Neural Tangent Kernel at Initialization: Linear Width Suffices

Arindam Banerjee[1] **Pedro Cisneros-Velarde**[1] **Libin Zhu**[2] **Mikhail Belkin**[2]

[1]University of Illinois at Urbana-Champaign
[2]University of California, San Diego

Abstract

In this paper we study the problem of lower bounding the minimum eigenvalue of the neural tangent kernel (NTK) at initialization, an important quantity for the theoretical analysis of training in neural networks. We consider feedforward neural networks with smooth activation functions. Without any distributional assumptions on the input, we present a novel result: we show that for suitable initialization variance, $\widetilde{\Omega}(n)$ width, where n is the number of training samples, suffices to ensure that the NTK at initialization is positive definite, improving prior results for smooth activations under our setting. Prior to our work, the sufficiency of linear width has only been shown either for networks with ReLU activation functions, and sublinear width has been shown for smooth networks but with additional conditions on the distribution of the data. The technical challenge in the analysis stems from the layerwise inhomogeneity of smooth activation functions and we handle the challenge using *generalized* Hermite series expansion of such activations.

1 INTRODUCTION

Recent years have seen advances in understanding convergence of gradient descent (GD) and variants for the training of deep learning models (Du et al., 2019; Allen-Zhu et al., 2019; Zou and Gu, 2019; Zou et al., 2020; Liu et al., 2022; Ji and Telgarsky, 2019; Oymak and Soltanolkotabi, 2020; Nguyen, 2021). Despite the fact that such optimization problems are non-convex, a series of recent results have shown that GD has geometric convergence and finds near global solution "near initialization" for wide networks. Such analysis is typically done based on the Neural Tangent Kernel (NTK) (Jacot et al., 2018). The NTK is positive definite "near initialization," the optimization problem then satisfies a condition closely related to the Polyak-Łojasiewicz (PL) condition, which in turn implies geometric convergence to the global minima (Liu et al., 2022; Nguyen, 2021). A very important step in the analysis is to derive a condition on the required network's width to ensure the *NTK condition* is satisfied at initialization, i.e., that the minimum eigenvalue of the NTK is lower bounded at initialization by a positive constant.

Much of the theoretical convergence analysis of deep models has focused on ReLU networks (Allen-Zhu et al., 2019; Nguyen, 2021). While handling the non-smoothness of ReLU activation presents unique challenges, the homogeneity of ReLU helps the analysis (Ji and Telgarsky, 2019; Zou and Gu, 2019; Zou et al., 2020; Allen-Zhu et al., 2019; Nguyen and Mondelli, 2020; Nguyen et al., 2021b). On the other hand, some progress has also been made for deep models with smooth activations, where such homogeneity property does not generally hold (Du et al., 2019; Huang and Yau, 2020). However, many existing results for smooth networks have a high requirement on the width of the models; e.g., as polynomial powers of the number of training samples (Du et al., 2019). Recently Bombari et al. (2022) have shown sublinear width on the number of training samples; however, they do require additional assumptions on the nature of the input data such as (i) scaling on the first two moments and on a variance-related quantity, as well as a (ii) Lipschitz concentration assumption on the distribution.

Consider a feedforward neural network model with L hidden layers of width m, and σ_0^2 initialization variance; trained with n samples. Recent literature indicates that the NTK condition at initialization for deep networks: (i) requires $m = \widetilde{\Omega}(n)$ with ReLU activation functions (Nguyen et al., 2021b); (ii) and for smooth activation functions requires $m = \Omega(\sqrt{n})$ under some distributional assumptions on the input data (Bombari et al., 2022) and $m = \Omega(n^2)$ without such assumptions (Du et al., 2019). Then, the motivating question for our work is: can we improve the dependence to linear width for smooth activation functions under different

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:110–118.

or weaker assumptions than distributional ones?

Our main contribution is to illustrate that $m = \widetilde{\Omega}(n)$ suffices for the NTK condition at initialization without strictly requiring additional assumptions on the distribution of the input data, such as the data distribution conditions stipulated in (Bombari et al., 2022). Instead, our analysis relies on a basic data scaling assumption and other algebraic or geometric conditions present in the existing literature (see Remark 4.2). However, our work assumes a neural network where all the layers have the same width, whereas Bombari et al. (2022) consider a challenging pyramidal topology since they study the question of achieving the minimum possible over-parameterization in neural networks.

Our analysis builds on prior work on ReLU networks based on Hermite series expansions (Oymak and Soltanolkotabi, 2020; Nguyen and Mondelli, 2020; Nguyen et al., 2021b), which however critically relies on the homogeneity of ReLU activations. We substantially generalize such analysis to handle the inhomogeneity of multiple layers of smooth activations based on *generalized* Hermite series expansions, yielding the desired sharper result. To the best of our knowledge, our work is the first in using this mathematical framework in the analysis of neural networks. Our analysis extends to general depth on the network, but does not improve depth dependence of prior work (Du et al., 2019). We also remark that our analysis technique is of a different nature than the one by (Bombari et al., 2022), since they use tools such as restricted isometry properties for random matrices.

Finally, our analysis also reveals a possible trade-off between the constants involved in (i) the Hessian spectral norm bound used in the recently introduced restricted strong convexity (RSC) based optimization analysis for linear convergence (Banerjee et al., 2023) of gradient descent for feedforward smooth networks and (ii) the minimum eigenvalue of the NTK as we consider here used in NTK based optimization analysis. In simple terms, a small variance reduces the Hessian bound and benefits convergence using the RSC condition, but such small variance can adversely affect (exponentially decrease) the constants in the NTK minimum eigenvalue lower bound; and vice versa.

The rest of the paper is organized as follows. We present related work in Section 2. We discuss the problem setup in Section 3. We analyze the NTK minimum eigenvalue lower bound in Section 4. We provide a discussion on the initialization variance in Section 5. We empirically verify our analysis on the lower bound of NTK minimum eigenvalue in Section 6. Conclusion is in Section 7. Technical proofs are in the supplementary material.

