Memorization and Optimization in Deep Neural Networks with Minimum Over-parameterization

Simone Bombari[*], **Mohammad Hossein Amani**[†], **Marco Mondelli**[*]

Abstract

The Neural Tangent Kernel (NTK) has emerged as a powerful tool to provide memorization, optimization and generalization guarantees in deep neural networks. A line of work has studied the NTK spectrum for two-layer and deep networks with at least a layer with $\Omega(N)$ neurons, N being the number of training samples. Furthermore, there is increasing evidence suggesting that deep networks with sub-linear layer widths are powerful memorizers and optimizers, as long as the number of parameters exceeds the number of samples. Thus, a natural open question is whether the NTK is well conditioned in such a challenging sub-linear setup. In this paper, we answer this question in the affirmative. Our key technical contribution is a lower bound on the smallest NTK eigenvalue for deep networks with the *minimum possible over-parameterization*: up to logarithmic factors, the number of parameters is $\Omega(N)$ and, hence, the number of neurons is as little as $\Omega(\sqrt{N})$. To showcase the applicability of our NTK bounds, we provide two results concerning memorization capacity and optimization guarantees for gradient descent training.

1 Introduction

Training a neural network is a non-convex problem that exhibits disconnected local minima [8, 57, 70]. Yet, in practice gradient descent (GD) and its variants routinely find solutions with zero training loss [72]. A framework to understand this phenomenon comes from the Neural Tangent Kernel (NTK). This quantity was introduced in [33], where it was proved that, during GD training, the network follows the kernel gradient of the functional cost with respect to the NTK. Furthermore, as the layer widths go large, the NTK converges to a deterministic limit which stays constant during training. Hence, in this infinite-width limit, it suffices that the smallest eigenvalue of the NTK is bounded away from 0 for gradient descent to reach zero loss. Going to finite widths, a recipe to prove GD convergence can be summarized as follows: show that *(i)* the NTK is well conditioned at initialization, and *(ii)* the NTK has not changed significantly by the time GD has reached zero loss (see e.g. [49, 17, 11]). A number of papers have exploited this recipe for networks with progressively smaller over-parameterization: two-layer networks [22, 50, 63, 69, 62], deep networks with polynomially wide layers [5, 21, 73, 74], and deep networks with a single wide layer [47, 43]. Besides optimization, showing that the NTK is well conditioned directly implies a result on memorization capacity [42, 48], and the smallest eigenvalue of the NTK has also been related to generalization [6].

In [48], it is shown that, given N training samples, a single layer with $\Omega(N)$ neurons suffices for the NTK to be well conditioned in networks of arbitrary depth. However, there is increasing evidence that, in the challenging setup in which the layer widths are *sub-linear* in N, neural networks still memorize the training data [14, 71, 66], reach zero loss under GD training in the two-layer setting [61, 50, 42], and in the deep case, GD explores a nicely behaved region of the loss landscape [44]. This is in agreement with simple back-of-the-envelope calculations on CIFAR-10 and ImageNet: CIFAR-10 has $N = 50000$ images and roughly 10^6 parameters suffice to fit random labels [72];

[*]Institute of Science and Technology Austria (ISTA). Emails: {simone.bombari, marco.mondelli}@ist.ac.at.

[†]EPFL, Switzerland. Email: mh.amani1998@gmail.com.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

furthermore, in order to fit random labels to a subset of $1.2 \cdot 10^6$ ImageNet data points, $2.4 \cdot 10^7$ parameters are enough [72]. These numbers suggest that having a number of parameters of the same order as the dataset size is much closer to practice than having a number of neurons of that order. We also note that, by counting degrees of freedom or bounding the VC dimension [10], $\Omega(N)$ parameters (and, therefore, $\Omega(\sqrt{N})$ neurons) are in general necessary to fit N data points. This naturally brings forward the following open question:

Is the NTK well conditioned for deep networks with the minimum possible over-parameterization (i.e., containing $\Omega(N)$ parameters corresponding to $\Omega(\sqrt{N})$ neurons)?

Main contributions. In this paper, *we settle this open question* for a large class of deep networks. We consider *(i)* a smooth activation function, *(ii)* i.i.d. data satisfying Lipschitz concentration (e.g., data with a Gaussian distribution, uniform on the sphere/hypercube, or obtained via a Generative Adversarial Network), *(iii)* a standard initialization of the weights (e.g., He's or LeCun's initialization), and *(iv)* a loose pyramidal topology in which the layer widths can increase by at most a multiplicative constant as the network gets deep. Then, in Theorem 3.1 we show that the NTK is well conditioned under the *minimum possible over-parameterization* requirement: the number of parameters between the last two layers has to be $\tilde{\Omega}(N)$ or, equivalently, the number of neurons has to be $\tilde{\Omega}(\sqrt{N})$, where $\tilde{\Omega}$ includes extra logarithmic factors. We achieve this goal by giving a lower bound on the smallest eigenvalue of the NTK. This lower bound is tight when all the layer widths are of the same order.

Our NTK bounds open the way towards understanding the behavior of deep networks with minimum over-parameterization. In particular, an immediate consequence of the fact that the NTK is well conditioned is a result on the memorization capacity (Corollary 4.1). Furthermore, by suitably choosing the initialization, we provide convergence guarantees for gradient descent training (Theorem 4.2). Finally, we highlight that, in order to obtain our bounds on the smallest NTK eigenvalue, we give a number of tight estimates on the ℓ_2 norms of feature vectors and of their centered counterparts, which may be of independent interest.

Proof ideas. To prove Theorem 3.1, we restrict to the kernel $K_{L-2} = J_{L-2} J_{L-2}^\top$, where J_{L-2} is the Jacobian of the output w.r.t. the parameters between the last two hidden layers. This suffices as K_{L-2} is a lower bound on the NTK in the positive semi-definite (PSD) sense. We note that the i-th row ($i \in \{1, \dots, N\}$) of J_{L-2} is given by the Kronecker product $\otimes$ between the feature vector at layer $L-2$ and the backpropagation term from the same layer. One key technical hurdle is to center J_{L-2}, so that its rows have the form $u \otimes v - \mathbb{E}[u \otimes v]$, where $\mathbb{E}[u] = \mathbb{E}[v] = 0$, and all expectations are taken with respect to the (random) training data. To do so, we perform three steps of centering: *(i)* we center the feature vectors (corresponding to u), *(ii)* we center the backpropagation terms (corresponding to v), and *(iii)* we center again the whole row (corresponding to $u \otimes v$). These centering steps are approximate in the sense that the centered matrix is *not necessarily* a lower bound (in the PSD sense) on the original one. However, we are able to control the operator norm of the difference, and show that it scales slower than the smallest eigenvalue of the centered kernel. At this point, we leverage the structure of the rows of the centered Jacobian to bound their sub-exponential norm via a version of the Hanson-Wright inequality for (weakly) correlated random vectors [1]. Finally, after providing also a tight estimate on the ℓ_2 norms of such rows, we can exploit a result from [2] to lower bound the smallest singular value of a matrix whose rows are independent random vectors with well controlled sub-exponential and ℓ_2 norms.

Existing work bounds the smallest NTK eigenvalue for networks with two layers [61, 42], or deep networks with a layer containing $\Omega(N)$ *neurons* [48]. In particular, [61] also exploits the results [1, 2]. However, the centering of the Jacobian is achieved via a combination of whitening and dropping rows, which appears to be difficult to generalize to deep networks. In contrast, the 3-step centering we described above applies to networks of arbitrary depth L. A different approach is put forward in [42], and it is based on a decomposition of the kernel via spherical harmonics. This technique allows to obtain the exact limit of the smallest NTK eigenvalue, it has been used to analyze random feature models [26, 41, 25, 40] and to obtain generalization bounds for two-layer networks (see again [42]). However, understanding how to carry out such a decomposition in the multi-layer setup is an open problem. Finally, [48] considers the deep case, and it relates the smallest eigenvalue of the NTK to the smallest singular value of a feature matrix. As feature matrices are full rank only when the number of neurons is $\Omega(N)$, this approach is inherently limited to networks with a linear-width layer.

The rest of the paper is organized as follows: Section 2 discusses the problem setup and our model assumptions; Section 3 presents our main result on the smallest NTK eigenvalue and gives a roadmap of the argument; Section 4 provides two applications of our NTK bounds: memorization capacity and gradient descent training; Section 5 discusses additional related work, and Section 6 provides some concluding remarks. The details of the proofs are deferred to the appendices.

2 Preliminaries

Neural network setup. We consider an L-layer neural network with feature maps $f_l : \mathbb{R}^d \to \mathbb{R}^{n_l}$ defined for every $x \in \mathbb{R}^d$ as

$$f_l(x) = \begin{cases} x, & l = 0, \\ \phi(W_l^\top f_{l-1}), & l \in [L-1], \\ W_L^\top f_{L-1}, & l = L. \end{cases} \tag{1}$$

Here, $W_l \in \mathbb{R}^{n_{l-1} \times n_l}$ is the weight matrix at layer l, ϕ is the activation function and, given an integer n, we use the shorthand $[n] = \{1, \ldots, n\}$. We assume that the network has a single output, *i.e.* $n_L = 1$ and $W_L \in \mathbb{R}^{n_{L-1}}$, and for consistency we have $n_0 = d$. Let $g_l : \mathbb{R}^d \to \mathbb{R}^{n_l}$ be the pre-activation feature map so that $f_l(x) = \phi(g_l(x))$ for $l \in [L]$. We define $g_0(x) = f_0(x) = x$. Let $X = [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N \times d}$ be the data matrix containing N samples in $\mathbb{R}^d$, $\theta = [\mathrm{vec}(W_1), \ldots, \mathrm{vec}(W_L)]$ be the vector of the parameters of the network, and $F_L(\theta) = [f_L(x_1), \ldots, f_L(x_N)]^\top$ be the network output. We denote by J the Jacobian of F_L with respect to all the parameters of the network:

$$J = \left[\frac{\partial F_L}{\partial \,\mathrm{vec}(W_1)}, \ldots, \frac{\partial F_L}{\partial \,\mathrm{vec}(W_L)} \right] \in \mathbb{R}^{N \times \sum_{l=1}^{L} n_{l-1} n_l}. \tag{2}$$

Our key object of interest is the *empirical Neural Tangent Kernel (NTK) Gram matrix*, denoted by $K \in \mathbb{R}^{N \times N}$ and defined as:

$$K = JJ^T = \sum_{l=1}^{L} \left[\frac{\partial F_L}{\partial \,\mathrm{vec}(W_l)} \right] \left[\frac{\partial F_L}{\partial \,\mathrm{vec}(W_l)} \right]^\top. \tag{3}$$

In [33], it is shown that, as $n_l \to \infty$ for all $l \in [L-1]$, K converges to a deterministic *limit*, which stays constant during gradient descent training. The focus of this paper is on the *finite-width* behavior of the empirical NTK (3). Quantitative bounds for the NTK convergence rate can be obtained from [7, 15]. However, these bounds lead to a significant over-parameterization requirement, see the discussion at the end of Section 3 in [48]. Here, our main result consists in showing that the NTK is well conditioned for a class of neural networks with the *minimum possible over-parameterization*, i.e., $\Omega(N)$ parameter or, equivalently, $\Omega(\sqrt{N})$ neurons.

Weight and data distribution. We consider the following initialization of the weight matrices: $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/n_{l-1})$ for $l \in [L-1], i \in [n_{l-1}], j \in [n_l]$, where β_l is a numerical constant independent of the layer widths. This covers the popular cases of He's and LeCun's initialization [27, 30, 36]. For the last layer, we assume that $(W_L)_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_L^2)$ for $i \in [n_{L-1}]$. Throughout the paper, we let $(x_1, \ldots, x_N)$ be N i.i.d. samples from the data distribution P_X, which satisfies the conditions below.

Assumption 2.1 (Data scaling)**.** *The data distribution P_X satisfies the following properties:*

1. $\int \|x\|_2 \, dP_X(x) = \Theta(\sqrt{d})$.
2. $\int \|x\|_2^2 \, dP_X(x) = \Theta(d)$.
3. $\int \left\| x - \int x' \, dP_X(x') \right\|_2^2 dP_X(x) = \Omega\left(d\right)$.

Assumption 2.2 (Lipschitz concentration)**.** *The data distribution P_X satisfies the* Lipschitz concentration property. *Namely, there exists an absolute constant $c > 0$ such that, for every Lipschitz continuous function $\varphi : \mathbb{R}^d \to \mathbb{R}$, we have $\mathbb{E}|\varphi(X)| < +\infty$, and for all $t > 0$,*

$$\mathbb{P}\left(\left| \varphi(x) - \int \varphi(x') \, dP_X(x') \right| > t \right) \leq 2e^{-ct^2/\|\varphi\|_{\mathrm{Lip}}^2}.$$

Assumption 2.1 is simply a scaling of the training data points and their centered counterparts. Assumption 2.2 covers a number of important cases, e.g., standard Gaussian distribution [65], uniform distributions on the sphere and on the unit (binary or continuous) hypercube [65], data produced via a Generative Adversarial Network (GAN)[3] [59], and more generally any distribution satisfying the log-Sobolev inequality with a dimension-independent constant. We also remark that Assumption 2.2 is rather common in the related literature [48], or it is even replaced by a stronger requirement (e.g., Gaussian distribution or uniform on the sphere) [42].

Assumption 2.3 (Activation function)**.** *The activation function ϕ satisfies the following properties:*

1. *ϕ is a non-linear (and therefore also non-constant) M-Lipschitz function;*
2. *its derivative ϕ' is a M'-Lipschitz function;*

These requirements are satisfied by common activations, e.g. smoothed ReLU, sigmoid, or $\tanh$.

Assumption 2.4 (Network topology)**.** *The network satisfies a loose pyramidal topology condition, i.e. $n_l = \mathcal{O}(n_{l-1})$, for all $l \in [L-1]$.*

A *strict* pyramidal topology (namely, non-increasing layer widths) has been considered in prior work concerning the loss landscape [45, 46] and gradient descent training [47]. Our Assumption 2.4 requires a *loose* pyramidal topology in the sense that, as we go deep, the layer widths are allowed to increase by a constant multiplicative factor. We also note that the widths of neural networks used in practice are often large in the first layers and then start decreasing [28, 60].

Assumption 2.5 (Over-parameterization)**.** *We have that*

$$N \cdot \log^8 N = o(n_{L-2} n_{L-1}). \tag{4}$$

Furthermore, there exists $\gamma > 0$ such that

$$N^\gamma = \mathcal{O}(n_{L-1}). \tag{5}$$

Condition (4) requires the number of parameters between the last two hidden layers to be linear in the number of samples, up to logarithmic factors. This represents our *key over-parameterization condition*. When all the widths have the same scaling, (4) reduces to $N = \tilde{o}(d^2)$. This is satisfied by several "standard" datasets, such as MNIST ($N = 60000$, $d = 784$), CIFAR-10 ($N = 5 \cdot 10^4$, $d = 3 \cdot 32^2$), and ImageNet ($N = 1.4 \cdot 10^7$, $d = 2 \cdot 10^5$). We also note that, if $N \gg d^2$, the NTK is low-rank and $\lambda_{\min}(K) = 0$. Furthermore, Corollary 4.1 and Theorem 4.2 cannot generally hold when $N \gg d^2$, as there are more data points to fit than parameters to help with the fitting. A milder requirement is possible for classification tasks, see the discussion at the end of Section 5. The second condition (5) is rather mild, as γ can be taken to be arbitrarily small, and it avoids an exponential bottleneck in the last hidden layer.

Notation. The feature matrix at layer l is $F_l = [f_l(x_1), \ldots, f_l(x_N)]^\top \in \mathbb{R}^{N \times n_l}$, and $\Sigma_l(x) = \mathrm{diag}([\phi'(g_{l,j}(x))]_{j=1}^{n_l})$ for $l \in [L-1]$, where $g_{l,j}(x)$ is the pre-activation neuron. Given two matrices $F, B \in \mathbb{R}^{m \times n}$, we denote by $F \circ B$ their Hadamard product, and by $F * B = [(F_{1:} \otimes B_{1:}), \ldots, (F_{m:} \otimes B_{m:})]^T \in \mathbb{R}^{m \times n^2}$ their row-wise Kronecker product (also known as Khatri-Rao product). Given a random vector v, let $\|v\|_{\psi_2}$ and $\|v\|_{\psi_1}$ denote its sub-Gaussian and sub-exponential norm, respectively (see also Appendix A for a detailed definition). Given a matrix A, let $\|A\|_{\mathrm{op}}$ be its operator norm, $\|A\|_F$ its Frobenius norm, $\lambda_{\min}(A)$ its smallest eigenvalue, and $\sigma_{\min}(A)$ its smallest singular value. We denote by $\|\varphi\|_{\mathrm{Lip}}$ the Lipschitz constant of the function φ. All the complexity notations $\Omega(\cdot)$, $\mathcal{O}(\cdot)$, $o(\cdot)$ and $\Theta(\cdot)$ are understood for sufficiently large $N, d, n_1, n_2, \ldots, n_{L-1}$. Tildes on such symbols are meant to neglect logarithmic factors.

3 NTK Bounds with Minimum Over-parameterization

Our main technical contribution on the smallest eigenvalue of the empirical NTK (3) is stated below.

Theorem 3.1 (Smallest NTK eigenvalue under minimum over-parameterization)**.** *Consider an L-layer neural network* (1)*, where the activation function satisfies Assumption 2.3 and the layer widths satisfy Assumptions 2.4 and 2.5. Let $\{x_i\}_{i=1}^N \sim_{\text{i.i.d.}} P_X$, where P_X satisfies the Assumptions 2.1-2.2,*

[3] By applying a Lipschitz map to a standard Gaussian distribution, the map output satisfies Assumption 2.2.

and let K be the empirical NTK Gram matrix (3). *Assume that the weights of the network are initialized as $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/n_{l-1})$ for $l \in [L-1]$, $i \in [n_{l-1}]$, $j \in [n_l]$ and $(W_L)_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_L^2)$ for $i \in [n_{L-1}]$. Then, we have*

$$\lambda_{\min}(K) = \Omega(n_{L-2} n_{L-1}), \tag{6}$$

with probability at least $1 - C\,N e^{-c \log^2 n_{L-1}} - C e^{-c \log^2 N}$ over $(x_i)_{i=1}^N$ and $(W_k)_{k=1}^L$, where c and C are numerical constants. Moreover, we have that

$$\lambda_{\min}(K) = \mathcal{O}(d n_{L-1}), \tag{7}$$

with probability at least $1 - C e^{-c n_{L-1}}$, over $(x_i)_{i=1}^N$ and $(W_k)_{k=1}^L$.

This result implies that the NTK is well conditioned for a class of networks in which the number of parameters $n_{L-2} n_{L-1}$ between the last two hidden layers is linear in the number of data samples N, up to logarithmic factors. This means that the total number of neurons of the network can be as little as $\tilde{\Omega}(\sqrt{N})$, which meets the *minimum possible amount of over-parameterization*. The lower bound (6) and the upper bound (7) match when all the widths (up to layer $L-2$) have the same scaling, i.e., $d = \Theta(n_{L-2})$. Throughout the paper, we do not explicitly track the dependence of our bounds on L, in the sense that the numerical constants c, C may depend on L.

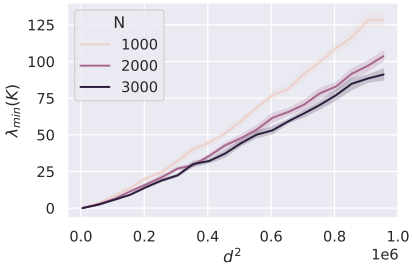

Figure 1: $\lambda_{\min}(K)$ as a function of d^2 in a 3-layer neural network, with sigmoid activation and $d = n_1 = n_2$.

In Figure 1, we consider a 3-layer neural network with $d = n_1 = n_2$, and we plot $\lambda_{\min}(K)$ as a function of d^2, for three different values of N. The inputs are sampled from a standard Gaussian distribution, the activation function is the sigmoid $\sigma(x) = (1 + e^{-x})^{-1}$, and we set $\beta_l = 1$ for all $l \in [L]$. We repeat the experiment 10 times, and report average and confidence interval at 1 standard deviation. The linear scaling of $\lambda_{\min}(K)$ in d^2 is in agreement with the result of Theorem 3.1. The code used to obtain the results of Figure 1 (and Figure 2 as well) is available at `https://github.com/simone-bombari/smallest-eigenvalue-NTK/`.

3.1 Roadmap of the proof of Theorem 3.1

After an application of the chain rule and some standard manipulations, we have

$$J J^\top = \sum_{k=0}^{L-1} K_k, \qquad \text{with } K_k = F_k F_k^\top \circ B_{k+1} B_{k+1}^\top \in \mathbb{R}^{N \times N}, \tag{8}$$

where $B_k \in \mathbb{R}^{N \times n_k}$ is a matrix whose i-th row is given by

$$(B_k)_{i:} = \begin{cases} \Sigma_k(x_i) \left(\prod_{l=k+1}^{L-1} W_l \Sigma_l(x_i) \right) W_L, & k \in [L-2], \\ \Sigma_{L-1}(x_i) W_L, & k = L-1, \\ 1, & k = L. \end{cases} \tag{9}$$

From the decomposition (8), one readily obtains that $K \succeq K_{L-2}$, since K_k is PSD for all $k \in \{0, \dots, L-1\}$. Hence, it suffices to prove a lower bound on $\lambda_{\min}(K_{L-2})$. We denote by J_{L-2} the Jacobian obtained by computing the gradient only over the parameters between layer $L-2$ and layer $L-1$. Then, $K_{L-2} = J_{L-2} J_{L-2}^\top$ and, by using (8) and (9), the i-th row of J_{L-2} can be expressed as

$$(J_{L-2})_{i:} = f_{L-2}(x_i) \otimes (D_L \phi'(W_{L-1}^\top f_{L-2}(x_i))), \qquad i \in [N], \tag{10}$$

where D_L is a diagonal matrix with the entries of W_L on its diagonal. Note that, if we fix the weight matrices $(W_l)_{l=1}^L$, the rows $((J_{L-2})_{i:})_{i=1}^N$ are i.i.d., as the training data $(x_i)_{i=1}^N$ are i.i.d. too.

The proof consists of two main parts. First, we construct the centered Jacobian $\tilde{J}_{L-2}$, which is obtained from J_{L-2} by iteratively removing expectations with respect to X to its parts, and we show that its smallest singular value is *close* to the smallest singular value of J_{L-2}. The details of this part are contained in Appendix D. Second, we bound the smallest eigenvalue of the kernel obtained from

the centered Jacobian $\tilde{J}_{L-2}$ by providing an accurate estimate of the ℓ_2 and sub-exponential norms of its rows. The details of this part are contained in Appendix E. In order to carry out this program, we exploit a number of concentration results on the ℓ_2 norms of feature and backpropagation vectors, and on the ℓ_2 norms of their centered counterparts. These results are contained in Appendix C, and they could be of independent interest. In particular, we provide tight high-probability estimates on *(i)* $\|f_l(x)\|_2$, *(ii)* $\mathbb{E}_x[\|f_l(x)\|_2^2]$, *(iii)* $\mathbb{E}_x[\|f_l(x)\|_2]$, *(iv)* $\mathbb{E}_x[\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2^2]$, $\mathbb{E}_x[\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2]$, and $\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2$, and *(v)* $\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x[D_L\phi'(g_{L-1}(x))]\|_2$. Some preliminary calculations are also contained in Appendix B.

Part 1: Centering. We consider the centered Jacobian $\tilde{J}_{L-2}$, whose i-th row is defined as

$$(\tilde{J}_{L-2})_{i:} = \tilde{f}_{L-2}(x_i) \otimes (D_L\tilde{\phi}'(g_{L-1}(x_i))) - \mathbb{E}_{x_i}\left[\tilde{f}_{L-2}(x_i) \otimes D_L\tilde{\phi}'(g_{L-1}(x_i))\right], \quad i \in [N], \tag{11}$$

where $\tilde{f}_{L-2}(x_i) = \tilde{\phi}(g_{L-2}(x_i))$, with $\tilde{\phi}(g_{L-2}(x_i)) = \phi(g_{L-2}(x_i)) - \mathbb{E}_{x_i}[\phi(g_{L-2}(x_i))]$ and $\tilde{\phi}'(g_{L-1}(x_i)) = \phi'(g_{L-1}(x_i)) - \mathbb{E}_{x_i}[\phi'(g_{L-1}(x_i))]$ being the centered versions of the activation function and of its derivative, respectively. Our strategy is to relate $\lambda_{\min}(J_{L-2}J_{L-2}^\top)$ to $\lambda_{\min}(\tilde{J}_{L-2}\tilde{J}_{L-2}^\top)$ via the following two steps.

Step (a): Centering F_{L-2} and B_{L-1}. We show that $\lambda_{\min}(J_{L-2}J_{L-2}^\top) \geq \lambda_{\min}(\tilde{J}_{FB}\tilde{J}_{FB}^\top) - o(n_{L-2}n_{L-1})$, where $\tilde{J}_{FB}\tilde{J}_{FB}^\top = \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top$ and $\tilde{F}_{L-2} = F_{L-2} - \mathbb{E}_X[F_{L-2}]$ and $\tilde{B}_{L-1} = B_{L-1} - \mathbb{E}_X[B_{L-1}]$ (Lemma D.1 in Appendix D.1). Our strategy differs from existing work (e.g., [61, 48]), where either only the backpropagation term is centered (cf. Proposition 7.1 of [61]) or only the features are centered (cf. Lemma 5.4 of [48]). We remark that, in order to handle deep networks with minimum overparameterization[4], the centering of F_{L-2} and B_{L-1} is crucially performed *at the same time*. In fact, by doing so, certain terms containing an eigenvalue which is negative and large in modulus suitably cancel. As a result, the difference between the original kernel $J_{L-2}J_{L-2}^\top$ and the centered one $\tilde{J}_{FB}\tilde{J}_{FB}^\top$ can be bounded by a matrix whose operator norm is $o(n_{L-2}n_{L-1})$.

