) - 1} \right) \right] \qquad \text{(since } e^z \leq 1 + 2z \text{ when } z \in [0,1]\text{).} \end{aligned}$$

With this choice of ρ, since c_1 is a large enough absolute constant, Lemma D.9 guarantees that

$$\mathbb{P}\left[\mathcal{E}_a\right] \geq 1 - 2\delta, \tag{47}$$

where the probability is over the randomness in the initialization. Continue by setting

$$\tau = 3L\rho = \frac{3c_1 L}{\sqrt{p}\gamma} \left[\sqrt{\log\left(\frac{n}{\delta}\right)} + \log\left(6n^{(2+24L)}\right) \right].$$

Since $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)/\gamma^2$ for a large enough polynomial it is guaranteed that

$$\tau = \Omega\left(\frac{\log^2\left(\frac{nL}{\delta}\right)}{p^{\frac{3}{2}} L^3} \right) \quad \text{and} \quad \tau \leq \frac{c_3}{(p\log(p))^{\frac{3}{8}} L^{\frac{15}{4}}}$$

where c_3 is the positive absolute constant from the statement of Lemma D.8. Also recall the value of $h = h_{\mathsf{NT}}$ from equation (5)

$$h = h_{\mathsf{NT}} = \frac{(1+24L)\log(n)}{6(6p)^{\frac{L+1}{2}} L^3} \overset{(i)}{\leq} \frac{3c_1 L\left[\sqrt{\log(n/\delta)} + \log\left(6n^{(2+24L)}\right)\right]}{p\gamma} = \frac{\tau}{\sqrt{p}},$$

where (i) follows since $\gamma \in (0,1]$ by assumption and because p is large enough. Under these choices of τ and h along with the choice of the step-size $\alpha_t = \Theta\left(\frac{1}{pL^5}\right)$, and number of steps T, Lemma D.8 guarantees that

$$\mathbb{P}\left[\mathcal{E}_b\right] \geq 1 - \delta. \tag{48}$$

A union bound over the events (47) and (48) proves the Claim (46), which completes the proof of this first part.

Proof of Part 2: To prove this part of the lemma, we will invoke Theorem 3.1 to guarantee that the loss decreases in the steps $t \in \{T+1, \ldots\}$. We defined $V^{(T+1)} = V^{(s)}$, where $s \in$

$\arg\min_{t\in[T]} J(V^{(t)})$, thus we are guaranteed to have $J(V^{(T+1)}) \le \frac{1}{n^{2+24L}} < \frac{1}{n^{1+24L}}$, if event $\mathcal{E}_1$ defined above occurs. Define another event

$$\mathcal{E}_2 := \left\{ \|V^{(1)}\| \le \sqrt{5pL} \right\}.$$

Lemma E.12 guarantees that $\mathbb{P}\left[\mathcal{E}_2\right] \ge 1-\delta$. Define the "good event" $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2$. A simple union bound shows that

$$\mathbb{P}[\mathcal{E}] \ge 1 - 4\delta.$$

Assume that this event $\mathcal{E}$ occurs for the remainder of this proof. This also establishes that the success probability of gradient descent is at least $1-4\delta$ as mentioned in the theorem statement.

To invoke Theorem 3.1 we need to ensure that $h < h_{\max}$. Recall that, in equation (3a), we defined

$$h_{\max} := \min\left\{ \frac{L^{\frac{L}{2}-3}\log(1/J_{T+1})}{24\sqrt{p}\|V^{(T+1)}\|^L}, 1 \right\}.$$

For all $\ell \in [L+1]$, $\|V_\ell^{(T+1)} - V^{(1)}\| \le \tau$ (this fact is implicit in the proof of Lemma D.8). By the triangle inequality

$$\|V^{(T+1)}\| \le \|V^{(1)}\| + \|V^{(T+1)} - V^{(1)}\| \le \sqrt{5pL} + \sqrt{L+1}\tau \le \sqrt{6pL}$$

by the choice of τ above and since $p \ge \frac{\mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)}{\gamma^2}$ for a large enough polynomial. This means that

$$h = h_{\mathsf{NT}} = \frac{(1+24L)\log(n)}{6^{\frac{L+3}{2}} p^{\frac{L+1}{2}} L^3} \le \frac{L^{\frac{L}{2}-3}(1+24L)\log(n)}{24\sqrt{p}(\sqrt{6pL})^L} \le h_{\max}.$$

Thus, our choice of h is valid. In this second stage the step-size is chosen to be

$$\begin{aligned}\alpha_{\max}(h) &= \min\left\{ \frac{h}{1024\,(L+1)^2\, pJ_{T+1}\|V^{(T+1)}\|^{3L+5}}, \frac{(L+\frac{1}{2})\|V^{(T+1)}\|^2}{2L(L+\frac{3}{4})^2 J_{T+1}\log^{2/L}(1/J_{T+1})} \right\} \\ &= \frac{h}{1024\,(L+1)^2\, pJ_{T+1}\|V^{(T+1)}\|^{3L+5}},\end{aligned}$$

where the first term of the minima wins out above by our choice of h and because $\|V^{(T+1)}\| \le \sqrt{6pL}$. Thus Theorem 3.1 guarantees that

$$J(V^{(t)}) \le \frac{J(V^{(T+1)})}{\widetilde{Q}(\alpha_{\max}(h)) \cdot (t-T-1)+1},$$

where $\widetilde{Q}(\cdot)$ was defined in equation (3c). Thus,

$$\begin{aligned}
\widetilde{Q}(\alpha_{\max}(h)) &= \frac{L(L+\frac{3}{4})^2\alpha_{\max}(h)J_{T+1}\log^{2/L}(1/J_{T+1})}{(L+\frac{1}{2})\|V^{(T+1)}\|^2} \\
&= \frac{L(L+\frac{3}{4})^2 J_{T+1}\log^{2/L}(1/J_{T+1})}{(L+\frac{1}{2})\|V^{(T+1)}\|^2} \times \frac{h}{1024\,(L+1)^2\,pJ_{T+1}\|V^{(T+1)}\|^{3L+5}} \\
&= \frac{hL(L+\frac{3}{4})^2}{1024(L+1)^2(L+\frac{1}{2})p\|V^{(T+1)}\|^{3L+7}} \\
&\geq \frac{(L+\frac{3}{4})^2\log(n)}{50L(L+1)^2(L+\frac{1}{2})(6p)^{\frac{L+3}{2}}(6pL)^{\frac{3L+7}{2}}} \\
&\geq \frac{\log(n)}{50L^{\frac{3L+7}{2}}(L+1)^2(6p)^{2L+5}} \\
&\geq \frac{1}{50L^{\frac{3L+7}{2}}(L+1)^2(6p)^{2L+5}}.
\end{aligned}$$

Thus, for all $t \geq T+1$

$$\begin{aligned}
J(V^{(t)}) &\leq \frac{J(V^{(T+1)})}{Q\cdot(t-T-1)+1} \\
&\leq \frac{1}{n^{1+24L}}\frac{1}{Q\cdot(t-T-1)+1} \\
&\leq \frac{1}{n^{1+24L}}\frac{50L^{\frac{3L+7}{2}}(L+1)^2(6p)^{2L+5}}{(t-T-1)+50L^{\frac{3L+7}{2}}(L+1)^2(6p)^{2L+5}} \\
&< \frac{50L^{\frac{3L+7}{2}}(L+1)^2(6p)^{2L+5}}{n^{1+24L}(t-T-1)} \\
&= O\left(\frac{L^{\frac{3L+11}{2}}(6p)^{2L+5}}{n^{1+24L}\cdot(t-T-1)}\right),
\end{aligned}$$

this completes the proof. ■

Appendix E. Proof of Lemma D.7

In this section we prove Lemma D.7 that controls the approximation error $\varepsilon_{\mathsf{app}}(V^{(1)},\tau)$ and establishes a bound on the maximum norm of the gradient $\Gamma(V^{(1)},\tau)$. The proof of this lemma requires analogs of several lemmas from (Allen-Zhu et al., 2019; Zou et al., 2020) adapted to our setting. In Appendix E.1 we prove that several useful properties hold at initialization with high probability. In Appendix E.2 we show that some of these properties extend to weight matrices close to initialization and in Appendix E.3 we prove Lemma D.7.

Throughout this section we analyze the initialization scheme described in Section 3.2. This scheme is as follows: for all $\ell \in [L]$ the entries of $V_\ell^{(1)}$ are drawn independently from $\mathcal{N}(0,2/p)$ and the entries of $V_{L+1}^{(1)}$ are drawn independently from $\mathcal{N}(0,1)$. Again, the results of this appendix apply only to the Huberized ReLU (see definition in (1)).

E.1. Properties at Initialization

In the next lemma we show that several useful properties hold with high probability at initialization.

Lemma E.1 *For any $\delta > 0$, suppose that $h < \frac{1}{50\sqrt{p}L}$, $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial and $\tau = \Omega\left(\frac{\log^2(\frac{nL}{\delta})}{p^{\frac{3}{2}}L^3}\right)$. Then with probability at least $1-\delta$ over the randomness in $V^{(1)}$ we have the following:*

(a) For all $s \in [n]$ and all $\ell \in [L]$:

$$\|x_{\ell,s}^{V^{(1)}}\| \in \left[\frac{9}{10}, \frac{11}{10}\right].$$

(b) For all all $\ell \in [L]$, $\|V_\ell^{(1)}\|_{op} \leq O(1)$, and $\|V_{L+1}^{(1)}\| \leq O(\sqrt{p})$.

(c) For all $s \in [n]$ and all $1 \leq \ell_1 \leq \ell_2 \leq L$,

$$\left\|V_{\ell_2}^{(1)}\Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)}\right\|_{op} \leq O(L),$$

and for any $1 \leq \ell_1 \leq L$

$$\left\|V_{L+1}^{(1)}\Sigma_{L,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)}\right\|_{op} \leq O(\sqrt{p}L).$$

(d) For all $s \in [n]$ and all $1 \leq \ell_1 \leq \ell_2 \leq L$,

$$\left\|V_{\ell_2}^{(1)}\Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)} a\right\| \leq 3\|a\|$$

for all vectors a with $\|a\|_0 \leq k = \frac{cp}{\log(p)L^2}$, where c is a small enough positive absolute constant.

(e) For all $s \in [n]$ and all $1 \leq \ell_1 \leq \ell_2 \leq L$,

$$\left\|a^\top V_{\ell_2}^{(1)}\Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)}\right\| \leq O(\|a\|)$$

for all vectors a with $\|a\|_0 \leq k = \frac{cp}{\log(p)L^2}$, where c is a small enough positive absolute constant.

(f) For all $s \in [n]$ and all $1 \leq \ell_1 \leq \ell_2 \leq L$,

$$|a^\top V_{\ell_2}^{(1)}\Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)} b| \leq O\left(\|a\|\|b\|\frac{\sqrt{k\log(p)}}{\sqrt{p}}\right)$$

for all vectors a, b with $\|a\|_0, \|b\|_0 \leq k = \frac{cp}{\log(p)L^2}$, where c is a small enough positive absolute constant.

(g) For all $s \in [n]$ and all $1 \leq \ell \leq L$,

$$|V_{L+1}^{(1)}\Sigma_{L,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_\ell^{(1)} a| \leq O\left(\|a\|\sqrt{k\log(p)}\right)$$

for all vectors a with $\|a\|_0 \leq k = \frac{cp}{\log(p)L^2}$, where c is a small enough positive absolute constant.

(h) For $\beta = O\left(\frac{L^2\tau^{2/3}}{\sqrt{p}}\right)$ *and*

$$\mathcal{S}_{\ell,s}(\beta) := \left\{j \in [p] : |V_{\ell,j}^{(1)} x_{\ell,s}^{V^{(1)}}| \le \beta\right\},$$

where $V_{\ell,j}^{(1)}$ *refers to the* j*th row of* $V_\ell^{(1)}$*, for all* $\ell \in [L]$ *and all* $s \in [n]$*:*

$$|\mathcal{S}_{\ell,s}(\beta)| \le O(p^{3/2}\beta) = O(pL^2\tau^{2/3}).$$

We will prove this lemma part by part and show that each of the eight properties holds with probability at least $1 - \delta/8$ and take a union bound at the end. We show that each of the parts hold with this probability in the eight lemmas (Lemmas E.2-E.9) that follow.

E.1.1. PROOF OF PART (A)

Lemma E.2 *For any* $\delta > 0$*, suppose that* $h < \frac{1}{50\sqrt{p}L}$ *and* $p \ge \text{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ *for a large enough polynomial, then with probability at least* $1 - \delta/8$ *over the randomness in* $V^{(1)}$ *we have that for all* $s \in [n]$ *and all* $\ell \in [L]$*:*

$$\|x_{\ell,s}^{V^{(1)}}\| \in \left[\frac{9}{10}, \frac{11}{10}\right].$$

Proof Fix any layer $\ell \in [L]$ and any sample $s \in [n]$. We will prove the result for this layer and sample, and apply a union bound at the end. To ease notation we drop $V^{(1)}$ from the superscript of $x_{\ell,s}^{V^{(1)}}$ and refer to $V_\ell^{(1)}$ as simply V_ℓ.

By definition

$$x_{\ell,s} = \phi\left(V_\ell x_{\ell-1,s}\right).$$

Conditioned on $x_{\ell-1,s}$, each coordinate of $V_\ell x_{\ell-1,s}$ is distributed as $\mathcal{N}\left(0, \frac{2\|x_{\ell-1,s}\|^2}{p}\right)$, since each entry of V_ℓ is drawn independently from $\mathcal{N}(0, \frac{2}{p})$. Let $\bar{\phi}(z) = \max\{0, z\}$ denote the ReLU activation function. Then we know that $\bar{\phi}(z) - \frac{h}{2} \le \phi(z) \le \bar{\phi}(z)$ for any $z \in \mathbb{R}$. Let $(x_{\ell,s})_i$ denote the ith coordinate of x_ℓ and let $V_{\ell,i}$ denote the ith row of V_ℓ. Therefore, conditioned on $x_{\ell-1,s}$,

$$\begin{aligned}
\mathbb{E}\left[(x_{\ell,s})_i^2 \middle| x_{\ell-1,s}\right] &= \mathbb{E}\left[\phi^2\left(V_{\ell,i}x_{\ell-1,s}\right)\middle| x_{\ell-1,s}\right] \\
&\ge \mathbb{E}\left[\bar{\phi}^2\left(V_{\ell,i}x_{\ell-1,s}\right)\middle| x_{\ell-1,s}\right] - h\mathbb{E}\left[\bar{\phi}\left(V_{\ell,i}x_{\ell-1,s}\right)\middle| x_{\ell-1,s}\right] + \frac{h^2}{4} \\
&\overset{(i)}{=} \frac{1}{2}\mathbb{E}\left[\left(V_{\ell,i}x_{\ell-1,s}\right)^2\middle| x_{\ell-1,s}\right] - \frac{h\mathbb{E}\left[|V_{\ell,i}x_{\ell-1,s}|\middle| x_{\ell-1,s}\right]}{2} + \frac{h^2}{4} \\
&= \frac{\|x_{\ell-1,s}\|^2}{p} - \frac{h\|x_{\ell-1,s}\|}{\sqrt{2p\pi}} + \frac{h^2}{4},
\end{aligned}$$

where (i) follows since $\bar{\phi}(z) = 0$ if $z < 0$ and the distribution of $V_{\ell,i}x_{\ell-1,s}$ is symmetric about the origin. Therefore summing up over all $i \in [p]$ we find

$$\begin{aligned}
\mathbb{E}\left[\|x_{\ell,s}\|^2 \middle| x_{\ell-1,s}\right] = \sum_{i\in[p]} \mathbb{E}\left[(x_{\ell,s})_i^2 \middle| x_{\ell-1,s}\right] &\ge \|x_{\ell-1,s}\|^2 - \frac{\sqrt{p}h\|x_{\ell-1,s}\|}{\sqrt{2\pi}} + \frac{h^2 p}{4} \\
&\ge \left(\|x_{\ell-1,s}\| - \frac{h\sqrt{p}}{2}\right)^2. \qquad (49)
\end{aligned}$$

Similarly we can also demonstrate an upper bound of $\mathbb{E}\left[\|x_{\ell,s}\|^2 \mid x_{\ell-1,s}\right] \leq \|x_{\ell-1,s}\|^2$ since $\phi(z) \leq \bar{\phi}(z)$ for any z as stated previously.

Let $\|\cdot\|_{\psi_2}$ denote the sub-Gaussian norm of a random variable (see Definition F.1) and let $\|\cdot\|_{\psi_1}$ denote the sub-exponential norm (see Definition F.2). Since the function ϕ is 1-Lipschitz, conditioned on $x_{\ell-1,s}$,

$$\begin{aligned}\|(x_{\ell,s})_i\|_{\psi_2} &= \|\phi(V_{\ell,i}x_{\ell-1,s})\|_{\psi_2} \\ &\leq \|\phi(V_{\ell,i}x_{\ell-1,s}) - \mathbb{E}\left[\phi(V_{\ell,i}x_{\ell-1,s})|x_{\ell-1,s}\right]\|_{\psi_2} + \|\mathbb{E}\left[\phi(V_{\ell,i}x_{\ell-1,s})|x_{\ell-1,s}\right]\|_{\psi_2} \\ &\overset{(i)}{\leq} c\frac{\|x_{\ell-1,s}\|}{\sqrt{p}} + \|\mathbb{E}\left[\phi(V_{\ell,i}x_{\ell-1,s})|x_{\ell-1,s}\right]\|_{\psi_2} \overset{(ii)}{\leq} c_1\frac{\|x_{\ell-1,s}\|}{\sqrt{p}} \end{aligned} \tag{50}$$

where (i) follows by invoking Lemma F.4, and (ii) follows since we showed above that

$$\|\mathbb{E}\left[\phi(V_{\ell,i}x_{\ell-1,s})|x_{\ell-1,s}\right]\|_{\psi_2} = |\mathbb{E}\left[\phi(V_{\ell,i}x_{\ell-1,s})|x_{\ell-1,s}\right]| \leq \sqrt{\mathbb{E}\left[\phi^2(V_{\ell,i}x_{\ell-1,s})|x_{\ell-1,s}\right]} \leq \frac{\|x_{\ell-1,s}\|}{\sqrt{p}}.$$

Therefore $\|(x_{\ell,s})_i^2\|_{\psi_1} \leq \|(x_{\ell,s})_i\|_{\psi_2}^2 \leq \frac{c_2\|x_{\ell-1,s}\|^2}{p}$ by Lemma F.3. Since the random variables $(x_{\ell,s})_1^2, \ldots, (x_{\ell,s})_p^2$ are conditionally independent given $x_{\ell-1,s}$, applying Bernstein's inequality (see Theorem F.6) we get that for any $\eta \in (0,1]$

$$\begin{aligned}&\mathbb{P}\left(\left|\|x_{\ell,s}\|^2 - \mathbb{E}\left[\|x_{\ell,s}\|^2 \,|x_{\ell-1,s}\right] \right| \leq \eta\|x_{\ell-1,s}\|^2 \,\middle|x_{\ell-1,s}\right) \\ &\quad\geq 1 - 2\exp\left(-c\min\left\{\frac{\eta^2\|x_{\ell-1,s}\|^4}{p\times\left(c_2^2\|x_{\ell-1,s}\|^4/p^2\right)}, \frac{\eta\|x_{\ell-1,s}\|^2}{c_2\|x_{\ell-1,s}\|^2/p}\right\}\right) \\ &\quad\geq 1 - 2\exp\left(-c_3\min\left\{\eta^2 p, \eta p\right\}\right) \\ &\quad\geq 1 - 2\exp\left(-c_3 p\eta^2\right).\end{aligned}$$

We established above that the expected value satisfies the following bounds:

$$\left(\|x_{\ell-1,s}\| - \frac{h\sqrt{p}}{2}\right)^2 \leq \mathbb{E}\left[\|x_{\ell,s}\|^2 \,|x_{\ell-1,s}\right] \leq \|x_{\ell-1,s}\|^2.$$

Thus

$$\begin{aligned}&\mathbb{P}\left(\|x_{\ell,s}\|^2 \in \left[\left(\|x_{\ell-1,s}\| - \frac{h\sqrt{p}}{2}\right)^2 - \eta\|x_{\ell-1,s}\|^2, \|x_{\ell-1,s}\|^2(1+\eta)\right] \,\middle|\, x_{\ell-1,s}\right) \\ &\quad\geq 1 - 2\exp\left(-c_3 p\eta^2\right).\end{aligned}$$

Taking a union bound over all samples and all hidden layers we find that

$$\begin{aligned}&\mathbb{P}\left(\forall s \in [n], \ell \in [L],\ \|x_{\ell,s}\|^2 \in \left[\left(\|x_{\ell-1,s}\| - \frac{h\sqrt{p}}{2}\right)^2 - \eta\|x_{\ell-1,s}\|^2, \|x_{\ell-1,s}\|^2(1+\eta)\right]\right) \\ &\quad\geq 1 - 2nL\exp\left(-c_3 p\eta^2\right).\end{aligned}$$

This implies that

$$\mathbb{P}\left(\forall s \in [n], \ell \in [L],\ \left|\|x_{\ell,s}\|^2 - \|x_{\ell-1,s}\|^2\right| \le \eta\|x_{\ell-1,s}\|^2 + h\sqrt{p}\|x_{\ell-1,s}\| + \frac{h^2 p}{4}\right)$$
$$\ge 1 - 2nL\exp\left(-c_3 p\eta^2\right).$$

Setting $\eta = \frac{1}{50L}$ and because by assumption $h\sqrt{p} < \frac{1}{50L} = \eta$ we get that

$$\mathbb{P}\left(\forall s \in [n], \ell \in [L],\ \left|\|x_{\ell,s}\|^2 - \|x_{\ell-1,s}\|^2\right| \le \eta\|x_{\ell-1,s}\|^2 + \eta\|x_{\ell-1,s}\| + \frac{\eta^2}{4}\right)$$
$$\ge 1 - 2nL\exp\left(-\frac{c_4 p}{L^2}\right). \tag{51}$$

Let us assume that the event of (51) holds for the rest of this proof. Starting with $\ell = 1$ we know that $\|x_{0,s}\| = \|x_s\| = 1$, thus if the event in the previous display holds then by the choice of $\eta = 1/(50L)$ we have that

$$\|x_{1,s}\|^2 \in [1 - 3\eta, 1 + 3\eta].$$

For any $z \in [0,1]$ we have that $(1+z)^{1/2} \le 1 + z$ and $(1-z)^{1/2} \ge 1 - z$. Thus, by taking square roots

$$\|x_{1,s}\| \in [1 - 3\eta, 1 + 3\eta].$$

We will now prove that $\|x_{\ell,s}\| \in [1 - 3\ell\eta, 1 + 3\ell\eta]$ using an inductive argument over $\ell = 1, \dots, L$. The base case when $\ell = 1$ of course holds by the display above. Now let us prove it for a layer $\ell > 1$ assuming it holds at layer $\ell - 1$.