2 RELATED WORK

The literature on gradient descent and variants for deep learning optimization typically uses the NTK condition at initialization or variations of it as an integral part of their analysis. Basically, the idea is that if the minimum eigenvalue of the NTK is bounded away from zero at initialization, then under suitable conditions, it is possible to show that this property also holds during training (in a local neighborhood around initialization). The literature on deep learning optimization is increasingly large, and we refer the readers to the following surveys for an overview of the field (Fan et al., 2021; Bartlett et al., 2021). For example, among the theoretical works on the analysis of multi-layer neural networks, we refer to the works (Du et al., 2019; Allen-Zhu et al., 2019; Zou and Gu, 2019; Zou et al., 2020; Liu et al., 2022; Banerjee et al., 2023). For a literature review on shallow and/or linear networks, we refer to the recent survey (Fang et al., 2021). Due to the rapidly growing related literature, we only mention the most related or recent work.

The works (Zou and Gu, 2019; Zou et al., 2020; Allen-Zhu et al., 2019; Nguyen and Mondelli, 2020; Nguyen, 2021; Nguyen et al., 2021b) analyzed deep ReLU networks, whereas (Du et al., 2019; Liu et al., 2022) consider smooth activation functions. The convergence analysis of the gradient descent in (Du et al., 2019; Allen-Zhu et al., 2019; Zou and Gu, 2019; Zou et al., 2020; Liu et al., 2022) relied on the near constancy of NTK for wide neural networks (Jacot et al., 2018; Lee et al., 2019; Arora et al., 2019b; Liu et al., 2020), which yield certain desirable properties for their training using gradient descent based methods. One such property is related to the PL condition (Karimi et al., 2016; Nguyen, 2021), formulated as PL* condition in (Liu et al., 2022). For models with $L = O(1)$ layers, existing results need $m = \Omega(n^4)$ for smooth activations (Du et al., 2019) to ensure convergence. The recent work (Banerjee et al., 2023) uses a different optimization analysis based on the restricted strong convexity (RSC) condition, which they relate to a restricted version of the PL condition, and compare such condition to the widely used NTK one. Finally, shallow ReLU networks (only one hidden layer) were studied by Ji and Telgarsky (2019). They showed that a data separability assumption along with m having a polylogarithmic dependence on n allows gradient descent to provide training and testing guarantees. Interestingly, data separability assumptions can be incorporated in our results to establish further lower bounds to the minimum eigenvalue of the NTK – see Remark 4.2.

We also remark that the smallest eigenvalue of the NTK at initialization plays a crucial role including fitting capacity, as well as generalization behavior (Arora et al., 2019a; Montanari and Zhong, 2020; Liu et al., 2020; Nguyen, 2021; Nguyen et al., 2021b; Oymak and Soltanolkotabi, 2020).

3 PROBLEM SETUP: DEEP LEARNING WITH SMOOTH ACTIVATIONS

Consider a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n, \mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d, y_i \in \mathcal{Y} \subseteq \mathbb{R}$. We will denote by $X \in \mathbb{R}^{n \times d}$ the matrix whose ith row is $\mathbf{x}_i^\top$. In our setting f is a feed-forward multi-layer (fully-connected) neural network with depth L [1] and widths $m_l, l \in [L] := \{1, \dots, L\}$ given by

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

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

$$\begin{aligned}
\theta := (\mathrm{vec}(W^{(1)})^\top, \dots, \mathrm{vec}(W^{(L)})^\top, \mathbf{v}^\top)^\top \\
\in \mathbb{R}^{\sum_{k=1}^L m_k m_{k-1} + m_L} \,,
\end{aligned} \tag{2}$$

with $m_0 = d$. For simplicity, we consider deep models with only one output, i.e., $f(\theta; \mathbf{x}) \in \mathbb{R}$ as in (Du et al., 2019), but our results can be extended to multi-dimension outputs as in (Zou and Gu, 2019), using $\mathbf{V} \in \mathbb{R}^{k \times m_L}$ for k outputs at the last layer. We use the notation $\alpha^{(l)}(\mathbf{x}) = \phi(\tilde{\alpha}^{(l)}(\mathbf{x}))$, with $\alpha^{(l)}$ being the output and $\tilde{\alpha}^{(l)}$ the pre-activation at later l. We also let $A^{(l)} \in \mathbb{R}^{n \times m_l}$ be such that the ith row is defined as $A^{(l)}_{i,:} := \alpha^{(l)}(x_i)$, i.e., $A^{(l)}$ is the output (matrix) of layer $l \in [L]$ for input dataset $x_i, i \in [n]$ – the weight vector θ under which this is evaluated will be understood by the context. Likewise, we let $A^{(L+1)} \in \mathbb{R}^n$ be the vector of outputs for the input dataset. Let $\mathbf{0}_p$ be the zero vector of dimension p and $\mathbb{I}_p$ the $p \times p$ identity matrix.

We denote the gradient and Hessian of $f(\cdot; \mathbf{x}_i) : \mathbb{R}^p \to \mathbb{R}$ as $\nabla_i f := \frac{\partial f(\theta; \mathbf{x}_i)}{\partial \theta}$, and $\nabla_i^2 f := \frac{\partial^2 f(\theta; \mathbf{x}_i)}{\partial \theta^2}$. The *neural tangent kernel* (NTK) $K_{\mathrm{ntk}}(\cdot; \theta) \in \mathbb{R}^{n \times n}$ corresponding to parameter θ is defined as

$$K_{\mathrm{ntk}}(\mathbf{x}_i, \mathbf{x}_j; \theta) = \langle \nabla_i f, \nabla_j f \rangle. \tag{3}$$

We make the following assumption regarding the activation function ϕ:

Assumption 1 (**Activation function**). *The activation ϕ is 1-Lipschitz, i.e., $|\phi'| \leq 1$, and β_ϕ-smooth, i.e., $|\phi''_l| \leq \beta_\phi$.*

Remark 3.1. Our analysis holds for any ς_ϕ-Lipchitz smooth activations, with a dependence on ς_ϕ on most key results. The main (qualitative) conclusions stay true if $\varsigma_\phi \leq 1 + o(1)$ or $\varsigma_\phi = \mathrm{poly}(L)$, which is typically satisfied for commonly used smooth activations and moderate values of L. □

Assumption 2 (Input data scaling). *Every input data $\mathbf{x}_i \in \mathbb{R}^d$, $i \in [n]$, has norm $\|\mathbf{x}_i\|_2^2 = d$.*

The previous assumption is done for convenience. Scaling assumptions are common in the literature (Allen-Zhu et al., 2019; Oymak and Soltanolkotabi, 2020; Nguyen et al., 2021b).