Step (b): Centering everything. We show that $\lambda_{\min}(\tilde{J}_{FB}\tilde{J}_{FB}^\top) \geq \lambda_{\min}(\tilde{J}_{L-2}\tilde{J}_{L-2}^\top) - o(n_{L-2}n_{L-1})$, where the rows of $\tilde{J}_{L-2}$ are given by (11) (Lemma D.2 in Appendix D.2). The idea is to decompose the difference $\tilde{J}_{FB}\tilde{J}_{FB}^\top - \tilde{J}_{L-2}\tilde{J}_{L-2}^\top$ into a rank-1 matrix plus a PSD matrix. Then, in order to bound the operator norm of the rank-1 term, we leverage the general version of the Hanson-Wright inequality given by Theorem 2.3 in [1] (see the tail bound on quadratic forms that we provide in Lemma B.5). This strategy avoids whitening and dropping rows (as done in [61]) and, hence, appears to be better suited to the deep case.

The main result of this part is stated below, and it is proved by combining Lemmas D.1 and D.2.

Theorem 3.2 (Jacobian centering)**.** *Consider the setting of Theorem 3.1, and let the rows of J_{L-2} and $\tilde{J}_{L-2}$ be given by* (10) *and* (11)*, respectively. Then, we have*

$$\lambda_{\min}\left(J_{L-2}J_{L-2}^\top\right) \geq \lambda_{\min}\left(\tilde{J}_{L-2}\tilde{J}_{L-2}^\top\right) - o(n_{L-2}n_{L-1}), \tag{12}$$

with probability at least $1 - C\exp(-c\log^2 n_{L-1}) - C\exp(-c\log^2 N)$ *over* $(x_i)_{i=1}^N$ *and* $(W_k)_{k=1}^L$*, where* c, C *are numerical constants.*

Part 2: Bounding the centered Jacobian. If we fix the weight matrices $(W_l)_{l=1}^L$, the rows of $\tilde{J}_{L-2}$ are i.i.d. vectors of the form $u \otimes v - \mathbb{E}[u \otimes v]$, where both u and v are centered. By exploiting this structure, in Appendix E.1 we bound the ℓ_2 and sub-exponential norms of these rows. The ℓ_2 norm is bounded in Lemma E.2, and this result relies on the tight estimates on the ℓ_2 norms of centered features and centered backpropagation terms given by Lemma C.4 and C.5, respectively. The sub-exponential norm is bounded in Lemma E.3, and we use again the tail bound of Lemma B.5, which exploits the version of the Hanson-Wright inequality in [1]. At this point, the problem is reduced to bounding the smallest eigenvalue of a matrix such that its rows are i.i.d. and we have a good control on their ℓ_2 and sub-exponential norms. This goal can be achieved via the following

[4] In contrast, [61] considers shallow networks, and [48] requires the existence of a layer with roughly $\Omega(N)$ neurons.

result, whose proof follows from Theorem 3.2 in [2] and it is presented for completeness in Appendix E.2.

Proposition 3.3. *Let $\tilde{J}$ be a matrix with rows $\tilde{J}_{i:} \in \mathbb{R}^{n_{L-2}n_{L-1}}$, for $i \in [N]$. Assume that $\{\tilde{J}_{i:}\}_{i\in[N]}$ are independent sub-exponential random vectors with $\psi = \max_i \|\tilde{J}_{i:}\|_{\psi_1}$. Let $\eta_{\min} = \min_i \|\tilde{J}_{i:}\|_2$ and $\eta_{\max} = \max_i \|\tilde{J}_{i:}\|_2$. Set $\xi = \psi K + K'$, and $\Delta := C\xi^2 N^{1/4}(n_{L-1}n_{L-2})^{3/4}$. Then, we have that*

$$\lambda_{\min}\left(\tilde{J}\tilde{J}^\top\right) \geq \eta_{\min}^2 - \Delta \tag{13}$$

holds with probability at least

$$1 - \exp\left(-cK\sqrt{N}\log\left(\frac{2n_{L-1}n_{L-2}}{N}\right)\right) - \mathbb{P}\left(\eta_{max} \geq K'\sqrt{n_{L-1}n_{L-2}}\right), \tag{14}$$

where c and C are numerical constants.

The main result of this part is stated below, and it is proved in Appendix E.3.

Theorem 3.4 (Bounding the centered Jacobian)**.** *Consider the setting of Theorem 3.1, and let the rows of $\tilde{J}_{L-2}$ be given by* (11)*. Then, we have*

$$\lambda_{\min}\left(\tilde{J}_{L-2}\tilde{J}_{L-2}^\top\right) = \Omega(n_{L-2}n_{L-1}), \tag{15}$$

with probability at least $1 - Ce^{-c\sqrt{N}} - C\,Ne^{-c\log^2 n_{L-1}}$ over $(x_i)_{i=1}^N$ and $(W_k)_{k=1}^L$, where c and C are numerical constants.

Recall that $K \succeq K_{L-2}$, hence $\lambda_{\min}(K) \geq \lambda_{\min}(K_{L-2}) = \lambda_{\min}\left(J_{L-2}J_{L-2}^\top\right)$. Thus, the lower bound (6) follows by combining the results from Theorem 3.2 and Theorem 3.4.

Finally, the upper bound (7) is rather straightforward. First, we write

$$\lambda_{\min}(K) = \lambda_{\min}\left(JJ^\top\right) \leq (JJ^\top)_{11} = \sum_{l=0}^{L-1} \|(F_l)_{1:}\|_2^2 \|(B_{l+1})_{1:}\|_2^2. \tag{16}$$

Then, by combining the bound on $\|(F_l)_{1:}\|_2^2$ of Lemma C.1 with a direct calculation for $\|(B_{l+1})_{1:}\|_2^2$, the desired result (7) follows. The details are contained in Appendix F.

4 Two Applications: Memorization and Optimization

Memorization capacity. The fact that the NTK Gram matrix (3) is well conditioned readily implies a result on memorization capacity. This was already observed in [42] for two-layer networks and in [48] for deep networks with a layer containing $\Omega(N)$ neurons. Here, by using Theorem 3.1, we show that $\Omega(N)$ parameters between the last pair of hidden layers are enough for the network to fit N real-valued labels up to an arbitrarily small error. The result is stated below and, for completeness, we give the proof in Appendix G.

Corollary 4.1 (Memorization)**.** *Consider an L-layer neural network* (1)*, where the activation function satisfies Assumption 2.3 and the layer widths satisfy Assumptions 2.4 and 2.5. Let $\{x_i\}_{i=1}^N \sim_{\text{i.i.d.}} P_X$, where P_X satisfies the Assumptions 2.1-2.2. Then, it holds that for every $Y \in \mathbb{R}^N$, and for every $\varepsilon > 0$, there exists a set of parameters θ such that*

$$\|F_L(\theta) - Y\|_2 \leq \varepsilon, \tag{17}$$

with probability at least $1 - C\,Ne^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}$ over $(x_i)_{i=1}^N$, where c and C are numerical constants.

Gradient descent training. Theorem 3.1 has implications on the convergence of gradient descent algorithms. In particular, by choosing carefully the initialization, we show that gradient descent trained on the N samples $\{x_i\}_{i=1}^N$ with labels $Y = (Y_1, \ldots, Y_N)$ converges to zero loss. This is – at the best of our knowledge – the *first result of this kind for deep networks with minimum over-parameterization*.

To define our initialization, let us assume that the $(L-1)$-th layer has an even number of neurons. With a slight abuse of notation, we indicate its width as $2n_{L-1}$. Thus, W_{L-1} can be written in the

form $\left[W_{L-1}^{(1)}, W_{L-1}^{(2)}\right]$, where $W_{L-1}^{(1)}, W_{L-1}^{(2)} \in \mathbb{R}^{n_{L-2}\times n_{L-1}}$. Similarly, $W_L \in \mathbb{R}^{2n_{L-1}}$ is the concatenation of $W_L^{(1)}, W_L^{(2)} \in \mathbb{R}^{n_{L-1}}$. Then, we define the initialization $\theta_0 = [\mathrm{vec}(W_{1,0}), \dots, \mathrm{vec}(W_{L,0})]$ as follows:

$$\begin{aligned} (W_{l,0})_{i,j} &\sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/n_{l-1}),\ l \in [L-2],\ i \in [n_{l-1}],\ j \in [n_l], \\ (W_{L-1,0}^{(1)})_{i,j} &\sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_{L-1}^2/n_{L-2}),\ i \in [n_{L-2}],\ j \in [n_{L-1}],\ \text{and } W_{L-1,0}^{(2)} = W_{L-1,0}^{(1)}, \\ (W_{L,0}^{(1)})_i &\sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_L^2\gamma),\ i \in [n_{L-1}],\ \text{and } W_{L,0}^{(2)} = -W_{L,0}^{(1)}. \end{aligned} \tag{18}$$

As usual, $(\beta_l)_{l=1}^L$ are numerical constants independent of the layer widths. In contrast, the quantity γ will be set to a suitably large value depending on the number of training samples and on the layer widths. In words, the initialization is standard for $W_{1,0}, \dots, W_{L-2,0}$ and $W_{L-1,0}^{(1)}$; the variance of $W_{L,0}^{(1)}$ is boosted up by a factor γ; and we duplicate the neurons of layer $L-1$ so that $W_{L-1,0}^{(2)} = W_{L-1,0}^{(1)}$ and $W_{L,0}^{(2)} = -W_{L,0}^{(1)}$. At this point, we are ready to state our optimization result.

Theorem 4.2 (Optimization via gradient descent)**.** *Consider an L-layer neural network* (1)*, where the activation function satisfies Assumption 2.3 and the layer widths satisfy Assumptions 2.4 and 2.5. Consider training data $\{x_i\}_{i=1}^N \sim_{\text{i.i.d.}} P_X$, where P_X satisfies the Assumptions 2.1-2.2, with labels $Y \in \mathbb{R}^N$ such that $\|Y\|_2 = \Theta(\sqrt{N})$. We consider solving the least-squares optimization problem*

$$\min_\theta \mathcal{L}(\theta) := \frac{1}{2}\min_\theta \|F_L(\theta) - Y\|_2^2, \tag{19}$$

by running gradient descent updates of the form $\theta_{t+1} = \theta_t - \eta\nabla\mathcal{L}(\theta_t)$, where the initialization θ_0 is defined in (18) *with $\gamma = d^3N^2$ and $\eta \le C(\gamma N d n_{L-1})^{-1}$. Then, for all $t \ge 1$,*

$$\mathcal{L}(\theta_t) \le \left(1 - c\eta\gamma n_{L-2}n_{L-1}\right)^t \mathcal{L}(\theta_0), \tag{20}$$

with probability at least $1 - C\,Ne^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}$ over $(x_i)_{i=1}^N$ and the initialization θ_0, where c and C are numerical constants.

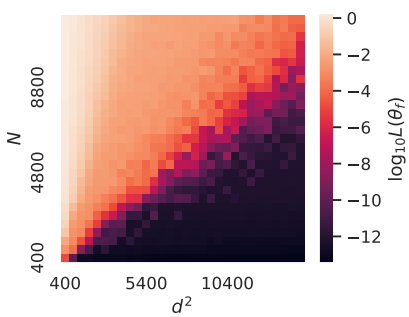

Figure 2: Value of the loss $\mathcal{L}(\theta_f)$ once gradient descent has converged for a 4-layer ReLU network with $d = n_1 = n_2 = n_3$.

In words, (20) guarantees that gradient descent *linearly* converges to a point with zero loss. We also note that (our upper bound on) the convergence rate can be expressed as a simple function of the learning rate η, the number of training samples N, and the layer widths $\{n_i\}_{i=0}^{L-1}$.

In Figure 2, we give an illustrative example that 4-layer networks achieve 0 loss when the number of parameters is at least linear in the number of training samples, i.e., under minimum over-parameterization. To ease the experimental setup, we use a ReLU activation, with Adam optimizer. We initialize the network as in the setting of Theorem 3.1, picking $\beta_l = 1$ for all $l \in [L]$. The inputs, as well as the targets, are sampled from a standard Gaussian distribution. The plot is averaged over 10 independent trials. As predicted by Theorem 4.2, the loss experiences a phase transition and it goes from 1 to 0 when the layer widths are of order $\sqrt{N}$.

We now provide a sketch of the argument of Theorem 4.2, deferring the detailed proof to Appendix H. As a consequence of Theorem 3.1, the NTK is well conditioned at the initialization θ_0, which implies that gradient descent makes progress at the beginning. The idea is that the NTK remains sufficiently well conditioned also during training and, hence, the loss vanishes with the linear rate in (20), since the trained weights remain confined in a ball of sufficiently small radius centered at θ_0. More specifically, we exploit the optimization framework developed in [49]. There, a convergence result is proved if the following condition holds: there exists α, β such that

$$\alpha \le \sigma_{\min}(J(\theta)) \le \|J(\theta)\|_{\text{op}} \le \beta, \quad \text{and} \quad \|J(\theta_1) - J(\theta_2)\|_{\text{op}} \le \frac{\alpha^2}{2\beta}, \tag{21}$$

for all $\theta, \theta_1, \theta_2 \in \mathcal{D} = \mathcal{B}(\theta_0, R)$, where $\mathcal{D}$ denotes a ball centered at θ_0 of radius $R := 4\|F_L(\theta_0) - Y\|_2/\alpha$ (cf. Proposition H.1). In order to control the radius R, we take advantage of the form (18) of the initialization θ_0. In particular, the spectrum of the Jacobian J (and, hence, α) scales linearly in $\sqrt{\gamma}$, and the network output at initialization $F_L(\theta_0)$ is equal to 0 (cf. Lemma H.2). Thus, as $\|Y\|_2 = \Theta(\sqrt{N})$, we have that $R = \Theta(\sqrt{N}/\alpha)$, which can be controlled by choosing suitably γ. At this point, the argument consists in providing a number of estimates on feature vectors, backpropagation terms, and finally the Jacobian $J(\theta)$ for all $\theta \in \mathcal{D}$ (cf. Lemmas H.3-H.9). This allows us to find α, β such that (21) holds and conclude.

5 Related Work and Discussion

Random matrices in deep learning. The limiting spectra of several random matrices related to neural networks have been the subject of a recent line of work. In particular, the Conjugate Kernel (CK) – namely, the Gram matrix of the features from the last hidden layer – has been studied for models with two layers [54, 37, 39, 51], with a bias term [3, 56], and with multiple layers [13]. The Hessian matrix has been considered in [52], and the closely related Fisher information matrix in [55]. Using tools from free probability, the input-output Jacobian (as opposed to the parameter-output Jacobian considered in this work) of deep networks is studied in [53] and the NTK of a two-layer model in [4]. In [23], the spectrum of NTK and CK for deep networks is characterized via an iterated Marchenko-Pastur map. The generalization error has also been studied via the spectrum of suitable random matrices, see [29, 41, 38, 42] and Section 6 of the review [11]. Most of the existing results focus on the linear-width asymptotic regime, where the widths of the various layers are linearly proportional. An exception is [68], which focuses on the ultra-wide two-layer case ($n_1 \gg N$).

Smallest eigenvalue of empirical kernels. In the line of work mentioned above, the limiting spectrum of the random matrix is characterized in terms of weak convergence, which does not lead to a direct implication on the behavior of its smallest eigenvalue. In fact, lower bounding the smallest eigenvalue often requires understanding the speed of convergence to the limit spectrum. Quantitative bounds for random Fourier features are obtained in [9]. In the two-layer setting, the smallest NTK eigenvalue is lower bounded in [61, 42, 68], and concentration bounds can also be obtained from [4, 63, 31]. In particular, [42] establishes a tight bound for two-layer networks with roughly as many parameters as training samples (hence, under minimum over-parameterization), and for a wide class of activations. However, this result still requires that $n_1 = \Omega(d)$. For deep networks, a convergence rate on the NTK can be obtained from [7] and potentially from [15], however these results require all the layers to be wide. In [48], it is proved that a single layer of linear width suffices for the NTK to be well conditioned. Here, Theorem 3.1 provides the first result for deep networks under minimum over-parameterization, therefore allowing for sub-linear layer widths.

Memorization capacity and gradient descent training. For binary classification, the memorization capacity of neural networks was first studied by Cover [18], who solved the single-neuron case. Later, Baum [12] proved that, for two-layer networks, the memorization capacity is lower bounded by the number of parameters, and upper bounds of the same order were given in [35, 58]. Recently, tight results for two-layer networks in the regression setting have been provided [14, 42]. In particular, [14] generalizes the construction of [12], while the memorization result of [42] comes from a bound on the smallest NTK eigenvalue. This same route is also followed to analyze deep networks in [48] and in this work. Results for multiple layers have been provided in [10, 71, 66, 24]. As concerns efficient algorithms achieving memorization, [19, 20] study classification with two-layer networks. A popular line of research has exploited the NTK view to give optimization guarantees on gradient descent. Existing work has focused on the two-layer case [61, 22, 50, 63, 69, 62], and optimal results in terms of over-parameterization have been obtained [42]. For deep networks, [21] assumes a lower bound on the smallest NTK eigenvalue, [5, 73, 74] require all the layer widths to be (rather large) polynomials in the number of samples, and [47, 43] reduce the over-parameterization to a single layer of linear width. In particular, in these last two papers, the analysis of the NTK is reduced to that of a certain feature matrix (the first one in [47] and the last one in [43]). Hence, these approaches do not seem suitable to tackle networks with sub-linear layer widths[5]. Finally, we remark that milder over-paramerization requirements are sufficient for classification problems. More specifically, it has been proved that polylogarithmic width suffices for two-layer [34] and deep [16]

[5] If the width is sub-linear, then the feature matrix is low-rank and, hence, its smallest eigenvalue is 0.

ReLU networks to converge and generalize. In fact, achieving memorization for regression is harder than for classification, where it suffices to satisfy inequality constraints [42].

6 Concluding Remarks

In this paper, we show that the NTK is well conditioned for deep networks with sub-linear layer widths. As consequences of our NTK bounds, we also prove results on memorization and optimization for such a class of networks. Our approach requires smooth activations and layer widths which do not significantly grow in size with the depth. We believe such assumptions to be purely technical, and we leave as an open question their removal. Our novel NTK analysis could potentially be applied beyond standard gradient descent, e.g., for optimization with momentum [67], or federated learning [32].

Acknowledgements

The authors were partially supported by the 2019 Lopez-Loreta prize, and they would like to thank Quynh Nguyen, Mahdi Soltanolkotabi and Adel Javanmard for helpful discussions.

References

[1] Radoslaw Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20:1–13, 2015.

[2] Radoslaw Adamczak, Alexander E Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88, 2011.

[3] Ben Adlam, Jake Levinson, and Jeffrey Pennington. A random matrix perspective on mixtures of nonlinearities for deep learning, 2019. `arXiv:1912.00827`.

[4] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning (ICML)*, 2020.

[5] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning (ICML)*, 2019.

[6] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

[7] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Rousong Wang. On exact computation with an infinitely wide neural net. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[8] P. Auer, M. Herbster, and M. Warmuth. Exponentially many local minima for single neurons. In *Neural Information Processing Systems (NIPS)*, 1996.

[9] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning (ICML)*, 2017.

[10] Peter Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research (JMLR)*, 20(63):1–17, 2019.

[11] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

[12] Eric B. Baum. On the capabilities of multilayer perceptrons. *Journal of Complexity*, 4, 1988.

[13] Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1 – 37, 2021.

[14] Sébastien Bubeck, Ronen Eldan, Yin Tat Lee, and Dan Mikulincer. Network size and weights size for memorization with two-layers neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[15] Sam Buchanan, Dar Gilboa, and John Wright. Deep networks and the multiple manifold problem. In *International Conference on Learning Representations (ICLR)*, 2021.

[16] 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 (ICLR)*, 2020.

[17] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

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

[19] Amit Daniely. Memorizing gaussians with no over-parameterizaion via gradient decent on neural networks, 2020. `arXiv:2003.12895`.

[20] Amit Daniely. Neural networks learning and memorization with (almost) no over-parameterization. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[21] 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 (ICML)*, 2019.

[22] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[23] Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Neural information processing systems (NeurIPS)*, 2020.

[24] Rong Ge, Runzhe Wang, and Haoyu Zhao. Mildly overparametrized neural nets can memorize training data efficiently, 2019. `arXiv:1909.11837`.

[25] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? In *Neural Information Processing Systems (NeurIPS)*, 2020.

[26] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

[27] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Machine Learning (ICML)*, 2010.

[28] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5927–5935, 2017.

[29] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[31] Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[32] Baihe Huang, Xiaoxiao Li, Zhao Song, and Xin Yang. Fl-ntk: A neural tangent kernel-based framework for federated learning analysis. In *International Conference on Machine Learning (ICML)*, 2021.

[33] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2018.

[34] 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 (ICLR)*, 2019.

[35] Adam Kowalczyk. Counting function theorem for multi-layer networks. In *Neural Information Processing Systems (NIPS)*, 1994.

[36] Yann LeCun, Léon Bottou, Genevieve Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[37] Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning (ICML)*, 2018.

[38] Zhenyu Liao, Romain Couillet, and Michael W. Mahoney. A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[39] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[40] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 2021.

[41] 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*, 75(4):667–766, 2022.

[42] Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training, 2020. `arXiv:2007.12826`.

[43] Quynh Nguyen. On the proof of global convergence of gradient descent for deep relu networks with linear widths. In *International Conference on Machine Learning (ICML)*, 2021.

[44] Quynh Nguyen, Pierre Bréchet, and Marco Mondelli. When are solutions connected in deep networks? In *Neural Information Processing Systems (NeurIPS)*, 2021.

[45] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

[46] Quynh Nguyen and Matthias Hein. Optimization landscape and expressivity of deep CNNs. In *International Conference on Machine Learning (ICML)*, 2018.

[47] Quynh Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[48] Quynh Nguyen, Marco Mondelli, and Guido Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks. In *International Conference on Machine Learning (ICML)*, 2021.

[49] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning (ICML)*, 2019.

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

[51] Sandrine Péché. A note on the Pennington-Worah distribution. *Electronic Communications in Probability*, 24:1–7, 2019.

[52] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning (ICML)*, 2017.

[53] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[54] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Neural Information Processing Systems (NeurIPS)*, 2017.

[55] Jeffrey Pennington and Pratik Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. In *Neural Information Processing Systems (NeurIPS)*, 2018.

[56] Vanessa Piccolo and Dominik Schröder. Analysis of one-hidden-layer neural networks via the resolvent method. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[57] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. In *International Conference on Machine Learning (ICML)*, 2018.

[58] Akito Sakurai. nh-1 networks store no less n* h+ 1 examples, but sometimes no more. In *IJCNN International Joint Conference on Neural Networks*, 1992.

[59] Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning representations of GAN-data behave as gaussian mixtures. In *International Conference on Machine Learning (ICML)*, 2020.

[60] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

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

[62] Chaehwan Song, Ali Ramezani-Kebrya, Thomas Pethick, Armin Eftekhari, and Volkan Cevher. Subquadratic overparameterization for shallow neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[63] Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound, 2020. `arXiv:1906.03593`.

[64] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012.

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

[66] Roman Vershynin. Memory capacity of neural networks with threshold and rectified linear unit activations. *SIAM Journal on Mathematics of Data Science*, 2(4):1004–1033, 2020.

[67] Jun-Kun Wang, Chi-Heng Lin, and Jacob D Abernethy. A modular analysis of provable acceleration via Polyak's momentum: Training a wide ReLU network and a deep linear network. In *International Conference on Machine Learning (ICML)*, 2021.

[68] Zhichao Wang and Yizhe Zhu. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks, 2021. `arXiv:2109.09304`.

[69] Xiaoxia Wu, Simon S. Du, and Rachel Ward. Global convergence of adaptive gradient methods for an over-parameterized neural network, 2019. `arXiv:1902.07111`.

[70] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small nonlinearities in activation functions create bad local minima in neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[71] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[72] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires re-thinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

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

[74] Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Assumptions 2.1, 2.2, 2.3, 2.4 and 2.5 in Section 2; see also l. 384-385 in Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Given the theoretical nature of the work, we do not see direct paths to applications with negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All the assumptions are clearly stated either in the description of the model in Section 2 or in the statements of the main results (Theorem 3.1, Corollary 4.1, and Theorem 4.2).
   (b) Did you include complete proofs of all theoretical results? [Yes] A sketch of the proof of Theorem 3.1 is provided in Section 3.1. The complete proof is contained in Appendices B-F: Appendix B contains preliminary calculations; Appendix C contains the high-probability estimates on the ℓ_2 norms of feature and backpropagation vectors; Appendix D contains the complete proofs for **Part 1: Centering**; Appendix E contains the complete proofs for **Part 2: Bounding the centered Jacobian**; and Appendix F contains the proof of the upper bound (7). The proof of Corollary 4.1 is contained in Appendix G. Finally, the proof of Theorem 4.2 is sketched in Section 4 right after the statement, and the full proof is contained in Appendix H.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The folder containing the code and the instructions is in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In Figure 1, we repeat the experiment 10 times, and we report the average together with the confidence interval at 1 standard deviation. In Figure 2, we repeat the experiment 10 times and we report the average of the final loss in the heat-map.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] We did not perform large scale experiments that require to specify this type of details.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We are not using existing assets or curating/releasing new assets.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use crowdsourcing or conducted research with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional Notations

Given a sub-exponential random variable X, let $\|X\|_{\psi_1} = \inf\{t > 0 \ : \ \mathbb{E}[\exp(|X|/t)] \leq 2\}$. Similarly, for a sub-Gaussian random variable, let $\|X\|_{\psi_2} = \inf\{t > 0 \ : \ \mathbb{E}[\exp(X^2/t^2)] \leq 2\}$.

We use the analogous definitions for vectors. In particular, let $X \in \mathbb{R}^n$ be a random vector, then $\|X\|_{\psi_2} := \sup_{\|u\|_2=1} \left\|u^\top X\right\|_{\psi_2}$ and $\|X\|_{\psi_1} := \sup_{\|u\|_2=1} \left\|u^\top X\right\|_{\psi_1}$.

We indicate with C and c absolute, strictly positive, numerical constants, that do not depend on the layer widths of the network $\{n_l\}_{l=0}^{L-1}$ or the number of training samples N. Their value may change from line to line.

