Tight Bounds on the Smallest Eigenvalue of the Neural Tangent Kernel for Deep ReLU Networks

Quynh Nguyen [1] **Marco Mondelli** [2] **Guido Montufar** [1 3]

Abstract

A recent line of work has analyzed the theoretical properties of deep neural networks via the Neural Tangent Kernel (NTK). In particular, the smallest eigenvalue of the NTK has been related to the memorization capacity, the global convergence of gradient descent algorithms and the generalization of deep nets. However, existing results either provide bounds in the two-layer setting or assume that the spectrum of the NTK matrices is bounded away from 0 for multi-layer networks. In this paper, we provide tight bounds on the smallest eigenvalue of NTK matrices for deep ReLU nets, both in the limiting case of infinite widths and for finite widths. In the finite-width setting, the network architectures we consider are fairly general: we require the existence of a wide layer with roughly order of N neurons, N being the number of data samples; and the scaling of the remaining layer widths is arbitrary (up to logarithmic factors). To obtain our results, we analyze various quantities of independent interest: we give lower bounds on the smallest singular value of hidden feature matrices, and upper bounds on the Lipschitz constant of input-output feature maps.

1. Introduction

Consider an L-layer ReLU network with feature maps $f_l : \mathbb{R}^d \to \mathbb{R}^{n_l}$ defined for every $x \in \mathbb{R}^d$ as

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

where $W_l \in \mathbb{R}^{n_{l-1} \times n_l}$, $\sigma(x) = \max(0, x)$ and, given an integer n, we use the shorthand $[n] = \{1, \dots, n\}$.

[1]MPI-MIS, Germany [2]IST, Austria [3]UCLA. Correspondence to: Quynh Nguyen <quynhnguyenngoc89@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

We assume that the network has a single output, namely $n_L = 1$ and $W_L \in \mathbb{R}^{n_{L-1}}$. For consistency, let $n_0 = d$. Let $g_l : \mathbb{R}^d \to \mathbb{R}^{n_l}$ be the pre-activation feature map so that $f_l(x) = \sigma(g_l(x))$. Let $(x_1, \dots, x_N)$ be N samples in $\mathbb{R}^d$, $\theta = [\mathrm{vec}(W_1), \dots, \mathrm{vec}(W_L)]$, and $F_L(\theta) = [f_L(x_1), \dots, f_L(x_N)]^T$. Let J be the Jacobian of F_L with respect to all the weights:

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

If not mentioned otherwise, we will assume throughout the paper that all the partial derivatives are computed by the standard back-propagation with the convention that $\sigma'(0) = 0$. The empirical Neural Tangent Kernel (NTK) Gram matrix, denoted by $\bar{K}^{(L)} \in \mathbb{R}^{N \times N}$, is defined as:

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

As shown in (Jacot et al., 2018), when $(W_l)_{ij} \sim \mathcal{N}(0,1)$ for all $l \in [L]$ and $\min\{n_1, \dots, n_{L-1}\} \to \infty$, the normalized NTK matrix converges in probability to a non-random limit, called the limiting NTK matrix:

$$\left(\prod_{l=1}^{L-1} \frac{2}{n_l} \right) \bar{K}^{(L)} \xrightarrow{p} K^{(L)}. \tag{4}$$

A quantitative bound for the convergence rate is provided in (Arora et al., 2019b). Several theoretical aspects of training neural networks have been related to the spectrum of the NTK matrices. For instance, considering the square loss $\Phi(\theta) = \frac{1}{2} \|F_L - Y\|_2^2$, then a simple calculation shows that

$$\|\nabla \Phi(\theta)\|_2^2 \geq \lambda_{\min}\left(\bar{K}^{(L)}\right) 2\Phi(\theta). \tag{5}$$

The idea is that, if the spectrum of $\bar{K}^{(L)}$ is bounded away from zero at initialization, then under suitable conditions, one can show that this property continues to hold during training. In that case, $\lambda_{\min}\left(\bar{K}^{(L)}\right)$ from (5) can be replaced by a positive constant, and thus minimizing the gradient on the LHS will drive the loss to zero. This property, together with other smoothness conditions of the loss, has

been used for proving the global convergence of gradient descent in many prior works: (Du et al., 2019b; Oymak & Soltanolkotabi, 2020; Song & Yang, 2020; Wu et al., 2019) consider two layer nets, (Allen-Zhu et al., 2019; Du et al., 2019a; Zou et al., 2020; Zou & Gu, 2019) consider deep nets with polynomially wide layers, and most recently (Nguyen & Mondelli, 2020) consider deep nets with one wide layer of linear width followed by a pyramidal shape. Beside optimization, the smallest eigenvalue of the NTK has been used to prove generalization bounds (Arora et al., 2019a; Montanari & Zhong, 2020) and memorization capacity (Montanari & Zhong, 2020). All these analyses show that understanding the scaling of the smallest eigenvalue of the NTK is a problem of fundamental importance.

The recent work (Fan & Wang, 2020) characterizes the full spectrum of the limiting NTK via an iterated Marchenko-Pastur map. Yet, this does not have implications on the scaling of any individual eigenvalue. (Montanari & Zhong, 2020) gives a quantitative lower bound on $\lambda_{\min}\left(\bar{K}^{(L)}\right)$ in a regime in which the number of parameters scales linearly with N. This result is particularly interesting but currently restricted to a two-layer setup. To our knowledge, for multi-layer architectures, the fact that the spectrum of the NTK is bounded away from zero is a typical working assumption (Du et al., 2019a; Huang & Yau, 2020).

Main contributions. The aim of this paper is to provide tight lower bounds on the smallest eigenvalues of the empirical NTK matrices for deep ReLU networks.

First, we consider the asymptotic setting. For i.i.d. data from a class of distributions that satisfy a Lipschitz concentration property and for $(W_l)_{ij} \sim \mathcal{N}(0,1)$, we show that the smallest eigenvalue of the limiting NTK matrix scales as

$$L\mathcal{O}(d) \geq \lambda_{\min}\left(K^{(L)}\right) \geq \Omega(d), \tag{6}$$

where d captures the scaling of the average L_2 norm of the data [1]. This result is proved in our Theorem 3.2.

Next, we consider networks with large but *finite* widths, and fixed depth. Let ξ_l be an auxiliary variable which takes value 1 if $n_l = \tilde{\Omega}(N)$ and 0 otherwise, where N is the number of data points and $\tilde{\Omega}$ neglects logarithmic factors. Then for $(W_l)_{ij} \sim \mathcal{N}(0, \beta_l^2)$, we show that

$$\mathcal{O}\left(\left(d\prod_{l=1}^{L-1} n_l\right)\left(\prod_{l=1}^{L}\beta_l^2\right)\left(\sum_{l=1}^{L}\beta_l^{-2}\right)\right) \geq \lambda_{\min}\left(\bar{K}^{L}\right)$$
$$\geq \Omega\left(\left(d\prod_{l=1}^{L-1} n_l\right)\left(\prod_{l=1}^{L}\beta_l^2\right)\left(\sum_{l=1}^{L}\xi_{l-1}\beta_l^{-2}\right)\right). \tag{7}$$

[1] As introduced later, d is also the input dimension. However, only the scaling of the data matters for our bounds.