4 NEURAL TANGENT KERNEL AT INITIALIZATION

In this section, we present a sharper analysis showing that effectively linear width, i.e., $m = \widetilde{\Omega}(n)$, suffices for smooth activations to ensure the NTK at initialization is positive definite. Our analysis builds on prior work on Hermite series expansion of activation functions (Oymak and Soltanolkotabi, 2020; Nguyen and Mondelli, 2020; Nguyen et al., 2021b), which has been however restricted to multi-layer ReLU networks using the homogeneity of ReLU activations. Smooth activations are typically inhomogeneous, so we develop a related but new analysis based on *generalized* Hermite polynomials which work for multiple layers of inhomogeneous activations, yielding Theorem 4.1. All detailed proofs are in Section A of the supplementary material.

Theorem 4.1 (**Linear width on the number of samples $m = \tilde{\Omega}(n)$ suffices for the NTK condition at initialization**). *Consider Assumptions 1 and 2. Assume that $L = O(1)$, $\phi(0) = 0$, and for $l \in [L]$,*

$$m_l = m = \Omega(n \log n \log(Ln/\delta)) \,.$$

Let $c_{\phi,\sigma_0} := \mathbb{E}_{z \sim \mathcal{N}(0,\sigma_0^2)}[\phi^2(z)]$ and $\nu_0^2 := \frac{\sigma_0^2}{c_{\phi,\sigma_0}}$. Then, assuming $w_{0,ij}^{(l)} \sim \mathcal{N}(0, \nu_0^2)$, $l \in [L]$, with probability at least $1 - \delta - \frac{4L}{m}$ over the draw of $\{W_0^{(l)}\}_{l \in [L]}$, we have that the minimum eigenvalue of the NTK at initialization satisfies

$$\lambda_{\min}(K_{\mathrm{ntk}}(\cdot; \theta_0)) \geq c_0 \lambda_1 \,,$$

for a suitable constant $c_0 > 0$ and $\lambda_1 := \lambda_{\min}(\mathbb{E}_{g \sim \mathcal{N}(\mathbf{0}_d, \nu_0^2 \mathbb{I}_d)}[\phi(\frac{1}{\sqrt{d}} X g) \phi(\frac{1}{\sqrt{d}} X g)^\top])$.

Remark 4.1 (**Extending to general depths**). Our results extend to the case of general depth L with essentially no changes in the analysis. For general L, the width needs to be relaxed to $m_l = m = \Omega(n h_C^4(L) \log n \log(n/\delta))$ where $h_C(L) = \sum_{i=1}^L \nu_0^{2i}$. Choosing $L = O(1)$ gives Theorem 4.1, and choosing $L = \log \log n$ also yields $m = \widetilde{\Omega}(n)$. More generally, if $\nu_0^2 \leq 1$, $h_C(L) = O(L)$, and the width

[1] The network has L hidden layers, and so has depth $L + 1$ if considering the output layer. However, since the term L appears more frequently in our results than $L + 1$, L will be referred as the *depth* for convenience.

m has poly(L) dependence; otherwise an $O(c^{O(L)})$ for some $c > 1$ dependence on L appears similar to (Du et al., 2019). □

Remark 4.2 (**Lower bound for λ_1 in Theorem 4.1**)**.** There are existing approaches in the literature for lower bounding λ_1 for specific (smooth) activation functions, using suitable (separability) assumptions on the input X (Oymak and Soltanolkotabi, 2020; Zou et al., 2020; Nguyen et al., 2021b; Du et al., 2019). To get an informal sense of a couple of such techniques, let $\bar{X} := \frac{1}{\sqrt{d}}X$ so that rows of $\bar{X}$ satisfy $\|\bar{x}_i\|_2 = 1$.

1. If $\lambda_{\min}(\bar{X}\bar{X}^T) > 0$, then the proof analysis of Theorem 4.1 can be extended to show $\lambda_1 > 0$, e.g., see Section A.6, also (Du et al., 2019).
2. For any unit vector v, let $\lambda_1(v) := v^\top \mathbb{E}_g[\phi(\bar{X}g)\phi(\bar{X}g)^\top]v = \mathbb{E}_g[\|\phi(\bar{X}g)^\top v\|_2^2]$. Note that with $\tilde{g} = \bar{X}g$, to show that $\lambda_1 > 0$, it suffices to show $\mathbb{E}_{\tilde{g}}[\langle\phi(\tilde{g}), v\rangle^2] = \mathbb{E}_{Z=\langle\phi(\tilde{g}),v\rangle}[Z^2] > 0$ for any unit vector v, which is violated only if $Z = 0$ a.s. This can be proved by using the fact that $g \sim \mathcal{N}(\mathbf{0}_d, \nu_0^2\mathbb{I}_d)$, properties of ϕ, Markov's inequality, and separability in X (Du et al., 2019; Oymak and Soltanolkotabi, 2020; Nguyen et al., 2021b).

We share additional remarks on λ_1 in Section A.6. Finally, we point out that, although our focus is on avoiding distributional assumptions on the data, it is possible to lower bound λ_1 under such type of assumptions too, e.g., (Nguyen and Mondelli, 2020, Theorem 3.1). □

The proof of Theorem 4.1 shown below is relatively standard in the existing literature with the particular exception of the crucial use of a new result we introduce in this paper: Theorem 4.2.