B Some Useful Estimates

Lemma B.1. *Under Assumption 2.5, we have that*

$$N \cdot \log^8 n_{L-1} = o(n_{L-1} n_{L-2}), \tag{22}$$

$$n_{L-2} \cdot \log^2 N \cdot \log^2 n_{L-1} = o(n_{L-1} n_{L-2}), \tag{23}$$

$$n_{L-1} \cdot \log^2 N \cdot \log^2 n_{L-1} = o(n_{L-1} n_{L-2}), \tag{24}$$

$$N \cdot \log^2 n_{L-1} \cdot \log^4 N = o(n_{L-1} n_{L-2}). \tag{25}$$

Proof. We start by proving (22). If $n_{L-1} = \mathcal{O}\left(N^2\right)$, then

$$N \cdot \log^8 n_{L-1} = \mathcal{O}\left(N \cdot \log^8 N\right) = o(n_{L-1} n_{L-2}), \tag{26}$$

where the last passage follows from (4). Conversely, if $n_{L-1} = \Omega(N^2)$, then

$$N \cdot \log^8 n_{L-1} = \mathcal{O}\left(\sqrt{n_{L-1}} \cdot \log^8 n_{L-1}\right) = o(n_{L-1}) = o(n_{L-1} n_{L-2}), \tag{27}$$

which concludes the proof of (22).

To obtain (23), we can exploit the second requirement of Assumption 2.5, which implies that $\log N = \mathcal{O}\left(\log n_{L-1}\right)$. This readily implies (23). Notice that (24) naturally follows since $n_{L-1} = \mathcal{O}\left(n_{L-2}\right)$ by Assumption 2.4.

Finally, to obtain (25), we write

$$N \cdot \log^2 n_{L-1} \cdot \log^4 N = \mathcal{O}\left(N \cdot \log^6 n_{L-1}\right) = o(n_{L-1} n_{L-2}), \tag{28}$$

where in the first passage we use that $\log N = \mathcal{O}\left(\log n_{L-1}\right)$ (from the second requirement of Assumption 2.5) and the last passage follows from (22). □

Lemma B.2 (Lipschitz constant of function of the features)**.** *For all $l \in [L-1]$, and for every Lipschitz function φ, we have*

$$\|\varphi(g_l(x))\|_{\text{Lip}} = \mathcal{O}\left(1\right), \tag{29}$$

with probability at least

$$1 - 2l \exp\left(-n_{L-1}\right), \tag{30}$$

over $(W_k)_{k=1}^l$. We recall that φ is applied component-wise to $g_l(x)$, and $\varphi(g_l(x)) : \mathbb{R}^d \to \mathbb{R}^{n_l}$ is intended as a function of x.

Proof. Note that $\varphi(g_l)$ is a composition of Lipschitz functions. Thus,

$$\begin{aligned} \|\varphi(g_l)\|_{\text{Lip}} &\leq \|\varphi\|_{\text{Lip}} \|g_l\|_{\text{Lip}} \leq \|\varphi\|_{\text{Lip}} \|W_l\|_{\text{op}} \prod_{k=1}^{l-1} \left(\|W_k\|_{\text{op}} \|\phi\|_{\text{Lip}}\right) \\ &\leq \|\varphi\|_{\text{Lip}} \|W_l\|_{\text{op}} M^{l-1} \prod_{k=1}^{l-1} \|W_k\|_{\text{op}}, \end{aligned} \tag{31}$$

where the last step is justified by Assumption 2.3.

Recall that, by the assumption on the initialization of the weights, $(W_k)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_k^2/n_{k-1})$, for some constant β_k which does not depend on the layer widths. Then, by Theorem 4.4.5 of [65], we have that, for any $k \in [l]$,

$$\|W_k\|_{\text{op}} \leq C \frac{\beta_k}{\sqrt{n_{k-1}}} (\sqrt{n_{k-1}} + 2\sqrt{n_k}), \tag{32}$$

with probability at least $1 - 2\exp(-n_k)$, C being a numerical constant. By Assumption 2.4 on the topology of the network, we can rewrite this result as

$$\|W_k\|_{\text{op}} = \mathcal{O}(1), \tag{33}$$

with probability at least $1 - 2\exp(-n_{L-1})$. To conclude, using a union bound over the layers up to layer l, we have that

$$\|\varphi(g_l)\|_{\text{Lip}} = \mathcal{O}(1), \tag{34}$$

with probability at least $1 - 2l\exp(-n_{L-1})$ over $(W_k)_{k=1}^l$. □

Lemma B.3. *We have that*

$$\|D_L\|_{\text{op}} \leq \log n_{L-1}, \tag{35}$$

with probability at least $1 - 2\exp(-c\log^2 n_{L-1})$ *over* W_L*, where* c *is a numerical constant.*

Proof. Recall that $D_L = \text{diag}(W_L)$ contains on the diagonal n_{L-1} independent Gaussian random variables $(D_L)_{ii} \sim \mathcal{N}(0, \beta_L^2)$. Thus, for any $i \in [n_{L-1}]$,

$$\mathbb{P}(|(D_L)_{ii}| > \log n_{L-1}) < 2\exp(-\log^2 n_{L-1}/(2\beta_L^2)), \tag{36}$$

which gives

$$\begin{aligned}\mathbb{P}(\|D_L\|_{\text{op}} > \log n_{L-1}) =& \mathbb{P}(\max_{i\in[n_{L-1}]} |(D_L)_{ii}| > \log n_{L-1}) \\ \leq& n_{L-1}\mathbb{P}(|(D_L)_{11}| > \log n_{L-1}) \\ <& 2\exp(\log n_{L-1} - \log^2 n_{L-1}/(2\beta_L^2)) \\ <& 2\exp(-c\log^2 n_{L-1}),\end{aligned} \tag{37}$$

where the second step is a union bound on the entries of D_L. This gives the desired result. □

Lemma B.4. *We have that*

$$\|D_L\phi'(g_{L-1}(x))\|_{\text{Lip}} = \mathcal{O}(\log n_{L-1}), \tag{38}$$

with probability at least $1 - 2\exp(-c\log^2 n_{L-1}) - C\exp(-n_{L-1})$ *over* $(W_k)_{k=1}^L$*. We recall that* ϕ' *is applied component-wise to* $g_l(x)$*,* $D_L\phi'(g_{L-1}(x)) : \mathbb{R}^d \to \mathbb{R}^{n_{L-1}}$ *is intended as a function of* x*, and* c *is a numerical constant.*

Proof. We know by composition of Lipschitz functions that

$$\|D_L\phi'(g_{L-1})\|_{\text{Lip}} \leq \|D_L\|_{\text{op}} \|\phi'(g_{L-1})\|_{\text{Lip}}. \tag{39}$$

By Assumption 2.3, ϕ' is Lipschitz. Hence, by combining Lemma B.2 (where we use $\varphi = \phi'$) and Lemma B.3, the result follows. □

Lemma B.5 (Exponential tails of quadratic forms)**.** *Let* $x \sim P_X$*. Let* $u : \mathbb{R}^d \to \mathbb{R}^{d_u}$ *and* $v : \mathbb{R}^d \to \mathbb{R}^{d_v}$ *be mean-0 Lipschitz functions with respect to* x*, i.e.,* $\mathbb{E}_x[u(x)] = 0$*,* $\mathbb{E}_x[v(x)] = 0$*,* $\|u\|_{\text{Lip}} = c_1$ *and* $\|v\|_{\text{Lip}} = c_2$*. Let* U *be a* $d_u \times d_v$ *matrix, and*

$$\Gamma(x) = u(x)^\top U v(x) - \mathbb{E}_x\left[u(x)^\top U v(x)\right]. \tag{40}$$

Then,

$$\|\Gamma\|_{\psi_1} < CK^2 \|U\|_F. \tag{41}$$

where $K = \sqrt{c_1^2 + c_2^2}$*, and* C *is a numerical constant.*

Proof. Consider the function $z(x) : \mathbb{R}^d \to \mathbb{R}^{d_u+d_v}$ obtained by concatenating the vectors u and v, i.e.,

$$z(x) := [u(x), v(x)]^\top. \tag{42}$$

One can readily verify that $\|z\|_{\text{Lip}} \leq \sqrt{c_1^2 + c_2^2} := K$. Let us set

$$M = \frac{1}{2}\left(\begin{array}{c|c} 0 & U \\ \hline U^\top & 0 \end{array}\right). \tag{43}$$

Then, we have that

$$\Gamma = z^\top M z - \mathbb{E}_x\left[z^\top M z\right]. \tag{44}$$

Since x satisfies Assumption 2.2 and $z(x)$ is Lipschitz, in order to obtain tail bounds on Γ, we can apply the version of the Hanson-Wright inequality given by Theorem 2.3 in [1]:

$$\begin{aligned} \mathbb{P}(|\Gamma| > t) =& \mathbb{P}(|z^\top M z - \mathbb{E}_x\left[z^\top M z\right] | > t) \\ <& 2\exp\left(-\frac{1}{C_1}\min\left(\frac{t^2}{K^4\|M\|_F^2}, \frac{t}{K^2\|M\|_{\text{op}}}\right)\right) \\ \leq& 2\exp\left(-\frac{1}{C_1}\min\left(\frac{t^2}{K^4\|U\|_F^2}, \frac{t}{K^2\|U\|_{\text{op}}}\right)\right), \end{aligned} \tag{45}$$

where C_1 is a numerical constant, and in the last step we use that $\|M\|_{\text{op}} = \|U\|_{\text{op}}$ and $\|M\|_F^2 = \|U\|_F^2/2 \leq \|U\|_F^2$. Thus, by Lemma 5.5 of [61], we conclude that

$$\|\Gamma\|_{\psi_1} < C_2 K^2 \|U\|_F, \tag{46}$$

for some numerical constant C_2, which gives the desired result. □

Lemma B.6. *Let $u \in \mathbb{R}^{d_u}$ and $v \in \mathbb{R}^{d_v}$ be two mean-0 sub-Gaussian vectors such that $\|u\|_{\psi_2} = c_1$ and $\|v\|_{\psi_2} = c_2$. Set $A_{uv} = \mathbb{E}\left[uv^\top\right]$. Then,*

$$\|A_{uv}\|_{\text{op}} \leq C(c_1 + c_2)^2, \tag{47}$$

where C is a numerical constant.

Proof. Consider the vector

$$z := [u, v]^\top. \tag{48}$$

Then, z is sub-Gaussian and, by triangle inequality on the vectors $[u, 0]$ and $[0, v]$, we have that $\|z\|_{\psi_2} \leq c_1 + c_2$. Since u and v are mean-0, then z is also mean-0 and its covariance matrix can be written as $A_z := \mathbb{E}\left[zz^\top\right]$. Furthermore, we can show that

$$\|A_z\|_{\text{op}} \leq C(c_1 + c_2)^2. \tag{49}$$

In fact, let w be the unitary eigenvector associated to the maximum eigenvalue of A_z. Then,

$$\|A_z\|_{\text{op}} = w^\top A_z w = \mathbb{E}\left[w^\top z z^\top w\right] = \mathbb{E}\left[(w^\top z)^2\right]. \tag{50}$$

Furthermore, we have that

$$\|z\|_{\psi_2} := \sup_{w' \text{s.t.} \|w'\|_2 = 1} \left\|(w')^\top z\right\|_{\psi_2} \geq \left\|w^\top z\right\|_{\psi_2} \geq \frac{1}{C}\sqrt{\mathbb{E}\left[(w^\top z)^2\right]}, \tag{51}$$

where C is a numerical constant, and the last inequality comes from Eq. (2.15) of [65]. By combining (50) and (51) with $\|z\|_{\psi_2} \leq c_1 + c_2$, (49) readily follows.

Finally, we have that

$$A_z = \left(\begin{array}{c|c} A_u & A_{uv} \\ \hline A_{uv}^\top & A_v \end{array}\right), \tag{52}$$

where $A_u := \mathbb{E}\left[uu^\top\right]$ and $A_v := \mathbb{E}\left[vv^\top\right]$. As A_u and A_v are PSD, we have that

$$\|A_{uv}\|_{\text{op}} \leq \|A_z\|_{\text{op}}. \tag{53}$$

Hence, the desired result follows from (49) and (53). □

Lemma B.7. *Let A be an $N \times n$ matrix whose rows A_i are i.i.d. mean-0 sub-Gaussian random vectors in $\mathbb{R}^n$. Let $K = \|A_i\|_{\psi_2}$ the sub-Gaussian norm of each row. Then, we have*

$$\left\|AA^\top\right\|_{\text{op}} = K^2 \mathcal{O}\left(N+n\right), \tag{54}$$

with probability at least $1 - 2\exp(-cn)$, for some numerical constant c.

Proof. Without loss of generality, we can assume $K = 1$ to simplify the proof. Let Σ be the second moment matrix of each of the rows of A. Then, $\Sigma = \mathbb{E}\left[A_i A_i^\top\right]$, since the rows are mean-0. Note that, as the rows of A are i.i.d., Σ is independent of i. Furthermore, Lemma B.6 implies that the covariance matrix $\mathbb{E}\left[A_i A_i^\top\right]$ has operator norm bounded by a constant, since the sub-Gaussian norm of the rows is 1. Then, by using Remark 5.40 in [64], we have that

$$\left\|\frac{A^\top A}{N} - \Sigma\right\|_{\text{op}} \leq \max(\delta, \delta^2), \qquad \text{where } \delta = C\sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}, \tag{55}$$

with probability at least $1 - 2\exp(-ct^2)$, where c and C are numerical constants. Setting $t = \sqrt{n}$ and using a triangular inequality gives that, with probability at least $1 - 2\exp(-ct^2)$,

$$\left\|AA^\top\right\|_{\text{op}} = \left\|A^\top A\right\|_{\text{op}} \leq N\left\|\Sigma\right\|_{\text{op}} + \max(C\sqrt{nN} + \sqrt{nN}, (C\sqrt{n} + \sqrt{n})^2) = \mathcal{O}\left(N+n\right), \tag{56}$$

which implies the desired result (after re-scaling by K). □

Lemma B.8. *Let $\tilde{F}_l = F_l - \mathbb{E}_x[F_l] \in \mathbb{R}^{N \times n_l}$ be the centered features matrix at layer l. Then, we have*

$$\left\|\tilde{F}_l \tilde{F}_l^\top\right\|_{\text{op}} = \mathcal{O}\left(N + n_l\right), \tag{57}$$

with probability at least $1 - C\exp\left(-cn_{L-1}\right)$ over $(W_k)_{k=1}^l$ and $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$.

Proof. From Lemma B.2, we have that

$$\left\|f_l(x)\right\|_{\text{Lip}} = \Theta(1), \tag{58}$$

with probability at least

$$1 - C'\exp\left(-n_{L-1}\right), \tag{59}$$

over $(W_k)_{k=1}^l$. We condition on this event in the rest of the proof.

Since $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$ and P_X satisfies Assumption 2.2, all the rows of $\tilde{F}_l$ are mean-0 sub-Gaussian vectors, with sub-Gaussian norm bounded by a numerical constant. Here, we fix $(W_k)_{k=1}^l$ s.t. (58) holds, and the "mean-0" and the "sub-Gaussian norm" is intended w.r.t. the probability space of $(x_i)_{i=1}^N$.

An application of Lemma B.7 gives that

$$\left\|\tilde{F}_l \tilde{F}_l^\top\right\|_{\text{op}} = \mathcal{O}\left(N + n_l\right), \tag{60}$$

with probability at least $1 - 2\exp\left(-cn_l\right)$ over $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$. Taking into account the previous conditioning, we conclude that

$$\left\|\tilde{F}_l \tilde{F}_l^\top\right\|_{\text{op}} = \mathcal{O}\left(N + n_l\right), \tag{61}$$

with probability at least $1 - (C'+2)\exp\left(-cn_{L-1}\right)$ over $(W_k)_{k=1}^l$ and $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$. □

Lemma B.9. *Let $\tilde{B}_{L-1} = B_{L-1} - \mathbb{E}_x[B_{L-1}] \in \mathbb{R}^{N \times n_{L-1}}$ be the centered back-propagation matrix at layer $L-1$. Then, we have*

$$\left\|\tilde{B}_{L-1} \tilde{B}_{L-1}^\top\right\|_{\text{op}} = \mathcal{O}\left(\left(N + n_{L-1}\right)\log^2 n_{L-1}\right), \tag{62}$$

with probability at least $1 - C\exp\left(-cn_{L-1}\right)$ over $(W_k)_{k=l}^L$ and $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$.

Proof. From Lemma B.4, we have that

$$\|D_L\phi'(g_{L-1}(x))\|_{\text{Lip}} = \mathcal{O}(\log n_{L-1}), \tag{63}$$

with probability at least

$$1 - 2\exp(-c\log^2 n_{L-1}) - C\exp(-n_{L-1}), \tag{64}$$

over $(W_k)_{k=1}^L$. We condition on this event in the rest of the proof.

Since $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$ and P_X satisfies Assumption 2.2, all the rows of $\tilde{B}_{L-1}$ are mean-0 sub-Gaussian vectors, with sub-Gaussian norm $\mathcal{O}(\log n_{L-1})$. Here, we fix $(W_k)_{k=1}^L$ s.t. (63) holds, and the "mean-0" and the "sub-Gaussian norm" is intended w.r.t. the probability space of $(x_i)_{i=1}^N$.

An application of Lemma B.7 gives that

$$\left\|\tilde{B}_{L-1}\tilde{B}_{L-1}^\top\right\|_{\text{op}} = \mathcal{O}\left((N + n_{L-1})\log^2 n_{L-1}\right), \tag{65}$$

with probability at least $1 - 2\exp(-cn_{L-1})$ over $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$. Taking into account the previous conditioning, we conclude that

$$\left\|\tilde{B}_{L-1}\tilde{B}_{L-1}^\top\right\|_{\text{op}} = \mathcal{O}\left((N + n_{L-1})\log^2 n_{L-1}\right), \tag{66}$$

with probability at least $1 - (C' + 2)\exp(-cn_{L-1})$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$. □

Lemma B.10. *Let ρ be a standard Gaussian random variable, then we have that*

$$\varphi_1(c) := \mathbb{E}_\rho\left[\phi(c\rho)\right], \tag{67}$$

and

$$\varphi_2(c) := \mathbb{E}_\rho\left[\phi^2(c\rho)\right], \tag{68}$$

are continuous functions in c. Furthermore, $\varphi_1(c)$ is Lipschitz in c, and

$$|\varphi_2(c_1) - \varphi_2(c_2)| \le C_1|c_1 - c_2| + C_2|c_1^2 - c_2^2|, \tag{69}$$

where C_1 and C_2 are numerical constants (independent of c_1, c_2).

Proof. Let $p(\rho) = \frac{1}{\sqrt{2\pi}}e^{-\rho^2/2}$. Then, we have

$$\begin{aligned}
|\varphi_1(c+\varepsilon) - \varphi_1(c)| &\le \int p(\rho)\,|\phi((c+\varepsilon)\rho) - \phi(c\rho)|\,d\rho \\
&\le \int p(\rho)\,|M\varepsilon\rho|\,d\rho \\
&= M\varepsilon\mathbb{E}_\rho\left[|\rho|\right] \\
&= C\varepsilon,
\end{aligned} \tag{70}$$

where in the second line we use that ϕ is M-Lipschitz by Assumption 2.3. Similarly, we have

$$\begin{aligned}
|\varphi_2(c+\varepsilon) - \varphi_2(c)| &\le \int p(\rho)\left|\phi^2((c+\varepsilon)\rho) - \phi^2(c\rho)\right|d\rho \\
&= \int p(\rho)\,|\phi((c+\varepsilon)\rho) - \phi(c\rho)|\,|\phi((c+\varepsilon)\rho) + \phi(c\rho)|\,d\rho \\
&\le \int p(\rho)\,|M\varepsilon\rho|\left(2|\phi(0)| + M(|c+\varepsilon| + |\varepsilon|)|\rho|\right)d\rho \\
&= C_1\varepsilon\mathbb{E}_\rho\left[|\rho|\right] + C_2\varepsilon|c|\mathbb{E}_\rho\left[\rho^2\right] + C_3\varepsilon^2\mathbb{E}_\rho\left[\rho^2\right] \\
&= C_4\varepsilon + C_2|c|\varepsilon + C_3\varepsilon^2 \\
&\le C_5\varepsilon + C_6\left|(c+\varepsilon)^2 - c^2\right|.
\end{aligned} \tag{71}$$

□

Lemma B.11. *Let ρ be a standard Gaussian distribution, and $c \neq 0$ be an absolute constant. Then, we have*

$$|\mathbb{E}_\rho [\phi(c\rho)]| = \mathcal{O}(1). \tag{72}$$

It also holds

$$|\mathbb{E}_\rho [\phi'(c\rho)]| = \mathcal{O}(1). \tag{73}$$

Proof. For the first statement, we exploit the fact that ϕ is Lipschitz:

$$|\mathbb{E}_\rho [\phi(c\rho)]| \leq \mathbb{E}_\rho [|\phi(0)| + M|c\rho|] = |\phi(0)| + M|c|\mathbb{E}_\rho [|\rho|] = C_1. \tag{74}$$

The statement on ϕ' is easily derived following the same proof and using that $\|\phi'\|_{\text{Lip}} \leq M'$. □

Lemma B.12. *Let ρ be a standard Gaussian distribution, and $c \neq 0$ be an absolute constant. Then, we have*

$$\mathbb{E}_\rho \left[\phi^2(c\rho)\right] = \Theta(1), \tag{75}$$

and

$$\mathbb{E}_\rho \left[(\phi'(c\rho))^2\right] = \Theta(1). \tag{76}$$

Proof. For the upper-bound of the first statement, we exploit the fact that ϕ is Lipschitz:

$$\mathbb{E}_\rho \left[\phi^2(c\rho)\right] \leq \mathbb{E}_\rho \left[(|\phi(0)| + M|c\rho|)^2\right] = \phi^2(0) + 2M|c||\phi(0)|\mathbb{E}_\rho [|\rho|] + M^2c^2\mathbb{E}_\rho \left[\rho^2\right] = C_1. \tag{77}$$

For the lower bound, since ϕ is non-zero and continuous, we have that there exist a strictly positive constant $c' > 0$ and an interval $[c_1, c_2]$ with $c_2 > c_1$ such that $\phi^2(x) \geq c'$ for each $x \in [c_1, c_2]$. Therefore, we have

$$\mathbb{E}_\rho \left[\phi^2(c\rho)\right] \geq c'\mathbb{P}(c_1 \leq c\rho \leq c_2) = C_2. \tag{78}$$

The second statements is proved in the same way, as ϕ' is a non-zero Lipschitz function. □

Lemma B.13. *Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ a Lipschitz function, and let $x \sim P_X$. Then,*

$$\mathbb{E}_x^2 [\varphi(x)] \geq \mathbb{E}_x \left[\varphi(x)^2\right] - c\|\varphi\|_{\text{Lip}}^2, \tag{79}$$

where c is a numerical constant.

Proof. We have

$$\begin{aligned}
\mathbb{E}_x^2 [\varphi(x)] &= \mathbb{E}_x \left[(\varphi(x))^2\right] - \mathbb{E}_x \left[(\varphi(x) - \mathbb{E}_x [\varphi(x)])^2\right] \\
&= \mathbb{E}_x \left[\varphi(x)^2\right] - \int_0^{+\infty} \mathbb{P}\left((\varphi(x) - \mathbb{E}_x [\varphi(x)])^2 > t\right) dt \\
&= \mathbb{E}_x \left[\varphi(x)^2\right] - \int_0^{+\infty} \mathbb{P}\left(|\varphi(x) - \mathbb{E}_x [\varphi(x)]| > \sqrt{t}\right) dt \\
&\geq \mathbb{E}_x \left[\varphi(x)^2\right] - \int_0^{+\infty} 2\exp\left(-Ct/\|\varphi\|_{\text{Lip}}^2\right) dt \\
&= \mathbb{E}_x \left[\varphi(x)^2\right] - 2\|\varphi\|_{\text{Lip}}^2 / C,
\end{aligned} \tag{80}$$

where the inequality is a consequence of Assumption 2.2. □

Lemma B.14. *Let $x \sim P_X$, and define $\tilde{c}_l(x) = \beta_l \|f_l(x)\| / \sqrt{n_l}$.*

Then, we have

$$\mathbb{E}_x [|\tilde{c}_l(x) - \mathbb{E}_x [\tilde{c}_l(x)]|] \leq C\frac{\|f_l\|_{\text{Lip}}}{\sqrt{n_l}}, \tag{81}$$

and

$$\mathbb{E}_x \left[(\tilde{c}_l(x) - \mathbb{E}_x [\tilde{c}_l(x)])^2\right] \leq C\frac{\|f_l\|_{\text{Lip}}^2}{n_l}, \tag{82}$$

where C is a numerical constant.

Proof. We have that

$$\begin{aligned}
\mathbb{E}_x\left[|\tilde{c}_l(x) - \mathbb{E}_x\left[\tilde{c}_l(x)\right]|\right] &= \int_0^{+\infty} \mathbb{P}\left(|\tilde{c}_l(x) - \mathbb{E}_x\left[\tilde{c}_l(x)\right]| > t\right) dt \\
&= \int_0^{+\infty} \mathbb{P}\left(|\|f_l(x)\| - \mathbb{E}_x\left[\|f_l(x)\|\right]| > \sqrt{n_l}t/\beta_l\right) dt \\
&\leq \int_0^{+\infty} 2\exp\left(-cn_l t^2/\left\|f_l\right\|_{\mathrm{Lip}}^2\right) dt \\
&= C\frac{\|f_l\|_{\mathrm{Lip}}}{\sqrt{n_l}},
\end{aligned} \tag{83}$$

where c and C are numerical constants, and the third line is justified by Assumption 2.2.