This is proved in Theorem 4.1. Our result directly implies that the spectrum of the NTK matrix is bounded away from zero whenever the network contains one wide layer of order N. This holds regardless of the position of the wide layer and the widths of the remaining ones (up to log factors). The last property allows for networks with bottleneck layers.

Comparing the lower and upper bounds of (7), we note that they only differ in the scaling of $\sum_{l=1}^{L}\beta_l^{-2}$ and $\sum_{l=1}^{L}\xi_{l-1}\beta_l^{-2}$. Let $k = \arg\min_{l\in[L-1]}\beta_l$. Then, as long as $\xi_{k-1} = 1$, both the sums will scale as β_k^{-2}. In that case, the lower bound in (7) is tight (up to a multiplicative constant). For instance, this occurs if *(i)* the network has one wide layer with $\tilde{\Omega}(N)$ neurons, and *(ii)* it is initialized under He's initialization (i.e., $\beta_l = \sqrt{2/n_{l-1}}$) or LeCun's initialization (i.e., $\beta_l = 1/\sqrt{n_{l-1}}$) (Glorot & Bengio, 2010; He et al., 2015; LeCun et al., 2012). Note also that our bound for finite widths is consistent with the asymptotic one in (6) (except that we do not track the dependence on L in (7)).

During the proof of our main theorems, we obtain other intermediate results which could be of independent interest:

- We give a tight bound on the smallest singular value of the feature matrices $F_k = [f_k(x_1), \ldots, f_k(x_N)]^T \in \mathbb{R}^{N\times n_k}$, for $k \in [L-1]$. Our analysis requires only a single wide layer, i.e. $n_k = \tilde{\Omega}(N)$, while all the previous layers can have *sublinear* widths.
- We obtain a new bound on the Lipschitz constant of the feature maps f_k's for random Gaussian weights. This bound is tighter than the one typically appearing in the literature (as given by the product of the operator norms of all the layers). The proof exploits a novel characterization of the Lipschitz constant of these maps, and leverages existing bounds on the number of activation patterns of deep ReLU nets.

This analysis allows us to prove the main results for a fairly general class of network shapes: there exists a layer with order of N neurons in an *arbitrary* position, and all the remaining layers can have *sublinear* widths, see Figure 1. No special ordering or relation between the scalings of these layers is needed. This goes beyond the setting of the typical NTK regime, where all the layers of the network have $\mathrm{poly}(N)$ neurons.

2. Preliminaries

Notations. The following notations are used throughout the paper: given two integers $n < m$, let $[n, m] = \{n, n+1, \ldots, m\}$; $X = [x_1, \ldots, x_N]^T \in \mathbb{R}^{N\times d}$; the feature matrix at layer l is $F_l = [f_l(x_1), \ldots, f_l(x_N)]^T \in \mathbb{R}^{N\times n_l}$; the centered feature matrices are $\tilde{F}_l = F_l - \mathbb{E}_X[F_l]$ for $l \in [L-1]$, where the expectation is taken over all the sam-

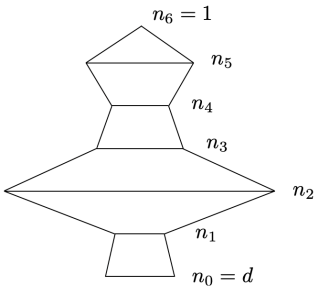

Figure 1. Illustration of a network architecture to which our results can be applied (and that does not fall in the typical NTK regime).

ples; $\Sigma_l(x) = \mathrm{diag}([\sigma'(g_{l,j}(x))]_{j=1}^{n_l})$ for $l \in [L-1]$, where $g_{l,j}(x)$ is the pre-activation neuron. Given two matrices $A, B \in \mathbb{R}^{m\times n}$, we denote by $A \circ B$ their Hadamard product, and by $A * B = [(A_{1:} \otimes B_{1:}), \dots, (A_{m:} \otimes B_{m:})]^T \in \mathbb{R}^{m\times n^2}$ their row-wise Khatri-Rao product. Let $\|A\|_{\mathrm{op}}$ be the operator norm of the matrix A. Given a p.s.d. matrix A, we denote by $\sqrt{A}$ its square root (i.e. $\sqrt{A} = \sqrt{A}^T$ and $\sqrt{A}\sqrt{A} = A$). We denote by $\|f\|_{\mathrm{Lip}}$ the Lipschitz constant of the function f. All the complexity notations $\Omega(\cdot)$ and $\mathcal{O}(\cdot)$ are understood for sufficiently large $N, d, n_1, n_2, \dots, n_{L-1}$. If not mentioned otherwise, the depth L is considered a constant.

Hermite expansion. Our bounds depend on the r-th Hermite coefficient of the ReLU activation function σ. Let us denote it by $\mu_r(\sigma)$. By standard calculations, we have for any even integer $r \geq 2$,

$$\mu_r(\sigma) = \frac{1}{\sqrt{2\pi}}(-1)^{\frac{r-2}{2}}\frac{(r-3)!!}{\sqrt{r!}}. \tag{8}$$

Weight and data distribution. We consider the setting where both the weights of the network and the data are random. In particular, $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2)$ for all $l \in [L], i \in [n_{l-1}], j \in [n_l]$, where the variable β_l may depend on layer widths. Throughout the paper, we let $(x_1, \dots, x_N)$ be N i.i.d. samples from a data distribution, say P_X, such that the following conditions are satisfied.

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. $\int \left\|x - \int x' \, dP_X(x')\right\|_2^2 dP_X(x) = \Omega(d)$.

These are just scaling conditions on the data vector x or its centered counterpart $x - \mathbb{E}x$. We remark that the data can have any scaling, but in this paper we fix it to be of order d for convenience. We further assume the following condition on the data distribution.

Assumption 2.2 (Lipschitz concentration) *The data distribution P_X satisfies the* Lipschitz concentration property. *Namely, for every Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$, there exists an absolute constant $c > 0$ such that, for all $t > 0$,*

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

In general, Assumption 2.2 covers the whole family of distributions which satisfies the log-Sobolev inequality with a dimension-independent constant (or distributions with log-concave densities). This includes, for instance, the standard Gaussian distribution, the uniform distribution on the sphere, or uniform distributions on the unit (binary or continuous) hypercube (Vershynin, 2018). Let us remark that the coordinates of a random sample need not be independent under the above assumptions. Note also that, by applying a Lipschitz map to the data, Assumption 2.2 still holds. Thus, data produced via a Generative Adversarial Network (GAN) fulfills our assumption, see (Seddik et al., 2020).