Proof of Theorem 4.1. Consider that $A^{(l)} \in \mathbb{R}^{n\times m_l}$ with $A^{(l)}_{i,:} = \alpha^{(l)}(\mathbf{x}_i)$, $i \in [n]$, is evaluated at the initialization vector θ_0. The corresponding Jacobian of the neural network is

$$J = \left[\frac{\partial A^{(L+1)}}{\partial \operatorname{vec}(W^{(1)})}, \ldots, \frac{\partial A^{(L+1)}}{\partial \operatorname{vec}(W^{(L)})}, \frac{\partial A^{(L+1)}}{\partial \mathbf{v}}\right],$$

of dimensions $n \times \sum_{l=1}^{L} m_{l-1}m_l + m_{L+1}$ where $m_0 = d$. Then, the kernel at initialization is

$$\begin{aligned} K_{\text{ntk}}(\cdot;\theta_0) &= JJ^\top \\ &= \sum_{l=1}^{L}\left[\frac{\partial A^{(L+1)}}{\partial \operatorname{vec}(W^{(1)})}\right]\left[\frac{\partial A^{(L+1)}}{\partial \operatorname{vec}(W^{(l)})}\right]^\top \\ &\quad + \left[\frac{\partial A^{(L+1)}}{\partial \mathbf{v}}\right]\left[\frac{\partial A^{(L+1)}}{\partial \mathbf{v}}\right]^\top . \end{aligned}$$

Note that

$$\left[\frac{\partial A^{(L+1)}}{\partial \mathbf{v}}\right]\left[\frac{\partial A^{(L+1)}}{\partial \mathbf{v}}\right]^\top = \frac{1}{m_L}A^{(L)}(A^{(L)})^\top .$$

By chain rule, it can be shown that, for any $l \in [L]$

$$\begin{aligned} \frac{\partial \alpha^{(L+1)}(\mathbf{x}_i)}{\partial \operatorname{vec}(W^{(l)})} &= \frac{1}{\sqrt{m_{l-1}}}\alpha^{(l-1)}(\mathbf{x}_i)D_l \\ &\quad \times \left(\prod_{l'=l+1}^{L}\frac{1}{\sqrt{m_{l'-1}}}W^{(l')}D_{l'}\right)\frac{1}{\sqrt{m_L}}\mathbf{v} . \end{aligned}$$

where $D_l = \operatorname{diag}(\phi'(\tilde{\alpha}^{(l)})(\mathbf{x}_i)) \in \mathbb{R}^{m_l\times m_l}$ is a diagonal matrix whose ith element is the derivative of the activation function evaluated at the ith preactivation of layer l. Then, in matrix notation

$$\begin{aligned} &\left[\frac{\partial A^{(L+1)}}{\partial \operatorname{vec}(W^{(l)})}\right]\left[\frac{\partial A^{(L+1)}}{\partial \operatorname{vec}(W^{(l)})}\right]^\top \\ &\qquad = \frac{1}{m_{l-1}}A^{(l-1)}(A^{(l-1)})^\top \odot B_lB_l^\top \end{aligned}$$

where for $l \in [L]$

$$B_l = D_l\left(\prod_{l'=l+1}^{L}\frac{1}{\sqrt{m_{l'-1}}}W^{(l')}D_{l'}\right)\frac{1}{\sqrt{m_L}}\mathbf{v} , \qquad (4)$$

where $\odot$ is the symbol for the Kronecker product operator. Now, note that in particular, note that $B_L = \frac{1}{\sqrt{m_L}}D_L\mathbf{v}$, and $B_{L-1} = \frac{1}{\sqrt{m_{L-1}m_L}}D_{L-1}W^{(L)}D_L\mathbf{v}$. As a result,

$$\begin{aligned} JJ^\top = \sum_{l=1}^{L}&\frac{1}{m_{l-1}}A^{(l-1)}(A^{(l-1)})^\top \odot B_lB_l^\top \\ &+ \frac{1}{m_L}A^{(L)}(A^{(L)})^\top . \end{aligned}$$

From the Schur product theorem (e.g., see (Oymak and Soltanolkotabi, 2020, Lemma 6.5),(Nguyen et al., 2021a, Theorem 3.2)), for positive semi-definite (PSD) matrices $P, Q \in \mathbb{R}^{n\times n}$, it holds that $\lambda_{\min}(P \odot Q) \geq \lambda_{\min}(P)\min_{i\in[n]} Q_{ii}$. Then,

$$\begin{aligned} \lambda_{\min}(JJ^\top) \geq \sum_{l=1}^{L}&\frac{1}{m_{l-1}}\lambda_{\min}\left(A^{(l-1)}(A^{(l-1)})^\top\right) \\ &\times \min_{i\in[n]}\|(B_l)_{i,:}\|_2^2 + \frac{1}{m_L}\lambda_{\min}\left(A^{(L)}(A^{(L)})^\top\right) , \end{aligned}$$

where the first inequality follows from the fact that the minimum eigenvalue of the sum of PSD matrices is lower bounded by the sum of the minimum eigenvalues of the matrices themselves. To lower bound $\lambda_{\min}(JJ^\top)$, because $A^{(l)}(A^{(l)})^\top$, $l \in [K]$, are positive semi-definite, it suffices to lower bound $\lambda_{\min}\left(A^{(L)}(A^{(L)})^\top\right)$. Following Theorem 4.2 and taking $m_l = m, l \in [L]$, with probability at least $1 - \delta - \frac{4L}{m}$, we have $\lambda_{\min}\left(A^{(L)}(A^{(L)})^\top\right) \geq c_0m\lambda_1$, where $c_0 = \max_{r>1} c_0^{(L-1,r)}$ and $c_0^{(l,r)}$ is as in Theorem 4.2. Plugging this back,

$$\begin{aligned} \lambda_{\min}(K_{ntk}(\cdot;\theta_0)) &= \lambda_{\min}(JJ^\top) \\ &\geq \frac{1}{m}\lambda_{\min}\left(A^{(L)}(A^{(L)})^\top\right) \geq c_0\lambda_1 , \end{aligned}$$

with probability at least $1 - \delta - \frac{4L}{m}$. This completes the proof. □

Remark 4.3 (About initialiazing the last layer weight). Notice that the result in Theorem 4.1 is independent on how we initialize the weights $\mathbf{v}$ of the last layer in the neural network. This follows from the fact that $\lambda_{\min}(K_{\text{ntk}}(\cdot;\theta_0)) \geq \frac{1}{m_L}\lambda_{\min}\left(A^{(L)}(A^{(L)})^\top\right)$ from the proof of Theorem 4.1. □

Next we focus our attention on Theorem 4.2, the main new result for smooth activations. The proof borrows ideas from existing related proofs for ReLU networks, however differs in an important way by handling inhomogeneity of smooth activations using *generalized* Hermite series expansions.