Similarly, we have

$$\begin{aligned}
\mathbb{E}_x\left[\left(\tilde{c}_l(x) - \mathbb{E}_x\left[\tilde{c}_l(x)\right]\right)^2\right] &= \int_0^{+\infty} \mathbb{P}\left(\left(\tilde{c}_l(x) - \mathbb{E}_x\left[\tilde{c}_l(x)\right]\right)^2 > t\right) dt \\
&= \int_0^{+\infty} \mathbb{P}\left(|\|f_l(x)\| - \mathbb{E}_x\left[\|f_l(x)\|\right]| > \sqrt{n_l t}/\beta_l\right) dt \\
&\leq \int_0^{+\infty} 2\exp\left(-cn_l t/\left\|f_l\right\|_{\mathrm{Lip}}^2\right) dt \\
&= C\frac{\|f_l\|_{\mathrm{Lip}}^2}{n_l},
\end{aligned} \tag{84}$$

where, again, c and C are numerical constants and the third line is justified by Assumption 2.2. □

Lemma B.15. *Let ρ_1 and ρ_2 be two standard Gaussian random variables, possibly correlated. Then, we have*

$$\begin{aligned}
&|\mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1 x_1)\phi(\rho_2 x_2) - \phi(\rho_1 y_1)\phi(\rho_2 y_2)\right]| \leq \\
&\qquad \leq C_1\left|x_1 - y_1\right| + C_2\left|x_2\right|\left|x_1 - y_1\right| + C_3\left|x_2 - y_2\right| + C_4\left|y_1\right|\left|x_2 - y_2\right|,
\end{aligned} \tag{85}$$

where C_1, C_2, C_3, C_4 are numerical constants (which do not depend on x_1, x_2, y_1, y_2). Furthermore, the same result holds with ϕ' instead of ϕ.

Proof. We have

$$\begin{aligned}
&|\mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1 x_1)\phi(\rho_2 x_2) - \phi(\rho_1 y_1)\phi(\rho_2 y_2)\right]| \\
&\leq |\mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1 x_1)\phi(\rho_2 x_2) - \phi(\rho_1 y_1)\phi(\rho_2 x_2)\right] + \mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1 y_1)\phi(\rho_2 x_2) - \phi(\rho_1 y_1)\phi(\rho_2 y_2)\right]| \\
&\leq \mathbb{E}_{\rho_1\rho_2}\left[|\phi(\rho_1 x_1) - \phi(\rho_1 y_1)|\,|\phi(\rho_2 x_2)|\right] + \mathbb{E}_{\rho_1\rho_2}\left[|\phi(\rho_2 x_2) - \phi(\rho_2 y_2)|\,|\phi(\rho_1 y_1)|\right] \\
&\leq \mathbb{E}_{\rho_1\rho_2}\left[|M\rho_1(x_1 - y_1)|\left(|\phi(0)| + M\left|\rho_2 x_2\right|\right)\right] + \mathbb{E}_{\rho_1\rho_2}\left[|M\rho_2(x_2 - y_2)|\left(|\phi(0)| + M\left|\rho_1 y_1\right|\right)\right] \\
&\leq C_1\left|x_1 - y_1\right|\mathbb{E}\left[|\rho_1|\right] + C_2\left|x_2\right|\left|x_1 - y_1\right|\mathbb{E}\left[|\rho_1||\rho_2|\right] \\
&\qquad + C_3\left|x_2 - y_2\right|\mathbb{E}\left[|\rho_2|\right] + C_4\left|y_1\right|\left|x_2 - y_2\right|\mathbb{E}\left[|\rho_1||\rho_2|\right] \\
&\leq C_1\left|x_1 - y_1\right| + C_2\left|x_2\right|\left|x_1 - y_1\right| + C_3\left|x_2 - y_2\right| + C_4\left|y_1\right|\left|x_2 - y_2\right|,
\end{aligned} \tag{86}$$

where in third inequality we use that ϕ is M-Lipschitz, and in the last inequality we use that the quantities $\mathbb{E}\left[|\rho_1|\right]$, $\mathbb{E}\left[|\rho_2|\right]$ and $\mathbb{E}\left[|\rho_1||\rho_2|\right]$ are all smaller than 1 (regardless of the correlation between ρ_1 and ρ_2). Since we only used the fact that ϕ is M-Lipschitz, the same result holds with ϕ' in place of ϕ.

□

C Concentration of ℓ_2 Norms

In this appendix, we state and prove a number of high-probability estimates on the ℓ_2 norms of feature and backpropagation vectors. More specifically, our results can be summarized as follows:

- Lemma C.1 gives tight bounds on $\|f_l(x)\|_2$, *i.e.* the ℓ_2 norm of the feature vector at layer l. The statement holds with high probability over x and $(W_k)_{k=1}^l$.
- Lemmas C.2 and C.3 give tight bounds on $\mathbb{E}_x\left[\|f_l(x)\|_2^2\right]$ and $\mathbb{E}_x\left[\|f_l(x)\|_2\right]$, respectively. These quantities represent the *expectation* with respect to x of the (squared) ℓ_2 norm of the feature vector at layer l. The statements hold with high probability over $(W_k)_{k=1}^l$.
- Lemma C.4 focuses on the *centered* feature vector $f_l(x) - \mathbb{E}_x\left[f_l(x)\right]$, and it gives tight bounds on *(i)* its expected (w.r.t. x) squared ℓ_2 norm $\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2^2\right]$, *(ii)* its expected (w.r.t. x) ℓ_2 norm $\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2\right]$, and *(iii)* its ℓ_2 norm $\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2$. The first two statements hold with high probability over $(W_k)_{k=1}^l$, and the probability in the last statement is also over x.
- Lemma C.5 focuses on the *centered* backpropagation vector at layer $L-1$, and it gives tight bounds on its ℓ_2 norm $\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2$. This statement holds with high probability over x and $(W_k)_{k=1}^l$.

Throughout this appendix, we always assume that P_X satisfies Assumptions 2.1 and 2.2, and that the layer widths satisfy Assumption 2.4. Furthermore, we use that the activation ϕ and its derivative ϕ' are Lipschitz (see Assumption 2.3).

Lemma C.1 (ℓ_2 norm of features)**.** *Let $x \sim P_X$. Then, for every $0 \le l \le L-1$,*

$$\|f_l(x))\|_2 = \Theta(\sqrt{n_l}), \tag{87}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over x and $(W_k)_{k=1}^l$. As usual, ϕ is applied component-wise on $g_l(x)$, and c and C are numerical constants.

Proof. We prove this by induction over l, and we start with the base case ($l = 0$). Recall that we have defined $f_0(x) := x$. As the ℓ_2 norm is a 1-Lipschitz function, by Assumption 2.2, we have that

$$\mathbb{P}\left(|\|x\|_2 - \mathbb{E}[\|x\|_2]| > t\right) \le 2e^{-ct^2}. \tag{88}$$

Furthermore, Assumption 2.1 implies that $\mathbb{E}[\|x\|_2] = \Theta(\sqrt{d})$, hence setting $t = \mathbb{E}[\|x\|_2]/2$ in (88) proves the desired result for the base case (recalling that $n_{L-1} = \mathcal{O}(d)$ by Assumption 2.4).

By inductive hypothesis, we have

$$\|f_{l-1}(x))\|_2 = \Theta(\sqrt{n_{l-1}}), \tag{89}$$

with probability at least $1 - C\exp(-cn_{L-1})$.

Define $\tilde{c} := \beta_l \|f_{l-1}(x)\|_2 / \sqrt{n_{l-1}}$. From now on, we condition on a realization of x and $(W_k)_{k=1}^{l-1}$ such that $\tilde{c} = \Theta(1)$. By (89), this happens with probability at least $1 - C\exp(-cn_{L-1})$.

To ease the notation, we use the shorthands $f := f_{l-1}(x)$ and $W := W_l$. Then, we can write

$$\|f_l(x)\|_2 = \left\|\phi(W^\top f)\right\|_2 = \sqrt{n_l}\sqrt{\frac{1}{n_l}\sum_{i=1}^{n_l}\phi^2((W^\top)_{i:}f)}. \tag{90}$$

Recall that $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/n_{l-1})$ and that the Gaussian distribution is rotationally invariant. Thus, the RHS of (90) has the same distribution as

$$\sqrt{n_l}\sqrt{\frac{1}{n_l}\sum_{i=1}^{n_l}\phi^2\left(\tilde{c}\rho_i\right)} = \sqrt{n_l}\sqrt{\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}\rho_1\right)\right] + \frac{1}{n_l}\sum_{i=1}^{n_l} Z_i}, \tag{91}$$

where $(\rho_i)_{i=1}^{n_l} \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ and also independent of f, and we have defined the independent, mean-0 random variables

$$Z_i = \phi^2\left(\tilde{c}\rho_i\right) - \mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}\rho_1\right)\right]. \tag{92}$$

Note that, in the definition of Z_i, the randomness comes only from ρ_i, since we are conditioning on $\tilde{c}$.

We have that

$$\|\phi(\tilde{c}\rho_i)\|_{\psi_2} \leq \|\phi(\tilde{c}\rho_i) - \mathbb{E}_{\rho_i}[\phi(\tilde{c}\rho_i)]\|_{\psi_2} + \|\mathbb{E}_{\rho_i}[\phi(\tilde{c}\rho_i)]\|_{\psi_2} \leq C_1 + C_2 = C_3, \tag{93}$$

where the first term is bounded by a constant by Theorem 5.2.2 in [65], and the bound on the second term follows from Lemma B.11. As a consequence, we have

$$\begin{aligned} \|Z_i\|_{\psi_1} =& \left\|\phi^2(\tilde{c}\rho_i) - \mathbb{E}_{\rho_i}\left[\phi^2(\tilde{c}\rho_i)\right]\right\|_{\psi_1} \\ \leq& C_4 \left\|\phi^2(\tilde{c}\rho_i)\right\|_{\psi_1} \\ =& C_4 \|\phi(\tilde{c}\rho_i)\|_{\psi_2}^2 \\ \leq& C_5, \end{aligned} \tag{94}$$

where the inequality in the second line follows from Exercise 2.7.10 of [65], the equality in the third line follows from Lemma 2.7.6 of [65], and the inequality in the last line follows from (93). Hence, the Z_i-s are i.i.d. sub-exponential random variables, with sub-exponential norm bounded by a numerical constant. An application of Bernstein inequality (cf. Corollary 2.8.3. in [65]) gives that

$$\mathbb{P}\left(\left|\frac{1}{n_l}\sum_{i=1}^{n_l} Z_i\right| > t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{C_6^2}, \frac{t}{C_6}\right) n_l\right), \tag{95}$$

where c, C_6 are numerical constants. Furthermore, by Lemma B.12, we have

$$\mathbb{E}_{\rho_1}\left[\phi^2(\tilde{c}\rho_1)\right] = \Theta(1). \tag{96}$$

By setting $t = \mathbb{E}_{\rho_1}\left[\phi^2(\tilde{c}\rho_1)\right]/2$ into (95) and using (91) and (96), we conclude that

$$\left\|\phi(W^\top f)\right\|_2 = \Theta(\sqrt{n_l}), \tag{97}$$

with probability at least $1 - C\exp(-cn_{L-1}) - 2\exp(-cn_l) \geq 1 - C_1\exp(-cn_{L-1})$, for some numerical constant c and C_1, which concludes the proof. □

Lemma C.2 (Expected squared ℓ_2 norm of features)**.** *Let $x \sim P_X$. Then, for every $0 \leq l \leq L-1$,*

$$\mathbb{E}_x\left[\|f_l(x)\|_2^2\right] = \Theta(n_l), \tag{98}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$. As usual, c and C are numerical constants.

Proof. The argument is by induction over l. The base case is a direct consequence of Assumption 2.1, since $f_0(x) = x$.

By inductive hypothesis, we have

$$\mathbb{E}_x\left[\|f_{l-1}(x))\|_2^2\right] = \Theta(n_{l-1}), \tag{99}$$

with probability at least $1 - C\exp(-cn_{L-1})$. Define $\tilde{c}(x) := \beta_l \|f_{l-1}(x)\|_2 / \sqrt{n_{l-1}}$. From now on, we condition on a realization of $(W_k)_{k=1}^{l-1}$ such that $\mathbb{E}_x\left[\tilde{c}^2(x)\right] = \Theta(1)$. By (99), this happens with probability at least $1 - C\exp(-cn_{L-1})$.

To ease the notation, we use the shorthands $f := f_{l-1}(x)$, $W := W_l$ and $w_i = W_{:i}$. Then, we can write

$$\begin{aligned} \mathbb{E}_x\left[\|f_l(x))\|_2^2\right] &= \mathbb{E}_x\left[\left\|\phi(W^\top f)\right\|_2^2\right] \\ &= n_l\left(\frac{1}{n_l}\sum_{i=1}^{n_l}\mathbb{E}_x\left[\phi^2((W^\top)_{i:}f)\right]\right) \\ &= n_l\left(\mathbb{E}_{w_1}\mathbb{E}_x\left[\phi^2(w_1^\top f)\right] + \frac{1}{n_l}\sum_{i=1}^{n_l} Z_i\right), \end{aligned} \tag{100}$$

where we use that the w_i-s are equally distributed and we have defined the independent, mean-0 random variables

$$Z_i = \mathbb{E}_x\left[\phi^2\left(w_i^\top f(x)\right)\right] - \mathbb{E}_{w_i}\mathbb{E}_x\left[\phi^2\left(w_i^\top f(x)\right)\right]. \tag{101}$$

Note that, in the definition of Z_i, the randomness comes only from w_i, since we are conditioning on $(W_k)_{k=1}^{l-1}$.

We have that

$$\begin{aligned}\|Z_i\|_{\psi_1} \leq & \mathbb{E}_x\left[\left\|\phi^2\left(w_i^\top f(x)\right) - \mathbb{E}_{w_i}\left[\phi^2\left(w_i^\top f(x)\right)\right]\right\|_{\psi_1}\right] \\ \leq & \mathbb{E}_x\left[C_1\left\|\phi^2\left(w_i^\top f(x)\right)\right\|_{\psi_1}\right] \\ = & C_1\mathbb{E}_x\left[\left\|\phi\left(w_i^\top f(x)\right)\right\|_{\psi_2}^2\right],\end{aligned} \tag{102}$$

where the first line follows from Jensen's inequality as $\|\cdot\|_{\psi_1}$ is convex, the inequality in the second line follows from Exercise 2.7.10 of [65], and the equality in the third line follows from Lemma 2.7.6 of [65].

Recall that $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/n_{l-1})$ and that the Gaussian distribution is rotationally invariant. Thus, $\phi\left(w_i^\top f(x)\right)$ has the same distribution as $\phi\left(\tilde{c}(x)\rho_i\right)$, where $(\rho_i)_{i=1}^{n_l} \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ and also independent of $\tilde{c}(x)$. We now condition on a realization of x and $(W_k)_{k=1}^{l-1}$ and provide an upper bound on the sub-Gaussian norm $\left\|\phi\left(w_i^\top f(x)\right)\right\|_{\psi_2}$, where the only randomness comes again from w_i (and, hence, from ρ_i). We have that

$$\begin{aligned}\left\|\phi\left(w_i^\top f(x)\right)\right\|_{\psi_2} &= \|\phi\left(\tilde{c}(x)\rho_i\right)\|_{\psi_2} \\ &\leq \|\phi\left(\tilde{c}(x)\rho_i\right) - \mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}(x)\rho_i\right)\right]\|_{\psi_2} + \|\mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}(x)\rho_i\right)\right]\|_{\psi_2} \\ &\leq C_1\tilde{c}(x) + C_2\tilde{c}(x) + C_3 = C_4(\tilde{c}(x)+1).\end{aligned} \tag{103}$$

where the first term in the RHS in the second line is bounded by $C_1\tilde{c}(x)$ by Theorem 5.2.2 in [65], and the second term is bounded by $C_2\tilde{c}(x) + C_3$ by following the same proof of Lemma B.11 as ϕ is Lipschitz. By combining (102) and (103), we get

$$\|Z_i\|_{\psi_1} \leq C_4^2\mathbb{E}_x\left[(\tilde{c}(x)+1)^2\right] \leq C_5, \tag{104}$$

where we use that $\mathbb{E}_x\left[\tilde{c}^2(x)\right] = \Theta(1)$.

Hence, the Z_i-s are i.i.d. sub-exponential random variables, with sub-exponential norm bounded by a numerical constant. An application of Bernstein inequality (cf. Corollary 2.8.3. in [65]) gives that

$$\mathbb{P}\left(\left|\frac{1}{n_l}\sum_{i=1}^{n_l} Z_i\right| > t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{C_5^2}, \frac{t}{C_5}\right)n_l\right), \tag{105}$$

where c, C_5 are numerical constants.

Let us consider the first term in (100):

$$\mathbb{E}_x\mathbb{E}_{w_1}\left[\phi^2(w_1^\top f)\right] = \mathbb{E}_x\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}(x)\rho_1\right)\right], \tag{106}$$

where the equality comes again from the rotational invariance of the Gaussian distribution of w_1. We will show that

$$\mathbb{E}_x\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}(x)\rho_1\right)\right] = \Theta(1), \tag{107}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{l-1}$.

The upper bound in (107) follows from the same passages in (77), as $\mathbb{E}_x[\tilde{c}^2(x)] = \Theta(1)$. We now prove the lower bound. By Lemma C.1, we have that there exist numerical constants $c_2 > c_1 > 0$ such that $\tilde{c}(x) \in [c_1, c_2]$ with probability at least $1 - C\exp(-cn_{L-1})$ over x and $(W_k)_{k=1}^{l-1}$. Hence, with probability at least $1 - 2C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{l-1}$, we have that

$$\mathbb{P}_x(\tilde{c}(x) \in [c_1, c_2]) \geq 1/2, \tag{108}$$

where we use the symbol $\mathbb{P}_x$ to highlight that this last probability is taken over x. Let us condition on a realization of $(W_k)_{k=1}^{l-1}$ s.t. (108) holds. Then, we have

$$\mathbb{E}_x\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}(x)\rho_1\right)\right] \geq \frac{1}{2}\inf_{c\in[c_1,c_2]}\mathbb{E}_{\rho_1}\left[\phi^2\left(c\rho_1\right)\right]. \tag{109}$$

By Lemma B.10, we have that $\varphi(c) = \mathbb{E}_{\rho_1}\left[\phi^2\left(c\rho_1\right)\right]$ is continuous in c. Therefore, by Weierstrass theorem, there exists a strictly positive $c^* \in [c_1, c_2]$ such that $\inf_{c\in[c_1,c_2]} \mathbb{E}_{\rho_1}\left[\phi^2\left(c\rho_1\right)\right] = \mathbb{E}_{\rho_1}\left[\phi^2\left(c^*\rho_1\right)\right]$. Thus,

$$\mathbb{E}_x\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}(x)\rho_1\right)\right] \geq \frac{1}{2}\mathbb{E}_{\rho_1}\left[\phi^2\left(c^*\rho_1\right)\right] = \Theta(1), \tag{110}$$

where the last equality is a consequence of Lemma B.12. This concludes the proof of the lower bound in (107).

By setting $t = \mathbb{E}_{\rho_1}\left[\phi^2\left(c^*\rho_1\right)\right]/4$ into (105) and using (107) and (100), we conclude that

$$\mathbb{E}_x\left[\|f_l(x))\|_2^2\right] = \Theta(n_l), \tag{111}$$

with probability at least $1 - C\exp(-cn_{L-1})$, for some numerical constants C and c, which concludes the proof. □

Lemma C.3 (Expected ℓ_2 norm of features)**.** *Let $x \sim P_X$. Then, for every $0 \leq l \leq L-1$,*

$$\mathbb{E}_x\left[\|f_l(x)\|_2\right] = \Theta(\sqrt{n_l}), \tag{112}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$. As usual, c and C are numerical constants.

Proof. We condition on the events

$$\mathbb{E}_x\left[\|f_l(x)\|_2^2\right] = \Theta(n_l), \tag{113}$$

and

$$\left\|\|f_l(x)\|_2\right\|_{\mathrm{Lip}} = \mathcal{O}\left(1\right), \tag{114}$$

which happen with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$ by Lemma C.2 and B.2.

The upper bound is a direct consequence of Jensen's inequality:

$$\mathbb{E}_x\left[\|f_l(x)\|_2\right] \leq \sqrt{\mathbb{E}_x\left[\|f_l(x)\|_2^2\right]} = \Theta(\sqrt{n_l}). \tag{115}$$

For the lower bound, we use Lemma B.13, and we obtain

$$\mathbb{E}_x\left[\|f_l(x)\|_2\right] \geq \sqrt{\mathbb{E}_x\left[\|f_l(x)\|_2^2\right] - c\left\|\|f_l(x)\|_2\right\|_{\mathrm{Lip}}^2} = \Theta(\sqrt{n_l}). \tag{116}$$

□

Lemma C.4 (ℓ_2 norms of centered features)**.** *Let $x \sim P_X$. Then, for every $0 \leq l \leq L-1$, the following results hold.*

1. *With probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$, we have that*

$$\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2^2\right] = \Theta(n_l). \tag{117}$$

2. *With probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$, we have that*

$$\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2\right] = \Theta(\sqrt{n_l}). \tag{118}$$

3. *With probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$ and x, we have that*

$$\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2 = \Theta(\sqrt{n_l}). \tag{119}$$

Proof. The argument is by induction over l. The base case for (117) follows directly from Assumption 2.1, since $f_0(x) = x$. Since the ℓ_2 norm is a 1-Lipschitz function, from Jensen inequality and Lemma B.13 we readily obtain the base case for (118). Note that $\|\|x - \mathbb{E}_x[x]\|_2\|_{\text{Lip}} \leq 1$. Then, the base case for (119) is a direct consequence of Assumption 2.2 on x and of the base case of (118), and it holds with probability at least $1 - C' \exp(-cd) \geq 1 - C' \exp(-cn_{L-1})$ over x.

By inductive hypothesis, we assume the three statements to be true for layer $l-1$, for $l \in [L-1]$. We will now prove (117) for layer l.

To ease the notation, we use the shorthands $f(x) := f_{l-1}(x)$, $f = \mathbb{E}_x[f(x)]$, $\tilde{f}(x) = f(x) - f$, $W := W_l$ and $w_i = W_{:i}$. We also define $\tilde{c}(x) = \beta_l \|f(x)\| / \sqrt{n_{l-1}}$ and $\tilde{c} = \mathbb{E}_x[\tilde{c}(x)]$.

We condition on the following events in the probability space of $(W_k)_{k=1}^{l-1}$:

(a) $\|f(x)\|_{\text{Lip}} = \mathcal{O}(1)$, which happens with probability at least $1 - C' \exp(-cn_{L-1})$ by Lemma B.2.

(b) $\tilde{c} = \Theta(1)$, which happens with probability at least $1 - C' \exp(-cn_{L-1})$ by Lemma C.3. Notice that by Jensen inequality this also implies $\|f\|_2^2 = \mathcal{O}(n_{l-1})$.

(c) By inductive hypothesis, we have that, with probability at least $1 - C' \exp(-cn_{n-1})$ over x and $(W_k)_{k=1}^{l-1}$,

$$\left\|\tilde{f}(x)\right\|_2^2 = \Theta(n_{l-1}). \tag{120}$$

Hence, with probability at least $1 - 2C' \exp(-cn_{n-1})$ over $(W_k)_{k=1}^{l-1}$, we have that

$$\mathbb{P}_x\left(c_1 n_{l-1} \leq \left\|\tilde{f}(x)\right\|_2^2 \leq c_2 n_{l-1}\right) \geq 1/2, \tag{121}$$

for some numerical constants $c_2 > c_1 > 0$. In (121), we use the symbol $\mathbb{P}_x$ to highlight that this last probability is taken over x. For the rest of the argument, we condition on a realization of $(W_k)_{k=1}^{l-1}$ s.t. (121) holds.

By taking a union bound, the events (a)-(c) happen with probability at least $1 - 4C' \exp(-cn_{n-1})$ over $(W_k)_{k=1}^{l-1}$.

Now, we can write

$$\begin{aligned}
\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2^2\right] &= \mathbb{E}_x\left[\left\|\phi(W^\top f(x)) - \mathbb{E}_x\left[\phi(W^\top f(x))\right]\right\|_2^2\right] \\
&= n_l\left(\frac{1}{n_l}\sum_{i=1}^{n_l}\mathbb{E}_x\left[\left(\phi((W^\top)_{i:}f(x)) - \mathbb{E}_x\left[\phi((W^\top)_{i:}f(x))\right]\right)^2\right]\right) \\
&= n_l\left(\mathbb{E}_{xw_1}\left[\left(\phi(w_1^\top f(x)) - \mathbb{E}_x\left[\phi(w_1^\top f(x))\right]\right)^2\right] + \frac{1}{n_l}\sum_{i=1}^{n_l} Z_i\right),
\end{aligned} \tag{122}$$

where we use that the w_i-s are identically distributed and we have defined the independent, mean-0 random variables

$$Z_i = \mathbb{E}_x\left[\left(\phi(w_i^\top f(x)) - \mathbb{E}_x\left[\phi(w_i^\top f(x))\right]\right)^2\right] - \mathbb{E}_{w_1 x}\left[\left(\phi(w_1^\top f(x)) - \mathbb{E}_x\left[\phi(w_1^\top f(x))\right]\right)^2\right]. \tag{123}$$

As in the proof of Lemma C.2, in the definition of Z_i, the randomness comes only from w_i, since we are conditioning on $(W_k)_{k=1}^{l-1}$.

We have that

$$
\begin{aligned}
\|Z_i\|_{\psi_1} &\leq C_0 \left\| \mathbb{E}_x \left[\left(\phi(w_i^\top f(x)) - \mathbb{E}_x \left[\phi(w_i^\top f(x)) \right] \right)^2 \right] \right\|_{\psi_1} \\
&\leq C_0 \mathbb{E}_x \left[\left\| \left(\phi(w_i^\top f(x)) - \mathbb{E}_x \left[\phi(w_i^\top f(x)) \right] \right)^2 \right\|_{\psi_1} \right] \\
&= C_0 \mathbb{E}_x \left[\left\| \phi(w_i^\top f(x)) - \mathbb{E}_x \left[\phi(w_i^\top f(x)) \right] \right\|_{\psi_2}^2 \right] \\
&\leq C_0 \mathbb{E}_x \left[\left(\left\| \phi(w_i^\top f(x)) \right\|_{\psi_2} + \left\| \mathbb{E}_x \left[\phi(w_i^\top f(x)) \right] \right\|_{\psi_2} \right)^2 \right] \\
&\leq C_0 \mathbb{E}_x \left[\left(\left\| \phi(w_i^\top f(x)) \right\|_{\psi_2} + \mathbb{E}_x \left[\left\| \phi(w_i^\top f(x)) \right\|_{\psi_2} \right] \right)^2 \right],
\end{aligned}
\tag{124}
$$

where C_0 is a numerical constant, the first inequality follows from Exercise 2.7.10 of [65], the second line follows from Jensen's inequality as $\|\cdot\|_{\psi_1}$ is convex, the equality follows from Lemma 2.7.6 of [65], and the last line follows from Jensen's inequality as $\|\cdot\|_{\psi_2}$ is convex.