3. Limiting NTK with All Wide Layers

This section provides tight bounds on the smallest eigenvalue of the *limiting* NTK matrix $K^{(L)} \in \mathbb{R}^{N\times N}$ from (4). As shown in (Jacot et al., 2018), one can compute this matrix recursively as follows, for all $l \in [2, L]$:

$$\begin{aligned} K_{ij}^{(1)} &= G^{(1)}, \\ K_{ij}^{(l)} &= K_{ij}^{(l-1)}\dot{G}_{ij}^{(l)} + G_{ij}^{(l)}, \\ \dot{G}_{ij}^{(l)} &= 2\,\mathbb{E}_{(u,v)\sim\mathcal{N}(0,A_{ij}^{(l)})}[\sigma'(u)\sigma'(v)], \end{aligned} \tag{9}$$

where the matrices $G^{(l)} \in \mathbb{R}^{N\times N}$ and $A_{ij}^{(l)} \in \mathbb{R}^{2\times 2}$ are given by, for all $l \in [2, L]$,

$$\begin{aligned} G_{ij}^{(1)} &= \langle x_i, x_j\rangle, \\ A_{ij}^{(l)} &= \begin{bmatrix} G_{ii}^{(l-1)} & G_{ij}^{(l-1)} \\ G_{ji}^{(l-1)} & G_{jj}^{(l-1)} \end{bmatrix}, \\ G_{ij}^{(l)} &= 2\,\mathbb{E}_{(u,v)\sim\mathcal{N}(0,A_{ij}^{(l)})}[\sigma(u)\sigma(v)], \end{aligned} \tag{10}$$

In order to prove our main result of this section, we first need to rewrite the entry-wise formula of the NTK (9) in a more compact form. In particular, the following lemma provides a helpful characterization of the limiting NTK matrix.

Lemma 3.1 *The following holds for the matrices (9)-(10):*

$$\begin{aligned} &G^{(1)} = XX^T, \\ &G^{(2)} = 2\,\mathbb{E}_{w\sim\mathcal{N}(0,\mathbb{I}_d)}\left[\sigma(Xw)\sigma(Xw)^T\right], \\ &G^{(l)} = 2\,\mathbb{E}_{w\sim\mathcal{N}(0,\mathbb{I}_N)}\left[\sigma\left(\sqrt{G^{(l-1)}}\,w\right)\sigma\left(\sqrt{G^{(l-1)}}\,w\right)^T\right], \\ &\textit{for } l \in [3, L]. \end{aligned} \tag{11}$$

$$K^{(1)} = G^{(1)},$$
$$K^{(l)} = K^{(l-1)} \circ \dot{G}^{(l)} + G^{(l)}, \quad \forall\, l \in [2, L], \tag{12}$$
$$\dot{G}^{(l)} = 2\,\mathbb{E}_{w\sim\mathcal{N}(0,\,\mathbb{I}_N)}\left[\sigma'\left(\sqrt{G^{(l-1)}}\,w\right)\sigma'\left(\sqrt{G^{(l-1)}}\,w\right)^T\right],$$

for $l \in [2, L]$.

Moreover, we have

$$K^{(L)} = G^{(L)} + \sum_{l=1}^{L-1} G^{(l)} \circ \dot{G}^{(l+1)} \circ \ldots \circ \dot{G}^{(L)}. \tag{13}$$

Proof: Fix $l \in [2, L]$, and let $B = \sqrt{G^{(l-1)}}$. Then, the equation (11) can be rewritten as

$$G_{ij}^{(l)} = 2\,\mathbb{E}_{w\sim\mathcal{N}(0,\,\mathbb{I}_N)}\left[\sigma\left(\langle B_{i:}, w\rangle\right)\sigma\left(\langle B_{j:}, w\rangle\right)\right].$$

Let $u = \langle B_{i:}, w\rangle$ and $v = \langle B_{j:}, w\rangle$. Then, one has $(u, v) \sim \mathcal{N}\left(0, \begin{bmatrix} G_{ii}^{(l-1)} & G_{ij}^{(l-1)} \\ G_{ji}^{(l-1)} & G_{jj}^{(l-1)} \end{bmatrix}\right)$, which suffices to prove the expressions for $G^{(l)}$. A similar argument applies to $\dot{G}^{(l)}$. The equation (13) is obtained by unrolling (12). □

We are now ready to state the main result of this section. For space reason, a proof sketch is given below, and the full proof is deferred to Appendix B.

Theorem 3.2 (Smallest eigenvalue of limiting NTK) *Let* $\{x_i\}_{i=1}^N$ *be a set of i.i.d. data points from* P_X*, where* P_X *has zero mean and satisfies the Assumptions 2.1 and 2.2. Let* $K^{(L)}$ *be the limiting NTK recursively defined in (9). Then, for any even integer constant* $r \geq 2$*, we have w.p. at least* $1 - Ne^{-\Omega(d)} - N^2 e^{-\Omega\left(dN^{-2/(r-0.5)}\right)}$ *that*

$$L\mathcal{O}(d) \geq \lambda_{\min}\left(K^{(L)}\right) \geq \mu_r(\sigma)^2\,\Omega(d), \tag{14}$$

where $\mu_r(\sigma)$ *is the* r*-th Hermite coefficient of the ReLU function given by (8).*

Proof: Recall that for two p.s.d. matrices P and Q, it holds $\lambda_{\min}(P \circ Q) \geq \lambda_{\min}(P)\min_{i\in[n]} Q_{ii}$ (Schur, 1911). By applying this inequality to the formula for the matrix K_L in Lemma 3.1, and exploiting the fact that $\dot{G}_{ii}^{(p)} = 1$ for all $p \in [2, L], i \in [N]$, we obtain that $\lambda_{\min}\left(K^{(L)}\right) \geq \sum_{l=1}^{L} \lambda_{\min}\left(G^{(l)}\right)$. By using the Hermite expansion and homogeneity of ReLU, one can bound $\lambda_{\min}\left(G^{(l)}\right)$ in terms of $\lambda_{\min}\left(\left((G^{(l-1)})^{*r}\right)\left((G^{(l-1)})^{*r}\right)^T\right)$, for any integer $r > 0$, where $(G^{(l-1)})^{*r}$ denotes the r-th Khatri Rao power of $G^{(l-1)}$. Iterating this argument, it suffices to bound $\lambda_{\min}\left((X^{*r})(X^{*r})^T\right)$. This can be done via the Gershgorin circle theorem, and by using Assumptions 2.1-2.2. □

Let us make a few remarks about the result of Theorem 3.2. First, the probability can be made arbitrarily close to 1 as long as N does not grow super-polynomially in d. Second, the Ω and $\mathcal{O}$ notations in (14) do not hide any other dependencies on the depth L. Finally, the proof of the theorem can be extended to other types of architectures, such as ResNet.

As mentioned in the introduction, non-trivial lower bounds on the smallest eigenvalue of the NTK have been used as a key assumption for proving optimization and generalization results in many previous works, see e.g. (Arora et al., 2019a; Chen et al., 2020; Du et al., 2019b) for shallow models and (Du et al., 2019a; Huang & Yau, 2020) for deep models. While quantitative lower bounds have been developed for shallow networks (Ghorbani et al., 2020), this is the first time, to the best of our knowledge, that these bounds are proved for deep ReLU models.