Theorem 4.2 (Bound on the minimum eigenvalue of activation matrices). *Consider Assumptions 1 and 2. Assume that $L = O(1)$, $\phi(0) = 0$, and for $l \in [L]$, $m_l = m = \Omega(n \log n \log(Ln/\delta))$. Let $c_{\phi,\sigma_0} := \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}[\phi^2(z)]$ and $\nu_0^2 := \frac{\sigma_0^2}{c_{\phi,\sigma_0}}$. Then, assuming $w_{0,ij}^{(l)} \sim \mathcal{N}(0,\nu_0^2), l \in [L]$, with probability at least $1 - \delta - \frac{4L}{m}$, uniformly over $l \in [L]$ over the draw of $\{W_0^{(l)}\}_{l\in[L]}$, for any integer $r > 1$ we have*

$$\lambda_{\min}(A^{(l)}(A^{(l)})^\top) \geq c_0^{(l-1,r)} m_l \lambda_1 ,$$

where $c_0^{(l-1,r)}$ is a positive constant and $\lambda_1 = \lambda_{\min}(\mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_d,\nu_0^2\mathbb{I}_d)}[\phi(\frac{1}{\sqrt{d}}Xg)\phi(\frac{1}{\sqrt{d}}Xg)^\top])$. Specifically, letting $c_{l,i} = \frac{\|\alpha^{(l)}(x_i)\|_2}{\sqrt{m_l}}$ and $(\mu_{r,0}^{(l)})^2 = \min_{i\in[n]}\left(\mu_r^{[c_{l,i}^2\sigma^2]}(\phi)\right)^2$ for any integer $r > 1$ and $l \in \{0,1,\ldots,L\}$, we have that $c_0^{(l,r)} = \left(\frac{(\mu_{r,0}^{(l)})^2}{6c_{\phi,\sigma_0}}\right)^l \left(\frac{\sigma_0^2}{2}\right)^{3rl}$, where $\mu_r^{[c_{l,i}^2\sigma^2]}(\phi)$ is the $(c_{l,i}^2\sigma^2)$-th generalized Hermite coefficient corresponding to the generalized Hermite series expansion of ϕ.

Remark 4.4 (The use of generalized Hermite polynomials). A key unique feature of our result and proof is the use of *generalized* Hermite coefficients, instead of standard Hermite coefficients in prior work (Oymak and Soltanolkotabi, 2020; Nguyen and Mondelli, 2020; Nguyen et al., 2021b). Since smooth activations are typically inhomogeneous, generalized Hermite coefficients help handle multiple layers of inhomogeneous activations which seems difficult with standard Hermite coefficients. Further, our proof technique, based on Hermite expansions, is different from prior related work on smooth activations (Du et al., 2019) and leads to a sharper sample dependence $\tilde{\Omega}(n)$ instead of $\tilde{\Omega}(n^2)$. □

Definition 4.1 (Generalized Hermite series expansion). For a given positive number $a \in \mathbb{R}_{++}$, we define the normalized *generalized Hermite polynomials* by

$$H_r^{[a]}(x) = \frac{(-1)^r}{\sqrt{r!}} e^{\frac{x^2}{2a}} \frac{d^r}{dx^r} e^{-\frac{x^2}{2a}} , \quad r = 0,1,\ldots . \tag{5}$$

For any function $g : \mathbb{R} \to \mathbb{R}$ such that $\int_{-\infty}^{+\infty} g^2(x) \frac{e^{\frac{-x^2}{2a}}}{\sqrt{2\pi a}} < \infty$, we define the *$r$-th generalized Hermite coefficient* by

$$\mu_r^{[a]}(g) = \int_{-\infty}^{+\infty} g(x) H_r^{[a]}(x) dx . \tag{6}$$

Finally, we define the *generalized Hermite series expansion* of g with respect to $H_r^{[q]}$ by

$$g(x) = \sum_{r=0}^{\infty} \mu_r^{[a]}(g) H_r^{[a]}(x) . \tag{7}$$

Remark 4.5. Since they are used in Theorem 4.2, we provide a self-contained gentle introduction to Hermite Polynomials and Hermite Series Expansions in Section A.4. □

We present the proof of Theorem 4.2, all missing proofs of auxiliary results are in the supplementary material.

Proof of Theorem 4.2. There are three key parts to the proof:

1. showing that under suitable conditions such as a requirement on the width of the network, the minimum eigenvalue of $A^{(l)}(A^{(l)})^\top$ for a model with width $\tilde{\Omega}(n)$ can be lower bounded by a constant scaled version of the minimum eigenvalue of the expectation $\mathbb{E}_{W_0^{(l)}}[A^{(l)}(A^{(l)})^\top]$ with high-probability, i.e., a matrix concentration result;

2. establishing suitable upper and lower bounds for $\|\alpha^{(l)}\|_2^2$, in particular $\|\alpha^{(l)}\|_2^2 = \Theta(m_l)$ with high probability, which let us further simplify the sufficient conditions for the matrix concentration result in (a) above.

3. lower bounding the minimum eigenvalue of the expectation $\mathbb{E}_{W_0^{(l)}}[A^{(l)}(A^{(l)})^\top]$ using generalized Hermite series expansion to handle multiple layers of inhomogenous activations and using the lower bounds on $\|\alpha^{(l)}\|_2^2$ as in (b) above.

Next we get into the details of each of these results.

1. Matrix Concentration. Note that by construction $A^{(l)} = \phi(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}(W_0^{(l)})^\top) \in \mathbb{R}^{n\times m_l}$, where $W_0^{(l)} \in \mathbb{R}^{m_l\times m_{l-1}}$, $w_{0,ij}^{(l)} \sim \mathcal{N}(0,\nu_0^2)$ with $\nu_0^2 = \frac{\sigma_0^2}{c_{\phi,\sigma_0}}$. Through a matrix concentration bound, the minimum eigenvalue of $A^l(A^l)^\top$ can be lower bounded by that of $\mathbb{E}_{W_0^{(l)}}[A^{(l)}(A^{(l)})^\top]$ with high probability, as shown in Lemma 4.1, whose proof is in Section A.1 of the supplementary material.