Let us first prove the upper bound on $\|x_{\ell,s}\|$, the lower bound will follow by the same logic. If the event in (51) holds then we know that

$$\|x_{\ell,s}\|^2 - \|x_{\ell-1,s}\|^2 \le \eta\|x_{\ell-1,s}\|^2 + \eta\|x_{\ell-1,s}\| + \frac{\eta^2}{4}$$

which implies that

$$\begin{aligned}
\|x_{\ell,s}\|^2 &\le \|x_{\ell-1,s}\|^2(1+\eta) + \eta\|x_{\ell-1,s}\| + \frac{\eta^2}{4} \\
&= \|x_{\ell-1,s}\|^2\left(1 + \eta + \frac{\eta}{\|x_{\ell-1,s}\|} + \frac{\eta^2}{4\|x_{\ell-1,s}\|^2}\right) \\
&\overset{(i)}{\le} \|x_{\ell-1,s}\|^2\left(1 + \eta + \frac{10\eta}{9} + \frac{25\eta^2}{81}\right) \\
&= \|x_{\ell-1,s}\|^2\left(1 + \frac{19\eta}{9} + \frac{25\eta^2}{81}\right) \overset{(ii)}{\le} \|x_{\ell-1,s}\|^2\left(1 + \frac{20\eta}{9}\right)
\end{aligned}$$

where (i) follows since by the inductive hypothesis $\|x_{\ell-1,s}\| \ge 1 - 3(\ell-1)\eta$ and because $\eta = \frac{1}{50L}$, therefore $\|x_{\ell-1,s}\| \ge 1 - \frac{3(\ell-1)}{100L} \ge \frac{97}{100} > \frac{9}{10}$, and (ii) again follows because $\eta = \frac{1}{50L}$ and $L \ge 1$.

Taking square roots we find that

$$\begin{aligned}\|x_{\ell,s}\| &\le \|x_{\ell-1,s}\|\sqrt{1+\frac{20\eta}{9}} \\ &\le (1+3(\ell-1)\eta)\sqrt{1+\frac{20\eta}{9}} \qquad \text{(by the IH)} \\ &\overset{(i)}{\le} (1+3(\ell-1)\eta)\left(1+\frac{20\eta}{9}\right) \\ &= 1+3(\ell-1)\eta+\frac{20\eta}{9}+\frac{60(\ell-1)\eta^2}{9} \overset{(ii)}{\le} 1+3\ell\eta,\end{aligned}$$

where (i) follows since $\sqrt{1+z} \le 1+z$ and (ii) follows since $\eta = \frac{1}{50L}$ and $L \ge 1$. This establishes the desired upper bound on $\|x_{\ell,s}\|$. As mentioned above, the lower bound $(1-3\ell\eta) \le \|x_{\ell,s}\|$ follows by mirroring the logic. This completes our induction and proves that for all s and all ℓ with probability at least $1-\delta/8$

$$\|x_{\ell,s}\| \in [1-3\ell\eta, 1+3\ell\eta].$$

Our choice of $\eta = \frac{1}{50L}$ establishes that

$$\|x_{\ell,s}\| \in \left[\frac{9}{10}, \frac{11}{10}\right] \tag{52}$$

for all $s \in [n]$ and $\ell \in [L]$ with probability at least $1-2nL\exp\left(-\frac{c_4 p}{L^2}\right) \ge 1-\delta/8$, which follows since $p \ge \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial. This wraps up our proof. ■

E.1.2. PROOF OF PART (B)

Lemma E.3 *For any $\delta > 0$ suppose that $p \ge \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial, then with probability at least $1-\delta/8$ over the randomness in $V^{(1)}$:*

$$\textit{for all } \ell \in [L], \quad \|V_\ell^{(1)}\|_{op} \le O(1), \quad \textit{and} \quad \|V_{L+1}^{(1)}\| \le O(\sqrt{p}).$$

Proof For any fixed $\ell \in [L]$ recall that each entry of $V_\ell^{(1)}$ is drawn independently from $\mathcal{N}\left(0, \frac{2}{p}\right)$. Thus, by invoking (Vershynin, 2018, Theorem 4.4.5) we know that

$$\|V_\ell^{(1)}\|_{op} \le O(1)$$

with probability at least $1-\exp(-\Omega(p))$. The entries of $V_{L+1}^{(1)}$ are drawn from $\mathcal{N}(0,1)$, therefore by Theorem F.7 we find that

$$\|V_{L+1}^{(1)}\|^2 \le 2p$$

with probability $1-\exp(-\Omega(p))$. By a union bound over the $L+1$ layers and noting that $p \ge \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ yields that

$$\text{for all } \ell \in [L], \quad \|V_\ell^{(1)}\|_{op} \le O(1) \quad \text{and} \quad \|V_{L+1}^{(1)}\| \le O(\sqrt{p})$$

with probability at least $1-\delta/8$ as claimed. ■

E.1.3. PROOF OF PART (C)

Lemma E.4 *For any $\delta > 0$ suppose that $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial, then with probability at least $1 - \delta/8$ over the randomness in $V^{(1)}$ we have that for all $s \in [n]$ and all $1 \leq \ell_1 \leq \ell_2 \leq L$,*

$$\left\| V_{\ell_2}^{(1)} \Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)} \right\|_{op} \leq O(L),$$

and all $1 \leq \ell_1 \leq L$

$$\left\| V_{L+1}^{(1)} \Sigma_{L,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)} \right\|_{op} \leq O(\sqrt{p}L).$$

Proof We begin by analyzing the case where $\ell_2 < L+1$. A similar analysis works to prove the claim when $\ell_2 = L+1$. This is because the variance of each entry of $V_{L+1}^{(1)}$ is 1, whereas when $\ell_2 < L+1$ the variance of each entry of $V_\ell^{(1)}$ is $2/p$. Therefore the bound is simply multiplied by a factor of $\sqrt{2p}$ in the case when $\ell_2 = L+1$.

Fix the layers $1 \leq \ell_1 \leq \ell_2 \leq L$ and fix the sample index s. At the end of the proof we shall take a union bound over all pairs of layers and all samples. Now to ease notation let us denote $V_\ell^{(1)}$ by simply V_ℓ and let $\Sigma_{\ell,s}^{V^{(1)}}$ be denoted by $\Sigma_{\ell,s}$.

To bound the operator norm

$$\left\| V_{\ell_2}^{(1)} \Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)} \right\|_{op} = \sup_{a:\|a\|=1} \left\| V_{\ell_2}^{(1)} \Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)} a \right\| \tag{53}$$

we will first consider a supremum over vectors that are non-zero only on an arbitrary fixed subset $S \subseteq [p]$ with cardinality $|S| \leq \left\lfloor \frac{c_1 p}{L^2} \right\rfloor$, where c_1 is small enough absolute constant. That is, we shall bound

$$\Xi := \sup_{a:\|a\|=1, \mathrm{supp}(a) \subseteq S} \left\| V_{\ell_2}^{(1)} \Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)} a \right\|.$$

Using this we will then bound the operator norm in (53) by decomposing any unit vector a into $\frac{p}{\left\lfloor \frac{c_1 p}{L^2} \right\rfloor}$ vectors that are non-zero only on subsets of size at most $\left\lfloor \frac{c_1 p}{L^2} \right\rfloor$.

Let us begin by first bounding Ξ. Part (b) of Lemma E.10, which is proved below, establishes that, for any fixed unit vector $z \in \mathbb{S}^{p-1}$

$$\|V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1} z\| \leq 2$$

with probability at least $1 - O(nL^3) e^{-\Omega\left(\frac{p}{L^2}\right)}$.

We take a $1/4$-net (see the definition of an ε-net in Definition F.8) of unit vectors $\{a_i\}_{i=1}^m$ whose coordinates are non-zero only on this particular subset S, with respect to the Euclidean norm. There exists such a $1/4$-net of size $m = 9^{c_1 p / L^2}$ (see Lemma F.9). By a union bound,

$$\forall i \in [m],\ \|V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1} a_i\| \leq 2 \tag{54}$$

with probability at least $1 - O(nL^3 \cdot 9^{c_1 p/L^2}) e^{-\Omega\left(\frac{p}{L^2}\right)} = 1 - e^{-\Omega\left(\frac{p}{L^2}\right)}$ since $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial, and because c_1 is a small enough constant. We will now proceed

to show that if the "good event" (54) regarding the $1/4$-net holds then we can use it to establish guarantees for all unit vectors a that are only non-zero on this subset S. To see this, if $\zeta(a)$ maps each unit vector a with support contained in S to its nearest neighbor in $\{a_1, \ldots, a_m\}$, then if the event in (54) holds then

$$\begin{aligned}
\Xi &= \sup_{a:\|a\|=1,\mathrm{supp}(a)\subseteq S} \|V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}a\| \\
&= \sup_{a:\|a\|=1,\mathrm{supp}(a)\subseteq S} \|V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}(a-\zeta(a)+\zeta(a))\| \\
&\leq \sup_{j\in[m]} \|V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}a_j\| + \sup_{a:\|a\|=1,\mathrm{supp}(a)\subseteq S} \|V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}(a-\zeta(a))\| \\
&\overset{(i)}{\leq} \sup_{j\in[m]} \|V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}a_j\| + \frac{1}{4}\sup_{a:\|a\|=1,\mathrm{supp}(a)\subseteq S} \left\|V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}\frac{(a-\zeta(a))}{\|a-\zeta(a)\|}\right\| \\
&\overset{(ii)}{\leq} 2+\frac{\Xi}{4}
\end{aligned}$$

where and (i) follows since $\|a-\zeta(a)\| \leq 1/4$, inequality (ii) follows since we assumed the event (54) to hold and by the definition of Ξ. By rearranging terms we find that, with the same probability that is at least $1-e^{-\Omega\left(\frac{p}{L^2}\right)}$, for any unit vector a that is only non-zero on subset S, we have that

$$\|V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}a\| \leq \frac{1}{1-\frac{1}{4}} \times 2 < 3. \tag{55}$$

As mentioned above we will now consider a partition of $[p] = S_1 \cup \ldots \cup S_q$, such that for all $i \in [q]$, $|S_i| \leq \left\lfloor \frac{c_1 p}{L^2} \right\rfloor$, and the number of sets (q) in the partition satisfies $q \leq \frac{p}{\left\lfloor \frac{c_1 p}{L^2} \right\rfloor} = \left\lceil \frac{L^2}{c_1} \right\rceil$. Given an arbitrary unit vector $b \in \mathbb{S}^{p-1}$, we can decompose it as $b = u_1 + \ldots + u_q$, where each u_i is non-zero only on the set S_i. Invoking the triangle inequality

$$\begin{aligned}
\|V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell}V_{\ell_1}b\| &\leq \sum_{i=1}^{q}\|V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}u_i\| \\
&= \sum_{i=1}^{q}\left\|V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}\frac{u_i}{\|u_i\|}\right\|\|u_i\|.
\end{aligned}$$

By applying the result of (55) to each term in the sum above along with a union bound over the q sets $S_1, \ldots, S_q$ we find the following: with probability at least $1 - qe^{-\Omega\left(\frac{p}{L^2}\right)} = 1 - O(L^2)e^{-\Omega\left(\frac{p}{L^2}\right)}$, for all unit vectors $b \in \mathbb{S}^{p-1}$

$$\|V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}b\| \leq 3\sum_{i=1}^{q}\|u_i\| \leq 3\sqrt{q}\left(\sum_{i=1}^{q}\|u_i\|^2\right)^{1/2} = 3\sqrt{q} = O(L).$$

The definition of the operator norm of a matrix $\|A\|_{op} = \sup_{v:\|v\|=1}\|Av\|$ along with the previous display establishes the claim for this particular pair of layers ℓ_1 and ℓ_2 and sample s. A union bound over pairs of layers and all samples to establish that, with probability at least $1 - O(nL^4)e^{-\Omega\left(\frac{p}{L^2}\right)}$, for all pairs $1 \leq \ell_1 \leq \ell_2 \leq L$ and all $s \in [n]$

$$\|V_{\ell_2}\Sigma_{\ell_2-1}\cdots\Sigma_{\ell}V_{\ell_1}\|_{op} \leq O(L). \tag{56}$$

As claimed above, a similar analysis shows that, with probability at least $1 - O(nL^4)e^{-\Omega\left(\frac{p}{L^2}\right)}$, for all $s \in [n]$ and all $\ell_1 \in [L]$, we have

$$\|V_{L+1}\Sigma_L \cdots \Sigma_\ell V_{\ell_1}\|_{op} \leq O(\sqrt{p}L). \tag{57}$$

Since $p \geq \text{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ we can ensure that both events in (56) and (57) occur simultaneously with probability at least $1 - \delta/8$. ■

E.1.4. PROOF OF PART (D)

Lemma E.5 *For any $\delta > 0$, suppose that $p \geq \text{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial, then with probability at least $1 - \delta/8$ over the randomness in $V^{(1)}$ we have that for all $s \in [n]$ and all $1 \leq \ell_1 \leq \ell_2 \leq L$,*

$$\left\| V_{\ell_2}^{(1)} \Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)} a \right\| \leq 3\|a\|$$

for all vectors a with $\|a\|_0 \leq k = \frac{cp}{\log(p)L^2}$, where c is a small enough positive absolute constant.

Proof We fix the layers $1 \leq \ell_1 \leq \ell_2 \leq L$ and fix the sample index s. At the end of the proof we shall take a union bound over all pairs of layers and all samples. Again, to ease notation, let us denote $V_\ell^{(1)}$ by simply V_ℓ and let $\Sigma_{\ell,s}^{V^{(1)}}$ be denoted by $\Sigma_{\ell,s}$.

For a fixed unit vector $z \in \mathbb{S}^{p-1}$ by Part (b) of Lemma E.10 that is proved below we have

$$\|V_{\ell_2}\Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s}V_{\ell_1}z\| \leq 2 \tag{58}$$

with probability at least $1 - O(nL^3)e^{-\Omega\left(\frac{p}{L^2}\right)}$. Consider a $1/4$-net of k-sparse unit vectors $\{a_i\}_{i=1}^m$, where $m = \binom{p}{k}9^k$ (such a net exists, see Lemma F.10).

Using (58) and taking a union bound, with probability at least $1 - O\left(\binom{p}{k}9^k \cdot nL^3\right) e^{-\Omega\left(\frac{p}{L^2}\right)}$, for all vectors $\{a_i\}_{i=1}^m$

$$\|V_{\ell_2}\Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s}V_{\ell_1}a_i\| \leq 2.$$

Now by mirroring the logic that lead from inequality (54) to inequality (55) in the proof of the previous lemma, we can establish that, again with probability that is at least $1 - \binom{p}{k}9^k \cdot O(nL^3)e^{-\Omega\left(\frac{p}{L^2}\right)}$, for any vector a that is k-sparse

$$\left\| V_{\ell_2}\Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s}V_{\ell_1} \frac{a}{\|a\|} \right\| \leq 3.$$

A union bound over all pairs of layers and all samples we find that, with probability at least $1 - O\left(\binom{p}{k}9^k \cdot n^2L^5\right) e^{-\Omega\left(\frac{p}{L^2}\right)}$, for all $1 \leq \ell_1 \leq \ell_2 \leq L$, for all $s \in [n]$ and for all vectors a that are k-sparse

$$\left\| V_{\ell_2}\Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s}V_{\ell_1} \frac{a}{\|a\|} \right\| \leq 3.$$

Moreover,

$$\begin{aligned}
1 - O\left(\binom{p}{k} 9^k \cdot n^2 L^5\right) e^{-\Omega\left(\frac{p}{L^2}\right)} &\geq 1 - O\left(\left(\frac{ep}{k}\right)^k 9^k \cdot n^2 L^5\right) e^{-\Omega\left(\frac{p}{L^2}\right)} \qquad \text{(since } \tbinom{p}{k} \leq \left(\tfrac{ep}{k}\right)^k\text{)} \\
&= 1 - O\left(\left(\frac{9ep}{k}\right)^k \cdot n^2 L^5\right) e^{-\Omega\left(\frac{p}{L^2}\right)} \\
&= 1 - O\left(n^2 L^5\right) e^{-\Omega\left(\frac{p}{L^2} - k\log(9ep)\right)} \\
&\geq 1 - \delta/8
\end{aligned}$$

where the last inequality follows since $k \leq \frac{cp}{\log(p)L^2}$ where c is a small enough absolute constant and $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial. This completes the proof. ■

E.1.5. PROOF OF PART (E)

Lemma E.6 *For any $\delta > 0$, suppose that $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial, then with probability at least $1 - \delta/8$ over the randomness in $V^{(1)}$ we have that for all $s \in [n]$ and all $1 \leq \ell_1 \leq \ell_2 \leq L$,*

$$\left\| a^\top V_{\ell_2}^{(1)} \Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)} \right\| \leq O(\|a\|)$$

for all vectors a with $\|a\|_0 \leq k = \frac{cp}{\log(p)L^2}$, where c is a small enough positive absolute constant.

Proof We fix the layers $1 \leq \ell_1 \leq \ell_2 \leq L$ and fix the sample index s. At the end of the proof we shall take a union bound over all pairs of layers and all samples. In the proof let us denote $V_\ell^{(1)}$ by simply V_ℓ and let $\Sigma_{\ell,s}^{V^{(1)}}$ be denoted by $\Sigma_{\ell,s}$.

For any fixed vector z we know from Part (a) of Lemma E.10 that with probability at least $1 - O(nL^3)e^{-\Omega\left(\frac{p}{L^2}\right)}$ over the randomness in $(V_{\ell_2-1}, \ldots, V_1)$

$$\|\Sigma_{\ell_2-1,s} V_{\ell_2-1} \ldots \Sigma_{\ell_1,s} V_{\ell_1} z\| \leq 2\|z\|. \tag{59}$$

Recall that the entries of V_{ℓ_2} are drawn independently from $\mathcal{N}(0, \frac{2}{p})$. Thus, conditioned on this event above, for any fixed vector w the random variable $w^\top V_{\ell_2} (\Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1} V_{\ell_1} z)$ is a mean-zero Gaussian with variance at most $\frac{8\|w\|^2\|z\|^2}{p}$. Thus over the randomness in V_{ℓ_2}

$$\mathbb{P}\left(\left|w^\top V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1} z\right| \leq \frac{4}{L}\|w\|\|z\| \middle| V_{\ell_2-1}, \ldots, V_1\right) \geq 1 - e^{-\Omega\left(\frac{p}{L^2}\right)}. \tag{60}$$

By union bound over the events in (59) and (60) we have

$$\mathbb{P}\left(\left|w^\top V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1} z\right| \leq \frac{4}{L}\|w\|\|z\|\right) \geq 1 - O(nL^3)e^{-\Omega\left(\frac{p}{L^2}\right)}. \tag{61}$$

Similar to the proof of Lemma E.4 our strategy will be to first bound

$$\sup_{a:\|a\|=1,\|a\|_0\leq k} \sup_{b:\|b\|=1,\mathrm{supp}(b)\in\subseteq S} \left|a^\top V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1} b\right|$$

where S is a fixed subset of $[p]$ with $|S| \leq \frac{c_1 p}{L^2}$, where c_1 is a small enough absolute constant. Let $\{z_i\}_{i=1}^r$ be a $1/4$-net of unit vectors with respect to the Euclidean norm whose coordinates are non-zero only on this subset S. There exists such a $1/4$-net of size $r = 9^{c_1 p/L^2}$ (see Lemma F.9). Let $\{w_i\}_{i=1}^m$ be a $1/4$-net of k-sparse unit vectors in Euclidean norm of size $m = \binom{p}{k} 9^k$ (Lemma F.10 guarantees the existence of such a net). Therefore by using (61) and taking a union bound we get that

$$\forall\, i \in [r], j \in [m], \quad \left| w_j^\top V_{\ell_2} \Sigma_{\ell_2 - 1, s} \cdots \Sigma_{\ell_1, s} V_{\ell_1} z_i \right| \leq \frac{4}{L} \tag{62}$$

with probability at least $1 - mrO(nL^3) e^{-\Omega\left(\frac{p}{L^2}\right)} = 1 - O(9^{c_1 p/L^2} \binom{p}{k} 9^k n L^3) e^{-\Omega\left(\frac{p}{L^2}\right)} = 1 - e^{-\Omega\left(\frac{p}{L^2}\right)}$, since $k = \frac{cp}{\log(p) L^2}$ where both c and c_1 are small enough absolute constants and because $p \geq \text{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial.