For finite-width networks, when all the layer widths are sufficiently large, one would expect that, at initialization, the smallest eigenvalue of the NTK matrix (3) has a scaling similar to that given by Theorem 3.2. A quantitative result can be obtained whenever the convergence rates of $\bar{K}^{(L)}$ to $K^{(L)}$ is available. For instance, by using Theorem 3.1 of (Arora et al., 2019b), one has that, for $(W_l)_{ij} \sim \mathcal{N}(0, 1)$,

$$\left|\left(\prod_{l=1}^{L-1}\frac{2}{n_l}\right)\bar{K}_{ij}^{(L)} - K_{ij}^{(L)}\right| \leq (L+1)\epsilon, \tag{15}$$

provided that $\min_{l\in[L-1]} n_l = \Omega\left(\epsilon^{-4}\text{poly}(L)\right)$. By taking $\epsilon = (2(L+1)N)^{-1}\lambda_{\min}\left(K^{(L)}\right)$, it follows that $\left\|\left(\prod_{l=1}^{L}\frac{2}{n_l}\right)\bar{K}^{(L)} - K^{(L)}\right\|_F \leq \lambda_{\min}\left(K^{(L)}\right)/2$, and thus

$$\lambda_{\min}\left(\left(\prod_{l=1}^{L}\frac{2}{n_l}\right)\bar{K}^{(L)}\right) \in \left[\frac{1}{2}, \frac{3}{2}\right]\lambda_{\min}\left(K^{(L)}\right). \tag{16}$$

By applying Theorem 3.2, one concludes that

$$\lambda_{\min}\left(\bar{K}^{(L)}\right) = \Theta\left(d\prod_{l=1}^{L-1} n_l\right) \tag{17}$$

if $\min_{l\in[L-1]} n_l = \Omega\left(N^4\right)$. This condition can be potentially improved if a better convergence rate of the NTK is available, e.g. plugging in the bounds of (Buchanan et al., 2021) may give $\Omega(N^2)$. Nevertheless, this still raises two questions: *(i)* can one further relax the current conditions on layer widths? And *(ii)* is it necessary to require all the layers to be wide to get a similar lower bound on the smallest eigenvalue? We address these questions in the next section.

4. NTK Matrix with a Single Wide Layer

In this section, we provide bounds on the smallest eigenvalue of the empirical NTK matrix for networks of finite widths and fixed depth. The networks we consider have a single wide layer (or more generally, any given subset of layers) with width linear in N (up to logarithmic factors),

while all the remaining layers can have poly-logarithmic scalings. Let us highlight that the position of the wide layer can be anywhere between the input and output layer of the network. This setting is more challenging and closer to practice than the typical NTK one where all the layers are often required to be very large in N. Our main result of this section is stated below. Its proof is given in Section 4.1.

Theorem 4.1 (Finite-width scaling of NTK eigenvalue) *Consider an L-layer ReLU network (1). Let $\{x_i\}_{i=1}^N$ be a set of i.i.d. data points from P_X, where P_X satisfies the Assumptions 2.1-2.2, and let $\bar{K}^{(L)}$ be the NTK Gram matrix, as defined in (3). Let the weights of the network be initialized as $[W_l]_{i,j} \sim \mathcal{N}(0, \beta_l^2)$, for all $l \in [L]$. Fix any $\delta > 0$ and any even integer $r \geq 2$. For $k \in [L-1]$, let ξ_k be 1 if the following condition holds:*

$$n_k = \Omega\left(N \log(N) \log\left(\frac{N}{\delta}\right)\right), \tag{18}$$

$$\prod_{l=1}^{k-2} \log(n_l) = o\left(\min_{l\in[0,k-1]} n_l\right), \tag{19}$$

and let ξ_k be 0 otherwise. Let $\mu_r(\sigma)$ be given by (8). Then,

$$\lambda_{\min}\left(\bar{K}^{(L)}\right) \geq \sum_{k=2}^{L} \xi_{k-1}\, \mu_r(\sigma)^2\, \Omega\left(d \prod_{l=1}^{L-1} n_l \prod_{\substack{l=1\\ l\neq k}}^{L} \beta_l^2\right) + \lambda_{\min}\left(XX^T\right) \Omega\left(\prod_{l=1}^{L-1} n_l \prod_{l=2}^{L} \beta_l^2\right) \tag{20}$$

w.p. at least

$$1-\delta-\sum_{k=1}^{L-1} \xi_k N^2 \exp\left(-\Omega\left(\frac{\min_{l\in[0,k-1]} n_l}{N^{2/(r/2-0.1)} \prod_{l=1}^{k-2} \log(n_l)}\right)\right) - N\sum_{l=1}^{L-1} \exp\left(-\Omega\left(n_l\right)\right) - N\exp(-\Omega\left(d\right)). \tag{21}$$

Moreover, we have that, w.p. at least $1 - \sum_{l=1}^{L-1} \exp\left(-\Omega\left(n_l\right)\right) - \exp(-\Omega\left(d\right))$,

$$\lambda_{\min}\left(\bar{K}^{(L)}\right) \leq \sum_{k=1}^{L} \mathcal{O}\left(d \prod_{l=1}^{L-1} n_l \prod_{\substack{l=1\\ l\neq k}}^{L} \beta_l^2\right). \tag{22}$$

The two plots in Figure 2 provide empirical evidence supporting our main results for $L = 3$. We perform 50 Monte-carlo trials, and report average and confidence interval at 1 standard deviation. On the left, we take $(W_l)_{i,j} \sim \mathcal{N}(0, 1)$, fix the parameters (N, n_1, n_2), scale the NTK matrix by $\frac{4}{n_1 n_2}$ (see (4)), and plot $\lambda_{\min}(\frac{4}{n_1 n_2}\bar{K}^{(L)})$ as a function of d. The three curves correspond to three different choices of (N, n_1, n_2). As predicted by our Theorem 3.2, the smallest eigenvalue of the NTK exhibits a linear dependence on d. On the right, we take $(W_l)_{i,j} \sim \mathcal{N}(0, 2/n_{l-1})$ (the popular He's initialization), fix (d, n_2), set $n_1 = 8N$, and plot $\lambda_{\min}(\bar{K}^{(L)})$ as a function of N. The three curves correspond to three different choices of (d, n_2). In this setting, there is a single wide layer and our Theorem 4.1 predicts that the smallest eigenvalue of the NTK scales linearly in the width of the wide layer (and hence linearly in N). This is in excellent agreement with the plot.

The results of both Theorem 3.2 and 4.1 rely on considering a single term in the sum over layers and a fixed r. However, we expect the gap due to this fact to be rather small: *(i)* the Hermite coefficients of the ReLU decay quite slowly (see (8)), so the dependence of the bounds in r is mild; *(ii)* we are mainly interested in networks with a single wide layer, and in this setting the sum is well approximated by the leading term. Taking into account more terms of the sum or more r is an interesting problem for future work. Unlike Theorem 3.2, we do not track the dependence on L in Theorem 4.1, and therefore the constants implicit in Ω and $\mathcal{O}$ may depend on L. One can see that the lower bound (20) and the upper bound (22) will have the same scaling, that is

$$\left(d \prod_{l=1}^{L-1} n_l\right)\left(\prod_{l=1}^{L} \beta_l^2\right)\left(\min_{l\in[L]} \beta_l\right)^{-2}, \tag{23}$$

provided that there exists a layer $k \in [L-1]$ such that $\xi_k = 1$ and $\beta_{k+1} = \min_{l\in[L]} \beta_l$. For instance, this holds if *(i)* the network contains one wide hidden layer with $\tilde{\Omega}(N)$ neurons, and *(ii)* it is initialized using the popular He's or LeCun's initialization (i.e., $\beta_l = c/\sqrt{n_{l-1}}$ for some constant c) (Glorot & Bengio, 2010; He et al., 2015; LeCun et al., 2012). In that case, the scaling of the lower bound (20) is tight (up to a multiplicative constant). Note also that the probability in (21) can be made arbitrarily close to 1 provided that all the layers before the wide layer k do not exhibit exponential bottlenecks in their widths.