Lemma 4.1 (Matrix Concentration). *Let $A^{(l)} = \phi(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}(W_0^{(l)})^\top) \in \mathbb{R}^{n\times m_l}$, where $W_0^{(l)} \in$*

$\mathbb{R}^{m_l \times m_{l-1}}$ and $w_{0,ij}^{(l)} \sim \mathcal{N}(0, \sigma^2)$. Let

$$\lambda_l := \lambda_{\min}\left(\mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_{m_{l-1}}, \sigma^2 \mathbb{I}_{m_{l-1}})}\left[\phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right) \times \phi\left(\frac{1}{\sqrt{m_{l-1}}}A^{(l-1)}g\right)^\top\right]\right) \;, \quad (8)$$

and $m_l \geq \max(n, c_2 v \max\left(1, \log(15v)\right) \log(Ln/\delta))$, where $v := \frac{2(\sqrt{\log n}+1)^2\sigma^2\|A^{(l-1)}\|_F^2}{c_3\lambda_l m_{l-1}}$, and c_2, c_3 are absolute constants. Then, with probability at least $(1 - \frac{\delta}{L})$ over the draw of $W_0^{(l)}$, we have

$$\lambda_{\min}(A^{(l)}(A^{(l)})^\top) \geq \frac{m_l\lambda_l}{4}. \quad (9)$$

Then, in order to choose $m_l, l \in [L]$, appropriately for (9), it suffices to upper bound $\|A^{(l-1)}\|_F^2$ and lower bound λ_l for $\sigma^2 = \nu_0^2 = \frac{\sigma_0^2}{c_{\phi,\sigma_0}}$ in (8).

2. Bounding $\|A^{(l)}\|_F^2$. To bound the squared Frobenius norm $\|A^{(l)}\|_F^2$, we focus on bounding the L_2-norm of each row of $A^{(l)}$ and show that $\|\alpha^l(\mathbf{x}_i)\|_2^2 = \Theta(m_l)$, $i \in [n]$. In Lemma 4.2 below, whose proof is in Section A.2 of the supplementary material, we show that the bound holds uniformly over the dataset $\{\mathbf{x}_i, i \in [n]\}$ with high probability.

Lemma 4.2 (**Bounding** $\|\alpha^{(l)}\|_2^2$)**.** *Let $\{\alpha^{(l)}(\mathbf{x}_i) \in \mathbb{R}^{m_l}, i \in [n]\}$ be the set of outputs at layer l at initialization for the set of inputs $\{\mathbf{x}_i, i \in [n]\}$. Let $c_{\phi,\sigma_0} := \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}[\phi^2(z)]$, $\nu_0^2 := \frac{\sigma_0^2}{c_{\phi,\sigma_0}}$, and $h_C(l) := \sum_{i=0}^{l} \nu_0^{2i}$. Let the depth L be such that $\max\left(\frac{8h_C^2(L)}{c_{\phi,\sigma_0}^2}, \frac{4h_C(L)}{c_{\phi,\sigma_0}}\right) \leq \sqrt{m_l}, l \in [L]$. Assuming the elements of $W_0^{(l)}, l \in [L]$ are drawn i.i.d. from $\mathcal{N}(0, \nu_0^2)$, with probability at least $1 - 2n\sum_{l=1}^{L} \frac{1}{m_l^2}$ over the draw of $\{W_0^{(l')}, l' \in [L]\}$, uniformly over $l \in [L]$ and $i \in [n]$, we have*

$$\frac{c_{\phi,\sigma_0}}{2} m_l \leq c_{\phi,\sigma_0}\left(1 - \frac{h_C(l)}{2h_C(L)}\right) m_l$$
$$\leq \|\alpha^{(l)}(x_i)\|_2^2 \leq c_{\phi,\sigma_0}\left(1 + \frac{h_C(l)}{2h_C(L)}\right) m_l \leq \frac{3c_{\phi,\sigma_0}}{2} m_l \;.$$

As a result, by union bound, with probability at least $(1 - 2n\sum_{l=1}^{L}\frac{1}{m_l^2}) \geq (1 - 2\sum_{l=1}^{L}\frac{1}{m_l})$ for $m_l \geq n$, uniformly over $l \in [L]$, we have $\|A^{(l)}\|_F^2 = \sum_{i=1}^{n} \|\alpha_i^{(l)}(\mathbf{x}_i)\|_2^2 \leq \frac{3c_{\phi,\sigma_0}}{2} n m_l$. Then, under the assumption $m_l = m$, taking $\sigma^2 = \nu_0^2 = \frac{\sigma_0^2}{c_{\phi,\sigma_0}}$ in Lemma 4.1, we have $v \leq c_2 \frac{\sigma_0^2 n \log n}{\lambda_l}$, for some constant $c_2 > 0$. For $L = O(1), h_C^4(L) = O(1) \leq m$. As a result, for $l \in [L]$, it suffices to have

$$m \geq \max\left(n \;,\; c_2\frac{\sigma_0^2 n \log n}{\lambda_l} \times \max\left(1, \log\left(c_3\frac{\sigma_0^2 n\log n}{\lambda_l}\right)\right)\log\frac{Ln}{\delta}\right)$$
$$\overset{(a)}{=} \tilde{\Omega}(n) \;, \quad (10)$$

where (a) holds as long as $\lambda_l = \Omega(1)$, which is the case with high probability as we show next.

3. Lower Bounding λ_l. Next, we focus on lower bounding λ_l (defined in equation (8)), for which we obtain the lemma below, whose proof is in Section A.3 of the supplementary material.