We will now demonstrate that if the "good event" in (62) holds then we can use this to establish a similar guarantee for all k-sparse unit vectors a and all unit vectors b that are only non-zero on the subset S. To see this, as before, suppose ζ maps any unit-length vector with support in S to its nearest neighbor in $\{z_1, \ldots, z_r\}$ and λ maps any k-sparse unit vector to its nearest neighbor in $\{w_1, \ldots, w_m\}$. Then if the event in (62) holds, we have

$$\begin{aligned}
\Xi &:= \sup_{a : \|a\| = 1, \|a\|_0 \leq k} \sup_{b : \|b\| = 1, \text{supp}(b) \subseteq S} \left| a^\top V_{\ell_2} \Sigma_{\ell_2 - 1, s} \cdots \Sigma_{\ell_1, s} V_{\ell_1} b \right| \\
&= \sup_{a : \|a\| = 1, \|a\|_0 \leq k} \sup_{b : \|b\| = 1, \text{supp}(b) \subseteq S} \left| (a - \lambda(a) + \lambda(a))^\top V_{\ell_2} \Sigma_{\ell_2 - 1, s} \cdots \Sigma_{\ell_1, s} V_{\ell_1} (b - \zeta(b) + \zeta(b)) \right| \\
&\leq \sup_{i \in [m], j \in [r]} \left| w_i^\top V_{\ell_2} \Sigma_{\ell_2 - 1, s} \cdots \Sigma_{\ell_1, s} V_{\ell_1} z_j \right| \\
&\quad + \sup_{a : \|a\| = 1, \|a\|_0 \leq k, j \in [r]} \left| (a - \lambda(a))^\top V_{\ell_2} \Sigma_{\ell_2 - 1, s} \cdots \Sigma_{\ell_1, s} V_{\ell_1} z_j \right| \\
&\quad + \sup_{i \in [m], b : \|b\| = 1, \text{supp}(b) \subseteq S} \left| w_i^\top V_{\ell_2} \Sigma_{\ell_2 - 1, s} \cdots \Sigma_{\ell_1, s} V_{\ell_1} (b - \zeta(b)) \right| \\
&\quad + \sup_{a : \|a\| = 1, \|a\|_0 \leq k} \sup_{i \in [m], b : \|b\| = 1, \text{supp}(b) \subseteq S} \left| (a - \lambda(a))^\top V_{\ell_2} \Sigma_{\ell_2 - 1, s} \cdots \Sigma_{\ell_1, s} V_{\ell_1} (b - \zeta(b)) \right| \\
&\overset{(i)}{\leq} \frac{4}{L} + \frac{\Xi}{4} + \frac{\Xi}{4} + \frac{\Xi}{16} \leq \frac{4}{L} + \frac{9}{16} \Xi
\end{aligned} \tag{63}$$

where (i) follows by the definition of Ξ along with Lemma F.10, because we assume that the event in (62) holds, and also because $\|a - \lambda(a)\| \leq 1/4$ and $\|b - \zeta(b)\| \leq 1/4$.

By rearranging terms in the previous display we can infer that

$$\Xi := \sup_{a : \|a\| = 1, \|a\|_0 \leq k} \sup_{b : \|b\| = 1, \text{supp}(b) \subseteq S} \left| a^\top V_{\ell_2} \Sigma_{\ell_2 - 1, s} \cdots \Sigma_{\ell_1, s} V_{\ell_1} b \right| \leq \frac{1}{\left(1 - \frac{9}{16}\right)} \frac{4}{L} < \frac{10}{L} \tag{64}$$

with probability at least $1 - e^{-\Omega\left(\frac{p}{L^2}\right)}$.

Finally, when b is an arbitrary unit vector we can partition $[p] = S_1 \cup \ldots \cup S_m$, such that for all $i \in [m]$, $|S_i| \leq \left\lfloor \frac{c_1 p}{L^2} \right\rfloor$ and the number of the sets in the partition $q \leq \frac{p}{\left\lfloor \frac{c_1 p}{L^2} \right\rfloor} = \left\lceil \frac{L^2}{c_1} \right\rceil$. Thus, given an

arbitrary unit vector $b \in \mathbb{S}^{p-1}$, we can decompose it as $b = u_1 + \ldots + u_q$, where each u_i is non-zero only on the set S_i. By invoking the triangle inequality

$$|a^\top V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell} V_{\ell_1} b| \leq \sum_{i=1}^{q} |a^\top V_{\ell_2} \Sigma_{\ell_2-1} \cdots \Sigma_{\ell_1,s} V_{\ell_1} u_i|.$$

By applying the result of (64) to each term in the sum above we find that: for all k-sparse unit vectors a and all unit vectors $b \in \mathbb{S}^{p-1}$

$$|a^\top V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1} b| \leq \frac{10}{L} \sum_{i=1}^{q} \|u_i\| \leq \frac{10}{L} \sqrt{q} \left(\sum_{i=1}^{q} \|u_i\|^2 \right)^{1/2} = \frac{10\sqrt{q}}{L} = O(1)$$

with probability at least $1 - qe^{-\Omega\left(\frac{p}{L^2}\right)} = 1 - O(L^2) e^{-\Omega\left(\frac{p}{L^2}\right)}$. In other words for all k-sparse unit vectors a

$$\|a^\top V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1}\| = \sup_{b:\|b\|=1} |a^\top V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1} b| \leq O(1)$$

with the same probability that is at least $1 - O(L^2) e^{-\Omega\left(\frac{p}{L^2}\right)}$. By a union bound over the pairs of layers ℓ_1 and ℓ_2 and all samples $s \in [n]$ we establish that, with probability at least $1 - O(nL^4) e^{-\Omega\left(\frac{p}{L^2}\right)}$, for all pairs $1 \leq \ell_1 \leq \ell_2 \leq L$, all $s \in [n]$ and all k-sparse vectors a

$$\left\| \frac{a^\top}{\|a\|} V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1} \right\| \leq O(1).$$

Since, $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ we can ensure that this happens with probability at least $1 - \delta/8$ which completes the proof. ■

E.1.6. PROOF OF PART (F)

Lemma E.7 *For any $\delta > 0$, if $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial, then with probability at least $1 - \delta/8$ over the randomness in $V^{(1)}$ we have that for all $s \in [n]$ and all $1 \leq \ell_1 \leq \ell_2 \leq L$,*

$$|a^\top V_{\ell_2}^{(1)} \Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)} b| \leq O\left(\|a\| \|b\| \sqrt{\frac{k \log(p)}{p}} \right)$$

for all vectors a, b with $\|a\|_0, \|b\|_0 \leq k = \frac{cp}{\log(p) L^2}$, where c is a small enough positive absolute constant.

Proof Fix the layers $1 \leq \ell_1 \leq \ell_2 \leq L$ and the sample index s. At the end of the proof we shall take a union bound over all pairs of layers and all samples. In the proof, let us denote $V_\ell^{(1)}$ by V_ℓ and $\Sigma_{\ell,s}^{V^{(1)}}$ by $\Sigma_{\ell,s}$.

For any fixed vector z we know from Part (a) of Lemma E.10 that with probability at least $1 - O(nL^3)e^{-\Omega\left(\frac{p}{L^2}\right)}$ over the randomness in $(V_{\ell_2-1}, \dots, V_1)$

$$\|\Sigma_{\ell_2-1,s} V_{\ell_2-1} \dots \Sigma_{\ell_1,s} V_{\ell_1} z\| \le 2\|z\|. \tag{65}$$

Recall that the entries of V_{ℓ_2} are drawn independently from $\mathcal{N}(0, \frac{2}{p})$. Thus, conditioned on this event above, for any fixed vector w the random variable $w^\top V_{\ell_2} \left(\Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1} V_{\ell_1} z\right)$ is a mean-zero Gaussian with variance at most $\frac{8\|w\|^2\|z\|^2}{p}$. Therefore over the randomness in V_{ℓ_2}

$$\mathbb{P}\left(\left|w^\top V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1} z\right| \ge \frac{1}{c_2}\sqrt{\frac{k\log(p)}{p}}\|w\|\|z\| \,\middle|\, V_{\ell_1-1}, \dots, V_1\right) \le e^{-\frac{k\log(p)}{128c_2^2}}, \tag{66}$$

where c_2 is a small enough positive absolute constant that will be chosen only as a function of the constant c. A union bound over the events in (65) and (66) yields

$$\begin{aligned}\mathbb{P}\left(\left|w^\top V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1} z\right| \le \frac{1}{c_2}\sqrt{\frac{k\log(p)}{p}}\|w\|\|z\|\right) &\ge 1 - O(nL^3)e^{-\Omega\left(\frac{p}{L^2}\right)} - e^{-\frac{k\log(p)}{128c_2^2}} \\ &= 1 - e^{-\Omega\left(\frac{p}{L^2}\right)} - e^{-\frac{k\log(p)}{128c_2^2}}\end{aligned} \tag{67}$$

where the last equality holds since $p \ge \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial.

Let $\{w_i\}_{i=1}^m$ be a $1/4$-net of k-sparse unit vectors in Euclidean norm of size $m = \binom{p}{k}9^k$ (such a net exists, see Lemma F.10). Therefore by using (67) and taking a union bound we find that

$$\forall\, i, j \in [m], \quad \left|w_i^\top V_{\ell_2} \Sigma_{\ell_2-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1} w_j\right| \le \frac{1}{c_2}\sqrt{\frac{k\log(p)}{p}} \tag{68}$$

with probability at least

$$\begin{aligned}1 - m^2\left(e^{-\Omega(p/L^2)} + e^{-\frac{k\log(p)}{128c_2^2}}\right) &= 1 - O\left(\left(\binom{p}{k}9^k\right)^2\right)\left(e^{-\Omega\left(\frac{p}{L^2}\right)} + e^{-\frac{k\log(p)}{128c_2^2}}\right) \\ &\overset{(i)}{\ge} 1 - O\left(\left(\frac{9ep}{k}\right)^{2k}\right)\left(e^{-\Omega(p/L^2)} + e^{-\frac{k\log(p)}{128c_2^2}}\right) \\ &= 1 - \left(e^{-\Omega\left(\frac{p}{L^2}\right)+2k\log(9ep)} + e^{-\frac{k\log(p)}{128c_2^2}+2k\log(9ep)}\right) \\ &\overset{(ii)}{=} 1 - \left(e^{-\Omega\left(\frac{p}{L^2}\right)} + e^{-\Omega(k\log(p))}\right) \overset{(iii)}{=} 1 - e^{-\Omega\left(\frac{p}{L^2}\right)}\end{aligned}$$

where (i) follows since $\binom{p}{k} \le \left(\frac{ep}{k}\right)^k$, (ii) follows since $p \ge \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$, $k = \frac{cp}{\log(p)L^2}$ and because c_2 is a small enough absolute constant (which can be chosen given the constant c), and (iii) again follows since $k = \frac{cp}{\log(p)L^2}$.

We will now demonstrate that if the "good event" in (68) holds then we can use this to establish a similar guarantee for all k-sparse unit vectors a and b. Suppose, for each k-sparse unit vector w,

that $\zeta(w)$ is its nearest neighbor in $\{w_1, \ldots, w_m\}$. Then, if the event in (68) holds,

$$\begin{aligned}
\Xi &:= \sup_{a:\|a\|=1,\|a\|_0\le k}\ \sup_{b:\|b\|=1,\|b\|_0\le k} \left|a^\top V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}b\right| \\
&= \sup_{a:\|a\|=1,\|a\|_0\le k}\ \sup_{b:\|b\|=1,\|b\|_0\le k} \left|(a-\zeta(a)+\zeta(a))^\top V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}(b-\zeta(b)+\zeta(b))\right| \\
&\le \sup_{i,j\in[m]} \left|w_i^\top V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}w_j\right| \\
&\quad + \sup_{a:\|a\|=1,\|a\|_0\le k,j\in[m]} \left|(a-\zeta(a))^\top V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}w_j\right| \\
&\quad + \sup_{i\in[m],b:\|b\|=1,\|b\|_0\le k} \left|w_i^\top V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}(b-\zeta(b))\right| \\
&\quad + \sup_{a:\|a\|=1,\|a\|_0\le k}\ \sup_{b:\|b\|=1,\|b\|_0\le k} \left|(a-\zeta(a))^\top V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}(b-\zeta(b))\right| \\
&\overset{(i)}{\le} \frac{1}{c_2}\sqrt{\frac{k\log(p)}{p}} + \frac{\Xi}{4} + \frac{\Xi}{4} + \frac{\Xi}{16} = \frac{1}{c_2}\sqrt{\frac{k\log(p)}{p}} + \frac{9}{16}\Xi
\end{aligned}$$

where (i) follows by the definition of Ξ along with Lemma F.10, because we assume that the event in (68) holds, and also since $\|a-\zeta(a)\| \le 1/4$ and $\|b-\zeta(b)\| \le 1/4$.

By rearranging terms in the previous display we can infer that

$$\begin{aligned}
\Xi = \sup_{a:\|a\|=1,\|a\|_0\le k}\ \sup_{b:\|b\|=1,\|b\|_0\le k} \left|a^\top V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}b\right| &\le \frac{1}{\left(1-\frac{9}{16}\right)}\frac{1}{c_2}\sqrt{\frac{k\log(p)}{p}} \\
&= O\left(\sqrt{\frac{k\log(p)}{p}}\right)
\end{aligned}$$

with probability at least $1 - e^{-\Omega\left(\frac{p}{L^2}\right)}$. Taking a union bound over all pairs of layers and all sample we find that, with probability at least $1 - O(nL^2)e^{-\Omega\left(\frac{p}{L^2}\right)}$, for all $1 \le \ell_1 \le \ell_2 \le L$, for all $s \in [n]$ and all k-sparse vectors a and b

$$\left|\frac{a^\top}{\|a\|}V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1}\frac{b}{\|b\|}\right| = O\left(\sqrt{\frac{k\log(p)}{p}}\right). \tag{69}$$

Since $p \ge \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial we can ensure that this probability is at least $1 - \delta/8$ which completes our proof. ■

E.1.7. PROOF OF PART (G)

Lemma E.8 *For any $\delta > 0$, if $p \ge \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial, then, with probability at least $1 - \delta/8$ over the randomness in $V^{(1)}$, for all $s \in [n]$ and all $1 \le \ell \le L$,*

$$|V_{L+1}^{(1)}\Sigma_{L,s}^{V^{(1)}}\cdots\Sigma_{\ell,s}^{V^{(1)}}V_\ell^{(1)}a| \le O\left(\|a\|\sqrt{k\log(p)}\right)$$

for all vectors a with $\|a\|_0 \le k = \frac{cp}{\log(p)L^2}$, where c is a small enough positive absolute constant.

Proof Fix the layer $1 \le \ell \le L$ and the sample index s. At the end of the proof we shall take a union bound over all layers and all samples. Let us denote $V_\ell^{(1)}$ by V_ℓ and $\Sigma_{\ell,s}^{V^{(1)}}$ by $\Sigma_{\ell,s}$.

For any fixed vector z we know from Part (a) of Lemma E.10 that with probability at least $1 - O(nL^3)e^{-\Omega\left(\frac{p}{L^2}\right)}$ over the randomness in $(V_L, \dots, V_1)$

$$\|\Sigma_{L,s} V_L \dots \Sigma_{\ell,s} V_\ell z\| \le 2\|z\|. \tag{70}$$

Recall that the entries of V_{L+1} are drawn independently from $\mathcal{N}(0,1)$. Thus, conditioned on this event above, for any fixed vector w the random variable $w^\top V_{L+1} (\Sigma_{L,s} \cdots \Sigma_\ell V_\ell z)$ is a mean-zero Gaussian with variance at most $4\|z\|^2$. Therefore over the randomness in V_{L+1}

$$\mathbb{P}\left(|V_{L+1}\Sigma_{L,s} \cdots \Sigma_{\ell,s} V_\ell z| \ge \frac{\sqrt{k\log(p)}}{c_2}\|z\| \middle| V_L, \dots, V_1 \right) \le e^{-\frac{k\log(p)}{32c_2^2}}, \tag{71}$$

where c_2 is a small enough positive absolute constant that will be chosen only as a function of the constant c. A union bound over the events in (70) and (71) yields

$$\begin{aligned}\mathbb{P}\left(|V_{L+1}\Sigma_{L,s} \cdots \Sigma_{\ell,s} V_\ell z| \le \frac{\sqrt{k\log(p)}}{c_2}\|z\| \right) &\ge 1 - O(nL^3)e^{-\Omega\left(\frac{p}{L^2}\right)} - e^{-\frac{k\log(p)}{32c_2^2}} \\ &= 1 - e^{-\Omega\left(\frac{p}{L^2}\right)} - e^{-\frac{k\log(p)}{32c_2^2}}\end{aligned} \tag{72}$$

where the last equality holds since $p \ge \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial.

Let $\{z_i\}_{i=1}^m$ be a $1/4$-net of k-sparse unit vectors in Euclidean norm of size $m = \binom{p}{k}9^k$ (such a net exists, see Lemma F.10). Therefore by using (72) and taking a union bound we find that

$$\forall\, i \in [m], \quad |V_{L+1}\Sigma_{L,s} \cdots \Sigma_{\ell,s} V_\ell z_i| \le \frac{\sqrt{k\log(p)}}{c_2} \tag{73}$$

with probability at least

$$\begin{aligned}1 - m\left(e^{-\Omega(p/L^2)} + e^{-\frac{k\log(p)}{128c_2^2}}\right) &= 1 - O\left(\binom{p}{k}9^k\right)\left(e^{-\Omega\left(\frac{p}{L^2}\right)} + e^{-\frac{k\log(p)}{32c_2^2}}\right) \\ &\overset{(i)}{\ge} 1 - O\left(\left(\frac{9ep}{k}\right)^k\right)\left(e^{-\Omega(p/L^2)} + e^{-\frac{k\log(p)}{32c_2^2}}\right) \\ &= 1 - \left(e^{-\Omega\left(\frac{p}{L^2}\right) + k\log(9ep)} + e^{-\frac{k\log(p)}{32c_2^2} + k\log(9ep)}\right) \\ &\overset{(ii)}{=} 1 - \left(e^{-\Omega\left(\frac{p}{L^2}\right)} + e^{-\Omega(k\log(p))}\right) \overset{(iii)}{=} 1 - e^{-\Omega\left(\frac{p}{L^2}\right)}\end{aligned}$$

where (i) follows since $\binom{p}{k} \le \left(\frac{ep}{k}\right)^k$, (ii) follows since $p \ge \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$, $k = \frac{cp}{\log(p)L^2}$ and because c_2 is a small enough absolute constant (which can be chosen given the constant c), and (iii) follows again since $k = \frac{cp}{\log(p)L^2}$.

We will now demonstrate that if the "good event" (73) holds then we can use this to establish a similar guarantee for all k-sparse unit vectors a. To see this, as before, suppose ζ maps any unit-length k-sparse vector to its nearest neighbor in $\{z_1, \ldots, z_r\}$. Suppose that the event in (73) holds then

$$\begin{aligned}
\Xi &:= \sup_{a:\|a\|=1,\|a\|_0\leq k} |V_{L+1}\Sigma_{L,s}\cdots\Sigma_{\ell,s}V_\ell a| \\
&= \sup_{a:\|a\|=1,\|a\|_0\leq k} |V_{L+1}\Sigma_{L,s}\cdots\Sigma_{\ell,s}V_\ell (a-\zeta(a)+\zeta(a))| \\
&\leq \sup_{i\in[m]} |V_{L+1}\Sigma_{L,s}\cdots\Sigma_{\ell,s}V_\ell z_j| + \sup_{a:\|a\|=1,\|a\|_0\leq k} |V_{L+1}\Sigma_{L,s}\cdots\Sigma_{\ell,s}V_\ell (a-\zeta(a))| \\
&\overset{(i)}{\leq} \frac{\sqrt{k\log(p)}}{c_2} + \frac{1}{4}\sup_{a:\|a\|=1,\|a\|_0\leq k} \left|V_{L+1}\Sigma_{L,s}\cdots\Sigma_{\ell,s}V_\ell \frac{a-\zeta(a)}{\|a-\zeta(a)\|}\right| \overset{(ii)}{\leq} \frac{\sqrt{k\log(p)}}{c_2} + \frac{\Xi}{4}
\end{aligned}$$

where (i) holds because we assume that the event in (73) holds and since $\|a-\zeta(a)\| \leq 1/4$, and (ii) follows by the definition of Ξ along with Lemma F.10.