In a nutshell, Theorem 4.1 shows (in a quantitative way) that the spectrum of the NTK matrix is bounded away from zero. The requirements on the network architecture are mild: *(i)* existence of a wide layer with $\tilde{\Omega}(N)$ neurons, and *(ii)* absence of exponential bottlenecks before the wide layer. This last condition means that after the wide layer(s), the widths of the network need not have any relation with each other, thus can scale differently. This is a more general setting than the one considered in (Nguyen, 2019; Nguyen & Hein, 2017; Nguyen & Mondelli, 2020) where the network has a single wide layer, which is then followed by a pyramidal shape (i.e. the widths are non-increasing towards the output layer). Here, the pyramidal constraint is not needed.

Let us make a few remarks about the case of shallow nets ($L = 2$) as tight lower bounds on $\lambda_{\min}\left(\bar{K}^{(L)}\right)$ have been also obtained in several recent works, albeit for a different setting than the one in Theorem 4.1. In particular, (Monta-

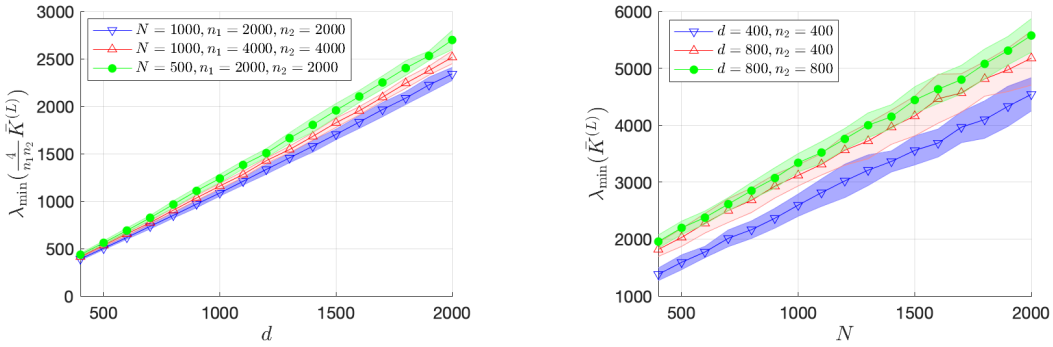

Figure 2. Scaling of the smallest eigenvalue of NTK matrices as a function of the input dimension d (on the left) and of the number of samples N (on the right). The theoretical results of Theorem 3.2 and 4.1 are in excellent agreement with the plot.

nari & Zhong, 2020) consider the regime where $n_0 = \Omega(n_1)$ and $n_0 n_1 = \Omega(N)$, whereas we consider $n_1 = \Omega(N)$ and have little restrictions on n_0. (Oymak & Soltanolkotabi, 2020) give bounds for a similar regime to ours, but a possible generalization of their proof to the case of multi-layer networks would require all the layers to be wide with at least $\tilde{\Omega}(N)$ neurons. In contrast, Theorem 4.1 essentially requires an *arbitrary* single wide layer of width $\Omega(N)$, while all the remaining layers can have almost any widths (up to log factors). To obtain this, the proof of Theorem 4.1 requires lower bounds on the smallest eigenvalue of the intermediate feature matrices F_k's for networks with a single wide layer, and the Lipschitz constant of the intermediate feature maps, which are not studied in the previous works.

Our Theorem 4.1 immediately implies that such a class of networks can fit N distinct data points arbitrarily well, for any *real* labels. The fact that the positive definiteness of the NTK implies a property on *memorization capacity* of neural nets has been already observed in (Montanari & Zhong, 2020), albeit for a two-layer model. The following corollary provides a formal connection between the two for the case of deep nets, and it should be seen as a proof of concept. Its proof is given in Appendix D.1.

Corollary 4.2 (Memorization capacity) *Consider an L-layer ReLU network (1). Let $\{x_i\}_{i=1}^N$ be a set of i.i.d. data points from P_X, where P_X satisfies the Assumptions 2.1-2.2. Fix any $\delta, \delta' > 0$. Assume that there exists a layer $k \in [L-1]$ such that $n_k = \Omega\left(N \log(N) \log\left(\frac{N}{\delta}\right)\right)$ and $\prod_{l=1}^{k-2} \log(n_l) = o\left(\min_{l\in[0,k-1]} n_l\right)$. Then, it holds*

$$\forall Y,\ \forall \epsilon > 0,\ \exists\, \theta: \quad \|F_L(\theta) - Y\|_2 \leq \epsilon$$

w.p. at least $1 - \delta - N^2 e^{-\Omega\left(\frac{\min_{l\in[0,k-1]} n_l}{N^{\delta'} \prod_{l=1}^{k-2} \log(n_l)}\right)} - N\sum_{l=1}^{L-1} e^{-\Omega(n_l)} - N e^{-\Omega(d)}$ over the data.

In words, Corollary 4.2 shows that if a deep ReLU network contains a wide layer of order $\tilde{\Omega}(N)$ neurons, then regardless of the position of this wide layer, and regardless of the widths of the remaining layers (up to log factors), the network can approximate N data points (with real labels) within arbitrary precision. Here, the network has $\tilde{\Omega}(N)$ total parameters, which is known to be (nearly) tight for memorization capacity. However, we remark that this is not optimal in terms of layer widths. In particular, several recent works (Bartlett et al., 2019; Ge et al., 2019; Vershynin, 2020; Yun et al., 2019) show that under some other mild conditions (without the existence of a wide layer as in Corollary 4.2), $\Omega(N)$ parameters suffice for the network to memorize N data points. Nevertheless, let us remark some differences in terms of the setting between these results and the one in Corollary 4.2: *(i)* prior works consider networks with biases while Corollary 4.2 consider nets with no biases, and *(ii)* prior works consider data with *bounded* labels while Corollary 4.2 applies to *arbitrary* real labels. For shallow networks (i.e. $L = 2$), stronger memorization results than Corollary 4.2 have been achieved. For instance, (Bubeck et al., 2020) show that width $\Omega(N/n_0)$ suffices for a two-layer ReLU net to memorize N arbitrary data points. (Montanari & Zhong, 2020) show a similar result under an additional assumption (i.e. $n_0 = \Omega(n_1)$ and $n_0 n_1 = \Omega(N)$), albeit for more general class of activations.

4.1. Proof of Theorem 4.1.

By chain rules and some standard manipulations, we have

$$JJ^T = \sum_{k=0}^{L-1} F_k F_k^T \circ B_{k+1} B_{k+1}^T$$

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

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

For PSD matrices $P, Q \in \mathbb{R}^{n\times n}$, it holds $\lambda_{\min}(P \circ Q) \geq \lambda_{\min}(P) \min_{i\in[n]} Q_{ii}$ (Schur, 1911). Thus,

$$\lambda_{\min}\left(JJ^T\right) \geq \sum_{k=0}^{L-1} \lambda_{\min}\left(F_k F_k^T\right) \min_{i\in[N]} \|(B_{k+1})_{i:}\|_2^2 . \tag{24}$$

We now bound every term on the RHS of (24). Doing so requires a careful analysis of various quantities involving the hidden layers. This includes the smallest singular value of the feature matrices $F_k \in \mathbb{R}^{N\times n_k}$, and the Lipschitz constant of the feature maps $f_k, g_k : \mathbb{R}^d \to \mathbb{R}^{n_k}$. As these results could be of independent interest, we put them separately in the following sections. In particular, our Theorem 5.1 from the next section proves bounds for $\lambda_{\min}\left(F_k F_k^T\right)$. To bound the norm of the rows of B_{k+1}, one can use the following lemma (for the proof, see Appendix D.2).