Recall that $(W)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/n_{L-1})$ and that the Gaussian distribution is rotationally invariant. Thus, $\phi'\left(w_i^\top f(x)\right)$ has the same distribution as $\phi'\left(\tilde{c}(x)\rho_i\right)$, where $(\rho_i)_{i=1}^{n_{L-1}} \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ and also independent of $\tilde{c}(x)$. Therefore,

$$
\begin{aligned}
\left\| \phi\left(w_i^\top f(x)\right) \right\|_{\psi_2} &= \left\| \phi\left(\tilde{c}(x)\rho_i\right) \right\|_{\psi_2} \\
&\leq \left\| \phi\left(\tilde{c}(x)\rho_i\right) - \mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}(x)\rho_i\right)\right] \right\|_{\psi_2} + \left\| \mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}(x)\rho_i\right)\right] \right\|_{\psi_2} \\
&\leq C_1 \tilde{c}(x) + C_2 \tilde{c}(x) + C_3 \leq C_4(\tilde{c}(x)+1),
\end{aligned}
\tag{125}
$$

where the first term in the RHS in the second line is bounded by $C_1\tilde{c}(x)$ for Theorem 5.2.2 in [65], and the second term is bounded by $C_2\tilde{c}(x) + C_3$ by following the same proof of Lemma B.11.

Merging together (125) and (124) we get

$$
\begin{aligned}
\|Z_i\|_{\psi_1} \leq & C_0 \mathbb{E}_x \left[\left(C_4 \tilde{c}(x) + C_4 \tilde{c} + C_5 \right)^2 \right] \\
= & C_6 \mathbb{E}_x [(\tilde{c}(x) - \tilde{c})^2] + C_4 \tilde{c}^2 + C_7 \tilde{c} + C_8 \\
= & C_6 \mathcal{O}\left(n_l^{-1}\right) + C_4 \tilde{c}^2 + C_7 \tilde{c} + C_8 \\
\leq & C_9,
\end{aligned}
\tag{126}
$$

where in the third line we use Lemma B.14.

Hence, the Z_i-s are i.i.d. sub-exponential random variables, with sub-exponential norm bounded by a numerical constant. An application of Bernstein inequality (cf. Corollary 2.8.3. in [65]) gives that

$$
\mathbb{P}\left(\left| \frac{1}{n_l} \sum_{i=1}^{n_l} Z_i \right| > t \right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{C^2}, \frac{t}{C} \right) n_l \right),
\tag{127}
$$

where c, C are numerical constants. We recall that this probability is intended over W_l.

Let's now focus on the first term in the last line of (122), using the shorthand $w = w_1$, to ease the notation. We can rewrite this term as

$$
\begin{aligned}
&\mathbb{E}_w \left[\mathbb{E}_x \left[\phi^2(w^\top f(x)) \right] - \mathbb{E}_{xy} \left[\phi(w^\top f(x)) \phi(w^\top f(y)) \right] \right] \\
&\quad = \mathbb{E}_x \mathbb{E}_w \left[\phi^2(w^\top f(x)) \right] - \mathbb{E}_{xy} \mathbb{E}_w \left[\phi(w^\top f(x)) \phi(w^\top f(y)) \right].
\end{aligned}
\tag{128}
$$

The aim of this part of the proof is to show that the quantity in (128) is $\Theta(1)$.

Recall that $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/n_{l-1})$ and that the Gaussian distribution is rotationally invariant. Thus, $\phi\left(w^\top f(x)\right)$ has the same distribution as $\phi\left(\tilde{c}(x)\rho\right)$, where $\rho \sim \mathcal{N}(0,1)$ is independent of $\tilde{c}(x)$.

We therefore have

$$\begin{aligned}
&\left|\mathbb{E}_x\mathbb{E}_w\left[\phi^2\left(w^\top f(x)\right)-\phi^2\left(w^\top f(x)\frac{\tilde{c}}{\tilde{c}(x)}\right)\right]\right| \\
&=\left|\mathbb{E}_x\mathbb{E}_\rho\left[\phi^2(\rho\tilde{c}(x))-\phi^2(\rho\tilde{c})\right]\right| \\
&\leq \mathbb{E}_x\left[C_1|\tilde{c}-\tilde{c}(x)|+C_2\left|\tilde{c}^2-\tilde{c}(x)^2\right|\right] \\
&\leq \mathbb{E}_x\left[C_1|\tilde{c}-\tilde{c}(x)|+C_2(\tilde{c}-\tilde{c}(x))^2+2C_2\tilde{c}\,|\tilde{c}-\tilde{c}(x)|\right] \\
&=\mathcal{O}\left(n_{l-1}^{-1/2}\right),
\end{aligned} \tag{129}$$

where the third line follows from Lemma B.10, and the last passage follows from Lemma B.14. Similarly, we have

$$\begin{aligned}
&\left|\mathbb{E}_{xy}\mathbb{E}_w\left[\phi(w^\top f(x))\phi(w^\top f(y))-\phi\left(w^\top f(x)\frac{\tilde{c}}{\tilde{c}(x)}\right)\phi\left(w^\top f(y)\frac{\tilde{c}}{\tilde{c}(y)}\right)\right]\right| \\
&=\left|\mathbb{E}_{xy}\mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1\tilde{c}(x))\phi(\rho_2\tilde{c}(y))-\phi(\rho_1\tilde{c})\phi(\rho_2\tilde{c})\right]\right| \\
&\leq \mathbb{E}_{xy}\left[C_1\left|\tilde{c}(x)-\tilde{c}\right|+C_2\tilde{c}(y)\left|\tilde{c}(x)-\tilde{c}\right|+C_3\left|\tilde{c}(y)-\tilde{c}\right|+C_4\tilde{c}\left|\tilde{c}(y)-\tilde{c}\right|\right] \\
&=\mathcal{O}\left(n_{l-1}^{-1/2}\right),
\end{aligned} \tag{130}$$

where ρ_1 and ρ_2 indicate two standard Gaussian random variables with correlation $f(x)^\top f(y)/(\|f(x)\|\,\|f(y)\|)$, the third line follows from B.15, and the last passage follows from Lemma B.14. By combining (129) and (130), we have that

$$\left|\mathbb{E}_w\left[\mathbb{E}_x\left[\phi^2(w^\top f(x))\right]-\mathbb{E}_{xy}\left[\phi(w^\top f(x))\phi(w^\top f(y))\right]\right]-\xi\right|=\mathcal{O}\left(n_{l-1}^{-1/2}\right), \tag{131}$$

with

$$\begin{aligned}
\xi :=&\mathbb{E}_x\mathbb{E}_w\left[\phi^2\left(w^\top f(x)\frac{\tilde{c}}{\tilde{c}(x)}\right)\right]-\mathbb{E}_{xy}\mathbb{E}_w\left[\phi\left(w^\top f(x)\frac{\tilde{c}}{\tilde{c}(x)}\right)\phi\left(w^\top f(y)\frac{\tilde{c}}{\tilde{c}(y)}\right)\right] \\
=&\mathbb{E}_{\rho_1}\left[\tilde{\phi}^2(\rho_1)\right]-\mathbb{E}_{xy}\left[\mathbb{E}_{\rho_1\rho_2}\left[\tilde{\phi}(\rho_1)\tilde{\phi}(\rho_2)\right]\right],
\end{aligned} \tag{132}$$

where we have set $\tilde{\phi}(t)=\phi(\tilde{c}t)$. Hence, in order to obtain that the quantity in (128) is $\Theta(1)$, it suffices to prove that $\xi=\Theta(1)$.

As ϕ is Lipschitz and $\tilde{c}$ is $\Theta(1)$, $\tilde{\phi}$ is also Lipschitz, which readily implies that $\xi=\mathcal{O}(1)$. We now prove that $\xi=\Omega(1)$. By exploiting the Hermite expansion of $\tilde{\phi}$, we have that

$$\xi=\sum_{i=0}^{\infty}\mu_i^2\left(1-\mathbb{E}_{xy}\left[\left(\frac{f(x)^\top f(y)}{\|f(x)\|\,\|f(y)\|}\right)^i\right]\right), \tag{133}$$

where μ_i is the i-th Hermite coefficient of $\tilde{\phi}$. Note that, since we conditioned on $\tilde{c}=\Theta(1)$, these coefficients are numerical constants. As ϕ (and therefore $\tilde{\phi}$) is non constant, there exist $j>0$ such that $\mu_j\neq 0$. Furthermore, we have that the sum in (133) contains only positive terms, as $|f(x)^\top f(y)|\leq\|f(x)\|\cdot\|f(y)\|$ by Cauchy-Schwarz. Therefore, in order to show that $\xi=\Omega(1)$, it suffices to prove that, for all $j>0$,

$$\mathbb{E}_{xy}\left[\left(\frac{\left|f(x)^\top f(y)\right|}{\|f(x)\|\,\|f(y)\|}\right)^j\right]\leq C_0<1, \tag{134}$$

where C_0 is an absolute constant strictly smaller than 1. Furthermore, (134) is implied by the following:

$$\mathbb{P}_{xy}\left(\frac{\left|f(x)^\top f(y)\right|}{\|f(x)\|\,\|f(y)\|}\leq C_1\right)\geq c_1, \tag{135}$$

where $C_1<1$ and $c_1>0$ are numerical constants.

By writing $f(x)$ as $\tilde{f}(x)+f$ and $f(y)$ as $\tilde{f}(y)+f$, we have

$$\begin{aligned}\frac{\left|f(x)^\top f(y)\right|}{\|f(x)\|\,\|f(y)\|} &= \frac{\left|(\tilde{f}(x)+f)^\top(\tilde{f}(y)+f)\right|}{\left\|\tilde{f}(x)+f\right\|\left\|\tilde{f}(y)+f\right\|} \\ &\le \frac{\left|\tilde{f}(x)^\top(\tilde{f}(y)+f)\right|+\left|f^\top\tilde{f}(y)\right|+\|f\|^2}{\min_{z\in\{x,y\}}\left\|\tilde{f}(z)+f\right\|^2} \\ &\le \frac{\left|\tilde{f}(x)^\top f(y)\right|+\left|f^\top\tilde{f}(y)\right|+\|f\|^2}{\min_{z\in\{x,y\}}\left(\left\|\tilde{f}(z)\right\|^2-2\left|f^\top\tilde{f}(z)\right|\right)+\|f\|^2}.\end{aligned} \tag{136}$$

Let us provide bounds on the various terms appearing in (136):

(i) Part (d) of the conditioning (cf. (121)) gives that

$$\mathbb{P}_{xy}\left(\min_{z\in\{x,y\}}\left\|\tilde{f}(z)\right\|_2^2 \ge cn_{l-1}\right) \ge \frac{1}{4}, \tag{137}$$

for some numerical constant $c>0$.

(ii) Part (b) of the conditioning gives that

$$\|f\|_2^2 \le C' n_{l-1}. \tag{138}$$

(iii) Part (a) of the conditioning gives that $\|f_{l-1}(x)\|_{\mathrm{Lip}} = \mathcal{O}(1)$, and part (b) of the conditioning gives that $\mathbb{E}_y\left[\|f_{l-1}(y)\|\right] = \Theta(\sqrt{n_{l-1}})$. Hence, as $y\sim P_X$, Assumption 2.2 implies that

$$\|f_{l-1}(y)\| = \Theta(\sqrt{n_{l-1}}), \tag{139}$$

with probability at least $1-2\exp(-cn_{l-1})$ over y, where c is a numerical constant. Furthermore, by recalling that $\mathbb{E}_x[\tilde{f}(x)]=0$ and using again Assumption 2.2, we have that, for any fixed vector u,

$$\mathbb{P}_x(|\tilde{f}(x)^\top u|>t) \le 2e^{-c_0 t^2/\|u\|_2^2}, \tag{140}$$

where c_0 is another numerical constant. Since x and y are independent, (140) implies that

$$\mathbb{P}_x(|\tilde{f}(x)^\top f(y)|>n_{l-1}^{3/4}) \le 2e^{-c_0 n_{l-1}^{3/2}/\|f(y)\|_2^2} \le 2e^{-c_2 n_{l-1}^{1/2}}, \tag{141}$$

where the first inequality holds for every y, and the second inequality holds with probability at least $1-2\exp(-cn_{l-1})$ over y by (139). As a result, we have

$$\mathbb{P}_{xy}(|\tilde{f}(x)^\top f(y)|>n_{l-1}^{3/4}) \le 2e^{-c_2\sqrt{n_{l-1}}}+2e^{-cn_{l-1}} \le 4e^{-c_3\sqrt{n_{l-1}}}. \tag{142}$$

(iv) By setting $t=n_{l-1}^{3/4}$ and $u=f$ into (140), we obtain

$$\mathbb{P}_y(|\tilde{f}(y)^\top f|>n_{l-1}^{3/4}) \le 2e^{-c_4\sqrt{n_{l-1}}}, \tag{143}$$

where c_4 is a numerical constant and we have also used (138).

(v) Finally, as x and y are independent, (143) implies that

$$\mathbb{P}_{xy}(\max_{z\in\{x,y\}}|\tilde{f}(z)^\top f|>n_{l-1}^{3/4}) \le 4e^{-c_4\sqrt{n_{l-1}}}. \tag{144}$$

By plugging into (136) the bounds (137), (142), (143) and (144), we obtain that

$$\mathbb{P}_{x,y}\left(\frac{\left|f(x)^\top f(y)\right|}{\|f(x)\|\,\|f(y)\|} \le \frac{2n_{l-1}^{3/4}+\|f\|_2^2}{c\,n_{l-1}-2n_{l-1}^{3/4}+\|f\|_2^2}\right) \ge 1/4-10e^{-c_5\sqrt{n_{l-1}}}. \tag{145}$$

By using also (138), we have that (135) readily follows from (145). From (135), we have that $\xi=\Theta(1)$. Hence, the quantity in (128) is $\Theta(1)$ and in particular it is lower bounded by a numerical constant, call it C_0.

By setting $t = C_0/2$ in (127), we conclude that

$$\mathbb{E}_x \left[\|f_l(x) - \mathbb{E}_x \left[f_l(x) \right] \|_2^2 \right] = \Theta(n_l), \tag{146}$$

with probability at least $1 - 2\exp(-cn_{l-1})$ over W_l, where c is an absolute constant. By taking into account the conditioning made at the beginning of the proof over the space $(W_k)_{k=1}^{l-1}$, we obtain that (146) holds with probability at least $1 - 5C' \exp(-cn_{n-1}) - 2\exp(-cn_{n-1}) \geq 1 - C\exp(-cn_{n-1})$ over $(W_k)_{k=1}^{l}$, where C is a numerical constant, which concludes the proof of (117).

Finally, we prove (118) and (119), again for layer l. By Lemma B.2, we have that $\|f_l(x)\|_{\mathrm{Lip}} = \mathcal{O}(1)$, with probability at least $1 - C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{l}$. By conditioning on this event, we also have that $\| \|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2 \|_{\mathrm{Lip}} = \mathcal{O}(1)$. Furthermore, we condition on a realization of $(W_k)_{k=1}^{l}$ such that (117) holds.

To obtain (118), we apply Jensen's inequality and Lemma B.13, which give that

$$\mathbb{E}_x \left[\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2 \right] = \Theta(\sqrt{n_l}), \tag{147}$$

with probability at least $1 - C'\exp(-cn_{L-1}) - 5C'\exp(-cn_{n-1}) - 2\exp(-cn_{n-1}) \geq 1 - C\exp(-cn_{n-1})$ over $(W_k)_{k=1}^{l}$.

To obtain (119), we condition on a realization of $(W_k)_{k=1}^{l}$ such that $\| \|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2 \|_{\mathrm{Lip}} = \mathcal{O}(1)$ and (118) holds. Then, by Assumption 2.2, we have that

$$\begin{aligned} &\mathbb{P}_x \left(\left| \|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2 - \mathbb{E}_x \left[\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2 \right] \right| > \mathbb{E}_x \left[\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2 \right] / 2 \right) \\ &\qquad \leq 2\exp(-c_1 n_l) \leq 2\exp(-cn_{L-1}), \end{aligned} \tag{148}$$

where c is a numerical constant. This gives that

$$\|f_l(x) - \mathbb{E}_x[f_l(x)]\| = \Theta(\sqrt{n_l}), \tag{149}$$

with probability at least $1 - 6C'\exp(-cn_{n-1}) - 2\exp(-cn_{n-1}) - 2\exp(-cn_{n-1}) \geq 1 - C\exp(-cn_{n-1})$ over x and $(W_k)_{k=1}^{l}$, which concludes the proof. □

Lemma C.5 (ℓ_2 norms of centered backpropagation)**.** *Let $x \sim P_X$. Then, we have*

$$\|D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[D_L \phi'(g_{L-1}(x)) \right] \|_2 = \Theta(\sqrt{n_{L-1}}), \tag{150}$$

with probability at least $1 - 10\exp(-c\log^2 n_{L-1}) - C\exp(-cn_{L-1})$ over x and $(W_k)_{k=1}^{L}$ over $(W_k)_{k=1}^{l}$ and x.

Proof. An application of Lemma C.4 for $l = L-2$ gives that

$$\|f_{L-2}(x) - \mathbb{E}_x \left[f_{L-2}(x) \right] \|_2 = \Theta(\sqrt{n_{L-2}}). \tag{151}$$

with probability at least $1 - C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L-2}$ and x.

To ease the notation, we use the shorthands $f(x) := f_{L-2}(x)$, $f = \mathbb{E}_x[f(x)]$, $\tilde{f}(x) = f(x) - f$, $W := W_{L-1}$ and $w_i = W_{:i}$. We also define $\tilde{c}(x) = \beta_l \|f(x)\| / \sqrt{n_{l-1}}$ and $\tilde{c} = \mathbb{E}_x[\tilde{c}(x)]$.

As in Lemma C.4, we condition on the 3 events (a)-(c), which jointly happen with probability at least $1 - 4C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L-2}$. Note that, to condition on the event (c), we use (151).

Now, we can write

$$\begin{aligned} &\mathbb{E}_x \left[\|D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[D_L \phi'(g_{L-1}(x)) \right] \|_2^2 \right] \\ &= \mathbb{E}_x \left[\left\| D_L \phi'(W^\top f_{L-2}(x)) - \mathbb{E}_x \left[D_L \phi'(W^\top f_{L-2}(x)) \right] \right\|_2^2 \right] \\ &= n_{L-1} \left(\frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} (D_L)_{ii}^2 \mathbb{E}_x \left[\left(\phi'(w_i^\top f(x)) - \mathbb{E}_x \left[\phi'(w_i^\top f(x)) \right] \right)^2 \right] \right) \\ &= n_{L-1} \left(\frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} (D_L)_{ii}^2 \mathbb{E}_{x w_1} \left[\left(\phi'(w_1^\top f(x)) - \mathbb{E}_x \left[\phi'(w_1^\top f(x)) \right] \right)^2 \right] + \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} Z_i \right), \end{aligned} \tag{152}$$

where we use that the w_i-s are identically distributed and we have defined the independent, mean-0 random variables

$$\begin{aligned} Z_i =& (D_L)_{ii}^2 \mathbb{E}_x \left[\left(\phi'(w_i^\top f(x)) - \mathbb{E}_x \left[\phi'(w_i^\top f(x)) \right] \right)^2 \right] \\ & - (D_L)_{ii}^2 \mathbb{E}_{w_1 x} \left[\left(\phi'(w_1^\top f(x)) - \mathbb{E}_x \left[\phi'(w_1^\top f(x)) \right] \right)^2 \right]. \end{aligned} \tag{153}$$

Note that in the definition of Z_i the randomness comes only from w_i and $(D_L)_{ii}$, since we are conditioning on $(W_k)_{k=1}^{L-2}$.

If we fix the $(D_L)_{ii}$-s and follow the same argument in (124)-(126) (cf. the proof of Lemma C.4), we have

$$\|Z_i\|_{\psi_1} \leq C_0 (D_L)_{ii}^2, \tag{154}$$

where C_0 is a numerical constant and we have used that ϕ' is Lipschitz. Let $\mathcal{E}_{\text{bad}}$ be the event s.t. $\max_i (D_L)_{ii}^2 > \log^2 n_{L-1}$. Then, by following the same argument as in Lemma B.3, we have that

$$\mathbb{P}(\mathcal{E}_{\text{bad}}) \leq 2 \exp(-c \log^2 n_{L-1}). \tag{155}$$

Hence, by conditioning on $\mathcal{E}_{\text{bad}}^c$, we have that

$$\max_i \|Z_i\|_{\psi_1} \leq C_0 \log^2 n_{L-1}. \tag{156}$$

By applying Bernstein inequality (cf. Corollary 2.8.3. in [65]), we get

$$\mathbb{P}\left(\left| \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} Z_i \right| > \frac{1}{\sqrt[4]{n_{L-1}}} \,\middle|\, \mathcal{E}_{\text{bad}}^c \right) \leq 2 \exp\left(-c \frac{\sqrt{n_{L-1}}}{\log^4 n_{L-1}} \right), \tag{157}$$

for some numerical constant c. By combining (155) and (157), we obtain that

$$\begin{aligned} \mathbb{P}\left(\left| \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} Z_i \right| > \frac{1}{\sqrt[4]{n_{L-1}}} \right) &\leq 2 \exp\left(-c \frac{\sqrt{n_{L-1}}}{\log^4 n_{L-1}} \right) + 2 \exp\left(-c \log^2 n_{L-1} \right) \\ &\leq 4 \exp\left(-c \log^2 n_{L-1} \right), \end{aligned} \tag{158}$$

where this probability is over W_{L-1} and W_L.

Let's now focus on the first term in the last line of (152). In particular, we have that

$$\xi = \mathbb{E}_{x w_1} \left[\left(\phi'(w_1^\top f(x)) - \mathbb{E}_x \left[\phi'(w_1^\top f(x)) \right] \right)^2 \right] = \Theta(1). \tag{159}$$

This can be proven by following the same argument in (129)-(145) (cf. the proof of Lemma C.4), as ϕ' is Lipschitz and non-constant.

Next, we re-write (152) as

$$\begin{aligned} & \mathbb{E}_x \left[\|D_L \phi'(g_{L-1}(x)) - \mathbb{E}_x \left[D_L \phi'(g_{L-1}(x)) \right] \|_2^2 \right] \\ &= n_{L-1} \left(\frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} (D_L)_{ii}^2 \mathbb{E}_{x w_1} \left[\left(\phi'(w_1^\top f(x)) - \mathbb{E}_x \left[\phi'(w_1^\top f(x)) \right] \right)^2 \right] + \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} Z_i \right) \\ &= n_{L-1} \left(\xi \mathbb{E}_{W_L} \left[(D_L)_{11}^2 \right] + \xi \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} \tilde{Z}_i + \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} Z_i \right) \\ &= n_{L-1} \left(\xi \beta_L^2 + \xi \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} \tilde{Z}_i + \frac{1}{n_{L-1}} \sum_{i=1}^{n_{L-1}} Z_i \right), \end{aligned} \tag{160}$$

where we have defined the independent, mean-0, sub-exponential random variables

$$\tilde{Z}_i = (D_L)_{ii}^2 - \mathbb{E}_{W_L} \left[(D_L)_{11}^2 \right]. \tag{161}$$

Since the $(D_L)_{ii}$-s are standard Gaussian, we have

$$\left\| \tilde{Z}_i \right\|_{\psi_1} \leq \left\| (D_L)_{ii}^2 \right\|_{\psi_1} = \|(D_L)_{ii}\|_{\psi_2}^2 = C_1. \tag{162}$$

Hence, another application of Bernstein inequality (cf. Corollary 2.8.3. in [65]) allows us to conclude that

$$\left|\frac{1}{n_{L-1}}\sum_{i=1}^{n_{L-1}}\tilde{Z}_i\right| = \mathcal{O}\left(n_{L-1}^{-1/4}\right), \tag{163}$$

with probability at least $1-2\exp(-c\sqrt{n_{L-1}})$ over W_L.

Thus, by using (158) and (163), and taking into account the initial conditioning over $(W_k)_{k=1}^{L-2}$, we conclude that

$$\mathbb{E}_x\left[\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2^2\right] = \Theta(n_{L-1}), \tag{164}$$

with probability at least $1-6\exp(-c\log^2 n_{L-1}) - 6C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L}$.

Proceeding in a similar fashion as in Lemma C.4, we apply Lemma B.4, which gives that $\|D_L\phi'(g_{L-1}(x))\|_{\text{Lip}} = \mathcal{O}(\log n_{L-1})$, with probability at least $1-2\exp(-c\log^2 n_{L-1}) - C'\exp(-n_{L-1})$ over $(W_k)_{k=1}^{L}$. By conditioning on this event, we also have that $\|\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2\|_{\text{Lip}} = \mathcal{O}(\log n_{L-1})$. Furthermore, we condition on a realization of $(W_k)_{k=1}^{L}$ such that (164) holds.

We can now apply Jensen's inequality and Lemma B.13, to obtain that

$$\mathbb{E}_x\left[\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2\right] = \Theta(\sqrt{n_l}), \tag{165}$$

with probability at least $1-8\exp(-c\log^2 n_{L-1}) - 7C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L}$.