By rearranging terms in the previous display we infer that

$$\Xi = \sup_{a:\|a\|=1,\|a\|_0\leq k} |V_{L+1}\Sigma_{L,s}\cdots\Sigma_{\ell,s}V_\ell a| \leq \frac{1}{\left(1-\frac{1}{4}\right)}\frac{\sqrt{k\log(p)}}{c_2} = O\left(\sqrt{k\log(p)}\right)$$

with probability at least $1 - e^{-\Omega\left(\frac{p}{L^2}\right)}$. Taking a union bound over all layers and all sample we find that, for all $1 \leq \ell \leq L$, for all $s \in [n]$ and all k-sparse vectors a

$$\left|V_{L+1}\Sigma_{L,s}\cdots\Sigma_{\ell,s}V_\ell \frac{a}{\|a\|}\right| = O\left(\sqrt{k\log(p)}\right) \tag{74}$$

with probability at least $1 - O(nL)e^{-\Omega\left(\frac{p}{L^2}\right)}$. Since $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial we can ensure that this probability is at least $1-\delta/8$ which completes our proof. ■

E.1.8. PROOF OF PART (H)

Lemma E.9 *For any $\delta > 0$, suppose that $h < \frac{1}{50\sqrt{p}L}$, $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial and $\tau = \Omega\left(\frac{\log^2(\frac{nL}{\delta})}{p^{\frac{3}{2}}L^3}\right)$. For $\beta = O\left(\frac{L^2\tau^{2/3}}{\sqrt{p}}\right)$, if*

$$\mathcal{S}_{\ell,s}(\beta) := \left\{j \in [p] : |V^{(1)}_{\ell,j}x^{V^{(1)}}_{\ell,s}| \leq \beta\right\}$$

where $V^{(1)}_{\ell,j}$ refers to the jth row of $V^{(1)}_\ell$, then with probability at least $1-\delta/8$ over the randomness in $V^{(1)}$ we have that for all $\ell \in [L]$ and all $s \in [n]$:

$$|\mathcal{S}_{\ell,s}(\beta)| \leq O(p^{3/2}\beta) = O(pL^2\tau^{2/3}).$$

Proof To ease notation let us refer to $V^{(1)}_\ell$ as V_ℓ and $x^{V^{(1)}}_{\ell,s}$ as $x_{\ell,s}$. For a fixed $\ell \in [L]$ and sample $s \in [n]$ define

$$Z(\ell, j, s) := \mathbb{I}\left[|V_{\ell,j}x_{\ell,s}| \leq \beta\right]$$

so that $|\mathcal{S}_{\ell,s}(\beta)| = \sum_{j=1}^{p} Z(j,\ell,s)$. Define $\mathcal{E}$ to be the event that $\|x_{\ell-1,s}\| \geq \frac{1}{2}$. By inequality (52) in the proof of Lemma E.2 above

$$\mathbb{P}\left[\mathcal{E}\right] \geq 1 - O(nL)\exp\left(-\Omega\left(\frac{p}{L^2}\right)\right). \tag{75}$$

Conditioned on $x_{\ell-1,s}$ since each entry of $V_{\ell,j}$ is drawn independently from $\mathcal{N}(0,\frac{2}{p})$ we know that the distribution of $V_{\ell,j}x_{\ell-1,s} \sim \mathcal{N}\left(0, \frac{2\|x_{\ell-1,s}\|^2}{p}\right)$. Thus, conditioned on the event $\mathcal{E}$, which is determined by the random weights *before* layer ℓ, we have that

$$\begin{aligned}\mathbb{E}\left[Z(j,\ell,s) \mid \mathcal{E}\right] = \mathbb{P}\left[j \in \mathcal{S}_{\ell,s}(\beta) \mid \mathcal{E}\right] &= \sqrt{\frac{p}{4\pi\|x_{\ell-1,s}\|^2}}\int_{-\beta}^{\beta}\exp\left(-\frac{x^2 p}{4\|x_{\ell-1,s}\|^2}\right)\mathrm{d}x \\ &\leq \sqrt{\frac{p}{\pi}}\int_{-\beta}^{\beta}\exp\left(-\frac{x^2 p}{4\|x_{\ell-1,s}\|^2}\right)\mathrm{d}x \leq 2\beta\sqrt{\frac{p}{\pi}}.\end{aligned}$$

On applying Hoeffding's inequality (see Theorem F.5) we find that

$$\begin{aligned}\mathbb{P}\Bigg[|\mathcal{S}_{\ell,s}(\beta)| \leq \mathbb{E}\left[\sum_{j=1}^{p} Z(j,\ell,s) \;\Big|\; \mathcal{E}\right] + p^{3/2}\beta \leq p\left(2\beta\sqrt{\frac{p}{\pi}}\right) + p^{3/2}\beta \leq 3p^{3/2}\beta \;\Big|\; \mathcal{E}\Bigg] \\ \geq 1 - \exp(-\Omega(p^{3/2}\beta)).\end{aligned} \tag{76}$$

Taking a union bound over the events in (75) and (76) we find that

$$\mathbb{P}\left[|\mathcal{S}_{\ell,s}(\beta)| \leq 3p^{3/2}\beta\right] \geq 1 - \exp(-\Omega(p^{3/2}\beta)) - O(nL)\exp\left(-\Omega\left(\frac{p}{L^2}\right)\right).$$

Applying a union bound over all samples and all layers we find that, with probability at least $1 - O(nL)\exp(-\Omega(p^{3/2}\beta)) - O(n^2L^2)\exp\left(-\Omega\left(\frac{p}{L^2}\right)\right)$, for all $\ell \in [L]$ and all $s \in [n]$,

$$|\mathcal{S}_{\ell,s}(\beta)| \leq 3p^{3/2}\beta = O(pL^2\tau^{2/3}).$$

We shall now demonstrate that this probability of success is at least $1 - \delta/8$. On substituting the value of $\beta = O(\frac{L^{8/3}\tau^{2/3}}{\sqrt{p}})$ we find that this probability is at least

$$\begin{aligned}&1 - O(nL)\exp(-\Omega(p^{3/2}\beta)) - O(n^2L^2)\exp\left(-\Omega\left(\frac{p}{L^2}\right)\right) \\ &\quad= 1 - O(nL)\exp(-\Omega(pL^2\tau^{2/3})) - O(n^2L^2)\exp\left(-\Omega\left(\frac{p}{L^2}\right)\right) \\ &\quad\overset{(i)}{=} 1 - O(nL)\exp(-\Omega(pL^2\tau^{2/3})) - \exp\left(-\Omega\left(\frac{p}{L^2}\right)\right) \\ &\quad\overset{(ii)}{=} 1 - O(nL)\exp\left(-\Omega\left(\log^{\frac{4}{3}}\left(\frac{nL}{\delta}\right)\right)\right) - \exp\left(-\Omega\left(\frac{p}{L^2}\right)\right) \\ &\quad\geq 1 - \delta/8,\end{aligned}$$

where (i) follows since $p \geq \text{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial and (ii) follows by assumption that $\tau = \Omega\left(\frac{\log^2(\frac{nL}{\delta})}{p^{\frac{3}{2}}L^3}\right)$. This completes our proof. ■

E.1.9. OTHER USEFUL CONCENTRATION LEMMAS

The following lemma is useful in the proofs of Lemmas E.4-E.8. It bounds the norm of an arbitrary unit vector z that is multiplied by alternating weight matrices $V_\ell^{(1)}$ and corresponding $\Sigma_{\ell,s}^{V^{(1)}}$.

Lemma E.10 *If $p \geq \text{poly}(L, \log(n))$ for a large enough polynomial, then given an arbitrary unit vector $z \in \mathbb{S}^{p-1}$, with probability at least $1 - O(nL^3)\exp\left(-\Omega\left(\frac{p}{L^2}\right)\right)$ over the randomness in $V^{(1)}$, for all $1 \leq \ell_1 \leq \ell_2 \leq L$ and for all $s \in [n]$,*

(a) $\left\|\Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)} z\right\| \leq 2$, *and*

(b) $\left\|V_{\ell_2}^{(1)} \Sigma_{\ell_2-1,s}^{V^{(1)}} \cdots \Sigma_{\ell_1,s}^{V^{(1)}} V_{\ell_1}^{(1)} z\right\| \leq 2.$

Proof We denote $V_\ell^{(1)}$ by V_ℓ, $\Sigma_{\ell,s}^{V^{(1)}}$ by $\Sigma_{\ell,s}$, and $x_{\ell,s}^{V^{(1)}}$ by $x_{\ell,s}$.

Proof of Part (a): For any layer $\ell \in \{\ell_1, \ldots, \ell_2 - 1\}$ define

$$z_{\ell,s} := \Sigma_{\ell,s} V_\ell \Sigma_{\ell-1,s} \cdots \Sigma_{\ell_1,s} V_{\ell_1} z$$

with the convention that $z_{\ell_1-1,s} := z$.

Conditioned on $z_{\ell-1,s}$ the distribution of $V_\ell z_{\ell-1,s} \sim \mathcal{N}\left(0, \frac{2\|z_{\ell-1,s}\|^2 I}{p}\right)$, since each entry of V_ℓ is drawn independently from $\mathcal{N}(0, \frac{2}{p})$. We begin by evaluating the expected value of its squared norm conditioned on the randomness in $V_{\ell-1}, \ldots, V_1$. Let $V_{\ell,j}$ denote the jth row of V_ℓ and let $(\Sigma_{\ell,s})_{jj}$ denote the jth element on the diagonal of $\Sigma_{\ell,s}$, then

$$\begin{aligned}\mathbb{E}\left[\|z_{\ell,s}\|^2 | V_{\ell-1}, \ldots, V_1\right] &= \mathbb{E}\left[\|\Sigma_{\ell,s} V_\ell z_{\ell-1,s}\|^2 | V_{\ell-1}, \ldots, V_1\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{p} \left((\Sigma_{\ell,s})_{jj} V_{\ell,j} z_{\ell-1,s}\right)^2 \Big| V_{\ell-1}, \ldots, V_1\right].\end{aligned}$$

By the definition of the Huberized ReLU observe that each entry

$$(\Sigma_{\ell,s})_{jj} = \phi'(V_{\ell,j} x_{\ell,s}) \leq \mathbb{I}\left[V_{\ell,j} x_{\ell-1,s} \geq 0\right]$$

and therefore

$$\mathbb{E}\left[\|z_{\ell,s}\|^2 | V_{\ell-1}, \ldots, V_1\right] \leq \mathbb{E}\left[\sum_{j=1}^{p} \mathbb{I}\left[V_{\ell,j} x_{\ell-1,s} \geq 0\right] \left(V_{\ell,j} z_{\ell-1,s}\right)^2 \Big| V_{\ell-1}, \ldots, V_1\right]. \tag{77}$$

Let us decompose $V_{\ell,j}$ into its component in the $x_{\ell-1,s}$ direction, and its a component that is perpendicular to $x_{\ell-1,s}$. That is, define

$$V_{\ell,j}^{\|} := \left(V_{\ell,j} \cdot \frac{x_{\ell-1,s}}{\|x_{\ell-1,s}\|}\right) \frac{x_{\ell-1,s}}{\|x_{\ell-1,s}\|} \quad \text{and} \quad V_{\ell,j}^{\perp} := V_{\ell,j} - V_{\ell,j}^{\|}.$$

Since $V_{\ell,j}$ is Gaussian, $V_{\ell,j}^{\|}$ and $V_{\ell,j}^{\perp}$ are conditionally independent given the previous layers.

Thus, continuing from inequality (77), we have

$$\begin{aligned}
&\mathbb{E}\left[\|z_{\ell,s}\|^2 | V_{\ell-1},\dots,V_1\right] \\
&\le \mathbb{E}\left[\sum_{j=1}^{p} \mathbb{I}\left[V_{\ell,j} x_{\ell-1,s}^{V^{(1)}} \ge 0\right] \left((V_{\ell,j}^{\|} + V_{\ell,j}^{\perp}) z_{\ell-1,s}\right)^2 | V_{\ell-1},\dots,V_1\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{p} \mathbb{I}\left[V_{\ell,j}^{\|} x_{\ell-1,s}^{V^{(1)}} \ge 0\right] \left((V_{\ell,j}^{\|} + V_{\ell,j}^{\perp}) z_{\ell-1,s}\right)^2 | V_{\ell-1},\dots,V_1\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{p} \mathbb{I}\left[V_{\ell,j}^{\|} x_{\ell-1,s}^{V^{(1)}} \ge 0\right] \left(V_{\ell,j}^{\|} z_{\ell-1,s}\right)^2 | V_{\ell-1},\dots,V_1\right] \\
&\quad + \mathbb{E}\left[\sum_{j=1}^{p} \mathbb{I}\left[V_{\ell,j}^{\|} x_{\ell-1,s}^{V^{(1)}} \ge 0\right] \left(V_{\ell,j}^{\perp} z_{\ell-1,s}\right)^2 | V_{\ell-1},\dots,V_1\right] \\
&\quad + 2\mathbb{E}\left[\sum_{j=1}^{p} \mathbb{I}\left[V_{\ell,j}^{\|} x_{\ell-1,s}^{V^{(1)}} \ge 0\right] \left(V_{\ell,j}^{\|} z_{\ell-1,s}\right) \left(V_{\ell,j}^{\perp} z_{\ell-1,s}\right) | V_{\ell-1},\dots,V_1\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{p} \mathbb{I}\left[V_{\ell,j} x_{\ell-1,s}^{V^{(1)}} \ge 0\right] \left(V_{\ell,j}^{\|} z_{\ell-1,s}\right)^2 | V_{\ell-1},\dots,V_1\right] \\
&\quad + \mathbb{E}\left[\sum_{j=1}^{p} \mathbb{I}\left[V_{\ell,j} x_{\ell-1,s}^{V^{(1)}} \ge 0\right] \left(V_{\ell,j}^{\perp} z_{\ell-1,s}\right)^2 | V_{\ell-1},\dots,V_1\right], \qquad (78)
\end{aligned}$$

since, after conditioning on $V_{\ell-1},\dots,V_1$, we have that $V_{\ell,j}^{\|}$ and $V_{\ell,j}^{\perp}$ independent and $V_{\ell,j}^{\perp} z_{\ell-1,s}$ is zero mean.

Now, decompose the vector $z_{\ell-1,s}$ into *its* component in the $x_{\ell-1,s}$ direction, which we refer to as $z_{\ell-1,s}^{\|}$, and a component that is perpendicular to $x_{\ell-1,s}$, which we refer to as $z_{\ell-1,s}^{\perp}$. That is, define

$$z_{\ell-1,s}^{\|} := \left(z_{\ell-1,s} \cdot \frac{x_{\ell-1,s}}{\|x_{\ell-1,s}\|}\right) \frac{x_{\ell-1,s}}{\|x_{\ell-1,s}\|} \qquad \text{and} \qquad z_{\ell-1,s}^{\perp} := z_{\ell-1,s} - z_{\ell-1,s}^{\|}.$$

Since $V_{\ell,j}^{\|} z_{\ell-1,s} = V_{\ell,j} z_{\ell-1,s}^{\|}$ and $V_{\ell,j}^{\perp} z_{\ell-1,s} = V_{\ell,j} z_{\ell-1,s}^{\perp}$, inequality (78) implies

$$\begin{aligned}
&\mathbb{E}\left[\|z_{\ell,s}\|^2 | V_{\ell-1},\dots,V_1\right] \\
&\quad \le \mathbb{E}\left[\sum_{j=1}^{p} \mathbb{I}\left[V_{\ell,j} x_{\ell-1,s}^{V^{(1)}} \ge 0\right] \left(V_{\ell,j} z_{\ell-1,s}^{\|}\right)^2 | V_{\ell-1},\dots,V_1\right] \\
&\qquad + \mathbb{E}\left[\sum_{j=1}^{p} \mathbb{I}\left[V_{\ell,j} x_{\ell-1,s}^{V^{(1)}} \ge 0\right] \left(V_{\ell,j} z_{\ell-1,s}^{\perp}\right)^2 | V_{\ell-1},\dots,V_1\right]. \qquad (79)
\end{aligned}$$

We begin by evaluating the term involving the parallel components. For any j, conditioned on $V_{\ell-1}, \ldots, V_1$, recalling that $V_{\ell,j}$ is the jth row of V_ℓ, the random variable $V_{\ell,j} x_{\ell-1,s} \sim \mathcal{N}\left(0, \frac{2\|x_{\ell-1,s}\|^2}{p}\right)$, and therefore

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}\left[V_{\ell,j} x_{\ell-1,s} \geq 0\right]\left(V_{\ell,j} z^{\|}_{\ell-1,s}\right)^2 \Big| V_{\ell-1}, \ldots, V_1\right] \\
&= \left(z_{\ell-1,s} \cdot \left(\frac{x_{\ell-1,s}}{\|x_{\ell-1,s}\|^2}\right)\right)^2 \mathbb{E}\left[\mathbb{I}\left[V_{\ell,j} x_{\ell-1,s} \geq 0\right]\left(V_{\ell,j} x_{\ell-1,s}\right)^2 \Big| V_{\ell-1}, \ldots, V_1\right] \\
&= \left(z_{\ell-1,s} \cdot \left(\frac{x_{\ell-1,s}}{\|x_{\ell-1,s}\|^2}\right)\right)^2 \frac{1}{2} \times \mathbb{E}\left[\left(V_{\ell,j} x_{\ell-1,s}\right)^2 \Big| V_{\ell-1}, \ldots, V_1\right] \\
&= \left(z_{\ell-1,s} \cdot \left(\frac{x_{\ell-1,s}}{\|x_{\ell-1,s}\|^2}\right)\right)^2 \frac{1}{2} \times \frac{2\|x_{\ell-1,s}\|^2}{p} = \frac{\|z^{\|}_{\ell-1,s}\|^2}{p}.
\end{aligned} \tag{80}$$

For the perpendicular component, notice that, conditioned on $(V_{\ell-1}, \ldots, V_1)$, we have that $V_{\ell,j} z^{\perp}_{\ell-1,s} = V^{\perp}_{\ell,j} z_{\ell-1,s}$ and $\mathbb{I}\left[V_{\ell,j} x_{\ell-1,s} \geq 0\right] = \mathbb{I}\left[V^{\|}_{\ell,j} x_{\ell-1,s} \geq 0\right]$ are independent, and hence

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}\left[V_{\ell,j} x_{\ell-1,s} \geq 0\right]\left(V_{\ell,j} z^{\perp}_{\ell-1,s}\right)^2 \Big| V_{\ell-1}, \ldots, V_1\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left(V_{\ell,j} z^{\perp}_{\ell-1,s}\right)^2 \Big| V_{\ell-1}, \ldots, V_1\right] = \frac{1}{2} \times \frac{2\|z^{\perp}_{\ell-1,s}\|^2}{p} = \frac{\|z^{\perp}_{\ell-1,s}\|^2}{p}.
\end{aligned} \tag{81}$$

By combining the results of (79)-(81) we find that

$$\mathbb{E}\left[\|z_{\ell,s}\|^2 | V_{\ell-1}, \ldots, V_1\right] \leq p\left(\frac{\|z^{\perp}_{\ell-1,s}\|^2 + \|z^{\|}_{\ell-1,s}\|^2}{p}\right) = \|z_{\ell-1,s}\|^2. \tag{82}$$

By symmetry among the p coordinates we can also infer that $\mathbb{E}\left[(z_{\ell,s})_i^2 | V_{\ell-1}, \ldots, V_1\right] \leq \|z_{\ell-1,s}\|^2/p$ for each $i \in [p]$. Thus, by the same argument as we used in Lemma E.2 to arrive at (50) we can show that conditioned on $V_{\ell-1}, \ldots, V_1$ the sub-Gaussian norm $\|(z_{\ell,s})_i\|_{\psi_2}$ is at most $c_1\|z_{\ell-1,s}\|/\sqrt{p}$ and hence the sub-exponential norm $\|(z_{\ell,s})_i^2\|_{\psi_1} \leq \|(z_{\ell,s})_i\|^2_{\psi_2} \leq c_2\|z_{\ell-1,s}\|^2/p$ (by Lemma F.3). Therefore by Bernstein's inequality (see Theorem F.6) for any $\eta \in (0, 1]$

$$\mathbb{P}\left[\|z_{\ell,s}\|^2 \leq \|z_{\ell-1,s}\|^2(1+\eta) \Big| V_{\ell-1}, \ldots, V_1\right] \geq 1 - \exp\left(-c_3 p \eta^2\right).$$

Setting $\eta = \frac{1}{50L}$ and taking a union bound we infer that

$$\begin{aligned}
&\mathbb{P}\left[\forall s \in [n], \ell \in \{\ell_2 - 1, \ldots, \ell_1\},\ \|z_{\ell,s}\| \leq \|z_{\ell-1,s}\|\sqrt{1+\eta}\right] \\
&\qquad \geq 1 - O(nL)\exp\left(-\Omega\left(\frac{p}{L^2}\right)\right).
\end{aligned} \tag{83}$$

We will now show by an inductive argument for the layers that if the "good event" in (83) holds then $\|z_{\ell,s}\| \leq 1 + 3(\ell - \ell_1 + 1)\eta$, for all $\ell \in \{\ell_1 - 1, \ldots, \ell_2 - 1\}$ and all $s \in [n]$. The base case

holds at $\ell_1 - 1$ since by definition $\|z_{\ell_1-1,s}\| = \|z\| = 1$. Now assume that the inductive argument holds at any layers $\ell_1, \dots, \ell - 1$. Then if the event in (83) holds we have

$$\begin{aligned}
\|z_{\ell,s}\| &\le \|z_{\ell-1,s}\|\sqrt{1+\eta} \\
&\le (1 + 3(\ell - \ell_1)\eta)\,(1+\eta) \qquad \text{(by the IH and because } \sqrt{1+\eta} \le 1+\eta) \\
&= 1 + 3\left(\ell - \ell_1 + \frac{1}{3}\right)\eta + 3(\ell-\ell_1)\eta^2 \\
&\le 1 + 3(\ell - \ell_1 + 1)\eta \qquad \text{(since } \eta = \tfrac{1}{50L} \text{ and } L \ge 1).
\end{aligned}$$

This completes the induction. Hence we have shown that for all

$$\mathbb{P}\left[\forall s \in [n],\ \|z_{\ell_2-1,s}\| \le 1 + \frac{3(\ell_2 - \ell_1 + 1)}{50L}\right] \ge 1 - O(nL)\exp\left(-\Omega\left(\frac{p}{L^2}\right)\right). \tag{84}$$

Recall that $z_{\ell_2-1,s} := \Sigma_{\ell_2-1}\cdots\Sigma_{\ell_1,s}V_{\ell_1}z$, therefore taking union bound over all pairs of layers we get that, with probability at least $1 - O(nL^3)\exp\left(-\Omega\left(\frac{p}{L^2}\right)\right)$, for all $1 \le \ell_1 \le \ell_2 \le L+1$ and all $s \in [n]$,

$$\|z_{\ell_2-1,s}\| \le 1 + \frac{3(\ell_2 - \ell_1 + 1)}{50L}.$$

This completes the proof of the first part of the lemma.

Proof of Part (b): For a fixed $s \in [n]$ we condition on $z_{\ell_2-1,s}$ and consider the random variable $a_s = V_{\ell_2}z_{\ell_2-1,s}$. Since $\ell_2 \in [L]$ each entry of V_ℓ is drawn independently from $\mathcal{N}(0, \frac{2}{p})$.