Lemma 4.3 *Fix any layer $k \in [L-2]$, and $x \sim P_X$. Then,*

$$\left\| \Sigma_{k+1}(x) \left(\prod_{l=k+2}^{L-1} W_l \Sigma_l(x) \right) W_L \right\|_2^2 = \Theta\left(\beta_L^2 \, n_{k+1} \prod_{l=k+2}^{L-1} n_l \beta_l^2 \right),$$

w.p. at least $1 - \sum_{l=1}^{L-1} \exp\left(-\Omega\left(n_l\right)\right) - \exp(-\Omega\left(d\right))$. Here, we assume by convention that the product term $\prod_{l=k+2}^{L-1}(\cdot)$ is inactive for $k = L-2$.

By plugging the bounds of Lemma 4.3 and Theorem 5.1 into (24), the lower bound in (20) immediately follows. For the upper bound, note that

$$\lambda_{\min}\left(JJ^T\right) \leq (JJ^T)_{11} = \sum_{k=0}^{L-1} \|(F_k)_{1:}\|_2^2 \, \|(B_{k+1})_{1:}\|_2^2 . \tag{25}$$

The second term in the RHS of (25) can be bounded by using Lemma 4.3 above. To bound the first term, we note that $(F_k)_{1:} = f_k(x_1)$ and that, for every $0 \leq k \leq L-1$,

$$\|f_k(x_1)\|_2^2 = \Theta\left(d \prod_{l=1}^{k} n_l \beta_l^2 \right), \tag{26}$$

w.p. at least $1 - \sum_{l=1}^{k} \exp\left(-\Omega\left(n_l\right)\right) - \exp(-\Omega\left(d\right))$. This last statement follows from Lemma C.1 in Appendix C. By plugging (26) and the bound of Lemma 4.3 into (25), the upper bound in (22) immediately follows.

5. Smallest Singular Values of Feature Matrices

As before, we assume throughout this section that $(W_l)_{ij} \sim \mathcal{N}(0, \beta_l^2)$ for $l \in [L]$, and the data points are i.i.d. from a distribution P_X satisfying Assumption 2.1 and 2.2. Let us recall the definition of the feature matrix at some hidden layer k: $F_k = [f_k(x_1), \ldots, f_k(x_N)]^T \in \mathbb{R}^{N\times n_k}$. Our main result of this section is the following tight bound on the smallest singular values of these matrices.

Theorem 5.1 (Smallest singular value of feature matrix) *Fix any $k \in [L-1]$ and any even integer constant $r \geq 2$. Let $\delta > 0$ be given. Assume that*

$$n_k = \Omega\left(N \log(N) \log\left(\frac{N}{\delta}\right) \right), \tag{27}$$

$$\prod_{l=1}^{k-2} \log(n_l) = o\left(\min_{l\in[0,k-1]} n_l\right). \tag{28}$$

Let $\mu_r(\sigma)$ be given by (8). Then, the smallest singular value of the feature matrix F_k satisfies

$$\mathcal{O}\left(d \prod_{l=1}^{k} n_l \beta_l^2 \right) \geq \sigma_{\min}\left(F_k\right)^2 \geq \mu_r(\sigma)^2 \, \Omega\left(d \prod_{l=1}^{k} n_l \beta_l^2 \right)$$

w.p. at least

$$1 - \delta - N^2 \exp\left(-\Omega\left(\frac{\min_{l\in[0,k-1]} n_l}{N^{2/(r/2-0.1)} \prod_{l=1}^{k-2} \log(n_l)} \right) \right) - N \sum_{l=1}^{k-1} \exp\left(-\Omega\left(n_l\right)\right) - N \exp(-\Omega\left(d\right)).$$

Proof of Theorem 5.1. First of all, the conditions of Theorem 5.1 imply that $n_k \geq N$, which further implies $\sigma_{\min}\left(F_k\right)^2 = \lambda_{\min}\left(F_k F_k^T\right)$. To bound this quantity, we first relate it to the smallest eigenvalue of the expected Gram matrix, namely $\mathbb{E}[F_k F_k^T]$, where the expectation is taken over W_k. Note that $\mathbb{E}[F_k F_k^T] = n_k \mathbb{E}[\sigma(F_{k-1}w)]\sigma(F_{k-1}w)^T]$, where w has the same distribution as any column of W_k. This is formalized in the following lemma, which is proved in Appendix E.1.

Lemma 5.2 *Let us define*

$$\lambda = \lambda_{\min}\left(\mathbb{E}_{w\sim\mathcal{N}(0,\beta_k^2 \mathbb{I}_{n_{k-1}})}[\sigma(F_{k-1}w)\sigma(F_{k-1}w)^T] \right). \tag{29}$$

Fix any $\delta > 0$. Assume that

$$n_k \geq \max\left(N, \;\; c\,Q \max\left(1, \log(4Q)\right) \log\frac{N}{\delta} \right),$$

where c is an absolute constant, and $Q := \frac{\beta_k^2 \|F_{k-1}\|_F^2}{\lambda}$. Then, we have w.p. at least $1-\delta$ over W_k that

$$\sigma_{\min}\left(F_k\right)^2 \geq \frac{n_k \lambda}{4}.$$

From here, it suffices to upper bound $\|F_{k-1}\|_F^2$ and lower bound λ. The first quantity can be bounded by using a standard induction argument over k. In particular, from Lemma C.1 in Appendix C, it follows that $\|F_{k-1}\|_F^2 = \Theta\left(Nd\prod_{l=1}^{k-1} n_l\beta_l^2\right)$ w.p. at least $1 - \sum_{l=1}^{k-1}\exp\left(-\Omega\left(n_l\right)\right) - \exp(-\Omega\left(d\right))$.

In the remainder of this section, we show how to lower bound λ. First, we relate λ to the smallest eigenvalue of (row-wise) Khatri-Rao powers of F_{k-1}. This is obtained via the following lemma, which is proved in Appendix E.2.

Lemma 5.3 *Fix any $k \in [L-1]$ and any integer $r > 0$. Then, we have*

$$\lambda_{\min}\left(\mathbb{E}_{w\sim\mathcal{N}(0,\beta_{k+1}^2\mathbb{I}_{n_k})}[\sigma(F_k w)\sigma(F_k w)^T]\right) \geq \beta_{k+1}^2\,\mu_r(\sigma)^2\frac{\lambda_{\min}\left((F_k^{*r})(F_k^{*r})^T\right)}{\max_{i\in[N]}\|(F_k)_{i:}\|_2^{2(r-1)}}.$$

Next, we relate the Khatri-Rao powers of F_k to a certain matrix involving the centered features $\tilde{F}_k = F_k - \mathbb{E}_X[F_k]$. This is formalized in the following lemma, which is proved in Appendix E.3.