Lemma 4.3. *Consider the same setting and assumptions as in Lemma 4.2. Let $\mu_r^{[q]}(\phi), q > 0$ be the r-th generalized Hermite coefficient corresponding to the generalized Hermite series expansion of ϕ w.r.t. $H_r^{[q]}$. Let $c_{l,i} = \frac{\left\|\alpha^{(l)}(\mathbf{x}_i)\right\|_2}{\sqrt{m_l}}$ and $(\mu_{r,0}^{(l)})^2 = \min_{i\in[n]}\left(\mu_r^{[c_{l,i}^2\nu_0^2]}(\phi)\right)^2$. For any integer $r > 0$, with probability at least $1 - 2n\sum_{l=1}^{L}\frac{1}{m_l}$, uniformly over $l \in [L]$ we have*

$$\lambda_{l+1} = \lambda_{\min}\left(\mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_{m_l},\nu_0^2\mathbb{I}_{m_l})}\left[\phi\left(\frac{1}{\sqrt{m_l}}A^{(l)}g\right) \times\phi\left(\frac{1}{\sqrt{m_l}}(A^{(l)}g)^\top\right)\right]\right)$$
$$\geq \left(\frac{(\mu_{r,0}^{(l)})^2}{6c_{\phi,\sigma_0}}\right)^l \left(\frac{\sigma_0^2}{2}\right)^{3rl} \lambda_1 \;,$$

with $\lambda_1 = \lambda_{\min}(\mathbb{E}_{g\sim\mathcal{N}(\mathbf{0}_d,\nu_0^2\mathbb{I}_d)}[\phi(\frac{1}{\sqrt{d}}Xg)\phi(\frac{1}{\sqrt{d}}Xg)^\top])$.

Finally, we have

$$\lambda_{\min}(A^{(l)}(A^{(l)})^\top) \overset{(a)}{\geq} \frac{m}{4}\lambda_l \overset{(b)}{\geq} c_0^{(l-1,r)} m\lambda_1 \;,$$

where (a) follows from Lemma 4.1, (b) from Lemma 4.3 with $c_0^{(l,r)} = \left(\frac{(\mu_{r,0}^{(l)})^2}{6c_{\phi,\sigma_0}}\right)^l \left(\frac{\sigma_0^2}{2}\right)^{3rl}$. As a result of a union bound, this expression, which holds for every $l \in [L]$ according to Lemma 4.1 and Lemma 4.3, holds with probability at least $1 - \delta - \frac{4L}{m}$. Lemma 4.3 also implies that (10) holds. This completes the proof for Theorem 4.2. □

Remark 4.6 (**Regarding our proof techniques**)**.** The proof technique used for Theorem 4.2 is general and quite different from influential prior work on multi-layer feedforward networks with smooth activations (Du et al., 2019). Indeed, our approach works for multiple layers of inhomogeneous smooth activation functions unlike prior work using basic Hermite expansions for homogenous activations especially ReLU (Nguyen and Mondelli, 2020; Nguyen et al., 2021b). To the best of our knowledge, our work represents the first use of generalized Hermite polynomials in such context. For the activation function ϕ, we assume $\phi(0) = 0$ for simplicity; however, this can be relaxed similar to the analysis done in (Banerjee et al., 2023, Section 4) for the derivation of the Hessian bound with an explicit dependence on $\phi(0)$. □

5 THE IMPORTANCE OF INITIALIZATION VARIANCE

Let us define the following type of ball over parameters.

Definition 5.1 **(Spectral ball).** *Given $\overline{\theta} \in \mathbb{R}^p$ of the form* (2) *with parameters $\overline{W}^{(l)}, l \in [L], \overline{\mathbf{v}}$ and with $\|\cdot\|_2$ denoting spectral norm for matrices and L_2-norm for vectors, we define*

$$B^{\mathrm{Spec}}_{\rho,\rho_1}(\bar{\theta}) := \{\theta \in \mathbb{R}^p \text{ as in } (2) \mid \|W^{(\ell)} - \overline{W}^{(\ell)}\|_2 \leq \rho,\\ \ell \in [L], \|\mathbf{v} - \bar{\mathbf{v}}\|_2 \leq \rho_1\} .$$

Proposition 5.1 **(Hessian Spectral Norm Bound).** *Consider Assumptions 1 and 2, and that the elements of $W_0^{(l)}$, $l \in [L]$, are drawn i.i.d from $\mathcal{N}(0, \nu_0^2)$, where $\nu_0^2 = \frac{\sigma_0^2}{c_{\phi,\sigma_0}}$ with $c_{\phi,\sigma_0} := \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_0^2)}[\phi^2(z)]$, and $\mathbf{v}_0$ is a random unit vector with $\|\mathbf{v}_0\|_2 = 1$. Then, for $\theta \in B^{\mathrm{Spec}}_{\rho,\rho_1}(\theta_0)$, with probability at least $1 - \frac{2(L+1)}{m}$, we have*

$$\max_{i\in[n]} \left\|\nabla^2_\theta f(\theta; \mathbf{x}_i)\right\|_2 \leq \frac{c_H}{\sqrt{m}} , \tag{11}$$

with $c_H = O(\mathrm{poly}(L)(1 + \gamma^{2L})(1 + \rho_1))$ where $\gamma := \frac{\rho}{\sqrt{m}} + 4\nu_0$.

Proof. The proof follows by a direct extension of (Banerjee et al., 2023, Theorem 4.1). Indeed, the original result in (Banerjee et al., 2023, Theorem 4.1) can be stated as $\max_{i\in[n]} \left\|\nabla^2_\theta f(\theta; \mathbf{x}_i)\right\|_2 \leq \frac{\tilde{c}_H}{\sqrt{m}}$, with $\tilde{c}_H = O(\mathrm{poly}(L)(1 + \tilde{\gamma}^{2L})(1 + \rho_1))$ where $\tilde{\gamma} := \frac{\rho}{\sqrt{m}} + 2\nu_0\left(1 + \frac{\sqrt{\log m}}{\sqrt{2m}}\right)$. We obtain (11) by upper bounding $\tilde{\gamma} \leq \frac{\rho}{\sqrt{m}} + 4\nu_0$ due to $\frac{\sqrt{\log m}}{\sqrt{2m}} \leq \frac{1}{\sqrt{2}} \leq 1$. Then $\tilde{c}_H \leq c_H$ since $L \geq 1$. □

A trade-off between the Hessian bound and the NTK condition at initialization. Smaller initial variance σ_0^2, based on $\sigma_1 \leq 1$ has a desirable effect on the Hessian bound, e.g., c_H in Theorem 5.1 has a $\mathrm{poly}(L)$ dependence (see Theorem 5.1) and thus is beneficial for the restricted strong convexity condition for geometric convergence in gradient descent (Banerjee et al., 2023). However $\sigma_1 \leq 1$ implies $\sigma_0^2 \leq \frac{1}{4}$, which may affect (exponentially decrease) the constant $c_0^{(l,r)}$ in Theorem 4.2 and thus likewise decrease the minimum eigenvalue of the NTK since $c_0 = \max_{r>1} c_0^{(L-1,r)}$ in Theorem 4.1. The subtlety here is that the dependence of $c_0^{(l,r)}$ on σ_0^2 is complex, involving both c_{ϕ,σ_0} and Hermite coefficient terms. This trade-off effect is not pronounced for small L, e.g., $L = O(1)$ or even $L = O(\log n)$. For general (large) L, the trade-off may be present since it would take m growing as $c^{O(L)}, c > 1$ to neutralize it.