The distribution of each entry of a_s conditioned on $z_{\ell_2-1,s}$ is $\mathcal{N}\left(0, \frac{2\|z_{\ell_2-1,s}\|^2}{p}\right)$. Therefore by the Gaussian-Lipschitz concentration inequality (see Theorem F.7) for any $\eta' > 0$

$$\mathbb{P}\left[\|a_s\| \le \sqrt{2}\|z_{\ell_2-1,s}\|(1+\eta') \mid z_{\ell_2-1,s}\right] \ge 1 - \exp\left(-c_4 p\eta'^2\right).$$

Setting $\eta' = \frac{1}{50L}$ and taking a union bound over all samples we get that

$$\mathbb{P}\left[\forall s \in [n],\ \|a_s\| \le \sqrt{2}\|z_{\ell_2-1,s}\|\left(1 + \frac{1}{50L}\right)\middle| z_{\ell_2-1,s}\right] \ge 1 - n\exp\left(-\frac{c_4 p}{L^2}\right). \tag{85}$$

By a union bound over the events in (84) and (85) we find that

$$\mathbb{P}\left[\forall s \in [n],\ \|a_s\| \le \sqrt{2}\left(1 + \frac{3(\ell_2 - \ell_1 + 1)}{50L}\right)\left(1 + \frac{1}{50L}\right)\right] \ge 1 - O(nL)\exp\left(-\Omega\left(\frac{p}{L^2}\right)\right).$$

The definition of $a_s = V_{\ell_2}z_{\ell_2-1} = V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1,s}$ and the previous display above yields that

$$\begin{aligned}
\mathbb{P}\Bigg[\forall s \in [n],\ \|V_{\ell_2}z_{\ell_2-1} = V_{\ell_2}\Sigma_{\ell_2-1,s}\cdots\Sigma_{\ell_1,s}V_{\ell_1,s}\| \le \sqrt{2}\left(1 + \frac{3L}{50L}\right)\left(1 + \frac{1}{50L}\right) \le 2\Bigg] \\
\ge 1 - O(nL)\exp\left(-\Omega\left(\frac{p}{L^2}\right)\right).
\end{aligned}$$

Finally a union bound over all pairs of $1 \le \ell_1 \le \ell_2 \le L$ completes the proof of the second part. ■

The next lemma bounds the magnitude of the initial function values with high probability.

Lemma E.11 *For any $\delta > 0$, suppose that $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial, then with probability at least $1 - \delta$ over the randomness in $V^{(1)}$ for all $s \in [n]$,*

$$|f_{V^{(1)}}(x_s)| \leq c\sqrt{\log(2n/\delta)}.$$

Proof By Lemma E.2, with probability at least $1 - \delta/8$

$$\|x_{L,s}^{V^{(1)}}\| \leq 2 \tag{86}$$

for all $s \in [n]$. Fix a sample with index $s \in [n]$. Conditioned on $x_{L,s}^{V^{(1)}}$, the random variable $V_{L+1}^{(1)} x_{L,s}^{V^{(1)}} \sim \mathcal{N}(0, \|x_{L,s}^{V^{(1)}}\|^2)$ since each entry of $V_{L+1}^{(1)}$ is drawn independently from $\mathcal{N}(0,1)$. Therefore for any $\eta > 0$

$$\mathbb{P}\left[|f_{V^{(1)}}(x_s)| \leq \eta \|x_{L,s}^{V^{(1)}}\|^2 \mid x_{L,s}^{V^{(1)}}\right] \geq 1 - 2\exp\left(-c_1\eta^2\right).$$

A union bound over all samples implies

$$\mathbb{P}\left[\forall\, s \in [n],\ |f_{V^{(1)}}(x_s)| \leq \eta \|x_{L,s}^{V^{(1)}}\|^2 \mid x_{L,s}^{V^{(1)}}\right] \geq 1 - 2n\exp\left(-c_1\eta^2\right).$$

Setting $\eta = c_2\sqrt{\log(n/\delta)}$ where c_2 is a large enough absolute constant we get that

$$\mathbb{P}\left[\forall\, s \in [n],\ |f_{V^{(1)}}(x_s)| \leq c_2\sqrt{\log(n/\delta)} \|x_{L,s}^{V^{(1)}}\|^2 \mid x_{L,s}^{V^{(1)}}\right] \geq 1 - \frac{7\delta}{8}. \tag{87}$$

Taking union bound over the events in (86) and (87) we find that

$$\mathbb{P}\left[\forall\, s \in [n],\ |f_{V^{(1)}}(x_s)| \leq c_3\sqrt{\log(n/\delta)}\right] \geq 1 - \delta$$

which completes the proof. ■

Lastly we prove a lemma that bounds the norm of the initial weight matrix with high probability.

Lemma E.12 *For any $\delta > 0$, suppose that $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial, then with probability at least $1 - \delta$ over the randomness in $V^{(1)}$*

$$\|V^{(1)}\| \leq \sqrt{5pL}.$$

Proof By definition

$$\|V^{(1)}\|^2 = \sum_{\ell \in [L+1]} \|V_\ell^{(1)}\|^2.$$

When $\ell \in [L]$, the matrix $V_\ell^{(1)}$ has its entries drawn from $\mathcal{N}(0, \frac{2}{p})$. Therefore by applying Theorem F.7 we find that for any fixed $\ell \in [L]$,

$$\mathbb{P}\left[\|V_\ell^{(1)}\|^2 \leq p^2 \times \frac{2}{p} \times \frac{5}{4} = \frac{5p}{2}\right] \leq \exp\left(-\Omega(p)\right).$$

While when $\ell = L+1$, the p-dimensional vector $V_{L+1}^{(1)}$ has its entries drawn from $\mathcal{N}(0,1)$. Hence, again applying Theorem F.7 we get

$$\mathbb{P}\left[\|V_{L+1}^{(1)}\|^2 \le p \times 1 \times \frac{5}{4} = \frac{5p}{4}\right] \le \exp\left(-\Omega(p)\right).$$

Taking a union bound over all $L+1$ layers we find that

$$\mathbb{P}\left[\forall\, \ell \in [L+1] : \|V_\ell^{(1)}\|^2 \le \frac{5p}{2}\right] \le (L+1)\exp\left(-\Omega(p)\right) \le 1-\delta$$

where the last inequality follows since $p \ge \text{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$. Therefore,

$$\|V^{(1)}\|^2 \le (L+1) \times \frac{5p}{2} \le 5pL$$

with probability at least $1-\delta$. Taking square roots establishes the claim. ■

E.2. Useful Properties in a Neighborhood Around the Initialization

In the next two lemmas we shall assume that the "good event" described in Lemma E.1 holds. We shall show that when the initial weight matrices satisfy those properties, we can also extend some of these properties to matrices in a neighborhood around the initial parameters.

Lemma E.13 *Let the event in Lemma E.1 hold and suppose that the conditions on h, p and τ described in that lemma hold. Let $\widetilde{V}$ be weights such that $\|\widetilde{V}_\ell - V_\ell^{(1)}\|_{op} \le \tau$ for all $\ell \in [L]$. For all $\ell \in [L]$ and $s \in [n]$, let $\widetilde{\Sigma}_{\ell,s}$ be diagonal matrices such that $\|\widetilde{\Sigma}_{\ell,s} - \Sigma_{\ell,s}^{V^{(1)}}\|_0 \le k$, and $(\widetilde{\Sigma}_{\ell,s})_{jj} \in [-3,3]$ for all $j \in [p]$. There is an absolute constant C' such that, for all small enough $c > 0$, if $\tau \le \sqrt{\frac{k\log(p)}{p}} \le \frac{c}{L^3}$ then, for all $1 \le \ell_1 \le \ell_2 \le L$,*

$$\left\| \prod_{j=\ell_1}^{\ell_2} \widetilde{V}_j^\top \widetilde{\Sigma}_j \right\|_{op} \le C' L^2.$$

Proof Fix an arbitrary sample index s. To ease notation let us refer to $V^{(1)}$ as V, $\Sigma_{\ell,s}^{V^{(1)}}$ as Σ_ℓ, and $\widetilde{\Sigma}_{\ell,s}$ as $\widetilde{\Sigma}_\ell$. Note that for any $j \in [L]$

$$\widetilde{V}_j^\top \widetilde{\Sigma}_j = V_j^\top \Sigma_j + \underbrace{V_j^\top\left(\widetilde{\Sigma}_j - \Sigma_j\right)}_{=:\Gamma_j} + \underbrace{(\widetilde{V}_j - V_j)^\top \widetilde{\Sigma}_j}_{=:\Delta_j}. \tag{88}$$

Let us refer to Γ_j and Δ_j as "flip matrices". Then, if we define the set $\mathcal{A}_j = \{V_j^\top \Sigma_j, \Gamma_j, \Delta_j\}$, expanding the product into a sum of terms yields

$$\prod_{j=\ell_1}^{\ell_2} \widetilde{V}_j^\top \widetilde{\Sigma}_j = \prod_{j=\ell_1}^{\ell_2} \left(V_j^\top \Sigma_{j,s} + \Gamma_{j,s} + \Delta_{j,s}\right) = \sum_{A_{\ell_1} \in \mathcal{A}_{\ell_1}, \ldots, A_{\ell_2} \in \mathcal{A}_{\ell_2}} \prod_{j=\ell_1}^{\ell_2} A_j. \tag{89}$$

Each term in the sum on the RHS of (89) is a product of $\ell_2 - \ell_1 + 1$ matrices (A_j), some of which are flip matrices. We will bound the operator norm of the sum by bounding the operator norms of each of the terms, and applying the triangle inequality. To bound the operator norms of the terms, we will decompose the terms into products of subsequences of matrices, and bound the operator norms of the subsequences. The subsequences will have at most two flip matrices, and will be determined by the positions of those flip matrices. One term in the sum has no flip matrices—it will have a single subsequence that is the entire term. Some terms have exactly one flip matrix. Those terms will be broken into two subsequences, one that ends at the flip matrix, and the other consisting of the rest of the term. The other terms in the sum have at least two flips. Each such term can be broken down as follows:

- one or more subsequences with at exactly two flip matrices ending in a flip matrix,
- possibly a subsequence with one flip matrix, ending with the flip matrix, and
- a (possibly empty) subsequence with no flip matrices.

In the calculations that follow the indices q_1, q_2 and q_3 satisfy: $1 \le \ell_1 \le q_1 \le q_2 \le q_3 \le \ell_2 \le L$. Let $C > 1$ be a large enough positive constant such that all the upper bounds in Lemma E.1 hold with this constant.

Subsequences with no flip matrices: First, subsequences for which which $A_j = V_j^\top \Sigma_j$ for all j can be bounded by Part (c) of Lemma E.1:

$$\left\| \prod_{j=q_1}^{q_2} V_j^\top \Sigma_j \right\|_{op} = \left\| \prod_{j=q_2}^{q_1} \Sigma_j V_j \right\|_{op} \le CL. \tag{90}$$

Subsequences with one flip matrix: There will be two types of sub-sequences with just one flip matrix. First, let us consider the following type of subsequence:

$$\begin{aligned} \left\| \left(\prod_{j=q_1}^{q_2-1} V_j^\top \Sigma_j \right) \Delta_{q_2} \right\|_{op} = \left\| \Delta_{q_2}^\top \left(\prod_{j=q_2-1}^{q_1} \Sigma_j V_j \right) \right\|_{op} &\le \left\| \Delta_{q_2}^\top \right\|_{op} \left\| \prod_{j=q_2-1}^{q_1} \Sigma_j V_j \right\|_{op} \\ &\overset{(i)}{\le} CL \left\| \Delta_{q_2}^\top \right\|_{op} \\ &= CL \left\| \widetilde{\Sigma}_{q_2} (\widetilde{V}_{q_2} - V_{q_2}) \right\|_{op} \\ &\overset{(ii)}{\le} CL \left\| \widetilde{V}_{q_2} - V_{q_2} \right\|_{op} \\ &\overset{(iii)}{\le} C\tau L \overset{(iv)}{\le} \frac{cC}{L^2}, \end{aligned} \tag{91}$$

where (i) follows by again invoking Part (c) of Lemma E.1, (ii) follows since by assumption the diagonal matrix $\widetilde{\Sigma}_{q_2}$ has its entries bounded between $[-3, 3]$, (iii) follows since by assumption $\left\| \widetilde{V}_{q_2} - V_{q_2} \right\|_{op} \le \tau$ and (iv) follows since by assumption $\tau = c/L^3$.

Next, let us consider the second type of subsequence with just one flip matrix:

$$\begin{aligned}\left\|\left(\prod_{j=q_1}^{q_2-1} V_j^\top \Sigma_j\right)\Gamma_{q_2}\right\|_{op} &= \left\|\left(\prod_{j=q_1}^{q_2-1} V_j^\top \Sigma_j\right) V_{q_2}^\top\left(\widetilde{\Sigma}_{q_2}-\Sigma_{q_2}\right)\right\|_{op} \\ &= \sup_{a:\|a\|=1}\left\|\left(\prod_{j=q_1}^{q_2-1} V_j^\top \Sigma_j\right) V_{q_2}^\top\left(\widetilde{\Sigma}_{q_2}-\Sigma_{q_2}\right) a\right\| \\ &= \sup_{a:\|a\|=1}\left\|a^\top\left(\widetilde{\Sigma}_{q_2}-\Sigma_{q_2}\right) V_{q_2}\prod_{j=q_2-1}^{q_1}\Sigma_j V_j\right\|.\end{aligned}$$

For each a let's define $b = \left(\widetilde{\Sigma}_{q_2}-\Sigma_{q_2}\right)a$. Since $\|\widetilde{\Sigma}_{q_2}-\Sigma_{q_2}\|_0 \le k$, therefore b is k-sparse. Also since the diagonal matrix $\widetilde{\Sigma}_{q_2}$ has entries in $[-3,3]$ and Σ_{q_2} has entries in $[0,1]$, therefore the entries of $\widetilde{\Sigma}_{q_2}-\Sigma_{q_2}$ lie in $[-4,4]$. This implies that $\|b\| \le 4\|a\| \le 4$. Applying Part (e) of Lemma E.1, we have

$$\left\|\left(\prod_{j=q_1}^{q_2-1} V_j^\top \Sigma_j\right)\Gamma_{q_2}\right\|_{op} \le C\|b\| = 4C. \tag{92}$$

Subsequences with two flip matrices: Now we continue to subsequences with two flip matrices. There shall be four types of such subsequences. We begin by consider subsequences of the type

$$\begin{aligned}&\left\|\left(\prod_{j=q_1}^{q_2-1} V_j^\top \Sigma_j\right)\Delta_{q_2}\left(\prod_{j=q_2+1}^{q_3-1} V_j^\top \Sigma_j\right)\Delta_{q_3}\right\|_{op} \\ &\quad\le \left\|\left(\prod_{j=q_1+1}^{q_2-1} V_j^\top \Sigma_j\right)\Delta_{q_2}\right\|_{op}\left\|\left(\prod_{j=q_2+1}^{q_3-1} V_j^\top \Sigma_j\right)\Delta_{q_3}\right\|_{op} \overset{(i)}{\le} (C\tau L)^2 \overset{(ii)}{\le} \frac{cC}{L^2}\cdot C\tau L,\end{aligned} \tag{93}$$

where (i) follows by (91) and (ii) follows since $\tau \le c/L^3$.

Next, we bound the operator norm of a subsequence of the type

$$\begin{aligned}&\left\|\left(\prod_{j=q_1}^{q_2-1} V_j^\top \Sigma_j\right)\Delta_{q_2}\left(\prod_{j=q_2+1}^{q_3-1} V_j^\top \Sigma_j\right)\Gamma_{q_3}\right\|_{op} \\ &\quad\le \left\|\left(\prod_{j=q_1}^{q_2-1} V_j^\top \Sigma_j\right)\Delta_{q_2}\right\|_{op}\left\|\left(\prod_{j=q_2+1}^{q_3-1} V_j^\top \Sigma_j\right)\Gamma_{q_3}\right\|_{op} \\ &\quad\overset{(i)}{\le} 4C\cdot C\tau L \\ &\quad= 4C^2\tau L,\end{aligned} \tag{94}$$

where (i) follows by invoking inequalities (91) and (92).

We continue to bound the operator norm of subsequences

$$\left\| \left(\prod_{j=q_1}^{q_2-1} V_j^\top \Sigma_j \right) \Gamma_{q_2} \left(\prod_{j=q_2+1}^{q_3-1} V_j^\top \Sigma_j \right) \Delta_{q_3} \right\|_{op} \leq \left\| \left(\prod_{j=q_1}^{q_2-1} V_j^\top \Sigma_j \right) \Gamma_{q_2} \right\|_{op} \left\| \left(\prod_{j=q_2}^{q_3-1} V_j^\top \Sigma_j \right) \Delta_{q_3} \right\|_{op}$$
$$\overset{(i)}{\leq} 4C^2 \tau L \tag{95}$$

where (i) follows again invoking inequalities (91) and (92).

Finally we bound the operator norm of subsequences of the type

$$\begin{aligned}
&\left\| \left(\prod_{j=q_1}^{q_2-1} V_j^\top \Sigma_j \right) \Gamma_{q_2} \left(\prod_{j=q_2+1}^{q_3-1} V_j^\top \Sigma_j \right) \Gamma_{q_3} \right\|_{op} \\
&\quad \leq \left\| \prod_{j=q_1}^{q_2-1} V_j^\top \Sigma_j \right\|_{op} \left\| V_{q_2}^\top \left(\widetilde{\Sigma}_{q_2} - \Sigma_{q_2} \right) \left(\prod_{j=q_2+1}^{q_3-1} V_j^\top \Sigma_j \right) V_{q_3}^\top \left(\widetilde{\Sigma}_{q_3} - \Sigma_{q_3} \right) \right\|_{op} \\
&\quad \overset{(i)}{\leq} CL \sup_{a:\|a\|=1} \left\| V_{q_2}^\top \left(\widetilde{\Sigma}_{q_2} - \Sigma_{q_2} \right) \left(\prod_{j=q_2+1}^{q_3-1} V_j^\top \Sigma_j \right) V_{q_3}^\top \left(\widetilde{\Sigma}_{q_3} - \Sigma_{q_3} \right) a \right\| \\
&\quad = CL \sup_{a:\|a\|=1} \sup_{b:\|b\|=1} \left| b^\top V_{q_2}^\top \left(\widetilde{\Sigma}_{q_2} - \Sigma_{q_2} \right) \left(\prod_{j=q_2+1}^{q_3-1} V_j^\top \Sigma_j \right) V_{q_3}^\top \left(\widetilde{\Sigma}_{q_3} - \Sigma_{q_3} \right) a \right| \\
&\quad = CL \sup_{a:\|a\|=1} \sup_{b:\|b\|=1} \left| a^\top \left(\widetilde{\Sigma}_{q_3} - \Sigma_{q_3} \right) V_{q_3} \left(\prod_{j=q_3-1}^{q_2+1} \Sigma_j V_j \right) \left(\widetilde{\Sigma}_{q_2} - \Sigma_{q_2} \right) V_{q_2} b \right| \\
&\quad = CL \sup_{a:\|a\|=1} \sup_{b:\|b\|=1} \left| a^\top \left(\widetilde{\Sigma}_{q_3} - \Sigma_{q_3} \right) V_{q_3} \left(\prod_{j=q_3-1}^{q_2+1} \Sigma_j V_j \right) \left(\widetilde{\Sigma}_{q_2} - \Sigma_{q_2} \right) \frac{V_{q_2} b}{\|V_{q_2} b\|} \right| \|V_{q_2} b\| \\
&\quad \overset{(ii)}{\leq} 4C^2 L \sqrt{\frac{k \log p}{p}} \sup_{b:\|b\|=1} \|V_{q_2} b\| = 4C^2 L \sqrt{\frac{k \log p}{p}} \|V_{q_2}\|_{op} \overset{(iii)}{\leq} 4C^3 L \sqrt{\frac{k \log p}{p}}
\end{aligned} \tag{96}$$

where (i) follows by invoking Part (c) of Lemma E.1. Inequality (ii) follows since the vectors $a^\top \left(\widetilde{\Sigma}_{q_2} - \Sigma_{q_2} \right)$ and $\left(\widetilde{\Sigma}_{q_1} - \Sigma_{q_1} \right) \frac{V_{q_1} b}{\|V_{q_1} b\|}$ are k-sparse and both have norm less than or equal to 4, thus we can apply Part (f) of Lemma E.1, and (iii) follows by applying Part (b) of Lemma E.1.

As stated above we can decompose each product in (89) that has at least two flips into subsequences that end in a flip and have exactly two flips, and subsequences that have at most one flip. The subsequences that have at most one flip have operator norm at most $4CL$ (by inequalities (90)-(92)).

The above logic (93)-(95) implies that subsequences with exactly two flips that have at least one Δ flip have operator norm at most $4C^2 \tau L \leq 4C^3 \sqrt{\frac{k \log p}{p}} L$ (since $\tau \leq \sqrt{\frac{k \log p}{p}}$ by assumption and $C > 1$). Subsequences with two Γ flips have operator norm at most $4C^3 \sqrt{\frac{k \log p}{p}} L$. Define

$\psi := 4C^3\sqrt{\frac{k\log p}{p}}L$. So, if a sequence has r flip matrices then its operator norm is bounded by

$$\psi^{\lfloor r/2 \rfloor} \times 4CL.$$

So, putting it together, by recalling the decomposition in (89) we have

$$\begin{aligned}
\left\| \prod_{j=\ell_1}^{\ell_2} \widetilde{V}_j^\top \widetilde{\Sigma}_j \right\|_{op} &= \left\| \sum_{A_{\ell_1} \in \mathcal{A}_{\ell_1}, \ldots, A_{\ell_2} \in \mathcal{A}_{\ell_2}} \prod_{j=\ell_1}^{\ell_2} A_j \right\|_{op} \\
&\overset{(i)}{\leq} (1+2L) \times 4CL + \sum_{A_{\ell_1} \in \mathcal{A}_{\ell_1}, \ldots, A_{\ell_2} \in \mathcal{A}_{\ell_2} : \geq 2 \text{ flips}} \left\| \prod_{j=\ell_1}^{\ell_2} A_j \right\|_{op} \\
&\overset{(ii)}{\leq} 12CL^2 + \sum_{r=2}^{L} \binom{L}{r} 2^r \cdot 4CL\psi^{\lfloor r/2 \rfloor} \\
&\leq 12CL^2 + 8CL \sum_{r=2}^{L} \binom{L}{r} (4\psi)^{\lfloor r/2 \rfloor} \\
&\leq 12CL^2 + 8CL \left[\sum_{r \in \{2,\ldots,L\}, r \text{ even}} \left(\binom{L}{r} + \binom{L}{r+1} \right) (\sqrt{4\psi})^r \right] \\
&= 12CL^2 + 8CL \left[\sum_{r \in \{2,\ldots,L\}, r \text{ even}} \binom{L}{r} \left(1 + \frac{\binom{L}{r+1}}{\binom{L}{r}} \right) (\sqrt{4\psi})^r \right] \\
&\leq 12CL^2 + 16CL^2 \left[\sum_{r \in \{2,\ldots,L\}, r \text{ even}} \binom{L}{r} (\sqrt{4\psi})^r \right] \\
&\leq 12CL^2 + 16CL^2 \sum_{r=0}^{L} \binom{L}{r} (\sqrt{4\psi})^r \\
&\overset{(iii)}{\leq} 12CL^2 + 16CL^2 \left(1 + \sqrt{4\psi} \right)^L \\
&\overset{(iv)}{=} 12CL^2 + 16CL^2 \left(1 + 4C^{3/2}\sqrt{L} \left(\frac{k\log(p)}{p} \right)^{1/4} \right)^L \\
&\overset{(v)}{\leq} 12CL^2 + 16CL \left(1 + \frac{4C^{3/2}\sqrt{c}}{L} \right)^L \overset{(vi)}{\leq} 30CL^2
\end{aligned}$$

where (i) follows since the number of terms with at most one flip matrix is $(1+2L)$ and the operator norm of each such term is upper bounded by $4CL$ by inequalities (90)-(92). Inequality (ii) is because the number of terms with r flip matrices is $\binom{L}{r}2^r$, (iii) is by the Binomial theorem, (iv) is by our definition of ψ. Inequality (v) follows since by assumption $\sqrt{\frac{k\log(p)}{p}} \leq \frac{c}{L^3}$, and (vi) follows since c is small enough and because there exists positive constants c_1 and $c_2(c_1)$ such that, for any $0 \leq z < \frac{c_1}{L}$, $(1+z)^L \leq 1 + c_2 Lz$. This completes our proof. ■

The following lemma bounds the difference between post-activation features at the ℓth layer when the weight matrix is perturbed from its initial value.