Finally, we condition on a realization of $(W_k)_{k=1}^{L}$ such that $\|\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2\|_{\text{Lip}} = \mathcal{O}(\log n_{L-1})$ and (165) hold. Then, by Assumption 2.2, we have that

$$\begin{aligned}\mathbb{P}_x\big(\big|\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2 - \mathbb{E}_x\left[\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2\right]\big| \\ > \mathbb{E}_x\left[\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2\right]/2\big) \\ \leq 2\exp^{(}-cn_{L-1}).\end{aligned} \tag{166}$$

This gives that

$$\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2 = \Theta(\sqrt{n_{L-1}}), \tag{167}$$

with probability at least $1-10\exp(-c\log^2 n_{L-1}) - 8C'\exp(-cn_{L-1})$ over x and $(W_k)_{k=1}^{L}$. This concludes the proof. □

D Proofs for Part 1: Centering

D.1 Step (a): Centering F_{L-2} and B_{L-1}

Lemma D.1 (Centering F_{L-2} and B_{L-1})**.** *Consider the setting of Theorem 3.1, let $F_{L-2} \in \mathbb{R}^{N\times n_{L-2}}$ be the feature matrix at layer $L-2$, and let B_{L-1} contain the backpropagation terms from layer $L-1$, i.e. $(B_{L-1})_{i:} = D_L\phi'(g_{L-1}(x_i))$. Let $J_{L-2}J_{L-2}^\top = F_{L-2}F_{L-2}^\top \circ B_{L-1}B_{L-1}^\top$ and $\tilde{J}_{FB}\tilde{J}_{FB}^\top = \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top$, where $\tilde{F}_{L-2} = F_{L-2} - \mathbb{E}_X[F_{L-2}]$ and $\tilde{B}_{L-1} = B_{L-1} - \mathbb{E}_X[B_{L-1}]$. Then, we have that*

$$\lambda_{\min}\left(J_{L-2}J_{L-2}^\top\right) \geq \lambda_{\min}\left(\tilde{J}_{FB}\tilde{J}_{FB}^\top\right) - o(n_{L-1}n_{L-2}), \tag{168}$$

with probability at least $1 - C\exp(-cn_{L-1}) - 4\exp(-c\log^2 N) - 8\exp(-c\log^2 n_{L-1})$ over $(W_k)_{k=1}^{L}$ and $(x_i)_{i=1}^{N} \sim_{\text{i.i.d.}} P_X$, where c and C are numerical constants.

Proof. By Lemma B.2 and Lemma B.4, we have that $\|f_{L-2}\|_{\text{Lip}} = \mathcal{O}(1)$ and $\|D_L\phi'(g_{L-1}(x))\|_{\text{Lip}} = \mathcal{O}(\log n_{L-1})$ with probability $1 - C\exp(-n_{L-1}) - 2\exp(-c\log^2 n_{L-1})$ over $(W_k)_{k=1}^{L}$. We will condition on these events for the rest of the proof.

Let's define $\tilde{J}_F\tilde{J}_F^\top = \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ B_{L-1}B_{L-1}^\top$. We can now re-write the quantity $J_{L-2}J_{L-2}^\top$ as follows:

$$\begin{aligned}
J_{L-2}J_{L-2}^\top &= \tilde{J}_F\tilde{J}_F^\top + \mathbb{E}[F_{L-2}]\mathbb{E}[F_{L-2}]^\top \circ B_{L-1}B_{L-1}^\top \\
&\quad + \left((F_{L-2} - \mathbb{E}[F_{L-2}])\mathbb{E}[F_{L-2}]^\top + \mathbb{E}[F_{L-2}](F_{L-2} - \mathbb{E}[F_{L-2}])^\top\right) \circ B_{L-1}B_{L-1}^\top \\
&= \tilde{J}_F\tilde{J}_F^\top + \|\nu\|_2^2 B_{L-1}B_{L-1}^\top + \left(\Lambda\mathbf{1}\mathbf{1}^\top + \mathbf{1}\mathbf{1}^\top\Lambda\right) \circ B_{L-1}B_{L-1}^\top \\
&= \tilde{J}_F\tilde{J}_F^\top + \left(\|\nu\|_2 \mathbf{1} + \frac{\Lambda\mathbf{1}}{\|\nu\|_2}\right)\left(\|\nu\|_2 \mathbf{1} + \frac{\Lambda\mathbf{1}}{\|\nu\|_2}\right)^\top \circ B_{L-1}B_{L-1}^\top \\
&\quad - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top \\
&\succeq \tilde{J}_F\tilde{J}_F^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top,
\end{aligned} \tag{169}$$

where $\nu = \mathbb{E}_{x_i}[(F_{L-2})_{i:}] \in \mathbb{R}^{n_{L-2}}$ (independent on i, since the x_i-s are i.i.d.), Λ is a diagonal matrix such that $\Lambda_{ii} = \nu^\top(F_{L-2})_{i:} - \|\nu\|_2^2 =: \mu(x_i)$, and $\mathbf{1} \in \mathbb{R}^N$ is a vector full of ones. The last step is justified since the Hadamard product of PSD matrices is PSD by the Schur product theorem. Notice that we are assuming $\|\nu\|_2 \neq 0$. In fact, if $\|\nu\|_2 = 0$, then we immediately have that $J = \tilde{J}_F$.

Expanding in an analogous way the term $\tilde{J}_F\tilde{J}_F^\top$, we get

$$\begin{aligned}
J_{L-2}J_{L-2}^\top &\succeq \tilde{J}_F\tilde{J}_F^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top \\
&= \tilde{J}_{FB}\tilde{J}_{FB}^\top + \left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)^\top \circ \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \\
&\quad - \frac{\Gamma\mathbf{1}\mathbf{1}^\top\Gamma}{\|\eta\|_2^2} \circ \tilde{F}_{L-2}\tilde{F}_{L-2}^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top \\
&\succeq \tilde{J}_{FB}\tilde{J}_{FB}^\top + \left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)^\top \circ \tilde{F}_{L-2}\left(\frac{\nu\nu^\top}{\|\nu\|_2^2}\right)\tilde{F}_{L-2}^\top \\
&\quad - \frac{\Gamma\mathbf{1}\mathbf{1}^\top\Gamma}{\|\eta\|_2^2} \circ \tilde{F}_{L-2}\tilde{F}_{L-2}^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top
\end{aligned} \tag{170}$$

where $\eta = \mathbb{E}_{x_i}[(B_{L-1})_{i:}] \in \mathbb{R}^{n_{L-1}}$ (independent on i, since the x_i-s are i.i.d.), Γ is a diagonal matrix such that $\Gamma_{ii} = \eta^\top(B_{L-1})_{i:} - \|\eta\|_2^2 =: \zeta(x_i)$. The last step is justified by the fact that the following matrix is PSD

$$\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)^\top \circ \tilde{F}_{L-2}\left(I - \frac{\nu\nu^\top}{\|\nu\|_2^2}\right)\tilde{F}_{L-2}^\top,$$

since it is the Hadamard product of two PSD matrices. Notice that we are assuming $\|\eta\|_2 \neq 0$. In fact, if $\|\eta\|_2 = 0$, then we immediately have that $\tilde{J}_F = \tilde{J}_{FB}$.

Taking into account the following relations

$$\Lambda\mathbf{1} = \tilde{F}_{L-2}\nu, \qquad \Gamma\mathbf{1} = \tilde{B}_{L-1}\eta, \qquad \mathbb{E}_X[B_{L-1}] = \mathbf{1}\eta^\top, \tag{171}$$

we can simplify the second and the fourth term of the RHS of equation (170) as follows

$$\begin{aligned}
&\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)^\top \circ \tilde{F}_{L-2}\left(\frac{\nu\nu^\top}{\|\nu\|_2^2}\right)\tilde{F}_{L-2}^\top - B_{L-1}B_{L-1}^\top \circ \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \\
&= \left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma\mathbf{1}}{\|\eta\|_2}\right)^\top \circ \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \\
&\quad - \left(\mathbf{1}\eta^\top + \tilde{B}_{L-1}\right)\left(\mathbf{1}\eta^\top + \tilde{B}_{L-1}\right)^\top \circ \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \\
&= \left(\frac{\Gamma\mathbf{1}\mathbf{1}^\top\Gamma}{\|\eta\|_2^2} - \tilde{B}_{L-1}\tilde{B}_{L-1}^\top\right) \circ \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \succeq -\tilde{B}_{L-1}\tilde{B}_{L-1}^\top \circ \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2}.
\end{aligned} \tag{172}$$

Merging this last relation with (170) we get

$$J_{L-2}J_{L-2}^\top \succeq \tilde{J}_{FB}\tilde{J}_{FB}^\top - \frac{\Gamma \mathbf{1}\mathbf{1}^\top \Gamma}{\|\eta\|_2^2} \circ \tilde{F}_{L-2}\tilde{F}_{L-2}^\top - \frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\|\nu\|_2^2} \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top$$
$$= \tilde{J}_{FB}\tilde{J}_{FB}^\top - \left(\frac{\Gamma}{\|\eta\|_2}\tilde{F}_{L-2}\right)\left(\frac{\Gamma}{\|\eta\|_2}\tilde{F}_{L-2}\right)^\top - \left(\frac{\Lambda}{\|\nu\|_2}\tilde{B}_{L-1}\right)\left(\frac{\Lambda}{\|\nu\|_2}\tilde{B}_{L-1}\right)^\top. \tag{173}$$

Note that $\|\mu\|_{\text{Lip}} \leq \|f_{L-2}\|_{\text{Lip}} \|\nu\|_2$, and that $\mathbb{E}_{x_i}[\mu(x_i)] = 0$ for all $i \in [N]$. Thus, by using Assumption 2.2 on x_i and exploiting the initial conditioning on the weights, we have that

$$\mathbb{P}(|\mu(x_i)|/\|\nu\|_2 > t) < 2\exp(-ct^2), \tag{174}$$

where the probability is intended over $x_i \sim P_X$. Thus, following the same argument of Lemma B.3, the last relation implies that

$$\|\Lambda/\|\nu\|_2\|_{\text{op}} = \mathcal{O}(\log N), \tag{175}$$

with probability at least $1 - 2\exp(-c\log^2 N)$, where c is a numerical constant, and the probability is intended over $\{x_i\}_{i=1}^N$. This implies that

$$\left\|\left(\frac{\Lambda}{\|\nu\|_2}\tilde{B}_{L-1}\right)\left(\frac{\Lambda}{\|\nu\|_2}\tilde{B}_{L-1}\right)^\top\right\|_{\text{op}}$$
$$\leq \left\|\frac{\Lambda}{\|\nu\|_2}\right\|_{\text{op}}^2 \left\|\tilde{B}_{L-1}\tilde{B}_{L-1}^\top\right\|_{\text{op}} = \mathcal{O}\left((N + n_{L-1}) \cdot \log^2 N \cdot \log^2 n_{L-1}\right) = o(n_{L-2}n_{L-1}), \tag{176}$$

where the second equality is justified by Lemma B.9, and the last by Lemma B.1. This result holds with probability $1 - C\exp(-cn_{L-1}) - 2\exp(-c\log^2 N) - 4\exp(-c\log^2 n_{L-1})$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^N$.

Note that $\|\zeta\|_{\text{Lip}} \leq \|D_L\phi'(g_{L-1}(x))\|_{\text{Lip}} \|\eta\|_2$, and that $\mathbb{E}_{x_i}[\zeta(x_i)] = 0$ for all $i \in [N]$. Thus, by using Assumption 2.2 on x_i and exploiting the initial conditioning on the weights, we have that

$$\mathbb{P}(|\zeta(x_i)|/\|\eta\|_2 > t \cdot \log n_{L-1}) < 2\exp(-ct^2), \tag{177}$$

where the probability is intended over $x_i \sim P_X$. Thus, following the same argument of Lemma B.3, the last relation implies that

$$\|\Gamma/\|\eta\|_2\|_{\text{op}} = \mathcal{O}(\log N \cdot \log n_{L-1}), \tag{178}$$

with probability at least $1 - 2\exp(-c\log^2 N)$, where c is a numerical constant, and the probability is intended over $\{x_i\}_{i=1}^N$. This implies that

$$\left\|\left(\frac{\Gamma}{\|\eta\|_2}\tilde{F}_{L-2}\right)\left(\frac{\Gamma}{\|\eta\|_2}\tilde{F}_{L-2}\right)^\top\right\|_{\text{op}}$$
$$\leq \left\|\frac{\Gamma}{\|\eta\|_2}\right\|_{\text{op}}^2 \left\|\tilde{F}_{L-2}\tilde{F}_{L-2}^\top\right\|_{\text{op}} = \mathcal{O}\left((N + n_{L-2}) \cdot \log^2 N \cdot \log^2 n_{L-1}\right) = o(n_{L-2}n_{L-1}), \tag{179}$$

where the second equality is justified by Lemma B.8, and the last by Lemma B.1. This result holds with probability $1 - C\exp(-cn_{L-1}) - 2\exp(-c\log^2 N) - 4\exp(-c\log^2 n_{L-1})$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^N$.

By merging (176) and (179) with (173), we readily obtain the desired result. □

D.2 Step (b): Centering everything

Lemma D.2 (Centering everything)**.** *Consider the setting of Theorem 3.1, let $F_{L-2} \in \mathbb{R}^{N \times n_{L-2}}$ be the feature matrix at layer $L-2$, and let B_{L-1} contain backpropagation terms from layer $L-1$, i.e. $(B_{L-1})_{i:} = D_L\phi'(g_{L-1}(x_i))$. Let $\tilde{J}_{FB}\tilde{J}_{FB}^\top = \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top$ and $\tilde{J}_{L-2}\tilde{J}_{L-2}^\top =$*

$\tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top - \mathbb{E}_X[\tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top]$, *where* $\tilde{F}_{L-2} = F_{L-2} - \mathbb{E}_X[F_{L-2}]$ *and* $\tilde{B}_{L-1} = B_{L-1} - \mathbb{E}_X[B_{L-1}]$. *Then, we have that*

$$\lambda_{\min}\left(\tilde{J}_{FB}\tilde{J}_{FB}^\top\right) \geq \lambda_{\min}\left(\tilde{J}_{L-2}\tilde{J}_{L-2}^\top\right) - o(n_{L-1}n_{L-2}), \tag{180}$$

with probability at least $1 - C\exp(-n_{L-1}) - 2\exp(-c\log^2 n_{L-1}) - 2\exp(-c\log^2 N)$ *over* $(W_k)_{k=1}^L$ *and* $(x_i)_{i=1}^N$.

Proof. Note that the i-th row of $\tilde{J}_{FB}$ is now in the form

$$(\tilde{J}_{FB})_{i:} = \tilde{f}_{L-2}(x_i) \otimes (D_L\tilde{\phi}'(g_{L-1}(x_i))), \tag{181}$$

where we recall that $\tilde{f}_{L-2}(x_i) = f_{L-2}(x_i) - \mathbb{E}_{x_i}[f_{L-2}(x_i)]$ and $\tilde{\phi}'(g_{L-1}(x_i)) = \phi'(g_{L-1}(x_i)) - \mathbb{E}[\phi'(g_{L-1}(x_i))]$. Furthermore, $(\tilde{J}_{L-2})_{i:} = (\tilde{J}_{FB})_{i:} - \mathbb{E}_{x_i}[(\tilde{J}_{FB})_{i:}]$. Then, by following similar passages as in (169), we have

$$\begin{aligned}
\tilde{J}_{FB}\tilde{J}_{FB}^\top &= \tilde{J}_{L-2}\tilde{J}_{L-2}^\top + \mathbb{E}[\tilde{J}_{FB}]\mathbb{E}[\tilde{J}_{FB}]^\top \\
&\quad + (\tilde{J}_{FB} - \mathbb{E}[\tilde{J}_{FB}])\mathbb{E}[\tilde{J}_{FB}]^\top + \mathbb{E}[\tilde{J}_{FB}](\tilde{J}_{FB} - \mathbb{E}[\tilde{J}_{FB}])^\top \\
&= \tilde{J}_{L-2}\tilde{J}_{L-2}^\top + \|A\|_F^2\,\mathbf{1}\mathbf{1}^\top + \Lambda\mathbf{1}\mathbf{1}^\top + \mathbf{1}\mathbf{1}^\top\Lambda \\
&= \tilde{J}_{L-2}\tilde{J}_{L-2}^\top + \left(\|A\|_F\,\mathbf{1} + \frac{\Lambda\mathbf{1}}{\|A\|_F}\right)\left(\|A\|_F\,\mathbf{1} + \frac{\Lambda\mathbf{1}}{\|A\|_F}\right)^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|A\|_F^2} \\
&\succeq \tilde{J}_{L-2}\tilde{J}_{L-2}^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|A\|_F^2},
\end{aligned} \tag{182}$$

where we have defined

$$A = \mathbb{E}_{x_i}\left[\tilde{f}_{L-2}(x_i)(D_L\tilde{\phi}'(g_{L-1}(x_i)))^\top\right], \tag{183}$$

which is independent on i (as the x_i-s are i.i.d.), and Λ is a diagonal matrix that contains in the i-th position

$$\Lambda_{ii} = \tilde{f}_{L-2}(x_i)^\top A(D_L\tilde{\phi}'(g_{L-1}(x_i))) - \mathbb{E}_{x_i}\left[\tilde{f}_{L-2}(x_i)^\top A(D_L\tilde{\phi}'(g_{L-1}(x_i)))\right]. \tag{184}$$

The last passage of (182) is true since we are subtracting a PSD matrix.

An application of Lemma B.2 gives that $\|f_{L-2}(x_i)\|_{\text{Lip}}$ and $\|\phi'(g_{L-1}(x_i))\|_{\text{Lip}}$ are upper bounded by a numerical constant both with probability at least $1 - C\exp(-n_{L-1})$ over $(W_k)_{k=1}^{L-1}$. Let us condition on this event on the probability space of $(W_k)_{k=1}^{L-1}$. Then, we can apply Lemma B.5 with $u(x) := \tilde{f}_{L-2}(x)$ and $v(x) := \tilde{\phi}'(g_{L-1}(x))$, which implies that

$$\|\Lambda_{ii}\|_{\psi_1} < C\,\|AD_L\|_F \leq C\,\|A\|_F\,\|D_L\|_{\text{op}} = \mathcal{O}\left(\log n_{L-1}\right)\|A\|_F. \tag{185}$$

In (185), C is a numerical constant and the last equality holds with probability at least $1 - 2\exp(-c\log^2 n_{L-1})$ over W_L by Lemma B.3. Thus,

$$\mathbb{P}(|\Lambda_{ii}|/\|A\|_F > t\cdot\log n_{L-1}) < 2\exp(-ct), \tag{186}$$

where the probability is intended over $x_i \sim P_X$ and c is a numerical constant. Thus, following the same argument of Lemma B.3, the last relation implies that

$$\|\Lambda/\|A\|_F\|_{\text{op}} = \mathcal{O}\left(\log^2 N\cdot\log n_{L-1}\right), \tag{187}$$

with probability at least $1 - 2\exp(-c\log^2 N)$, where c is a numerical constant, and the probability is intended over $\{x_i\}_{i=1}^N$.

Thus, with probability $1 - C\exp(-n_{L-1}) - 2\exp(-c\log^2 n_{L-1}) - 2\exp(-c\log^2 N)$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^N$, we have

$$\left\|\frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|A\|_F^2}\right\|_{\text{op}} \leq \|\mathbf{1}\mathbf{1}\|_{\text{op}}\left\|\frac{\Lambda}{\|A\|_F}\right\|_{\text{op}}^2 = \mathcal{O}\left(N\cdot\log^4 N\cdot\log^2 n_{L-1}\right) = o(n_{L-2}n_{L-1}), \tag{188}$$

where the last equality is a consequence of Lemma B.1.

The desired result follows from merging (188) with (182). □

E Proofs for Part 2: Bounding the Centered Jacobian

E.1 ℓ_2 and sub-exponential norms of centered Jacobian

We start by providing an upper bound on the quantity $\mathbb{E}_x\left[\tilde{f}_{L-2}(x)(D_L\tilde{\phi}'(g_{L-1}(x)))^\top\right]$. This preliminary result will be useful when bounding the ℓ_2 norm of the rows of the centered Jacobian.

Lemma E.1. *Consider the setting of Theorem 3.1, let $x \sim P_X$, and let A be defined as*

$$A = \mathbb{E}_x\left[\tilde{f}_{L-2}(x)(D_L\tilde{\phi}'(g_{L-1}(x)))^\top\right]. \tag{189}$$

Then, we have

$$\|A\|_F = \mathcal{O}\left(\sqrt{n_{L-1}}\log(n_{L-1})\right), \tag{190}$$

with probability at least $1 - 2\exp(-c\log^2 n_{L-1}) - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^L$, where c is an absolute constant.

Proof. We condition on $\|f_{L-2}(x)\|_{\text{Lip}} = \mathcal{O}(1)$ and on $\|\phi'(g_{L-1}(x))\|_{\text{Lip}} = \mathcal{O}(1)$. By Lemma B.2, these two conditions hold with probability at least $1 - C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L-1}$. Then, as P_X satisfies Assumption 2.2, we have that $\left\|\tilde{f}_{L-2}(x)\right\|_{\psi_2} = \mathcal{O}(1)$ and $\left\|\tilde{\phi}'(g_{L-1}(x))\right\|_{\psi_2} = \mathcal{O}(1)$. Hence, an application of Lemma B.6 gives that

$$\left\|\mathbb{E}_x\left[\tilde{f}_{L-2}(x)\tilde{\phi}'(g_{L-1}(x))^\top\right]\right\|_{\text{op}} \leq C_1, \tag{191}$$

where C_1 is a numerical constant.

The following chain of inequalities holds:

$$\begin{aligned}
\|A\|_F &\leq \left\|\mathbb{E}_x\left[\tilde{f}_{L-2}(x)\tilde{\phi}'(g_{L-1}(x))^\top\right]\right\|_F \|D_L\|_{\text{op}} \\
&\leq \sqrt{n_{L-1}}\left\|\mathbb{E}_x\left[\tilde{f}_{L-2}(x)\tilde{\phi}'(g_{L-1}(x))^\top\right]\right\|_{\text{op}} \|D_L\|_{\text{op}} \\
&\leq C_1\sqrt{n_{L-1}}\|D_L\|_{\text{op}} \\
&= \mathcal{O}\left(\sqrt{n_{L-1}}\log(n_{L-1})\right),
\end{aligned} \tag{192}$$

where the third line uses (191), and the last holds with probability $1 - 2\exp(-c\log^2 n_{L-1})$ over W_L, because of Lemma B.3. Taking into account the initial conditioning, we get the desired result. □

The next two results provide bounds on the ℓ_2 norm and on the sub-exponential ψ_1 norm of the rows of $\tilde{J}_{L-2}$, respectively.

Lemma E.2 (ℓ_2 norm of rows of centered Jacobian)**.** *Consider the setting of Theorem 3.1, let $x \sim P_X$, and let $\tilde{J}_x$ be defined as*

$$\tilde{J}_x = \tilde{f}_{L-2}(x) \otimes (D_L\tilde{\phi}'(g_{L-1}(x))) - \mathbb{E}_x\left[\tilde{f}_{L-2}(x) \otimes D_L\tilde{\phi}'(g_{L-1}(x))\right]. \tag{193}$$

Then, we have

$$\left\|\tilde{J}_x\right\|_2 = \Theta(\sqrt{n_{L-1}n_{L-2}}), \tag{194}$$

with probability at least $1 - C\exp(-cn_{L-1}) - 12\exp(-c\log^2 n_{L-1})$ over x and $(W_k)_{k=1}^L$ and x.