The motivation for studying this trade-off is as follows: for homogeneous activation functions (like ReLU), the effect of the choice for the initialization variance σ_0^2 is well understood (Allen-Zhu et al., 2019); however, such understanding is currently limited for smooth activation functions. Our discussion on the trade-off acknowledges the fact that the choice of the variance may imply whether the NTK based analysis (Liu et al., 2022) or RSC based analysis (Banerjee et al., 2023) is more appropriate to understand the optimization behavior with smooth activations.

6 NUMERICAL VERIFICATION

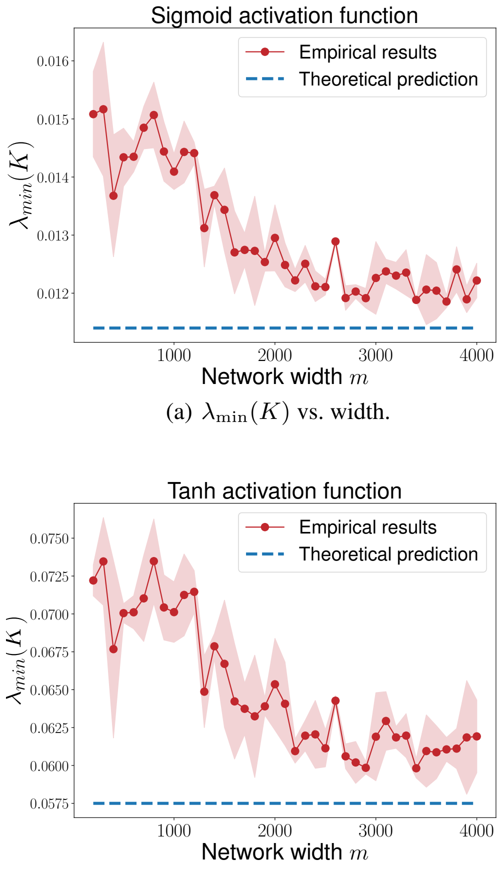
In this section, we conduct experiments to verify our theoretical results, i.e., Theorem 4.1. Specifically, we aim to verify that if the network width grows linearly with the number of training samples, the minimum eigenvalue of the NTK is bounded from a positive constant.

Our experimental setup is as follows: we consider a 3-layer (2-hidden layer) fully-connected neural network with Sigmoid/Tanh activation function, and whose hidden layers have the same width m. We consider $m \in \{200, 300, ..., 4000\}$. We train the network over n data points, with each data point being drawn i.i.d. from $\mathcal{N}(0, \mathbb{I}_{100})$. We let the number of data points be the same as the width, i.e., $n = m$. We report the average of the minimum eigenvalue of the NTK out of 3 independent runs (i.e., each run uses the same training data but has different random initialization of the weights).

From Figure 1 we can see that across all experiments, the minimum eigenvalue of NTK stays persistently positive. Furthermore, as m (as well as n) is made sufficiently large, the minimum eigenvalue of NTK shows only a mild decrease in the average value which mostly flattens out as m (and n) increases. In summary, the minimum eigenvalue of the NTK can be lower bounded by some constant.

7 CONCLUSIONS

In this paper, we revisit the NTK analysis with smooth activations and show that effectively linear width suffices for the NTK at initialization to be positive definite. Our analysis



(a) $\lambda_{\min}(K)$ vs. width.

(b) $\lambda_{\min}(K)$ vs. width.

Figure 1: Positive $\lambda_{\min}(K)$ with linear width. In the experiments, we train a 3-layer fully-connected neural network with Sigmoid/Tanh activation functions whose width has the same numerical value as the number of data points. Each curve is the average of 3 independent runs.

makes a novel use of generalized Hermite series expansion for smooth function activation. Though standard Hermite series expansion has been used for ReLU activation functions, such analysis relied heavily on the homogeneous assumption of ReLU functions — a property generally absent in smooth activation functions. Finally, our work highlights the importance of initialization variance in determining a trade-off between tighter Hessian bounds and larger lower bounds on the NTK condition. Given the growing literature on optimization of neural networks based on NTK analysis, we hope our work contributes by providing a better theoretical understanding on the performance of networks whose width may beneficially scale with the number of training samples.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 2019.

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

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b.

Arindam Banerjee, Pedro Cisneros-Velarde, Libin Zhu, and Misha Belkin. Restricted strong convexity of deep learning models with smooth activations. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Simone Bombari, Mohammad Hossein Amani, and Marco Mondelli. Memorization and optimization in deep neural networks with minimum over-parameterization. In *Advances in Neural Information Processing Systems*, 2022.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 09–15 Jun 2019.

Jianqing Fan, Cong Ma, and Yiqiao Zhong. A selective overview of deep learning. *Statistical Science*, 36(2):264 – 290, 2021.

Cong Fang, Hanze Dong, and Tong Zhang. Mathematical models of overparameterized neural networks. *Proceedings of the IEEE*, 109(5):683–703, 2021.

Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In *International conference on machine learning*, pages 4542–4551. PMLR, 2020.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

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*, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-Lojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

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

Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *arXiv preprint arXiv:2007.12826*, 2020.

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

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.

Quynh Nguyen, Pierre Bréchet, and Marco Mondelli. When are solutions connected in deep networks? In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021a.

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)*, 2021b.

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

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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