Lemma E.14 *Let the event in Lemma E.1 hold and suppose that the conditions on h, p and τ described in that lemma hold with the additional assumptions that $\tau \leq \frac{c}{L^{12}\log^{\frac{3}{2}}(p)}$, where c is a small enough constant, and $h \leq \frac{\tau}{\sqrt{p}}$. Let $V^{(1)}$ be the initial weights and $\widetilde{V}$, $\hat{V}$ be such that $\|\widetilde{V}_\ell - V_\ell^{(1)}\|_{op}, \|\hat{V}_\ell - V_\ell^{(1)}\|_{op} \leq \tau$ for all $\ell \in [L]$. Then*

1. $\|\Sigma_{\ell,s}^{\widetilde{V}} - \Sigma_{\ell,s}^{\hat{V}}\|_0 \leq O(pL^2\tau^{2/3})$;
2. $\|x_{\ell,s}^{\widetilde{V}} - x_{\ell,s}^{\hat{V}}\| \leq O(L^3\tau)$;

for all $\ell \in [L]$ and all $s \in [n]$.

Proof Fix the sample s In this proof, we will refer to $\Sigma_{\ell,s}^{V^{(1)}}, \Sigma_{\ell,s}^{\widetilde{V}}, \Sigma_{\ell,s}^{\hat{V}}, x_{\ell,s}^{V^{(1)}}, x_{\ell,s}^{\widetilde{V}}$ and $x_{\ell,s}^{\hat{V}}$ as Σ_ℓ, $\widetilde{\Sigma}_\ell$, $\hat{\Sigma}_\ell$, x_ℓ, $\widetilde{x}_\ell$ and $\hat{x}_\ell$ respectively.

Before the first layer (at layer 0) define $\Sigma_0 = \widetilde{\Sigma}_0 = \hat{\Sigma}_0 = I$ and recall that by definition for any sample $s \in [n]$, $x_{0,s}^{V^{(1)}} = x_{0,s}^{\widetilde{V}} = x_{0,s}^{\hat{V}} = x_s$.

For constants c_1, c_2 to be determined later, we will prove using induction that, for all $\ell \in [L]$,

1. $\|\Sigma_\ell - \widetilde{\Sigma}_\ell\|_0, \|\Sigma_\ell - \hat{\Sigma}_\ell\|_0 \leq c_1 pL^2\tau^{2/3}$,
2. $\|\widetilde{\Sigma}_\ell - \hat{\Sigma}_\ell\|_0 \leq 2c_1 pL^2\tau^{2/3}$, and
3. $\|x_\ell - \widetilde{x}_\ell\|, \|x_\ell - \hat{x}_\ell\|, \|\widetilde{x}_\ell - \hat{x}_\ell\| \leq c_2 L^3\tau$.

The base case, where $\ell = 0$, is trivially true since $x_0 = \hat{x}_0 = \widetilde{x}_0$ and $\Sigma_0 = \widetilde{\Sigma}_0 = \hat{\Sigma}_0$.

Now let us assume that the inductive hypothesis holds for all layers $r = 1, \ldots, \ell - 1$. We shall prove that the inductive hypothesis holds at layer ℓ in two steps.

Step 1: By the triangle inequality

$$\|\widetilde{\Sigma}_\ell - \hat{\Sigma}_\ell\|_0 \leq \|\Sigma_\ell - \widetilde{\Sigma}_\ell\|_0 + \|\Sigma_\ell - \hat{\Sigma}_\ell\|_0. \tag{97}$$

Note that showing that $\|\Sigma_\ell - \widetilde{\Sigma}_\ell\|_0$ and $\|\Sigma_\ell - \hat{\Sigma}_\ell\|_0$ are at most $c_1 pL^2\tau^{2/3}$ also proves the claim that $\|\widetilde{\Sigma}_\ell - \hat{\Sigma}_\ell\|_0 \leq 2c_1 pL^2\tau^{2/3}$.

We begin by bounding $\|\Sigma_\ell - \widetilde{\Sigma}_\ell\|_0$. Recall that by definition the diagonal matrix $(\Sigma_\ell)_{jj} = (\phi'(V_\ell x_{\ell-1}))_j$. So to bound the difference between $\Sigma_\ell - \widetilde{\Sigma}_\ell$ we characterize the difference between

$$V_\ell x_{\ell-1} - \widetilde{V}_\ell \widetilde{x}_{\ell-1} = (V_\ell - \widetilde{V}_\ell)x_{\ell-1} + \widetilde{V}_\ell(x_{\ell-1} - \widetilde{x}_{\ell-1}).$$

We know that, by assumption, $\|\widetilde{V}_\ell - V_\ell\|_{op} \leq \tau$, and that $\|x_{\ell-1}\| \leq 2$ by Part (a) of Lemma E.1, and $\|\widetilde{x}_{\ell-1} - x_{\ell-1}\| \leq c_2 L^3\tau$ by the inductive hypothesis. Therefore,

$$\begin{aligned}
\|V_\ell x_{\ell-1} - \widetilde{V}_\ell \widetilde{x}_{\ell-1}\| &\leq \|(\widetilde{V}_\ell - V_\ell)x_{\ell-1}\| + \|\widetilde{V}_\ell(\widetilde{x}_{\ell-1} - x_{\ell-1})\| \\
&\leq \|\widetilde{V}_\ell - V_\ell\|_{op}\|x_{\ell-1}\| + \|\widetilde{V}_\ell\|_{op}\|\widetilde{x}_{\ell-1} - x_{\ell-1}\| \\
&\leq 2\tau + c_2 L^3\tau\|\widetilde{V}_\ell\|_{op} \\
&\leq 2\tau + c_2 L^3\tau\left(\|\widetilde{V}_\ell - V_\ell\|_{op} + \|V_\ell\|_{op}\right) \\
&\overset{(i)}{\leq} 2\tau + c_2 L^3\tau\left(\tau + c_3\right) \overset{(ii)}{\leq} \tau\left(c_2 c_4 L^3 + 2\right),
\end{aligned}$$

where (i) follows since $\|V_\ell\|_{op} \leq c_3$ by Part (b) of Lemma E.1 and (ii) follows since τ is smaller than a constant by assumption.

Let $\beta = \frac{c_5 L^2 \tau^{\frac{2}{3}}}{\sqrt{p}} > 2h > 0$. The reason for this particular choice of the value of β shall become clear shortly, and $h \leq \beta/2$ since $h \leq \frac{\tau}{\sqrt{p}}$ by assumption. Define the set

$$\mathcal{S}_\ell(\beta) := \{j \in [p] : |V_{\ell,j} x_{\ell-1}| \leq \beta\}$$

where $V_{\ell,j}$ refers to the jth row of V_ℓ. Also define

$$s_\ell^{(1)}(\beta) := \left|\{j \in \mathcal{S}_\ell(\beta) : (\Sigma_\ell)_{jj} \neq (\widetilde{\Sigma}_\ell)_{jj}\}\right| \qquad \text{and}$$
$$s_\ell^{(2)}(\beta) := \left|\{j \in \mathcal{S}_\ell^c(\beta) : (\Sigma_\ell)_{jj} \neq (\widetilde{\Sigma}_\ell)_{jj}\}\right|.$$

Clearly we must have that

$$\|\Sigma_\ell - \widetilde{\Sigma}_\ell\|_0 = s_\ell^{(1)}(\beta) + s_\ell^{(2)}(\beta).$$

To bound $s_\ell^{(1)}(\beta)$ we note that $s_\ell^{(1)}(\beta) \leq |\mathcal{S}_\ell(\beta)| \leq c_6 p^{3/2}\beta$ by Part (h) of Lemma E.1. We focus on $s_\ell^{(2)}(\beta)$. For a $j \in \mathcal{S}_\ell^c(\beta)$ by the definition of the Huberized ReLU if $(\Sigma_\ell)_{jj} \neq (\widetilde{\Sigma}_\ell)_{jj}$ then we must have that

$$\left|\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1} - V_{\ell,j}x_{\ell-1}\right| \geq \beta - h.$$

This further implies that

$$(\beta-h)^2 s_\ell^{(2)}(\beta) \leq \sum_{j \in \mathcal{S}_\ell^c(\beta):(\Sigma_\ell)_{jj} \neq (\widetilde{\Sigma}_\ell)_{jj}} \left|\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1} - V_{\ell,j}x_{\ell-1}\right|^2 \leq \|V_\ell x_{\ell-1} - \widetilde{V}_\ell \widetilde{x}_{\ell-1}\|^2$$
$$\leq \tau^2 \left(c_2 c_4 L^3 + 2\right)^2.$$

Therefore, we find that

$$\begin{aligned}\|\Sigma_\ell - \widetilde{\Sigma}_\ell\|_0 &= s_\ell^{(1)}(\beta) + s_\ell^{(2)}(\beta) \\ &\leq \frac{\tau^2 \left(c_2 c_4 L^3 + 2\right)^2}{(\beta-h)^2} + c_6 p^{3/2}\beta \\ &\leq \frac{\left(c_2 c_7 L^3 \tau\right)^2}{\beta^2} + c_6 p^{3/2}\beta,\end{aligned}$$

if $c_2 \geq 1/c_4$, since $h \leq \beta/2$. Balancing both of these terms on the RHS leads to the choice $\beta = \frac{c_5 L^2 \tau^{2/3}}{\sqrt{p}}$. This choice of β shows that

$$\|\Sigma_\ell - \widetilde{\Sigma}_\ell\|_0 \leq 2c_6 p^{3/2}\beta = 2c_6 c_5 p L^2 \tau^{2/3} = c_1 p L^2 \tau^{2/3}.$$

Similarly we can also show that $\|\Sigma_\ell - \hat{\Sigma}_\ell\|_0 \leq c_1 p L^2 \tau^{2/3}$. These two bounds along with (97) proves the first part of the inductive hypothesis. This combined with inequality (97) also proves the second part of the inductive hypothesis.

Step 2: Now, for the third part we want to show that $\|\widetilde{x}_{\ell,s} - \hat{x}_{\ell,s}\|$ remains bounded. We can also show that $\|x_{\ell,s} - \widetilde{x}_{\ell,s}\|$ and $\|x_{\ell,s} - \widetilde{x}_{\ell,s}\|$ remain bounded by mirroring the logic that follows. Define a diagonal matrix $\check{\Sigma}_\ell$, whose diagonal entries are

$$(\check{\Sigma}_\ell)_{jj} := (\hat{\Sigma}_\ell - \widetilde{\Sigma}_\ell)_{jj} \left[\frac{\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}}{\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}} \right], \qquad \text{for all, } j \in [p].$$

In the definition above we use the convention that $0/0 = 0$. We will show that for any $j \in [p]$

$$|(\check{\Sigma}_\ell)_{jj}| = \left| (\hat{\Sigma}_\ell - \widetilde{\Sigma}_\ell)_{jj} \left[\frac{\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}}{\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}} \right] \right| \leq 1.$$

Firstly observe that the matrices $\widetilde{\Sigma}_\ell$ and $\hat{\Sigma}_\ell$ have entries between $[0, 1]$, therefore $\widetilde{\Sigma}_\ell - \Sigma_\ell$ has entries between -1 and 1. Also recall that by the definition of the Huberized ReLU,

$$(\hat{\Sigma}_\ell)_{jj} = \begin{cases} 1 & \text{if } \hat{V}_{\ell,j}\hat{x}_{\ell-1} > h, \\ \frac{\hat{V}_{\ell,j}\hat{x}_{\ell-1}}{h} & \text{if } \hat{V}_{\ell,j}\hat{x}_{\ell-1} \in [0, h], \\ 0 & \text{if } \hat{V}_{\ell,j}\hat{x}_{\ell-1} < 0. \end{cases}$$

Now we will analyze a few cases and show that the absolute values of the entries of $\check{\Sigma}_\ell$ are smaller than 1 in each case.

If the signs of $\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}$ and $\hat{V}_{\ell,j}\hat{x}_{\ell-1}$ are opposite then we must have that

$$\left| (\hat{\Sigma}_\ell - \widetilde{\Sigma}_\ell)_{jj} \frac{\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}}{\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}} \right| \overset{(i)}{\leq} \left| \frac{\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}}{\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}} \right| = \frac{|\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}|}{|\hat{V}_{\ell,j}\hat{x}_{\ell-1}| + |\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}|} \leq 1,$$

where (i) follows since $|(\hat{\Sigma}_\ell - \widetilde{\Sigma}_\ell)_{jj}| \leq 1$. If they have the same sign and are both negative then $(\widetilde{\Sigma}_\ell - \Sigma_\ell)_{jj} = 0$ in this case. The same is true when they are both positive and are bigger than h. Therefore, we are only left with the case when both are positive and one of them is smaller than h. If $\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1} > h$ and $\hat{V}_{\ell,j}\hat{x}_{\ell-1} \in [0, h]$ we have that

$$\left| (\hat{\Sigma}_\ell - \widetilde{\Sigma}_\ell)_{jj} \frac{\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}}{\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}} \right| = \left| \frac{\left(\frac{\hat{V}_{\ell,j}\hat{x}_{\ell-1}}{h} - 1 \right) \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}}{\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}} \right| = \left| \frac{\frac{\hat{V}_{\ell,j}\hat{x}_{\ell-1}}{h} - 1}{\frac{\hat{V}_{\ell,j}\hat{x}_{\ell-1}}{\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}} - 1} \right| \leq 1$$

where the last inequality follows since $\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1} > h$. And finally in the case where $\hat{V}_{\ell,j}\hat{x}_{\ell-1} > h$ and $\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1} \in [0, h]$ we have that

$$\left| (\hat{\Sigma}_\ell - \widetilde{\Sigma}_\ell)_{jj} \frac{\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}}{\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}} \right| = \left| \left(1 - \frac{\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}}{h} \right) \frac{\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}}{\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}} \right| = \frac{\left(h - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1} \right) \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}}{h \left(\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1} \right)}.$$

To show that this term of the RHS above is smaller than 1 it is sufficient to show that

$$(h - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1})\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1} \leq h(\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}),$$

in our case where $0 \leq \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1} \leq h < \hat{V}_{\ell,j}\hat{x}_{\ell-1}$. Consider the change of variables $a = \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}$ and $b = \hat{V}_{\ell,j}\hat{x}_{\ell-1}$, then it suffices to show that

$$(h-a)a \leq h(b-a) \Leftarrow 0 \leq a^2 - 2ah + hb.$$

The derivative of $a^2 - 2ah + hb$ with respect to a is $2(a-h)$, which is non-positive when $a \leq h$. Therefore the minimum of the quadratic when $a \in [0,h]$ is at $a = h$ and the minimum value is $h^2 - 2h^2 + hb = hb - h^2 = h(b-h) > 0$. This proves that $|(\check{\Sigma})_{jj}| \leq 1$ in this final case as well.

With this established we note that

$$\begin{aligned}
e_\ell := \hat{x}_\ell - \widetilde{x}_\ell &= \phi(\hat{V}_\ell \hat{x}_{\ell-1}) - \phi(\widetilde{V}_\ell \widetilde{x}_{\ell-1}) \\
&= \hat{\Sigma}_\ell \hat{V}_\ell \hat{x}_{\ell-1} - \widetilde{\Sigma}_\ell \widetilde{V}_\ell \widetilde{x}_{\ell-1} + \underbrace{\phi(\hat{V}_\ell \hat{x}_{\ell-1}) - \hat{\Sigma}_\ell \hat{V}_\ell \hat{x}_{\ell-1} - \phi(\widetilde{V}_\ell \widetilde{x}_{\ell-1}) + \widetilde{\Sigma}_\ell \widetilde{V}_\ell \widetilde{x}_{\ell-1}}_{=:\chi_\ell} \\
&\overset{(i)}{=} \left(\hat{\Sigma}_\ell + \check{\Sigma}_\ell\right)\left(\hat{V}_\ell \hat{x}_{\ell-1} - \widetilde{V}_\ell \widetilde{x}_{\ell-1}\right) + \chi_\ell \\
&= \underbrace{\left(\hat{\Sigma}_\ell + \check{\Sigma}_\ell\right)\hat{V}_\ell}_{=:A_\ell} \underbrace{(\hat{x}_{\ell-1} - \widetilde{x}_{\ell-1})}_{=:e_{\ell-1}} + \underbrace{\left(\hat{\Sigma}_\ell + \check{\Sigma}_\ell\right)\left(\widetilde{V}_\ell - V_\ell\right)\widetilde{x}_{\ell-1} + \chi_\ell}_{=:b_\ell} \\
&= A_\ell e_{\ell-1} + b_\ell
\end{aligned} \tag{98}$$

where (i) follows because by the definition of the matrix $\check{\Sigma}_\ell$ for each coordinate j we have

$$\begin{aligned}
\left(\hat{\Sigma}_\ell + \check{\Sigma}_\ell\right)_{jj}\left(\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}\right) &= \left(\hat{\Sigma}_\ell\right)_{jj}\left(\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}\right) + \left(\check{\Sigma}_\ell\right)_{jj}\left(\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}\right) \\
&= \left(\hat{\Sigma}_\ell\right)_{jj}\left(\hat{V}_{\ell,j}\hat{x}_{\ell-1} - \widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}\right) + (\hat{\Sigma}_\ell - \widetilde{\Sigma}_\ell)_{jj}\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1} \\
&= (\hat{\Sigma}_\ell)_{jj}\hat{V}_{\ell,j}\hat{x}_{\ell-1} - (\widetilde{\Sigma}_\ell)_{jj}\widetilde{V}_{\ell,j}\widetilde{x}_{\ell-1}.
\end{aligned}$$

In equation (98) above we have expressed the difference between the post-activation features at layer ℓ in terms of the difference at layer $\ell-1$ plus some error terms. Repeating this $\ell-1$ more times yields

$$e_\ell = A_\ell e_{\ell-1} + b_\ell = A_\ell(A_{\ell-1}e_{\ell-2} + b_{\ell-1}) + b_\ell = \prod_{j=\ell}^{1} A_j e_0 + \left(\sum_{r=1}^{\ell-1}\left[\prod_{j=\ell}^{r+1} A_j\right] b_r\right) + b_\ell.$$

Since $e_0 = \|\hat{x}_0 - \widetilde{x}_0\| = 0$, by re-substituting the values of A_ℓ and b_ℓ we find that

$$\begin{aligned}
\hat{x}_\ell - \widetilde{x}_\ell = \sum_{r=1}^{\ell-1}\left[\prod_{j=\ell}^{r+1}\left(\hat{\Sigma}_j + \check{\Sigma}_j\right)\hat{V}_j\right](\hat{\Sigma}_r + \check{\Sigma}_r)(\hat{V}_r - \widetilde{V}_r)\widetilde{x}_{r-1} + \left(\hat{\Sigma}_\ell + \check{\Sigma}_\ell\right)\left(\widetilde{V}_\ell - \hat{V}_\ell\right)\widetilde{x}_{\ell-1} \\
+ \sum_{r=1}^{\ell-1}\left[\prod_{j=\ell}^{r+1}(\hat{\Sigma}_j + \check{\Sigma}_j)\hat{V}_j\right]\chi_r + \chi_\ell,
\end{aligned} \tag{99}$$

and therefore by the triangle inequality

$$
\begin{aligned}
&\|\hat{x}_\ell - \widetilde{x}_\ell\| \\
&\leq \left\| \sum_{r=1}^{\ell-1} \left[\prod_{j=\ell}^{r+1} \left(\hat{\Sigma}_j + \check{\Sigma}_j \right) \hat{V}_j \right] (\hat{\Sigma}_r + \check{\Sigma}_r)(\hat{V}_r - \widetilde{V}_r)\widetilde{x}_{r-1} \right\| + \left\| \left(\hat{\Sigma}_\ell + \check{\Sigma}_\ell \right) \left(\widetilde{V}_\ell - \hat{V}_\ell \right) \widetilde{x}_{\ell-1} \right\| \\
&\qquad + \left\| \sum_{r=1}^{\ell-1} \left[\prod_{j=\ell}^{r+1} (\hat{\Sigma}_j + \check{\Sigma}_j) V_j \right] \chi_r \right\| + \|\chi_\ell\| \\
&\leq \sum_{r=1}^{\ell-1} \left\| \prod_{j=\ell}^{r+1} \left(\hat{\Sigma}_j + \check{\Sigma}_j \right) \hat{V}_j \right\|_{op} \left\| \hat{\Sigma}_r + \check{\Sigma}_r \right\|_{op} \left\| \hat{V}_r - \widetilde{V}_r \right\|_{op} \|\widetilde{x}_{r-1}\| + \left\| \hat{\Sigma}_\ell + \check{\Sigma}_\ell \right\|_{op} \left\| \widetilde{V}_\ell - \hat{V}_\ell \right\|_{op} \|\widetilde{x}_{\ell-1}\| \\
&\qquad + \sum_{r=1}^{\ell-1} \left\| \prod_{j=\ell}^{r+1} (\hat{\Sigma}_j + \check{\Sigma}_j) V_j \right\|_{op} \|\chi_r\| + \|\chi_\ell\|.
\end{aligned}
$$