Proof. We have that

$$\begin{aligned}
\left\|\tilde{J}_x\right\|_2 &= \left\|\tilde{f}_{L-2}(x) \otimes (D_L\tilde{\phi}'(g_{L-1}(x))) - \mathbb{E}_x\left[\tilde{f}_{L-2}(x) \otimes D_L\tilde{\phi}'(g_{L-1}(x))\right]\right\|_2 \\
&= \left\|\tilde{f}_{L-2}(x)(D_L\tilde{\phi}'(g_{L-1}(x)))^\top - \mathbb{E}_x\left[\tilde{f}_{L-2}(x)(D_L\tilde{\phi}'(g_{L-1}(x)))^\top\right]\right\|_F \\
&= \left\|\tilde{f}_{L-2}(x)^\top(D_L\tilde{\phi}'(g_{L-1}(x))) - A\right\|_F,
\end{aligned} \tag{195}$$

where A is defined in (189). The second equality is justified by the identity $\|u \otimes v\|_2 = \left\|uv^\top\right\|_F$ that holds for any vectors u, v. An application of the triangle inequality gives that

$$\eta - \|A\|_F \leq \left\|\tilde{J}_x\right\|_2 \leq \eta + \|A\|_F, \quad \text{with } \eta = \left\|\tilde{f}_{L-2}(x)\right\|_2 \left\|D_L \tilde{\phi}'(g_{L-1}(x))\right\|_2. \tag{196}$$

Lemma C.4 gives that

$$\left\|\tilde{f}_{L-2}(x)\right\|_2 = \|f_{L-2}(x) - \mathbb{E}_x\left[f_{L-2}(x)\right]\|_2 = \Theta(\sqrt{n_{L-2}}), \tag{197}$$

with probability at least $1 - C' \exp(-cn_{L-1})$ over $(W_k)_{k=1}^L$ and x. Furthermore, Lemma C.5 gives that

$$\left\|D_L \tilde{\phi}'(g_{L-1}(x))\right\|_2 = \|D_L\left(\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[\phi'(g_{L-1}(x))\right]\right)\|_2 = \Theta(\sqrt{n_{L-1}}), \tag{198}$$

with probability at least $1 - 10\exp(-c\log^2 n_{L-1}) - C'\exp(-cn_{L-1})$ over x and $(W_k)_{k=1}^L$. By combining (196), (197), (198) and the bound on $\|A\|_F$ provided by Lemma E.1, we conclude that

$$\left\|\tilde{J}_x\right\|_2 = \Theta(\sqrt{n_{L-1}n_{L-2}}), \tag{199}$$

with probability at least

$$\begin{aligned} &1 - C'\exp(-cn_{L-1}) - 10\exp(-c\log^2 n_{L-1}) \\ &\qquad - C'\exp(-cn_{L-1}) - 2\exp(-c\log^2 n_{L-1}) - C'\exp(-cn_{L-1}) \\ &\geq 1 - C\exp(-cn_{L-1}) - 12\exp(-c\log^2 n_{L-1}), \end{aligned} \tag{200}$$

over x and $(W_k)_{k=1}^L$, which gives the desired result. □

Lemma E.3 (Sub-exponential norm of rows of centered Jacobian)**.** *Consider the setting of Theorem 3.1, let $x \sim P_X$, and let $\tilde{J}_x$ be defined as in* (193)*. Fix a realization of $(W_k)_{k=1}^L$. Then, with probability at least $1 - 2\exp(-c\log^2 n_{L-1}) - C\exp(-cn_{L-1})$ over this realization (c being a numerical constant), we have that*

$$\left\|\tilde{J}_x\right\|_{\psi_1} = \mathcal{O}\left(\log n_{L-1}\right). \tag{201}$$

Proof. We condition on $\|f_{L-2}(x)\|_{\text{Lip}} = \mathcal{O}(1)$ and on $\|\phi'(g_{L-1}(x))\|_{\text{Lip}} = \mathcal{O}(1)$. By Lemma B.2, these two conditions hold with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L-1}$. Then, we have

$$\begin{aligned} \left\|\tilde{J}_x\right\|_{\psi_1} &= \sup_{u \text{ s.t. } \|u\|_2 = 1} \left\|u^\top \tilde{J}_x\right\|_{\psi_1} \\ &= \sup_{U \text{ s.t. } \|U\|_F = 1} \left\|\tilde{f}_{L-2}(x) U (D_L \tilde{\phi}'(g_{L-1}(x))) - \mathbb{E}_x\left[\tilde{f}_{L-2}(x) U D_L \tilde{\phi}'(g_{L-1}(x))\right]\right\|_{\psi_1} \\ &\leq C_0 \sup_{U \text{ s.t. } \|U\|_F = 1} \|U D_L\|_F \\ &\leq C_0 \|D_L\|_{\text{op}} \\ &\leq C_0 \log n_{L-1}, \end{aligned} \tag{202}$$

where the third line follows from Lemma B.5 and the last inequality holds with probability at least $1 - 2\exp(-c\log^2 n_{L-1})$ over W_L by Lemma B.3. Taking into account the initial conditioning, we get the desired result. □

E.2 Proof of Proposition 3.3

Proof of Proposition 3.3. Following the notation in [2], we define

$$B := \sup_{z \in \mathbb{R}^N : \|z\|_2 = 1} \left| \left\|\sum_{i=1}^N z_i \tilde{J}_{i:}\right\|_2^2 - \sum_{i=1}^N z_i^2 \left\|\tilde{J}_{i:}\right\|_2^2 \right|^{\frac{1}{2}}. \tag{203}$$

Then, for any $z \in \mathbb{R}^N$ with unit norm, we have that

$$\left\|\tilde{J}z\right\|_2^2 = \left\|\sum_{i=1}^N z_i \tilde{J}_{i:}\right\|_2^2 - \sum_{i=1}^N z_i^2 \left\|\tilde{J}_{i:}\right\|_2^2 + \sum_{i=1}^N z_i^2 \left\|\tilde{J}_{i:}\right\|_2^2 \geq \min_i \left\|\tilde{J}_{i:}\right\|_2^2 - B^2, \tag{204}$$

which implies that

$$\lambda_{\min}\left(\tilde{J}\tilde{J}^\top\right) = \inf_{z \in \mathbb{R}^N : \|z\|_2 = 1} \left\|\tilde{J}z\right\|_2^2 \geq \min_i \left\|\tilde{J}_{i:}\right\|_2^2 - B^2. \tag{205}$$

In our case, $\tilde{J}_{i:} \in \mathbb{R}^{n_{L-2}n_{L-1}}$. Notice that this dimension is indicated with n in Theorem 3.2 of [2]. In the statement of the mentioned Theorem, let's fix $r = 1$, $m = N$, and $\theta = (N/(n_{L-1}n_{L-2}))^{1/4} < 1/4$. Then, we have that the condition required to apply Theorem 3.2 is satisfied, *i.e.*

$$N \log^2\left(2\sqrt[4]{\frac{n_{L-2}n_{L-1}}{N}}\right) \leq \sqrt{N n_{L-2} n_{L-1}}, \tag{206}$$

where the inequality follows from Assumption 2.5. By combining (205) with the upper bound on B given by Theorem 3.2 of [2], the desired result readily follows. □

E.3 Proof of Theorem 3.4

Proof of Theorem 3.4. By Lemma E.3, we have that, with probability at least $1 - 2\exp(-c\log^2 n_{L-1}) - C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^L$, the rows of $\tilde{J}$ are sub-exponential (with respect to the randomness in $(x_i)_{i=1}^N$). In particular, we have that

$$\psi := \max_i \left\|\tilde{J}_{i:}\right\|_{\psi_1} \leq C_1 \log n_{L-1}. \tag{207}$$

Furthermore, by Lemma E.2, we have that

$$\left\|\tilde{J}_{i:}\right\|_2 = \Theta(\sqrt{n_{L-2}n_{L-1}}), \tag{208}$$

with probability at least $1 - p$ over x_i and $(W_k)_{k=1}^L$, where to ease the notation we have defined $p := C'\exp(-c_0 n_{L-1}) + 12\exp(-c_0 \log^2 n_{L-1})$. Hence, with probability at least $1 - \sqrt{p}$ over $(W_k)_{k=1}^L$, we have that

$$\mathbb{P}_{x_i}\left(c_1\sqrt{n_{L-2}n_{L-1}} \leq \left\|\tilde{J}_{i:}\right\|_2 \leq c_2\sqrt{n_{L-2}n_{L-1}}\right) \geq 1 - \sqrt{p}, \tag{209}$$

for some numerical constants $c_2 > c_1 > 0$. In (209), we use the symbol $\mathbb{P}_{x_i}$ to highlight that this probability is taken over x_i. For the rest of the argument, we condition on a realization of $(W_k)_{k=1}^L$ s.t. (207) and (209) hold. Then, by performing a union bound over the samples, we have that

$$\eta_{\min} = \min_i \left\|\tilde{J}_{i:}\right\|_2 \geq c_1\sqrt{n_{L-2}n_{L-1}}, \tag{210}$$

and

$$\eta_{\max} = \max_i \left\|\tilde{J}_{i:}\right\|_2 \leq c_2\sqrt{n_{L-2}n_{L-1}}, \tag{211}$$

with probability at least $1 - N\sqrt{p}$ over $(x_i)_{i=1}^N$.

Next, we apply Proposition 3.3 with $K = 1$, $K' = c_2$ and

$$\begin{aligned}\Delta &= C_1(\psi K + K')^2 N^{1/4}(n_{L-1}n_{L-2})^{3/4} \\ &\leq C_2 \log^2 n_{L-1} N^{1/4}(n_{L-1}n_{L-2})^{3/4} \\ &= o(n_{L-1}n_{L-2}).\end{aligned} \tag{212}$$

Note that (22) in Lemma B.1 gives that $N^{1/4} \cdot \log^2 n_{L-1} = o((n_{L-1}n_{L-2})^{1/4})$, which justifies the last line. Thus, (13) implies that

$$\lambda_{\min}\left(\tilde{J}\tilde{J}^\top\right) \geq \eta_{\min}^2 - \Delta \geq c_1 n_{L-2}n_{L-1} - o(n_{L-1}n_{L-2}) = \Theta(n_{L-2}n_{L-1}), \tag{213}$$

with probability at least

$$
\begin{aligned}
1-\exp\left(-cK\sqrt{N}\log\left(\frac{2n_{L-1}n_{L-2}}{N}\right)\right)-\mathbb{P}\left(\eta_{\max}\geq K'\sqrt{n_{L-1}n_{L-2}}\right)\\
\geq 1-\exp\left(-c\sqrt{N}\right)-\mathbb{P}\left(\eta_{\max}\geq c_2\sqrt{n_{L-1}n_{L-2}}\right)\\
\geq 1-\exp\left(-c\sqrt{N}\right)-N\sqrt{p},
\end{aligned}
\tag{214}
$$

where the last inequality follows from (211). By taking into account the conditioning over $(W_k)_{k=1}^L$ made in order to guarantee (207) and (209), we conclude that $\lambda_{\min}\left(\tilde{J}\tilde{J}^\top\right)=\Omega(n_{L-1}n_{L-2})$ with probability at least

$$
\begin{aligned}
&1-\exp\left(-c\sqrt{N}\right)-N\sqrt{p}-\sqrt{p}-2\exp(-c\log^2 n_{L-1})-C'\exp(-cn_{L-1})\\
&=1-\exp\left(-c\sqrt{N}\right)-(N+1)\sqrt{C'\exp(-c_0n_{L-1})+12\exp(-c_0\log^2 n_{L-1})}\\
&\qquad\qquad-2\exp(-c\log^2 n_{L-1})-C'\exp(-cn_{L-1})\\
&\geq 1-\exp\left(-c\sqrt{N}\right)-(N+1)\left(\sqrt{C'}\exp(-c_0n_{L-1}/2)+\sqrt{12}\exp(-c_0\log^2 n_{L-1}/2)\right)\\
&\qquad\qquad-2\exp(-c\log^2 n_{L-1})-C'\exp(-cn_{L-1})\\
&\geq 1-\exp\left(-c\sqrt{N}\right)-C''N\exp(-c_1n_{L-1})-C''N\exp(-c\log^2 n_{L-1}),
\end{aligned}
\tag{215}
$$

over $(x_i)_{i=1}^N$ and $(W_k)_{k=1}^L$, which gives the desired result. □

F Proof of the Upper Bound 7

Before giving the proof of the upper bound 7, we provide again its statement for the reader's convenience.

Lemma F.1 (Upper bound on the smallest NTK eigenvalue)**.** *Consider the setting of Theorem 3.1, and let K be the NTK Gram matrix* (3)*. Then, we have*

$$
\lambda_{\min}(K)=\mathcal{O}\left(dn_{L-1}\right), \tag{216}
$$

with probability at least $1-C\exp(-cn_{L-1})$ over $(x_i)_{i=1}^N$ and $(W_k)_{k=1}^L$, where c and C are numerical constants.

Proof. By using the expression in (8), we have that

$$
\lambda_{\min}(K)=\lambda_{\min}\left(JJ^\top\right)\leq(JJ^\top)_{11}=\sum_{l=0}^{L-1}\|(F_l)_{1:}\|_2^2\,\|(B_{l+1})_{1:}\|_2^2. \tag{217}
$$

An application of Lemma C.1 gives that

$$
\|(F_l)_{1:}\|_2^2=\|f_l(x_1)\|_2^2=\Theta(n_l), \tag{218}
$$

with probability at least $1-C'\exp(-cn_{L-1})$ over $(W_k)_{k=1}^l$ and x_1. We condition on the event such that (218) holds for all $l\in\{0,\dots,L-1\}$. This happens with probability at least $1-C''\exp(-cn_{L-1})$ over $(W_k)_{k=1}^L$ and x_1.

By definition, we have that $\|B_L\|_2=1$ and that, for $l\in[L-1]$,

$$
\|(B_l)_{1:}\|_2^2=\left\|\prod_{k=l}^{L-1}\Sigma_k(x_1)W_{k+1}\right\|_2^2. \tag{219}
$$

Since $\Sigma_k(x_1)=\mathrm{diag}([\phi'(g_{k,j}(x_1))]_{j=1}^{n_k})$, by Assumption 2.3, we have that

$$
\|\Sigma_k(x_1)\|_{\mathrm{op}}\leq M. \tag{220}
$$

Let us now condition on the following two events: *(i)* $\|W_k\|_{\text{op}} = \Theta(1)$, for all $k \in [L-1]$ (this happens with probability at least $1 - C' \exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L-1}$, see (33) in the proof of Lemma B.2), and *(ii)* $\|W_L\|_2 = \Theta(\sqrt{n_{L-1}})$ (this happens with probability at least $1 - \exp(-cn_{L-1})$ over W_L, by Theorem 3.1.1 in [65]). Then, we readily get

$$\|(B_l)_{1:}\|_2^2 = \mathcal{O}\left(n_{L-1}\right). \tag{221}$$

Taking the intersection of all the events over which we conditioned, we finally obtain

$$\sum_{l=0}^{L-1} \|(F_l)_{1:}\|_2^2 \, \|(B_{l+1})_{1:}\|_2^2 = \mathcal{O}\left(n_{L-1} \sum_{l=0}^{L-1} n_l\right) = \mathcal{O}\left(dn_{L-1}\right), \tag{222}$$

with probability at least $1 - (1 + C' + C'') \exp(-cn_{L-1})$ over $(W_k)_{k=1}^{L}$ and x_1, where in the last step we have used Assumption 2.4. By combining (217) and (222), the desired result follows. □

G Proof of Corollary 4.1

Proof of Corollary 4.1. By Theorem 3.1, we have that the smallest eigenvalue of $JJ^\top$ is bounded away from zero with probability at least $1 - p$ over $(x_i)_{i=1}^N$ and $(W_k)_{k=1}^L$, where

$$p := C\,Ne^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}. \tag{223}$$

Hence, with probability at least $1 - p$ over $(x_i)_{i=1}^N$, there exists a set of parameters θ_0 such that $J(\theta_0)$ has a right inverse. Thus, for all $Y \in \mathbb{R}^N$, there exists θ' such that

$$J(\theta_0)\theta' = \left.\frac{\partial F_L(\theta)}{\partial \theta}\right|_{\theta=\theta_0} \theta' = Y. \tag{224}$$

This can also be written, for all $i \in [N]$, as

$$y_i = \left.\frac{\partial f_L(\theta, x_i)}{\partial \theta}\right|_{\theta=\theta_0}^{\top} \theta' = \lim_{h \to 0} \frac{f_L(\theta_0 + h\theta', x_i) - f_L(\theta_0, x_i)}{h}. \tag{225}$$

Then, for all $\varepsilon > 0$, there exists h^* such that, for all $i \in [N]$,

$$|y_i - f^*(x_i)| \leq \frac{\varepsilon}{\sqrt{N}}, \tag{226}$$

where

$$f^*(x_i) := \frac{f_L(\theta_0 + h^*\theta', x_i) - f_L(\theta_0, x_i)}{h^*}. \tag{227}$$

Finally, the desired result follows by noticing that f^* can be implemented by a network with the same depth and twice more neurons at every hidden layer. □

H Proof of Theorem 4.2

Notation for this appendix. In this appendix, we use $J(\theta)$ to denote the Jacobian of the network output F_L, evaluated in θ. We recall that $J(\theta)$ is a matrix with N rows and $\sum_{l=0}^{L-3} n_l n_{l+1} + 2n_{L-2}n_{L-1} + 2n_{L-1}$ columns (for the optimization result, we assume that the $(L-1)$-th layer has an even number of neurons and denote its width as $2n_{L-1}$). Let $K(\theta) = J(\theta)(J(\theta))^\top$ be the associated empirical NTK Gram matrix, and let θ_0 be the initialization defined in (18). We also make the dependence on θ explicit for feature vectors and backpropagation terms: the feature vector at the l-th layer with input x_i and network parameter θ is denoted by $f_l(\theta, x_i)$, and the corresponding backpropagation term is denoted by $b_l(\theta, x_i)$, where $b_l(\theta, x_i) = (B_l(\theta))_{i:}$. Finally, we use $W_l(\theta)$ to denote the weights of the l-th layer evaluated at the parameter θ.

A straightforward application of Theorem 2.1 in [49] gives the following proposition.

Proposition H.1. *Consider solving the least-squares optimization problem*

$$\min_\theta \mathcal{L}(\theta) := \frac{1}{2} \min_\theta \|F_L(\theta) - Y\|_2^2, \tag{228}$$

by running gradient descent updates of the form $\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}(\theta_t)$, with some initialization $\tilde{\theta}_0$. Assume there exists $\alpha, \beta \in \mathbb{R}$, such that, if we define $\mathcal{D} = \mathcal{B}(\tilde{\theta}_0, R)$ as the ℓ_2 ball centered in $\tilde{\theta}_0$ with radius R, with

$$R := \frac{4 \left\| F_L(\tilde{\theta}_0) - Y \right\|_2}{\alpha}, \tag{229}$$

the following holds

$$\forall \theta \in \mathcal{D} : \ \alpha \leq \sigma_{\min}(J(\theta)) \leq \|J(\theta)\|_{\text{op}} \leq \beta, \tag{230}$$

$$\forall \theta_1, \theta_2 \in \mathcal{D} : \ \|J(\theta_1) - J(\theta_2)\|_{\text{op}} \leq \frac{\alpha^2}{2\beta}. \tag{231}$$

Then, by setting $\eta \leq 1/(2\beta^2)$, we have that, for all $t \geq 1$,

$$\mathcal{L}(\theta_t) \leq \left(1 - \frac{\eta \alpha^2}{2}\right)^t \mathcal{L}(\tilde{\theta}_0). \tag{232}$$

In order to apply this proposition with initialization $\tilde{\theta}_0 = \theta_0$, we need to prove that the necessary assumptions hold. We will do so by showing the following intermediate results:

- Lemma H.2 shows that, at the initial point θ_0, the network output is 0 and the smallest NTK eigenvalue is lower bounded.
- Lemma H.3 gives a tight estimate on the operator norm of the weights inside a ball $\mathcal{D}$ centered at θ_0 and with radius $R = o(1)$.
- Lemma H.4 gives an upper bound on the distance between a feature vector in $\mathcal{D}$ and the feature vector at θ_0.
- Lemma H.5 gives upper bounds on the ℓ_2 norm and the ℓ_2 distance between feature vectors in $\mathcal{D}$.
- Lemmas H.6 and H.7 give upper bounds on the ℓ_2 norm and the ℓ_2 distance of backpropagation terms in $\mathcal{D}$, respectively.
- Lemma H.8 gives an upper bound on the difference in operator norm between Jacobians in $\mathcal{D}$.
- Finally, Lemma H.9 gives upper and lower bounds on the NTK spectrum in $\mathcal{D}$.

Lemma H.2 (Network output and smallest NTK eigenvalue at initialization)**.** *Let θ_0 be defined in* (18)*. Then, we have that, for all $x \in \mathbb{R}^d$,*

$$f_L(x, \theta_0) = 0. \tag{233}$$

Furthermore, we have that

$$\sigma_{\min}(J(\theta_0)) \geq c_1 \sqrt{\gamma n_{L-2} n_{L-1}}), \tag{234}$$

with probability at least $1 - C\,N e^{-c \log^2 n_{L-1}} - C e^{-c \log^2 N}$ over $(x_i)_{i=1}^N \sim_{\text{i.i.d.}} P_X$ and θ_0, where c, c_1 and C are numerical constants.

Proof. By definition (18) of the initialization θ_0, we have that

$$\begin{aligned} f_L(x, \theta_0) &= (W_L^{(1)}(\theta_0))^\top \phi((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x)) \\ &\quad + (W_L^{(2)}(\theta_0))^\top \phi((W_{L-1}^{(2)}(\theta_0))^\top f_{L-2}(\theta_0, x)) \\ &= (W_L^{(1)}(\theta_0))^\top \phi((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x)) \\ &\quad + (-W_L^{(1)}(\theta_0))^\top \phi((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x)) \\ &= 0, \end{aligned} \tag{235}$$

where in the second equality we use that $W_{L-1}^{(2)}(\theta_0) = W_{L-1}^{(1)}(\theta_0)$ and that $W_L^{(2)}(\theta_0) = -W_L^{(1)}(\theta_0)$. This proves (233).

Let us now compute the Jacobian at initialization $J(\theta_0)$. For $l \in [L-2]$, we have that

$$\begin{aligned}
\frac{\partial f_L(x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0} &= (W_L^{(1)}(\theta_0))^\top \Bigg(\phi'\left((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x) \right) \left((W_{L-1}^{(1)}(\theta_0))^\top \frac{\partial f_{L-2}(\theta, x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0} \right) \Bigg) \\
&+ (W_L^{(2)}(\theta_0))^\top \left(\phi'\left((W_{L-1}^{(2)}(\theta_0))^\top f_{L-2}(\theta_0, x) \right) \left((W_{L-1}^{(2)}(\theta_0))^\top \frac{\partial f_{L-2}(\theta, x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0} \right) \right) \\
&= (W_L^{(1)}(\theta_0))^\top \Bigg(\phi'\left((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x) \right) \left((W_{L-1}^{(1)}(\theta_0))^\top \frac{\partial f_{L-2}(\theta, x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0} \right) \Bigg) \\
&- (W_L^{(1)}(\theta_0))^\top \left(\phi'\left((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x) \right) \left((W_{L-1}^{(1)}(\theta_0))^\top \frac{\partial f_{L-2}(\theta, x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0} \right) \right) \\
&= 0,
\end{aligned} \tag{236}$$

where in the second equality we use again that $W_{L-1}^{(2)}(\theta_0) = W_{L-1}^{(1)}(\theta_0)$ and that $W_L^{(2)}(\theta_0) = -W_L^{(1)}(\theta_0)$.

Let us define $f_{L-1}^{(k)}(\theta, x) := \phi((W_{L-1}^{(k)}(\theta))^\top f_{L-2}(\theta, x))$ for $k \in \{1, 2\}$. Then, for the $(L-1)$-th layer, by isolating the computation over $W_{L-1}^{(1)}$, we have that

$$\begin{aligned}
\frac{\partial f_L(\theta, x)}{\partial (W_{L-1}^{(1)})_{ij}}\bigg|_{\theta=\theta_0} &= (W_L^{(1)}(\theta_0))^\top \frac{\partial f_{L-1}^{(1)}(\theta, x)}{\partial (W_{L-1}^{(1)})_{ij}}\bigg|_{\theta=\theta_0} + (W_L^{(2)}(\theta_0))^\top \frac{\partial f_{L-1}^{(2)}(\theta, x)}{\partial (W_{L-1}^{(1)})_{ij}}\bigg|_{\theta=\theta_0} \\
&= (W_L^{(1)}(\theta_0))^\top \frac{\partial f_{L-1}^{(1)}(\theta, x)}{\partial (W_{L-1}^{(1)})_{ij}}\bigg|_{\theta=\theta_0} \\
&=: J^{(1)}(\theta_0),
\end{aligned} \tag{237}$$

where we use that $f_{L-1}^{(2)}(\theta, x)$ does not depend on the parameters $W_{L-1}^{(1)}$. Proceeding in the same way and using that $W_{L-1}^{(2)}(\theta_0) = W_{L-1}^{(1)}(\theta_0)$ and $W_L^{(2)}(\theta_0) = -W_L^{(1)}(\theta_0)$, we also obtain that

$$\frac{\partial f_L(\theta, x)}{\partial (W_{L-1}^{(2)})_{ij}}\bigg|_{\theta=\theta_0} = -J^{(1)}(\theta_0). \tag{238}$$

Finally, by observing that $f_L(\theta, x) = (W_L^{(1)})^\top f_{L-1}^{(1)}(\theta, x) + (W_L^{(2)})^\top f_{L-1}^{(2)}(\theta, x)$, we deduce

$$\frac{\partial f_L(\theta, x)}{\partial (W_L^{(k)})_i}\bigg|_{\theta=\theta_0} = \left(f_{L-1}^{(k)}(\theta_0, x) \right)_i, \quad \text{for } k \in \{1, 2\}. \tag{239}$$

Hence, the NTK at initialization $K(\theta_0)$ can be expressed as

$$\begin{aligned}
K(\theta_0) = J^{(1)}(\theta_0)(J^{(1)}(\theta_0))^\top + J^{(1)}(\theta_0)(J^{(1)}(\theta_0))^\top \\
+ F_{L-1}^{(1)}(\theta_0)(F_{L-1}^{(1)}(\theta_0))^\top + F_{L-1}^{(2)}(\theta_0)(F_{L-1}^{(2)}(\theta_0))^\top.
\end{aligned} \tag{240}$$

Note that, by construction, $J^{(1)}(\theta_0)$ has the same distribution of J_{L-2}, whose rows are given by (10). Therefore, by combining the results from Theorem 3.2 and Theorem 3.4, we conclude that

$$\sigma_{\min}(J(\theta_0)) \geq 2\sqrt{\gamma}\sigma_{\min}(J_{L-2}) \geq c_1\sqrt{\gamma n_{L-2} n_{L-1}}, \tag{241}$$

with probability at least $1 - C\,N e^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}$ over $(x_i)_{i=1}^N$ and θ_0, where c_1, C are numerical constants. This proves (234), and concludes the proof of the lemma. □

Lemma H.3 (Operator norm of weights in $\mathcal{D}$)**.** *Let θ_0 be defined in* (18)*, let $\mathcal{D} = \mathcal{B}(\theta_0, R)$ and assume that $R = o(1)$. Then, for any $l \in [L-1]$,*

$$\sup_{\theta \in \mathcal{D}} \|W_l(\theta)\|_{\text{op}} = \mathcal{O}(1), \tag{242}$$

with probability at least $1 - 2\exp(-cn_{L-1})$ over $W_l(\theta_0)$.

Proof. By Weyl's theorem, we have that, for all $l \in [L-2]$,

$$\begin{aligned}
\sup_{\theta\in\mathcal{D}} \|W_l(\theta)\|_{\text{op}} &\leq \|W_l(\theta_0)\|_{\text{op}} + \sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_{\text{op}} \\
&\leq \|W_l(\theta_0)\|_{\text{op}} + \sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_F \\
&\leq \|W_l(\theta_0)\|_{\text{op}} + \sup_{\theta\in\mathcal{D}} \|\theta - \theta_0\|_2 \\
&= \|W_l(\theta_0)\|_{\text{op}} + o(1) \\
&= \mathcal{O}(1),
\end{aligned} \tag{243}$$

where in the fourth line we use that $R = o(1)$, and the result of the last line holds with probability at least $1 - 2\exp(-cn_{L-1})$ over $W_l(\theta_0)$ by Theorem 4.4.5 of [65]. By following the same argument, we have that, with probability at least $1 - 2\exp(-cn_{L-1})$ over $W_{L-1}(\theta_0)$,

$$\sup_{\theta\in\mathcal{D}} \left\|W_{L-1}^{(k)}(\theta)\right\|_{\text{op}} = \mathcal{O}(1), \quad \text{for } k \in \{1,2\}, \tag{244}$$

which readily implies that $\sup_{\theta\in\mathcal{D}} \|W_{L-1}(\theta)\|_{\text{op}} = \mathcal{O}(1)$ and concludes the proof. □

Lemma H.4 (Distance of features in $\mathcal{D}$ from initialization)**.** *Let θ_0 be defined in* (18)*, $x \sim P_X$, $\mathcal{D} = \mathcal{B}(\theta_0, R)$ and assume that $R = o(1)$. Then, for any $0 \leq l \leq L-1$, we have*

$$\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x) - f_l(\theta_0, x)\|_2 \leq C\, R\sqrt{d}, \tag{245}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^l$ and x, where c, C are numerical constants.

Proof. We prove the claim by induction over l. For the base case, we have $f_0(\theta, x) = f_0(\theta_0, x)$, hence (245) holds with probability 1.