Lemma 5.4 (Centering features) *Fix any $k \in [L-1]$, and any integer $r > 0$. Let $\mu = \mathbb{E}_x[f_k(x)] \in \mathbb{R}^{n_k}$ and $\Lambda = \mathrm{diag}(F_k\mu - \|\mu\|_2^2 1_N)$, where $1_N \in \mathbb{R}^N$ is the all-one vector. Then, we have*

$$(F_k^{*r})(F_k^{*r})^T = (F_kF_k^T)^{\circ r} \succeq \left(\tilde{F}_k\tilde{F}_k^T - \frac{\Lambda 1_N 1_N^T\Lambda}{\|\mu\|_2^2}\right)^{\circ r}, \tag{30}$$

where $M^{\circ r}$ denotes the r-th Hadamard power of the matrix M.

The last step is to bound the smallest eigenvalue of the matrix $\left(\tilde{F}_k\tilde{F}_k^T - \frac{\Lambda 1_N 1_N^T\Lambda}{\|\mu\|_2^2}\right)^{\circ r}$, as done in the following lemma which is proved in Appendix E.4.

Lemma 5.5 (Hadamard powers of centered features) *Fix any $k \in [L-1]$ and any even integer $r \geq 2$. Assume $\prod_{l=1}^{k-1}\log(n_l) = o\left(\min_{l\in[0,k]} n_l\right)$. Then, we have*

$$\lambda_{\min}\left(\left(\tilde{F}_k\tilde{F}_k^T - \frac{\Lambda 1_N 1_N^T\Lambda}{\|\mu\|_2^2}\right)^{\circ r}\right) = \Theta\left(\left(d\prod_{l=1}^{k} n_l\beta_l^2\right)^r\right) \tag{31}$$

w.p. at least

$$1 - N^2\exp\left(-\Omega\left(\frac{\min_{l\in[0,k]} n_l}{N^{2/(r/2-0.1)}\prod_{l=1}^{k-1}\log(n_l)}\right)\right) - N\sum_{l=1}^{k}\exp\left(-\Omega\left(n_l\right)\right). \tag{32}$$

Combining these lemmas, one gets the desired lower bound of $\sigma_{\min}\left(F_k\right)^2$. For the upper bound: $\lambda_{\min}\left(F_kF_k^T\right) \leq \min_{i\in[N]}\|(F_k)_{i:}\|_2^2 = \mathcal{O}\left(d\prod_{l=1}^{k} n_l\beta_l^2\right)$, where we use Lemma C.1 in Appendix C.

6. Lipschitz Constant of Feature Maps

The Lipschitz constants of the feature maps $g_k : \mathbb{R}^d \to \mathbb{R}^{n_k}$ are critical to several proofs of this paper, including Lemma 5.4 and Lemma 5.5. A simple upper bound is given by $\|g_k\|_{\mathrm{Lip}} \leq \prod_{l=1}^{k}\|W_l\|_{\mathrm{op}}$. From standard bounds on the operator norm of Gaussian matrices (see Theorem 2.12 of (Davidson & Szarek, 2001)), one obtains that $\prod_{l=1}^{k}\|W_l\|_{\mathrm{op}}$ scales as $\prod_{l=1}^{k}\beta_l\max(\sqrt{n_{l-1}},\sqrt{n_l})$. However, this simple estimate leads to restrictions on the network architectures for which our Theorem 4.1 holds. The product of many large random matrices is also studied in (Hanin & Nica, 2019), where it is shown that the logarithm of the ℓ_2 norm between the Jacobian of deep networks and any fixed vector is asymptotically Gaussian. However, the findings of (Hanin & Nica, 2019) are not applicable to our setting, which would require bounds that hold with probability exponentially close to 1.

As usual, let $(W_l)_{ij} \sim \mathcal{N}(0,\beta_l^2)$ for $l \in [L]$. For every $z \in \mathbb{R}^d$, denote its activation pattern up to layer k by

$$\mathcal{A}_{1\to k}(z) = [\mathrm{sign}(g_{lj}(z))]_{l\in[k],j\in[n_l]} \in \{-1,0,1\}^{\sum_{l=1}^{k} n_l},$$

where $\mathrm{sign}(g_{lj}(z)) = 1$ if $g_{lj}(z) > 0$, -1 if $g_{lj}(z) < 0$ and 0 otherwise. For every differentiable point of g_k, we denote by $J(g_k)(z) \in \mathbb{R}^{n_k\times d}$ the corresponding Jacobian matrix.

Our starting point is to relate the Lipschitz constant of g_k with the operator norm of its Jacobian. First, we have via the Rademacher theorem that $\|g_k\|_{\mathrm{Lip}} = \sup_{z\in\mathbb{R}^d\setminus\Omega_{g_k}}\|J(g_k)(z)\|_{\mathrm{op}}$, where Ω_{g_k} is the set of non-differentiable points of g_k which has measure zero. The issue here is that even if we restrict ourself to the "good" set $\mathbb{R}^d\setminus\Omega_{g_k}$, the formula of the Jacobian matrix as computed by the standard back-propagation algorithm[2] (which is also the object that we know how to handle analytically) may not represent the true Jacobian of g_k. This happens, for example, when the input to any of the ReLU activations is 0. The following lemma circumvents this problem by restricting the supremum to the set of inputs where the two Jacobian matrices agree. Its proof is deferred to Appendix F.2.

Lemma 6.1 *Fix any $k \in [L]$. Then w.p. 1 over $(W_l)_{l=1}^{k-1}$, the following holds for all choices of W_k:*

$$\|g_k\|_{\mathrm{Lip}} = \max_{z\in\mathbb{R}^d:\ \mathcal{A}_{1\to k-1}(z)\in\{-1,+1\}^{\sum_{l=1}^{k-1} n_l}}\|J(g_k)(z)\|_{\mathrm{op}}. \tag{33}$$

[2] using a convention that $\sigma'(0) = 0$

In words, Lemma 6.1 shows that the Lipschitz constant of g_k is given by the maximum operator norm of its Jacobian over all the inputs z's which fulfill $g_{lj}(z) \neq 0$ for all $l \in [k-1], j \in [n_l]$. This has two implications. First, g_k is differentiable at every such input, and chain rules can be applied through all the layers to compute the true Jacobian. In particular, we have for all such z's that:

$$J(g_k)(z) = W_k^T \prod_{l=1}^{k-1} \Sigma_{k-l}(z) W_{k-l}^T. \tag{34}$$

Second, one observes that $J(g_k)(z) = J(g_k)(z')$ for all z, z' with $\mathcal{A}_{1\to k-1}(z) = \mathcal{A}_{1\to k-1}(z')$. Thus, the number of Jacobian matrices that one needs to bound in (33) is at most the number of activation patterns, which has been studied in (Hanin & Rolnick, 2019; Montufar et al., 2014; Serra et al., 2018). By exploiting these facts via a careful induction argument, we obtain the following result.