Recall that the diagonal matrices $(\Sigma_\ell - \hat{\Sigma}_\ell - \check{\Sigma}_\ell)$ are $3c_1 p L^2 \tau^{2/3}$ sparse by the inductive hypothesis. Also the matrices $(\Sigma_\ell - \hat{\Sigma}_\ell - \check{\Sigma}_\ell)$ have entries in $[-3, 3]$. Therefore by applying Lemma E.13 (note that since $\tau \leq \frac{c}{L^{12} \log^{\frac{3}{2}}(p)}$ therefore Lemma E.13 applies at this level of sparsity) we find that for a constant c_8 (that does not depend on c_1), we have

$$
\begin{aligned}
\|\hat{x}_\ell - \widetilde{x}_\ell\| &\leq c_8 L^2 \left[\sum_{r=1}^{\ell} \|\hat{V}_r - \widetilde{V}_r\|_{op} \|\widetilde{x}_{r-1}\| + \sum_{r=1}^{\ell} \|\chi_r\| \right] \\
&\overset{(i)}{\leq} c_8 L^2 \left[\sum_{r=1}^{\ell} \|\hat{V}_r - \widetilde{V}_r\|_{op} \|\widetilde{x}_{r-1}\| + \ell h \sqrt{p} \right] \\
&\leq c_8 L^2 \left[\sum_{r=1}^{\ell} \|\hat{V}_r - \widetilde{V}_r\|_{op} \left(\|\widetilde{x}_{r-1} - x_{r-1}\| + \|x_{r-1}\| \right) + \ell h \sqrt{p} \right] \\
&\overset{(ii)}{\leq} c_8 L^2 \left[\sum_{r=1}^{\ell} \|\hat{V}_r - \widetilde{V}_r\|_{op} \left(c_1 L^3 \tau + 2 \right) + \ell h \sqrt{p} \right] \\
&\overset{(iii)}{\leq} c_9 L^2 \left[\sum_{r=1}^{\ell} \|\hat{V}_r - \widetilde{V}_r\|_{op} + L\tau \right] \leq 2 c_9 L^3 \tau = c_2 L^3 \tau,
\end{aligned}
$$

where inequality (i) follows since by definition of the Huberized ReLU for any $z \in \mathbb{R}$ we have that $\phi(z) \leq \phi'(z) z \leq \phi(z) + h/2$, therefore

$$
\|\chi_r\|_\infty = \left\| \phi(V_\ell x_{\ell-1}) - \Sigma_\ell V_\ell x_{\ell-1} - \phi(\widetilde{V}_\ell \widetilde{x}_{\ell-1}) + \widetilde{\Sigma}_\ell \widetilde{V}_\ell \widetilde{x}_{\ell-1} \right\|_\infty \leq 2 \cdot \frac{h}{2} = h \tag{100}
$$

which implies that $\|\chi_r\| \leq h\sqrt{p}$. Next (ii) follows by bound on $\|\widetilde{x}_{r-1} - x_{r-1}\|$ due to the inductive hypothesis and because $\|x_{r-1}\|_2 \leq 2$ by Part (a) of Lemma E.1. Finally (iii) follows by assumption

$\tau \leq O(1/L^3)$ and $h < \frac{\tau}{\sqrt{p}}$. This establishes a bound on $\|\hat{x}_\ell - \widetilde{x}_\ell\|$. We can also mirror the logic to bound $\|x_\ell - \widetilde{x}_\ell\|$ and $\|x_\ell - \widetilde{x}_\ell\|$. This completes the induction and the proof of the lemma. ■

Lemma E.15 *Let the event in Lemma E.1 hold and suppose that the conditions on h, p and τ described in that lemma hold with the additional assumptions that, for a sufficient small constant $c > 0$, $\tau \leq \frac{c}{L^{12}\log^{\frac{3}{2}}(p)}$, and $h \leq \frac{\tau}{\sqrt{p}}$. Let $V^{(1)}$ be the initial weight matrix and $\widetilde{V}$, $\hat{V}$ be weight matrices such that $\|\widetilde{V}_\ell - V_\ell^{(1)}\|_{op}, \|\hat{V}_\ell - V_\ell^{(1)}\|_{op} \leq \tau$ for all $\ell \in [L]$. Also let $\bar{\Sigma}_{\ell,s}$ be $O(pL^2\tau^{2/3})$-sparse diagonal matrices with entries in $[-1,1]$ for all $\ell \in [L]$ and $s \in [n]$. Then*

$$\left\| \widetilde{V}_{L+1} \prod_{r=L}^{\ell} \left(\Sigma_{r,s}^{\widetilde{V}} + \bar{\Sigma}_{r,s} \right) \widetilde{V}_r - \hat{V}_{L+1} \prod_{r=L}^{\ell} \Sigma_{r,s}^{\hat{V}} \hat{V}_r \right\|_{op} \leq O\left(\sqrt{p\log(p)} L^4 \tau^{1/3} \right).$$

for all $\ell \in [L]$ and all $s \in [n]$.

Proof We want to bound the operator norm of

$$\begin{aligned}
&\widetilde{V}_{L+1} \prod_{r=L}^{\ell} \left(\Sigma_{r,s}^{\widetilde{V}} + \bar{\Sigma}_{r,s} \right) \widetilde{V}_r - \hat{V}_{L+1} \prod_{r=L}^{\ell} \Sigma_{r,s}^{\hat{V}} \hat{V}_r \\
&= \underbrace{\widetilde{V}_{L+1} \prod_{r=L}^{\ell} \left(\Sigma_{r,s}^{\widetilde{V}} + \bar{\Sigma}_{r,s} \right) \widetilde{V}_r - V_{L+1}^{(1)} \prod_{r=L}^{\ell} \Sigma_{r,s}^{V^{(1)}} V_r^{(1)}}_{=:\chi_1} + \underbrace{V_{L+1}^{(1)} \prod_{r=L}^{\ell} \Sigma_{r,s}^{V^{(1)}} V_r^{(1)} - \hat{V}_{L+1} \prod_{r=L}^{\ell} \hat{\Sigma}_{r,s}^{V} \hat{V}_r}_{=:\chi_2}.
\end{aligned} \tag{101}$$

We shall instead bound the operator norm of χ_1 and χ_2. Let us proceed to bound the operator norm of χ_1 (the bound on χ_2 will hold using exactly the same logic). Now to ease notation let us fix a sample index $s \in [n]$ and drop it from all subscripts. Also to simplify notation let us refer to $\Sigma_{r,s}^{\widetilde{V}}$ as $\widetilde{\Sigma}_r$, $\Sigma_{r,s}^{\hat{V}}$ as $\hat{\Sigma}_r$, $\bar{\Sigma}_{r,s}$ as $\bar{\Sigma}_r$, and $\Sigma_{r,s}^{V^{(1)}}$ as Σ_r. We shall also refer to $V^{(1)}$ as simply V.

By assumption the diagonal matrix $\bar{\Sigma}_r$ is $O(pL^2\tau^{2/3})$-sparse with entries in $[-1,1]$. Also the matrix $\Sigma_r - \widetilde{\Sigma}_r$ is $O(pL^2\tau^{2/3})$-sparse by Lemma E.14. Therefore the matrix $\check{\Sigma}_r := \widetilde{\Sigma}_r + \bar{\Sigma}_r - \Sigma_r$ is also $O(pL^2\tau^{2/3})$-sparse and has entries in $[-2,2]$. Thus,

$$\begin{aligned}
\chi_1 &= \widetilde{V}_{L+1} \prod_{r=L}^{\ell} \left(\widetilde{\Sigma}_r + \bar{\Sigma}_r \right) \widetilde{V}_r - V_{L+1} \prod_{r=L}^{\ell} \Sigma_r V_r \\
&= \widetilde{V}_{L+1} \prod_{r=L}^{\ell} \left(\Sigma_r + \check{\Sigma}_r \right) \widetilde{V}_r - V_{L+1} \prod_{r=L}^{\ell} \Sigma_r V_r \\
&= \underbrace{(\widetilde{V}_{L+1} - V_{L+1}) \prod_{r=L}^{\ell} \left(\Sigma_r + \check{\Sigma}_r \right) \widetilde{V}_r}_{=:\spadesuit} \\
&\quad + \underbrace{V_{L+1} \left(\prod_{r=L}^{\ell} \left(\Sigma_r + \check{\Sigma}_r \right) \widetilde{V}_r - \prod_{r=L}^{\ell} \Sigma_r V_r \right)}_{=:\clubsuit}.
\end{aligned} \tag{102}$$

The operator norm of ♠ is easy to bound by invoking Lemma E.13

$$\|\spadesuit\|_{op} \leq \|\widetilde{V}_{L+1} - V_{L+1}\|_{op} \left\| \prod_{r=L}^{\ell} \left(\Sigma_r + \check{\Sigma}_r\right) \widetilde{V}_r \right\|_{op} \leq O(\tau L^2). \tag{103}$$

To bound the operator norm of ♣ we will decompose the difference of the products of matrices terms into a sum. Each term in this sum corresponds to either a flip from V_r to $\tilde{V}_r$ or from Σ_r to $\Sigma_r + \check{\Sigma}_r$. That is,

$$\begin{aligned}
\clubsuit &= -\left(V_{L+1} \prod_{r=L}^{\ell} \Sigma_r V_r - V_{L+1} \prod_{r=L}^{\ell} \left(\Sigma_r + \check{\Sigma}_r\right) \widetilde{V}_r \right) \\
&= -\left[\sum_{q=L}^{\ell} \underbrace{V_{L+1} \overbrace{\left(\prod_{r=L}^{q+1} \left(\Sigma_r V_r\right) \right)}^{=:\omega_{1,q}} \left(\check{\Sigma}_q\right) \left(\widetilde{V}_q \overbrace{\prod_{r=q-1}^{\ell} \left(\Sigma_r + \check{\Sigma}_r\right) \widetilde{V}_r}^{=:\omega_{2,q}} \right)}_{=:\blacklozenge_q} \right] \\
&\quad - \left[\sum_{q=L}^{\ell} \underbrace{V_{L+1} \overbrace{\left(\prod_{r=L}^{q+1} \left(\Sigma_r V_r\right) \right)}^{=:\omega_{3,q}} \Sigma_q \left(V_q - \widetilde{V}_q\right) \overbrace{\prod_{r=q-1}^{\ell} \left(\Sigma_r + \check{\Sigma}_r\right) \widetilde{V}_r}^{=:\omega_{4,q}}}_{=:\blacktriangledown_q} \right]
\end{aligned} \tag{104}$$

where in the previous equality above, the indices in the products "count down", so that cases in which $q = L$ include "empty products", and we adopt the convention that, in such cases,

$$\omega_{1,q} = \omega_{3,q} = I,$$

and when $q = \ell$

$$\omega_{2,q} = \omega_{4,q} = I.$$

We begin by bounding the operator norm of $\blacklozenge_q$ (for a q that is not ℓ or L, the exact same bound follows in these boundary cases):

$$\begin{aligned}
\|\blacklozenge_q\|_{op} &= \left\| V_{L+1} \prod_{r=L}^{q-1} (\Sigma_r V_r) \left(\check{\Sigma}_q\right) \left(\widetilde{V}_q \prod_{r=q+1}^{\ell} \left(\Sigma_r + \check{\Sigma}_{r,s}\right) \widetilde{V}_r \right) \right\|_{op} \\
&\overset{(i)}{=} \left\| V_{L+1} \prod_{r=L}^{q-1} (\Sigma_r V_r) \, \Sigma_q^{0/1} \check{\Sigma}_q \Sigma_q^{0/1} \widetilde{V}_\ell \prod_{r=q+1}^{\ell} \left(\Sigma_r + \check{\Sigma}_r\right) \widetilde{V}_r \right\|_{op} \\
&\leq \left\| V_{L+1} \prod_{r=L}^{q-1} (\Sigma_r V_r) \, \Sigma_q^{0/1} \right\|_{op} \|\check{\Sigma}_q\|_{op} \left\| \Sigma_q^{0/1} \widetilde{V}_\ell \prod_{r=q+1}^{\ell} \left(\Sigma_r + \check{\Sigma}_r\right) \widetilde{V}_r \right\|_{op} \\
&\overset{(ii)}{\leq} 2 \left\| V_{L+1} \prod_{r=L}^{q-1} (\Sigma_r V_r) \, \Sigma_q^{0/1} \right\|_{op} \left\| \Sigma_q^{0/1} \widetilde{V}_\ell \prod_{r=q+1}^{\ell} \left(\Sigma_r + \check{\Sigma}_r\right) \widetilde{V}_r \right\|_{op} \\
&\overset{(iii)}{\leq} O\left(\sqrt{pL^2\tau^{2/3}\log(p)}\right) \left\|\Sigma_q^{0/1}\right\|_{op} \left\| \widetilde{V}_\ell \prod_{r=q+1}^{\ell} \left(\Sigma_r + \check{\Sigma}_r\right) \widetilde{V}_r \right\|_{op} \\
&\overset{(iv)}{\leq} O\left(\sqrt{pL^2\tau^{2/3}\log(p)}\right) \times O(L^2) = O\left(\sqrt{p\log(p)} L^3 \tau^{1/3}\right) \qquad (105)
\end{aligned}$$

where in (i) we define $\Sigma_q^{0/1}$ to be a diagonal matrix with $(\Sigma_q^{0/1})_{jj} := \mathbb{I}\left[(\widetilde{\Sigma}_{q,s})_{jj} \neq 0\right]$. Note that since $\check{\Sigma}_q$ is $O(pL^2\tau^{2/3})$ sparse, therefore $\Sigma_q^{0/1}$ is also $O(pL^2\tau^{2/3})$ sparse. Inequality (ii) follows since the entries of $\check{\Sigma}_{q,s}$ lie between $[-2,2]$, (iii) follows by applying Part (g) of Lemma E.1. Finally, (iv) by applying Lemma E.13 since the matrix $\check{\Sigma}_r$ is $O(pL^2\tau^{2/3})$-sparse and has entries in $[-2,2]$.

To control the operator norm of $\blacktriangledown_q$ (again for a $q \neq \ell$ or L, the exact same bound follows in these boundary cases):

$$\begin{aligned}
\|\blacktriangledown_q\|_{op} &= \left\| V_{L+1} \prod_{r=L}^{q-1} (\Sigma_r V_r) \, \Sigma_q \left(\widetilde{V}_q - V_q\right) \left(\prod_{r=q+1}^{\ell} \left(\Sigma_r + \check{\Sigma}_r\right) \widetilde{V}_r \right) \right\|_{op} \\
&\leq \left\| V_{L+1} \prod_{r=L}^{q-1} (\Sigma_r V_r) \right\|_{op} \|\Sigma_q\|_{op} \left\| \widetilde{V}_q - V_q \right\|_{op} \left\| \prod_{r=q+1}^{\ell} \left(\Sigma_r + \check{\Sigma}_r\right) \widetilde{V}_r \right\|_{op} \\
&\leq 2\tau \left\| V_{L+1} \prod_{r=L}^{q-1} (\Sigma_r V_r) \right\|_{op} \left\| \prod_{r=q+1}^{\ell} \left(\Sigma_r + \check{\Sigma}_r\right) \widetilde{V}_r \right\|_{op} \\
&\overset{(i)}{\leq} O(\tau L^2) \left\| V_{L+1} \prod_{r=L}^{q-1} (\Sigma_r V_r) \right\|_{op} \overset{(ii)}{\leq} O(\sqrt{p}\tau L^3)
\end{aligned}$$

where (i) follows by applying Lemma E.13 and (ii) follows by Part (c) of Lemma E.1.

With these bounds on $\blacklozenge_q$ and $\heartsuit_q$ along with the decomposition in (104) we find that

$$\|\clubsuit\|_{op} \leq L \times \left(O\left(\sqrt{p\log(p)}L^3\tau^{1/3}\right) + O(\sqrt{p}\tau L^3)\right) \leq O\left(\sqrt{p\log(p)}L^4\tau^{1/3}\right).$$

Thus by using this bound on $\|\clubsuit\|_{op}$ along with (102) and (103) we get that

$$\|\chi_1\|_{op} \leq O(\tau L^2) + O\left(\sqrt{p\log(p)}L^4\tau^{1/3}\right) = O\left(\sqrt{p\log(p)}L^4\tau^{1/3}\right).$$

As mentioned above we can also bound $\|\chi_2\|_{op}$ using the exact same logic to get that

$$\|\chi_2\|_{op} \leq O\left(\sqrt{p\log(p)}L^4\tau^{1/3}\right).$$

Thus, the decomposition in (101) along with an application of the triangle inequality proves the claim of the lemma. ■

E.3. The Proof

With these various lemmas in place we are now finally ready to prove Lemma D.7

Lemma D.7 *For any $\delta > 0$, suppose that $\tau = \Omega\left(\frac{\log^2\left(\frac{nL}{\delta}\right)}{p^{\frac{3}{2}}L^3}\right)$ and, for a sufficiently small positive constant c, we have $\tau \leq \frac{c}{L^{12}\log^{\frac{3}{2}}(p)}$, $h \leq \frac{\tau}{\sqrt{p}}$ and $p = \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for some sufficiently large polynomial. Then, with probability at least $1-\delta$ over the random initialization $V^{(1)}$, we have*

(a) $\varepsilon_{\mathsf{app}}(V^{(1)}, \tau) \leq O(\sqrt{p\log(p)}L^5\tau^{4/3})$, *and*

(b) $\Gamma(V^{(1)}, \tau) \leq O(\sqrt{p}L^2)$.

Proof Note that since $\tau = \Omega\left(\frac{\log^2\left(\frac{nL}{\delta}\right)}{p^{\frac{3}{2}}L^3}\right)$ and $\tau \leq \frac{c}{L^{12}\log^{\frac{3}{2}}(p)}$, $h \leq \frac{\tau}{\sqrt{p}} \leq \frac{1}{50\sqrt{p}L}$, and because $p \geq \mathsf{poly}\left(L, \log\left(\frac{n}{\delta}\right)\right)$ for a large enough polynomial all the conditions required to invoke Lemma E.1 are satisfied. Let us assume that the event in Lemma E.1 which occurs with probability at least $1-\delta$ holds in the rest of this proof.

Proof of Part (a): Recall the definition of the approximation error

$$\varepsilon_{\mathsf{app}}(V^{(1)}, \tau) := \sup_{s\in[n]} \sup_{\hat{V},\widetilde{V}\in\mathcal{B}(V^{(1)},\tau)} \left|f_{\hat{V}}(x_s) - f_{\widetilde{V}}(x_s) - \nabla f_{\widetilde{V}}(x_s)\cdot\left(\hat{V}-\widetilde{V}\right)\right|.$$

Fix a $\hat{V}, \widetilde{V} \in \mathcal{B}(V^{(1)}, \tau)$ and a sample $s \in [n]$. To ease notation denote $\Sigma_{\ell,s}^{\hat{V}}$ by $\hat{\Sigma}_\ell$, $\Sigma_{\ell,s}^{\widetilde{V}}$ by $\widetilde{\Sigma}_\ell$, $x_{\ell,s}^{\hat{V}}$ by $\hat{x}_\ell$, $x_{\ell,s}^{\widetilde{V}}$ by $\widetilde{x}_\ell$ and $x_{\ell,s}^{V^{(1)}}$ by $x_{\ell,s}$. We know that $f_{\widetilde{V}}(x_s) = \widetilde{V}_{L+1}\widetilde{x}_L$ and $f_{\hat{V}}(x_s) = \hat{V}_{L+1}\hat{x}_L$. Also since $\nabla_{\hat{V}_{L+1}} f_{\hat{V}}(x_s) = \hat{x}_L$ we have

$$\begin{aligned}
&f_{\hat{V}}(x_s) - f_{\widetilde{V}}(x_s) - \nabla f_{\widetilde{V}}(x_s)\cdot\left(\hat{V}-\widetilde{V}\right) \\
&\quad= \hat{V}_{L+1}\hat{x}_L - \widetilde{V}_{L+1}\widetilde{x}_L - (\hat{V}_{L+1}-\widetilde{V}_{L+1})\cdot\widetilde{x}_L - \sum_{\ell=1}^{L}\nabla_{\widetilde{V}_\ell} f_{\widetilde{V}}(x_s)\cdot(\hat{V}_\ell - \widetilde{V}_\ell) \\
&\quad= \hat{V}_{L+1}(\hat{x}_L - \widetilde{x}_L) - \sum_{\ell=1}^{L}\nabla_{\widetilde{V}_\ell} f_{\widetilde{V}}(x_s)\cdot(\hat{V}_\ell - \widetilde{V}_\ell). \qquad (106)
\end{aligned}$$