For the induction case, let $l > 0$. Then,

$$\begin{aligned}
\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x) - f_l(\theta_0, x)\|_2 &= \sup_{\theta\in\mathcal{D}} \left\|\phi\left((W_l(\theta))^\top f_{l-1}(\theta, x)\right) - \phi\left((W_l(\theta_0))^\top f_{l-1}(\theta_0, x)\right)\right\|_2 \\
&\leq M \sup_{\theta\in\mathcal{D}} \left\|(W_l(\theta))^\top f_{l-1}(\theta, x) - (W_l(\theta_0))^\top f_{l-1}(\theta_0, x)\right\|_2 \\
&\leq M \sup_{\theta\in\mathcal{D}} \left\|(W_l(\theta))^\top f_{l-1}(\theta, x) - (W_l(\theta))^\top f_{l-1}(\theta_0, x)\right\|_2 \\
&\quad + M \sup_{\theta\in\mathcal{D}} \left\|(W_l(\theta))^\top f_{l-1}(\theta_0, x) - (W_l(\theta_0))^\top f_{l-1}(\theta_0, x)\right\|_2 \\
&\leq M \sup_{\theta\in\mathcal{D}} \|W_l(\theta)\|_{\text{op}} \sup_{\theta\in\mathcal{D}} \|f_{l-1}(\theta, x) - f_{l-1}(\theta_0, x)\|_2 \\
&\quad + M \sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_{\text{op}} \|f_{l-1}(\theta_0, x)\|_2 .
\end{aligned} \tag{246}$$

By Lemma H.3, we have that

$$\sup_{\theta\in\mathcal{D}} \|W_l(\theta)\|_{\text{op}} = \mathcal{O}(1), \tag{247}$$

with probability at least $1 - 2\exp(-cn_{L-1})$ over $W_l(\theta_0)$. By inductive hypothesis, we have

$$\sup_{\theta\in\mathcal{D}} \|f_{l-1}(\theta, x) - f_{l-1}(\theta_0, x)\|_2 \leq C\, R\sqrt{d}, \tag{248}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l-1}$ and x. Clearly, we also have that

$$\sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_{\text{op}} \leq \sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_F \leq \sup_{\theta\in\mathcal{D}} \|\theta - \theta_0\| \leq R. \tag{249}$$

Furthermore, an application of Lemma C.1 gives that

$$\|f_{l-1}(\theta_0, x)\|_2 = \Theta(\sqrt{n_{l-1}}) = \mathcal{O}\left(\sqrt{d}\right), \tag{250}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l-1}$ and x. By combining (246), (247), (248), (249) and (250), we obtain that

$$\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x) - f_l(\theta_0, x)\|_2 \leq C\,R\sqrt{d}, \tag{251}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and x, which completes the proof. □

Lemma H.5 (ℓ_2 norm and ℓ_2 distance of features in $\mathcal{D}$)**.** *Let θ_0 be defined in* (18)*, $x \sim P_X$, $\mathcal{D} = \mathcal{B}(\theta_0, R)$ and assume that $R = o(1)$. Then, for any $0 \leq l \leq L-1$, we have*

$$\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x)\|_2 = \mathcal{O}\left(\sqrt{d}\right), \tag{252}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and x, where c, C are numerical constants. Furthermore,

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1, x) - f_l(\theta_2, x)\|_2 \leq C\,R\,\sqrt{d}, \tag{253}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and x.

Proof. The first statement follows from the chain of inequalities below:

$$\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x)\|_2 \leq \|f_l(\theta_0, x)\|_2 + \sup_{\theta\in\mathcal{D}} \|f_l(\theta) - f_l(\theta_0)\|_2 \leq C\sqrt{n_l} + C\,R\sqrt{d}, \tag{254}$$

where the second inequality holds with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and x by combining Lemma C.1 and Lemma H.4.

We prove the second statement by induction over l. For the base case, we have $f_0(\theta_1, x) = f_0(\theta_2, x)$, hence (253) holds with probability 1.

For the induction case, let $l > 0$. Then,

$$\begin{aligned}
\sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1, x) - f_l(\theta_2, x)\|_2 &= \sup_{\theta_1,\theta_2\in\mathcal{D}} \left\|\phi\left((W_l(\theta_1))^\top f_{l-1}(\theta_1, x)\right) - \phi\left((W_l(\theta_2))^\top f_{l-1}(\theta_2, x)\right)\right\|_2 \\
&\leq M \sup_{\theta_1,\theta_2\in\mathcal{D}} \left\|(W_l(\theta_1))^\top f_{l-1}(\theta_1, x) - (W_l(\theta_2))^\top f_{l-1}(\theta_2, x)\right\|_2 \\
&\leq M \sup_{\theta_1,\theta_2\in\mathcal{D}} \left\|(W_l(\theta_1))^\top f_{l-1}(\theta_1, x) - (W_l(\theta_1))^\top f_{l-1}(\theta_2, x)\right\|_2 \\
&\quad + M \sup_{\theta_1,\theta_2\in\mathcal{D}} \left\|(W_l(\theta_1))^\top f_{l-1}(\theta_2, x) - (W_l(\theta_2))^\top f_{l-1}(\theta_2, x)\right\|_2 \\
&\leq M \sup_{\theta_1\in\mathcal{D}} \|W_l(\theta_1)\|_{\text{op}} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_{l-1}(\theta_1, x) - f_{l-1}(\theta_2, x)\|_2 \\
&\quad + M \sup_{\theta_1,\theta_2\in\mathcal{D}} \|W_l(\theta_1) - W_l(\theta_2)\|_{\text{op}} \sup_{\theta_2\in\mathcal{D}} \|f_{l-1}(\theta_2, x)\|_2 .
\end{aligned} \tag{255}$$

By Lemma H.3, we have that

$$\sup_{\theta_1\in\mathcal{D}} \|W_l(\theta_1)\|_{\text{op}} = \mathcal{O}(1), \tag{256}$$

with probability at least $1 - 2\exp(-cn_{L-1})$ over $W_l(\theta_0)$. By inductive hypothesis, we have

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_{l-1}(\theta_1, x) - f_{l-1}(\theta_2, x)\|_2 \leq C\,R\,\sqrt{d}, \tag{257}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l-1}$ and x. Clearly, we also have that

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|W_l(\theta_1) - W_l(\theta_2)\|_{\text{op}} \leq \sup_{\theta_1,\theta_2\in\mathcal{D}} \|W_l(\theta_1) - W_l(\theta_2)\|_F \leq \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\theta_1 - \theta_2\| \leq R. \tag{258}$$

Furthermore, by using (252), we have that

$$\sup_{\theta_2\in\mathcal{D}} \|f_{l-1}(\theta_2, x)\|_2 \leq C\,R\,\sqrt{d}, \tag{259}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l-1}$ and x. By combining (255), (256), (257), (258) and (259), we obtain that

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1, x) - f_l(\theta_2, x)\|_2 \le C\,R\sqrt{d}, \tag{260}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and x, which completes the proof. □

Lemma H.6 (ℓ_2 norm of backpropagation in $\mathcal{D}$)**.** *Let θ_0 be defined in* (18)*, $x \sim P_X$, and $\mathcal{D} = \mathcal{B}(\theta_0, R)$. Assume that $R = o(1)$ and that $\gamma > 1$. Then, for any $l \in [L]$, we have*

$$\sup_{\theta\in\mathcal{D}} \|b_l(\theta, x)\|_2 \le C\sqrt{\gamma \cdot n_{L-1}}, \tag{261}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^{L}$, where c, C are numerical constants.

Proof. We prove the claim by induction on $l \in \{L, L-1, \dots, 1\}$. For the base case, we have that $\|b_L(\theta, x)\|_2 = 1$, hence (261) clearly holds.

For the induction case, pick $l \in [L-1]$. Then,

$$\begin{aligned}
\sup_{\theta\in\mathcal{D}} \|b_l(\theta, x)\|_2 &= \sup_{\theta\in\mathcal{D}} \left\| \left(\prod_{k=l}^{L-2} \Sigma_k(\theta, x) W_{k+1}(\theta) \right) \Sigma_{L-1}(\theta, x) W_L(\theta) \right\|_2 \\
&\le \sup_{\theta\in\mathcal{D}} \left\| \left(\prod_{k=l}^{L-2} \Sigma_k(\theta, x) W_{k+1}(\theta) \right) \Sigma_{L-1}(\theta, x) \right\|_{\text{op}} \sup_{\theta\in\mathcal{D}} \|W_L(\theta)\|_2 \\
&\le \left(\prod_{k=l}^{L-2} \sup_{\theta\in\mathcal{D}} \|\Sigma_k(\theta, x)\|_{\text{op}} \sup_{\theta\in\mathcal{D}} \|W_{k+1}(\theta)\|_{\text{op}} \right) \sup_{\theta\in\mathcal{D}} \|\Sigma_{L-1}(\theta, x)\|_{\text{op}} \sup_{\theta\in\mathcal{D}} \|W_L(\theta)\|_2 \\
&\le M^{L-l} \left(\prod_{k=l+1}^{L-1} \sup_{\theta\in\mathcal{D}} \|W_k(\theta)\|_{\text{op}} \right) \left(\|W_L(\theta_0)\|_2 + \sup_{\theta\in\mathcal{D}} \|W_L(\theta) - W_L(\theta_0)\|_2 \right) \\
&\le C\,M^{L-l} (\|W_L(\theta_0)\|_2 + \sup_{\theta\in\mathcal{D}} \|\theta - \theta_0\|_2) \\
&\le C\,M^{L-l} (\sqrt{\gamma n_{L-1}} + \sup_{\theta\in\mathcal{D}} \|\theta - \theta_0\|_2) \\
&= C\sqrt{\gamma n_{L-1}}.
\end{aligned} \tag{262}$$

Here, the fourth line follows from Assumption 2.3, which gives $\sup_{\theta\in\mathcal{D}} \|\Sigma_k(\theta, x)\|_{\text{op}} \le M$; the fifth line holds with probability $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^{L-1}$ by Lemma H.3; the sixth line holds with probability at least $1 - \exp(-cn_{L-1})$ over $W_L(\theta_0)$ by Theorem 3.1.1 in [65]; and the last line follows from $R = o(1)$. Taking the intersection of these events gives the desired result. □

Lemma H.7 (ℓ_2 distance of backpropagation in $\mathcal{D}$)**.** *Let θ_0 be defined in* (18)*, $x \sim P_X$, and $\mathcal{D} = \mathcal{B}(\theta_0, R)$. Assume that $R = o(1)$ and that $\gamma > 1$. Then, for any $l \in [L]$, we have*

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_l(\theta_1, x) - b_l(\theta_2, x)\|_2 \le C\,R\,\sqrt{\gamma d n_{L-1}}, \tag{263}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^{L}$ and x, where c, C are numerical constants.

Proof. We prove the claim by induction on $l \in \{L, L-1, \ldots, 1\}$. For the base case, $b_L(\theta, x)$ does not depend on θ, hence (263) clearly holds. For the induction case, pick $l \in [L-1]$. Then,

$$
\begin{aligned}
&\sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_l(\theta_1,x) - b_l(\theta_2,x)\|_2 \\
&\quad= \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_1,x)W_{l+1}(\theta_1)b_{l+1}(\theta_1,x) - \Sigma_l(\theta_2,x)W_{l+1}(\theta_2)b_{l+1}(\theta_2,x)\|_2 \\
&\quad\le \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_1,x)W_{l+1}(\theta_1)b_{l+1}(\theta_1,x) - \Sigma_l(\theta_1,x)W_{l+1}(\theta_1)b_{l+1}(\theta_2,x)\|_2 \\
&\qquad+ \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_1,x)W_{l+1}(\theta_1)b_{l+1}(\theta_2,x) - \Sigma_l(\theta_2,x)W_{l+1}(\theta_2)b_{l+1}(\theta_2,x)\|_2 \\
&\quad\le \sup_{\theta_1\in\mathcal{D}} \|\Sigma_l(\theta_1,x)\|_{\text{op}} \sup_{\theta_1\in\mathcal{D}} \|W_{l+1}(\theta_1)\|_{\text{op}} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_1,x) - b_{l+1}(\theta_2,x)\|_2 \\
&\qquad+ \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_1,x)W_{l+1}(\theta_1) - \Sigma_l(\theta_2,x)W_{l+1}(\theta_2)\|_{\text{op}} \sup_{\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_2,x)\|_2 \\
&\quad\le \sup_{\theta_1\in\mathcal{D}} \|\Sigma_l(\theta_1,x)\|_{\text{op}} \sup_{\theta_1\in\mathcal{D}} \|W_{l+1}(\theta_1)\|_{\text{op}} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_1,x) - b_{l+1}(\theta_2,x)\|_2 \\
&\qquad+ \sup_{\theta_1,\theta_2\in\mathcal{D}} \|(\Sigma_l(\theta_1,x) - \Sigma_l(\theta_2,x))\, W_{l+1}(\theta_2)\|_{\text{op}} \sup_{\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_2,x)\|_2 \\
&\qquad+ \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_2,x)\,(W_{l+1}(\theta_1) - W_{l+1}(\theta_2))\|_{\text{op}} \sup_{\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_2,x)\|_2 .
\end{aligned}
\tag{264}
$$

Furthermore, we have that the following results hold.

(i) By Assumption 2.3 and Lemma H.3,

$$\sup_{\theta_1\in\mathcal{D}} \|\Sigma_l(\theta_1,x)\|_{\text{op}} \sup_{\theta_1\in\mathcal{D}} \|W_{l+1}(\theta_1)\|_{\text{op}} = \mathcal{O}(1),$$

with probability $1 - 2\exp(-cn_{L-1})$ over $W_{l+1}(\theta_0)$;

(ii) By inductive hypothesis,

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_1,x) - b_{l+1}(\theta_2,x)\|_2 \le C\,R\,\sqrt{\gamma d n_{L-1}},$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=l+2}^{L}$ and x;

(iii) By the same argument of the second statement in Lemma H.5 and again Lemma H.3,

$$
\begin{aligned}
&\sup_{\theta_1,\theta_2\in\mathcal{D}} \|(\Sigma_l(\theta_1,x) - \Sigma_l(\theta_2,x))\, W_{l+1}(\theta_2)\|_{\text{op}} \\
&\qquad\le \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\phi'(g_l(\theta_1,x)) - \phi'(g_l(\theta_2,x))\|_2 \sup_{\theta_2\in\mathcal{D}} \|W_{l+1}(\theta_2)\|_{\text{op}} \\
&\qquad\le C\,R\,\sqrt{d},
\end{aligned}
$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l+1}$ and x;

(iv) By Lemma H.6,

$$\sup_{\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_2,x)\|_2 \le C\,\sqrt{\gamma n_{L-1}},$$

with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^{L}$;

(v) By Assumption 2.3,

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_2,x)\,(W_{l+1}(\theta_1) - W_{l+1}(\theta_2))\|_{\text{op}} \le C\,R.$$

By combining (i)-(v) with (264), we conclude that

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_l(\theta_1,x) - b_l(\theta_2,x)\|_2 \le C\,R\,\sqrt{\gamma d n_{L-1}}, \tag{265}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over x and $(W_k(\theta_0))_{k=1}^{L}$, which concludes the proof. □

Lemma H.8 (Difference of Jacobians in $\mathcal{D}$)**.** *Let θ_0 be defined in* (18)*, $x \sim P_X$, and $\mathcal{D} = \mathcal{B}(\theta_0, R)$. Assume that $R = o(1)$ and that $\gamma > 1$. Then, we have*

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|J(\theta_1) - J(\theta_2)\|_{\text{op}} \leq C\, R\, d\sqrt{\gamma n_{L-1} N}, \tag{266}$$

with probability at least $1 - CN\exp(-cn_{L-1})$ over $(x_i)_{i=1}^N$ and θ_0, where c, C are numerical constants.

Proof. Pick $i \in [N]$. Then, we have

$$\begin{aligned}
&\sup_{\theta_1,\theta_2\in\mathcal{D}} \|(J(\theta_1))_{i:} - (J(\theta_2))_{i:}\|_2^2 \\
&\quad\leq \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|(F_l(\theta_1))_{:i} \otimes (B_{l+1}(\theta_1))_{:i} - (F_l(\theta_2))_{:i} \otimes (B_{l+1}(\theta_2))_{:i}\|_2^2 \\
&\quad= \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1, x_i) \otimes b_{l+1}(\theta_1, x_i) - f_l(\theta_2, x_i) \otimes b_{l+1}(\theta_2, x_i)\|_2^2 \\
&\quad\leq \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|(f_l(\theta_1, x_i) - f_l(\theta_2, x_i)) \otimes b_{l+1}(\theta_1, x_i)\|_2^2 \\
&\qquad+ \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_2, x_i) \otimes (b_{l+1}(\theta_1, x_i) - b_{l+1}(\theta_1, x_i))\|_2^2 \\
&\quad\leq \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1, x_i) - f_l(\theta_2, x_i)\|_2^2 \sup_{\theta_1\in\mathcal{D}} \|b_{l+1}(\theta_1, x_i)\|_2^2 \\
&\qquad+ \sum_{l=0}^{L-1} \sup_{\theta_2\in\mathcal{D}} \|f_l(\theta_2, x_i)\|_2^2 \sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_1, x_i) - b_{l+1}(\theta_1, x_i)\|_2^2 .
\end{aligned} \tag{267}$$

Since $x_i \sim P_X$, we can merge together the results from Lemmas H.4, H.5, H.6 and H.7 and obtain

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|(J(\theta_1))_{i:} - (J(\theta_2))_{i:}\|_2^2 \leq C\gamma R^2 d^2 n_{L-1}, \tag{268}$$

with probability at least $1 - C\exp(-cn_{L-1})$ over x_i and θ_0.

Therefore, we have

$$\begin{aligned}
\sup_{\theta_1,\theta_2\in\mathcal{D}} \|J(\theta_1) - J(\theta_2)\|_{\text{op}} &\leq \sup_{\theta_1,\theta_2\in\mathcal{D}} \|J(\theta_1) - J(\theta_2)\|_F \\
&\leq \sqrt{\sum_{i=1}^N \sup_{\theta_1,\theta_2\in\mathcal{D}} \|(J(\theta_1))_{i:} - (J(\theta_2))_{i:}\|_2^2} \\
&\leq C\, R\, d\sqrt{\gamma n_{L-1} N},
\end{aligned} \tag{269}$$

with probability $1 - CN\exp(-cn_{L-1})$ over $(x_i)_{i=1}^N$ and θ_0. □

Lemma H.9 (NTK spectrum in $\mathcal{D}$)**.** *Let θ_0 be defined in* (18)*, $x \sim P_X$, and $\mathcal{D} = \mathcal{B}(\theta_0, R)$. Assume that $R = o(1)$ and that $\gamma > 1$. Then, we have*

$$\sup_{\theta\in\mathcal{D}} \|K(\theta)\|_{\text{op}} \leq C\,\gamma\, N\, d\, n_{L-1}, \tag{270}$$

with probability at least $1 - C\,N\exp(-cn_{L-1})$ over θ_0 and $(x_i)_{i=1}^N$, where c, C are numerical constants. Furthermore,

$$\inf_{\theta\in\mathcal{D}} \sigma_{\min}(J(\theta)) \geq c_1\sqrt{\gamma n_{L-2} n_{L-1}} - C_1\, R\, d\sqrt{\gamma n_{L-1} N}, \tag{271}$$

with probability at least $1 - C\,Ne^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}$ over θ_0 and $(x_i)_{i=1}^N$, where c_1, C_1 are also numerical constants.

Proof. We have

$$\begin{aligned}
\sup_{\theta\in\mathcal{D}} \|K(\theta)\|_{\text{op}} &= \sup_{\theta\in\mathcal{D}} \left\| \sum_{l=0}^{L-1} F_l(\theta)F_l^\top(\theta) \circ B_{l+1}(\theta)B_{l+1}^\top(\theta) \right\|_{\text{op}} \\
&\leq \sum_{l=0}^{L-1} \sup_{\theta\in\mathcal{D}} \left\| F_l(\theta)F_l^\top(\theta) \circ B_{l+1}(\theta)B_{l+1}^\top(\theta) \right\|_{\text{op}} \\
&\leq \sum_{l=0}^{L-1} \sup_{\theta\in\mathcal{D}} \left\| F_l(\theta)F_l^\top(\theta) \right\|_{\text{op}} \sup_{\theta\in\mathcal{D}} \max_{i\in[N]} \|(B_{l+1}(\theta))_{i:}\|_2^2 \\
&\leq \sum_{l=0}^{L-1} \sup_{\theta\in\mathcal{D}} \|F_l(\theta)\|_F^2 \sup_{\theta\in\mathcal{D}} \max_{i\in[N]} \|b_{l+1}(\theta, x_i)\|_2^2 \\
&\leq \sum_{l=0}^{L-1} \left(\sum_{i=1}^{N} \sup_{\theta\in\mathcal{D}} \|f_l(\theta, x_i)\|_2^2 \right) \sup_{\theta\in\mathcal{D}} \max_i \|b_{l+1}(\theta, x_i)\|_2^2 .
\end{aligned} \tag{272}$$

By Lemma H.5, we have that $\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x_i)\|_2^2 \leq C\, d$ with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=1}^l$ and x_i, for any $0 \leq l \leq L-1$ and $i \in [N]$. By Lemma H.6, we have that $\sup_{\theta\in\mathcal{D}} \|b_l(\theta, x_i)\|_2^2 \leq C\gamma n_{L-1}$ with probability at least $1 - C\exp(-cn_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^L$, for any $l \in [L]$ and $i \in [N]$. Therefore, we obtain

$$\sup_{\theta\in\mathcal{D}} \|K(\theta)\|_{\text{op}} \leq C\,\gamma\, N\, d\, n_{L-1}, \tag{273}$$

with probability at least $1 - C\,N\exp(-cn_{L-1})$ over θ_0 and $(x_i)_{i=1}^N$, which gives the first statement of the lemma.

By using Weyl's inequality, we get

$$\begin{aligned}
\inf_{\theta\in\mathcal{D}} \sigma_{\min}(J(\theta)) &\geq \sigma_{\min}(J(\theta_0)) - \sup_{\theta\in\mathcal{D}} \|J(\theta_1) - J(\theta_0)\|_{\text{op}} \\
&\geq c_1\sqrt{\gamma n_{L-2}n_{L-1}} - C_1\, R\, d\, \sqrt{\gamma\, n_{L-1}\, N},
\end{aligned} \tag{274}$$

where the last inequality follows from Lemma H.2 and Lemma H.8, and it holds with probability $1 - C\,Ne^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}$ over $(x_i)_{i=1}^N$ and θ_0. This gives the second statement of the lemma and concludes the proof. □

Armed with Proposition H.1 and the intermediate estimates of Lemmas H.2-H.9, we are finally ready to prove Theorem 4.2.

Proof of Theorem 4.2. We show that there exist two absolute constants $\tilde{c}$ and $\tilde{C}$ such that

$$\alpha = \tilde{c}\sqrt{\gamma n_{L-2}n_{L-1}} \tag{275}$$

and

$$\beta = \tilde{C}\sqrt{\gamma N d n_{L-1}} \tag{276}$$

satisfy the two assumptions in Proposition H.1 with initialization $\tilde{\theta}_0 := \theta_0$, where θ_0 is defined in (18). This holds with probability at least $1 - C\,Ne^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}$ over $(x_i)_{i=1}^N$ and θ_0.

Recall from Proposition H.1 that R is defined as $4\,\|F_L(\theta_0) - Y\|_2\,/\alpha$, since we have set $\tilde{\theta}_0 = \theta_0$. For the moment, we assume that

$$R = \mathcal{O}\left(\sqrt{\frac{N}{\gamma n_{L-2}n_{L-1}}}\right), \tag{277}$$

and we will verify that this is the case later. Note that $\gamma = d^3N^2 > 1$ and, hence, (277) and Assumption 2.5 imply that $R = o(1)$. Thus, we can apply Lemma H.9 and obtain

$$\inf_{\theta\in\mathcal{D}} \sigma_{\min}(J(\theta)) \geq c_1\sqrt{\gamma n_{L-2}n_{L-1}} - C_1\, R\, d\sqrt{\gamma n_{L-1}N} \geq \tilde{c}\sqrt{\gamma n_{L-2}n_{L-1}}, \tag{278}$$

with probability at least $1 - C\,N e^{-c\log^2 n_{L-1}} - Ce^{-c\log^2 N}$ over $(x_i)_{i=1}^N$ and θ_0, where the last inequality uses (277). This shows that the lower bound in (230) holds.

Now, by using (278), we verify that (277) holds. Recall that, by assumption of the theorem, $\|Y\|_2 = \Theta(\sqrt{N})$. Furthermore, by Lemma H.2, $F_L(\theta_0)$ is a vector of all zeros. Then,

$$R = \frac{4\,\|F_L(\theta_0) - Y\|_2}{\alpha} = \frac{4\,\|Y\|_2}{\alpha} = \mathcal{O}\left(\sqrt{\frac{N}{\gamma n_{L-2} n_{L-1}}}\right). \tag{279}$$

By Lemma H.9, we have that

$$\sup_{\theta\in\mathcal{D}} \|J(\theta)\|_{\text{op}} \leq C\sqrt{\gamma N d n_{L-1}}, \tag{280}$$

with probability at least $1 - C\,N\exp(-cn_{L-1})$ over θ_0. Thus, by our choice (276) of β, we obtain that the upper bound in (230) holds.

Next, we verify the second assumption of Proposition H.1. To do so, let us write

$$\frac{\alpha^2}{2\beta} = \frac{\tilde{c}^2 n_{L-2} n_{L-1}\gamma}{2\tilde{C}\sqrt{\gamma N d n_{L-1}}} = \Omega(\sqrt{n_{L-2}}Nd), \tag{281}$$

where we have used Assumption 2.5. Thus,

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|J(\theta_1) - J(\theta_2)\|_{\text{op}} \leq C\,R\,d\sqrt{\gamma n_{L-1} N} = \mathcal{O}\left(\frac{dN}{\sqrt{n_{L-2}}}\right) \leq \frac{\alpha^2}{2\beta}, \tag{282}$$

with probability at least $1 - C\,N\exp(-cn_{L-1})$ over $(x_i)_{i=1}^N$ and θ_0. Here, the first passage follows from Lemma H.8, in the second passage we use (277), and in the last one we use (281). This completes the proof of (231) and also of the theorem, since the desired claim follows from an application of Proposition H.1. □