Theorem 6.2 (Lipschitz constant of feature maps) *Fix any $k \in [L-1]$. Then, we have w.p. at least $1 - \sum_{l=1}^{k} \exp\left(-\Omega\left(n_l\right)\right)$ that*

$$\|g_k\|_{\text{Lip}}^2 = \mathcal{O}\left(\frac{\prod_{l=0}^{k} n_l}{\min_{l\in[0,k]} n_l} \prod_{l=1}^{k-1} \log(n_l) \prod_{l=1}^{k} \beta_l^2\right). \tag{35}$$

The idea of the proof is to bound the operator norm of the Jacobian matrix from (34) for all inputs having a given activation pattern (via an ϵ-net argument and concentration inequalities), and then to do a union bound over all the possible patterns. The details are deferred to Appendix F.1.

7. Further Related Work

The spectrum of various random matrices arising from deep learning models has been the subject of recent investigations. Most of the existing results focus on the linear-width asymptotic regime, where the widths of the various layers are linearly proportional. In particular, the spectrum of the conjugate kernel (CK) is studied in the single-layer case for Gaussian i.i.d. data (Pennington & Worah, 2017), for Gaussian mixtures (Liao & Couillet, 2018), for general training data (Louart et al., 2018), and for a model with an additive bias (Adlam et al., 2019). The multi-layer case is tackled in (Benigni & Péché, 2019). The Hessian matrix of a two-layer network can be decomposed into two pieces, one coming from the second derivatives and the other of the form $J^T J$ (a.k.a. the Fisher information matrix). This second term is studied in (Pennington & Bahri, 2017; Pennington & Worah, 2018). Note that this is different from the NTK matrix, given by JJ^T, as analyzed in this paper. Typically, for an over-parameterized model, the Fisher information matrix is rank-deficient, whereas the NTK one is full-rank. The work (Pennington et al., 2018) uses tools from free probability to study the spectrum of the input-output Jacobian of the network. Again, this is different from the parameter-output Jacobian considered in this paper. Generalization error has been also studied via the spectrum of suitable random matrices: for linear regression (Hastie et al., 2019), random feature models (Mei & Montanari, 2019), random Fourier features (Liao et al., 2020), and most recently for a two-layer network (Montanari & Zhong, 2020).

Generally speaking, the line of literature reviewed above has studied the spectrum of various random matrices related to neural networks. Our work is complementary in the sense that it concerns the smallest eigenvalue of the NTK and the feature maps. We remark that obtaining an almost-sure convergence of the empirical spectral distribution of a random matrix in general does not have any implications on the limit of its individual eigenvalues. The closest existing work is (Montanari & Zhong, 2020), which focuses on a two-layer model and gives a lower bound on the smallest eigenvalue of the NTK matrix when the number of parameters of the network exceeds the number of training samples.

8. Conclusions and Open Problems

This paper provides tight bounds on the smallest eigenvalues of NTK matrices for deep ReLU networks. In the finite-width setting, our result holds for networks with a single wide layer, regardless of its position, as long as the wide layer has roughly order of N neurons. This gives hope that gradient descent methods will be successful in optimizing such architectures. However, we note that it is not possible to directly apply existing results in the literature such as (Chizat et al., 2019), since the Jacobian matrix is not Lipschitz with respect to the weights. Furthermore, to get optimization guarantees, one often has to track the movement of the NTK-related quantities during the course of training, which is not done in this paper. Providing rigorous convergence guarantees for deep ReLU networks with an *arbitrary* single wide layer of linear width is an exciting open problem. Other interesting extensions include the study of networks with biases and non-Gaussian initializations.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments. MM was partially supported by the 2019 Lopez-Loreta Prize. QN and GM acknowledge support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 757983). MM would like to thank Simone Bombari, Adel Javanmard and Mahdi Soltanolkotabi for helpful discussions concerning the edit of Lemmas 5.4 and 5.5.

References

Adlam, B., Levinson, J., and Pennington, J. A random matrix perspective on mixtures of nonlinearities for deep learning, 2019. arXiv:1912.00827.

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

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning (ICML)*, 2019a.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. On exact computation with an infinitely wide neural net. In *Neural Information Processing Systems (NeurIPS)*, 2019b.

Bartlett, P., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research (JMLR)*, 20(63):1–17, 2019.

Benigni, L. and Péché, S. Eigenvalue distribution of nonlinear models of random matrices, 2019. arXiv:1904.03090.

Bubeck, S., Eldan, R., Lee, Y. T., and Mikulincer, D. Network size and weights size for memorization with two-layers neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Buchanan, S., Gilboa, D., and Wrighta, J. Deep networks and the multiple manifold problem. In *International Conference on Learning Representations (ICLR)*, 2021.

Chen, Z., Cao, Y., Gu, Q., and Zhang, T. A generalized neural tangent kernel analysis fortwo-layer neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Davidson, K. R. and Szarek, S. J. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(140):317–366, 2001.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019a.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2019b.

Fan, Z. and Wang, Z. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Ge, R., Wang, R., and Zhao, H. Mildly overparametrized neural nets can memorize training data efficiently, 2019. arXiv:1909.11837.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension, 2020. arXiv:1904.12191.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Machine Learning (ICML)*, 2010.

Gorokhovik, V. V. Geometrical and analytical characteristic properties of piecewise affine mappings, 2011. arXiv:1111.1389.

Hanin, B. and Nica, M. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, pp. 1–36, 2019.

Hanin, B. and Rolnick, D. Deep relu networks have surprisingly few activation patterns. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation, 2019. arXiv:1903.08560.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Huang, J. and Yau, H.-T. Dynamics of deep neural networks and neural tangent hierarchy. In *International Conference on Machine Learning (ICML)*, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2018.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–48. Springer, 2012.

Liao, Z. and Couillet, R. On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning (ICML)*, 2018.

Liao, Z., Couillet, R., and Mahoney, M. W. 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.

Louart, C., Liao, Z., and Couillet, R. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve, 2019. arXiv:1908.05355.

Montanari, A. and Zhong, Y. The interpolation phase transition in neural networks: Memorization and generalization under lazy training, 2020. arXiv:2007.12826.

Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. On the number of linear regions of deep neural networks. In *Neural Information Processing Systems (NIPS)*, 2014.

Nguyen, Q. On connected sublevel sets in deep learning. In *International Conference on Machine Learning (ICML)*, 2019.

Nguyen, Q. and Hein, M. The loss surface of deep and wide neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

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

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

Pennington, J. and Bahri, Y. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning (ICML)*, 2017.

Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. In *Neural Information Processing Systems (NeurIPS)*, 2017.

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

Pennington, J., Schoenholz, S., and Ganguli, S. The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Schur, J. Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1911(140):1–28, 1911.

Seddik, M. E. A., Louart, C., Tamaazousti, M., and Couillet, R. Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures. In *International Conference on Machine Learning (ICML)*, 2020.

Serra, T., Tjandraatmadja, C., and Ramalingam, S. Bounding and counting linear regions of deep neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Song, Z. and Yang, X. Quadratic suffices for overparametrization via matrix chernoff bound, 2020. arXiv:1906.03593.

Tropp, J. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, pp. 389–434, 2012.

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

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

Wu, X., Du, S. S., and Ward, R. Global convergence of adaptive gradient methods for an over-parameterized neural network, 2019. arXiv:1902.07111.

Yun, C., Sra, S., and Jadbabaie, A. Small relu networks are powerful memorizers: a tight analysis of memorization capacity. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Zou, D. and Gu, Q. An improved analysis of training overparameterized deep neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2019.

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