By equation (99) from the proof of Lemma E.14 above we can decompose the difference as follows,

$$\hat{x}_L - \widetilde{x}_L = \sum_{\ell=1}^{L-1} \left[\prod_{j=L}^{\ell+1} \left(\hat{\Sigma}_j + \check{\Sigma}_j \right) \hat{V}_j \right] (\hat{\Sigma}_\ell + \check{\Sigma}_\ell)(\hat{V}_\ell - \widetilde{V}_\ell)\widetilde{x}_{\ell-1} + \left(\hat{\Sigma}_L + \check{\Sigma}_L \right) \left(\hat{V}_L - \widetilde{V}_L \right) \widetilde{x}_{\ell-1} + \sum_{\ell=1}^{L-1} \left[\prod_{j=L}^{\ell+1} (\hat{\Sigma}_j + \check{\Sigma}_j)\hat{V}_j \right] \chi_\ell + \chi_L, \tag{107}$$

where the diagonal matrix $\check{\Sigma}_{j,s}$ is $O(pL^4\tau^{2/3})$-sparse and has entries in $[-1, 1]$, and the p-dimensional vectors χ_ℓ have infinity norm at most h (see inequality (100)). Now when $\ell \in [L]$, the formula for the gradient given in (2a), using this formula and because given two matrices A and B, $A \cdot B = \mathrm{Tr}(A^\top B)$ we get

$$\begin{aligned}
\nabla_{\widetilde{V}_\ell} f_{\widetilde{V}}(x_s) \cdot \left(\hat{V}_\ell - \widetilde{V}_\ell \right) &= \mathrm{Tr} \left[\nabla_{\widetilde{V}_\ell} f_{\widetilde{V}}(x_s)^\top \left(\hat{V}_\ell - \widetilde{V}_\ell \right) \right] \\
&= \mathrm{Tr} \left[\left(\left(\widetilde{\Sigma}_\ell \prod_{j=\ell+1}^{L} \widetilde{V}_j^\top \widetilde{\Sigma}_j \right) \widetilde{V}_{L+1}^\top \widetilde{x}_{\ell-1}^\top \right)^\top \left(\hat{V}_\ell - \widetilde{V}_\ell \right) \right] \\
&= \mathrm{Tr} \left[\widetilde{x}_{\ell-1} \widetilde{V}_{L+1} \left(\prod_{j=L}^{\ell+1} \widetilde{\Sigma}_j \widetilde{V}_j \right) \widetilde{\Sigma}_\ell \left(\hat{V}_\ell - \widetilde{V}_\ell \right) \right] \\
&= \widetilde{V}_{L+1} \left(\prod_{j=L}^{\ell+1} \widetilde{\Sigma}_{j,s} \widetilde{V}_j \right) \widetilde{\Sigma}_{\ell,s} \left(\hat{V}_\ell - \widetilde{V}_\ell \right) \widetilde{x}_{\ell-1}.
\end{aligned} \tag{108}$$

Using (106)-(108), and noting that, here $\prod_{j=L}^{\ell+1} A_j$ denotes $A_L A_{L-1} \dots A_{\ell+1}$, i.e. the indices "count down", we find

$$
\begin{aligned}
& f_{\hat{V}}(x_s) - f_{\widetilde{V}}(x_s) - \nabla f_{\widetilde{V}}(x_s) \cdot \left(\hat{V} - \widetilde{V}\right) \\
& \overset{(i)}{=} \sum_{\ell=1}^{L} \left(\hat{V}_{L+1} \left[\prod_{j=L}^{\ell+1} \left(\hat{\Sigma}_j + \check{\Sigma}_j \right) \widetilde{V}_j \right] (\hat{\Sigma}_\ell + \check{\Sigma}_\ell)(\hat{V}_\ell - \widetilde{V}_\ell)\widetilde{x}_{\ell-1} \right. \\
& \qquad \left. - \widetilde{V}_{L+1} \left(\prod_{j=L}^{\ell+1} \widetilde{\Sigma}_j \widetilde{V}_j \right) \widetilde{\Sigma}_\ell \left(\hat{V}_\ell - \widetilde{V}_\ell \right) \widetilde{x}_{\ell-1} \right) + \sum_{\ell=1}^{L} \hat{V}_{L+1} \left[\prod_{j=L}^{\ell+1} (\hat{\Sigma}_j + \check{\Sigma}_j) \hat{V}_j \right] \chi_\ell \\
& = \sum_{\ell=1}^{L} \left(\hat{V}_{L+1} \left[\prod_{j=L}^{\ell+1} \left(\hat{\Sigma}_j + \check{\Sigma}_j \right) \hat{V}_j \right] (\hat{\Sigma}_\ell + \check{\Sigma}_\ell) - \widetilde{V}_{L+1} \left(\prod_{j=L}^{\ell+1} \widetilde{\Sigma}_j \widetilde{V}_j \right) \widetilde{\Sigma}_\ell \right) (\hat{V}_\ell - \widetilde{V}_\ell)\widetilde{x}_{\ell-1} \\
& \qquad + \sum_{\ell=1}^{L} \hat{V}_{L+1} \left[\prod_{j=L}^{\ell+1} (\hat{\Sigma}_j + \check{\Sigma}_j) \hat{V}_j \right] \chi_\ell \\
& = \sum_{\ell=1}^{L} \underbrace{\left(\hat{V}_{L+1} \left[\prod_{j=L}^{\ell+1} \left(\hat{\Sigma}_j + \check{\Sigma}_j \right) \hat{V}_j \right] - \widetilde{V}_{L+1} \left(\prod_{j=L}^{\ell+1} \widetilde{\Sigma}_j \widetilde{V}_j \right) \right) \widetilde{\Sigma}_\ell (\hat{V}_\ell - \widetilde{V}_\ell)\widetilde{x}_{\ell-1}}_{=:\spadesuit_\ell} \\
& \qquad + \sum_{\ell=1}^{L} \underbrace{\hat{V}_{L+1} \left[\prod_{j=L}^{\ell+1} \left(\hat{\Sigma}_j + \check{\Sigma}_j \right) \hat{V}_j \right] \left(\hat{\Sigma}_\ell + \check{\Sigma}_\ell - \widetilde{\Sigma}_\ell \right) (\hat{V}_\ell - \widetilde{V}_\ell)\widetilde{x}_{\ell-1}}_{=:\clubsuit_\ell} \\
& \qquad + \sum_{\ell=1}^{L} \underbrace{\hat{V}_{L+1} \left[\prod_{j=L}^{\ell+1} (\hat{\Sigma}_j + \check{\Sigma}_j) \hat{V}_j \right] \chi_\ell}_{=:\heartsuit_\ell}
\end{aligned}
\tag{109}
$$

where in (i), we adopt the convention that when $\ell = L$, the "empty products" $\prod_{j=L}^{\ell+1} \left(\hat{\Sigma}_j + \check{\Sigma}_j \right) \hat{V}_j$ and $\prod_{j=L}^{\ell+1} \widetilde{\Sigma}_j \widetilde{V}_j$ are interpreted as I. Let us bound the norm of $\spadesuit_\ell$ in the case where $\ell \neq L$ (the

bound in the boundary case when $\ell = L$ follows by exactly the same logic):

$$\begin{aligned}
\|\spadesuit_\ell\| &= \left\| \left(\hat{V}_{L+1} \left[\prod_{j=L}^{\ell+1} \left(\hat{\Sigma}_j + \check{\Sigma}_j \right) \hat{V}_j \right] - \widetilde{V}_{L+1} \left(\prod_{j=L}^{\ell+1} \widetilde{\Sigma}_j \widetilde{V}_j \right) \right) \widetilde{\Sigma}_\ell (\hat{V}_\ell - \widetilde{V}_\ell) \widetilde{x}_{\ell-1} \right\| \\
&\leq \left\| \hat{V}_{L+1} \left[\prod_{j=L}^{\ell+1} \left(\hat{\Sigma}_j + \check{\Sigma}_j \right) \hat{V}_j \right] - \widetilde{V}_{L+1} \left(\prod_{j=L}^{\ell+1} \widetilde{\Sigma}_j \widetilde{V}_j \right) \right\|_{op} \left\| \widetilde{\Sigma}_\ell (\hat{V}_\ell - \widetilde{V}_\ell) \widetilde{x}_{\ell-1} \right\| \\
&\overset{(i)}{\leq} O(\sqrt{p \log(p)} L^4 \tau^{1/3}) \left\| \widetilde{\Sigma}_\ell \right\|_{op} \|\hat{V}_\ell - \widetilde{V}_\ell\| \|\widetilde{x}_{\ell-1}\| \\
&\overset{(ii)}{\leq} O(\sqrt{p \log(p)} L^4 \tau^{4/3}) \left(\|\widetilde{x}_{\ell-1} - x_{\ell-1}\| + \|x_{\ell-1}\| \right) \\
&\overset{(iii)}{\leq} O(\sqrt{p \log(p)} L^4 \tau^{4/3}) \left(2 + O(L^3 \tau) \right) \leq O(\sqrt{p \log(p)} L^4 \tau^{4/3})
\end{aligned} \tag{110}$$

where (i) follows by invoking Lemma E.15, (ii) is because the entries of $\widetilde{\Sigma}_\ell$ lie between 0 and 1 and because $\|\hat{V}_\ell - \widetilde{V}_\ell\| \leq 2\tau$ since both $\hat{V}$ and $\widetilde{V}$ are in $\mathcal{B}(V^{(1)}, \tau)$. Inequality (iii) is because $\|\widetilde{x}_{\ell-1} - x_{\ell-1}\| \leq O(L^3 \tau)$ by Lemma E.14 and $\|x_{\ell-1}\| \leq 2$ by Part (a) of Lemma E.1.

Moving on to $\clubsuit_\ell$ (again consider the case where $\ell \neq L$, the bound in the boundary case when $\ell = L$ follows by exactly the same logic),

$$\begin{aligned}
\|\clubsuit_\ell\| &= \left\| \hat{V}_{L+1} \left[\prod_{j=L}^{\ell+1} \left(\hat{\Sigma}_j + \check{\Sigma}_j \right) \hat{V}_j \right] \left(\hat{\Sigma}_\ell + \check{\Sigma}_\ell - \widetilde{\Sigma}_\ell \right) (\hat{V}_\ell - \widetilde{V}_\ell) \widetilde{x}_{\ell-1} \right\| \\
&\leq \left\| \hat{V}_{L+1} \left[\prod_{j=L}^{\ell+1} \left(\hat{\Sigma}_j + \check{\Sigma}_j \right) \hat{V}_j \right] \right\|_{op} \left\| \hat{\Sigma}_\ell + \check{\Sigma}_\ell - \widetilde{\Sigma}_\ell \right\|_{op} \left\| \hat{V}_\ell - \widetilde{V}_\ell \right\| \|\widetilde{x}_{\ell-1}\| \\
&\overset{(i)}{\leq} O(\sqrt{p} L^2) \times \tau \times \left(2 + O(L^3 \tau) \right) = O(\sqrt{p} \tau L^2)
\end{aligned} \tag{111}$$

where (i) follows by invoking Lemma E.13, since the diagonal matrix $\hat{\Sigma}_\ell + \check{\Sigma}_\ell - \widetilde{\Sigma}_\ell$ have entries between -3 and 3 and by bounding $\|\widetilde{x}_{\ell-1}\|$ as we did above. Finally, we bound the norm of $\blacktriangledown_\ell$ (again in the case where $\ell \neq L$, the bound when $\ell = L$ follows by exactly the same logic)

$$\begin{aligned}
\|\blacktriangledown_\ell\| &= \left\| \hat{V}_{L+1} \prod_{j=L}^{\ell+1} (\hat{\Sigma}_j + \check{\Sigma}_j) \hat{V}_j \chi_\ell \right\| \\
&\leq \left\| \hat{V}_{L+1} \prod_{j=L}^{\ell+1} (\hat{\Sigma}_j + \check{\Sigma}_j) \hat{V}_j \right\| \|\chi_\ell\| \\
&\overset{(i)}{\leq} O(\sqrt{p} L^2) \|\chi_\ell\| \leq O(\sqrt{p} L^2) \sqrt{p} \|\chi_\ell\|_\infty \overset{(ii)}{\leq} O(\sqrt{p} L^2) \sqrt{p} h \overset{(iii)}{\leq} O(\sqrt{p} \tau L^2)
\end{aligned} \tag{112}$$

where (i) is by invoking Lemma E.13, (ii) is due to a bound on the $\|\chi_\ell\|_\infty \leq h$ derived in inequality 100 and (iii) is by the assumption that $h < \frac{\tau}{\sqrt{p}}$. The bounds on the norms of $\spadesuit_\ell, \clubsuit_\ell$ and $\blacktriangledown_\ell$

along with the decomposition in (109) reveals that for any $s \in [n]$, $\hat{V}, \widetilde{V} \in \mathcal{B}(V^{(1)}, \tau)$:

$$\begin{aligned}\left|f_{\hat{V}}(x_s) - f_{\widetilde{V}}(x_s) - \nabla f_{\widetilde{V}}(x_s) \cdot \left(\hat{V} - \widetilde{V}\right)\right| &\le L\left(O(\sqrt{p\log(p)}L^4\tau^{4/3}) + O(\sqrt{p}\tau L^2)\right) \\ &\le O(\sqrt{p\log(p)}L^5\tau^{4/3}).\end{aligned}$$

This completes the proof of the first part.

Proof of Part (b): Recall the definition of $\Gamma(V^{(1)}, \tau)$

$$\Gamma(V^{(1)}, \tau) = \sup_{s\in[n]} \sup_{\ell\in[L+1]} \sup_{V\in\mathcal{B}(V^{(1)},\tau)} \|\nabla_{V_\ell} f_V(x_s)\|.$$

Fix a sample $s \in [n]$. First let us bound the Frobenius norm of the gradient when $\ell \in [L]$. By the formula in (2a) we have

$$\begin{aligned}\|\nabla_{V_\ell} f_V(x_s)\| &= \left\|\left(\Sigma_{\ell,s}^V \prod_{j=\ell+1}^{L}\left(V_j^\top \Sigma_{j,s}^V\right)\right) V_{L+1}^\top x_{\ell-1,s}^{V\top}\right\| \\ &\le \left\|\left(\Sigma_{\ell,s}^V \prod_{j=\ell+1}^{L}\left(V_j^\top \Sigma_{j,s}^V\right)\right) V_{L+1}^\top\right\|_{op} \|x_{\ell-1,s}^{V\top}\| \\ &\le \left\|\Sigma_{\ell,s}^V\right\|_{op} \left\|\prod_{j=\ell+1}^{L} V_j^\top \Sigma_{j,s}^V\right\|_{op} \|V_{L+1}\|\|x_{\ell-1,s}\| \\ &\le \left\|\prod_{j=\ell+1}^{L} V_j^\top \Sigma_{j,s}^V\right\|_{op} \|V_{L+1}\|\|x_{\ell-1,s}\| \qquad \text{(since } \|\Sigma_{\ell,s}^V\|_{op} \le 1\text{)} \\ &\le \left\|\prod_{j=\ell+1}^{L} V_j^\top \Sigma_{j,s}^V\right\|_{op} \left(\|V_{L+1}^{(1)}\| + \|V_{L+1}^{(1)} - V_{L+1}\|\right)\left(\|x_{\ell-1,s}^{V^{(1)}}\| + \|x_{\ell-1,s}^{V} - x_{\ell-1,s}^{V^{(1)}}\|\right) \\ &\overset{(i)}{\le} \left\|\prod_{j=\ell+1}^{L} V_j^\top \Sigma_{j,s}^V\right\|_{op} \left(O(\sqrt{p}) + \tau\right)\left(2 + O(L^3\tau)\right) \\ &\overset{(ii)}{\le} O(\sqrt{p})\left\|\prod_{j=\ell+1}^{L} V_j^\top \Sigma_{j,s}^V\right\|_{op} \overset{(iii)}{\le} O(\sqrt{p}L^2)\end{aligned} \tag{113}$$

where (i) follows since $\|V_{L+1}^{(1)}\| \le O(\sqrt{p})$ by Part (b) of Lemma E.1, $\|V_{L+1}^{(1)} - V_{L+1}\| \le \tau$, $\|x_{\ell-1,s}^{V^{(1)}}\| \le 2$ by Part (a) of Lemma E.1 and $\|x_{\ell-1,s}^{V} - x_{\ell-1,s}^{V^{(1)}}\| \le O(L^3\tau)$ by Lemma E.14. Next (ii) follows since $\tau = O(1/L^3)$. Finally, (iii) follows since the matrix $\Sigma_{j,s}^V - \Sigma_{j,s}^{V^{(1)}}$ is $O(pL^2\tau^{2/3})$ sparse by Lemma E.14, therefore we can apply Lemma E.13 to bound the operator norm of the product of the matrices (since $\tau = O\left(\frac{1}{L^{12}\log^{\frac{3}{2}}(p)}\right)$, that Lemma applies that this level of sparsity).

If $\ell = L+1$, then the gradient at V is $x_{L,s}^V$, therefore

$$\begin{aligned}\sup_{s\in[n]}\sup_{V\in\mathcal{B}(V^{(1)},\tau)}\|\nabla_{V_{L+1}} f_V(x_s)\| &= \sup_{s\in[n]}\sup_{V\in\mathcal{B}(V^{(1)},\tau)}\|x_{L,s}^V\| \\ &\le \sup_{s\in[n]}\sup_{V\in\mathcal{B}(V^{(1)},\tau)}\Big(\|x_{L,s}^{V^{(1)}}\| + \|x_{L,s}^{V^{(1)}} - x_{L,s}^V\|\Big) \\ &\le \sup_{s\in[n]}\sup_{V\in\mathcal{B}(V^{(1)},\tau)}\big(2 + O(L^3\tau)\big) \le O(1),\end{aligned}$$

where above we used the fact that $\|x_{L,s}^{V^{(1)}}\| \le 2$ by Part (a) of Lemma E.1 and $\|x_{L,s}^{V^{(1)}} - x_{L,s}^V\| \le O(L^3\tau)$ by Lemma E.14 along with the fact that $\tau \le O(1/L^3)$. Combining the conclusions in the two cases when $\ell \in [L]$ and $\ell \in [L+1]$ establishes our second claim. ■

Now that we have proved Lemma D.7, the reader can next jump to Appendix D.2.

Appendix F. Probabilistic Tools

For an excellent reference of sub-Gaussian and sub-exponential concentration inequalities we refer the reader to Vershynin (2018). We begin by defining sub-Gaussian and sub-exponential random variables.

Definition F.1 *A random variable θ is sub-Gaussian if*

$$\|\theta\|_{\psi_2} := \inf\big\{t > 0 : \mathbb{E}[\exp(\theta^2/t^2)] < 2\big\}$$

is bounded. Further, $\|\theta\|_{\psi_2}$ is defined to be its sub-Gaussian norm.

Definition F.2 *A random variable θ is said to be sub-exponential if*

$$\|\theta\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}[\exp(|\theta|/t) < 2]\}$$

is bounded. Further, $\|\theta\|_{\psi_1}$ is defined to be its sub-exponential norm.

Next we state a few well-known facts about sub-Gaussian random variables.

Lemma F.3 *(Vershynin, 2018, Lemma 2.7.6) If a random variable θ is sub-Gaussian then θ^2 is sub-exponential with $\|\theta^2\|_{\psi_1} = \|\theta\|_{\psi_2}^2$.*

Lemma F.4 *(Vershynin, 2018, Theorem 5.2.2) If a random variable $\theta \sim \mathcal{N}(0,1)$ and g is a 1-Lipschitz function then $\|g(\theta) - \mathbb{E}[g(\theta)]\|_{\psi_2} \le c$, for some absolute positive constant c.*

Let us state Hoeffding's inequality (see, e.g., Vershynin, 2018, Theorem 2.6.2), a concentration inequality for a sum of independent sub-Gaussian random variables.

Theorem F.5 *For independent mean-zero sub-Gaussian random variables $\theta_1, \ldots, \theta_m$, for every $\eta > 0$, we have*

$$\mathbb{P}\left[\Big|\sum_{i=1}^m \theta_i\Big| \ge \eta\right] \le 2\exp\left(-\frac{c\eta^2}{\sum_{i=1}^m \|\theta_i\|_{\psi_2}^2}\right),$$

where c is a positive absolute constant.

We shall also use Bernstein's inequality (see, e.g., Vershynin, 2018, Theorem 2.8.1) a concentration inequality for a sum of independent sub-exponential random variables.

Theorem F.6 *For independent mean-zero sub-exponential random variables $\theta_1, \ldots, \theta_m$, for every $\eta > 0$, we have*

$$\mathbb{P}\left[\left|\sum_{i=1}^{m} \theta_i\right| \geq \eta\right] \leq 2\exp\left(-c\min\left\{\frac{\eta^2}{\sum_{i=1}^{m}\|\theta_i\|_{\psi_1}^2}, \frac{\eta}{\max_i \|\theta_i\|_{\psi_1}}\right\}\right),$$

where c is a positive absolute constant.

Next is the Gaussian-Lipschitz contraction inequality applied to control the squared norm of a Gaussian random vector (see, e.g., Wainwright, 2019, Example 2.28).

Theorem F.7 *Let $\theta_1, \ldots, \theta_m$ be drawn i.i.d. from $\mathcal{N}(0, \sigma^2)$ then, for every $\eta > 0$, we have*

$$\mathbb{P}\left[\sum_{i=1}^{m} \theta_i^2 \geq \sigma^2 m(1+\eta)^2\right] \leq \exp\left(-cm\eta^2\right),$$

where c is a positive absolute constant.

Let us continue by defining an ε-net with respect to the Euclidean distance.

Definition F.8 *Let $S \subseteq \mathbb{R}^p$. A subset K is called an ε-net of S if every point in S is within a distance ε (in Euclidean distance) of some point in K.*

The following lemma bounds the size of a $1/4$-net of unit vectors in $\mathbb{R}^p$.

Lemma F.9 *Let S be the set of all unit vectors in $\mathbb{R}^p$. Then there exists a $1/4$-net of S of size 9^p.*

Proof Follows immediately by invoking (Vershynin, 2018, Corollary 4.2.13) with $\varepsilon = 1/4$. ■

Here is a bound on the size of a $1/4$-net of k-sparse unit vectors, along with a somewhat stronger property of the net.

Lemma F.10 *Let S be the set of all k-sparse unit vectors in $\mathbb{R}^p$. Then there exists a $1/4$-net N of S of size $\binom{p}{k}9^k$, and a mapping ζ from S to N such that, for all $s \in S$, in addition to $\|s - \zeta(s)\| \leq 1/4$, we have $\|s - \zeta(s)\|_0 \leq k$.*

Proof We construct a $1/4$-net as follows. The number of distinct k-sparse subsets of $[p]$ are $\binom{p}{k}$. Over each of these distinct subsets build a $1/4$-net of unit vectors of size 9^k, this is guaranteed by the preceding lemma. Thus by building a $1/4$-net for each of these subset and taking union of these nets we have built a $1/4$-net of k-sparse unit vectors of size $\binom{p}{k}9^k$ as claimed. ■