Towards moderate overparameterization: global convergence guarantees for training shallow neural networks

Samet Oymak and Mahdi Soltanolkotabi

Abstract

Many modern neural network architectures are trained in an overparameterized regime where the parameters of the model exceed the size of the training dataset. Sufficiently overparameterized neural network architectures in principle have the capacity to fit any set of labels including random noise. However, given the highly nonconvex nature of the training landscape it is not clear what level and kind of overparameterization is required for first order methods to converge to a global optima that perfectly interpolate any labels. A number of recent theoretical works have shown that for very wide neural networks where the number of hidden units is polynomially large in the size of the training data gradient descent starting from a random initialization does indeed converge to a global optima. However, in practice much more moderate levels of overparameterization seems to be sufficient and in many cases overparameterized models seem to perfectly interpolate the training data as soon as the number of parameters exceed the size of the training data by a constant factor. Thus there is a huge gap between the existing theoretical literature and practical experiments. In this paper we take a step towards closing this gap. Focusing on shallow neural nets and smooth activations, we show that (stochastic) gradient descent when initialized at random converges at a geometric rate to a nearby global optima as soon as the square-root of the number of network parameters exceeds the size of the training data. Our results also benefit from a fast convergence rate and continue to hold for non-differentiable activations such as Rectified Linear Units (ReLUs).

Department of Electrical and Computer Engineering, University of California, Riverside, CA
Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA

I. Introduction

A. Motivation

Modern neural networks typically have more parameters than the number of data points used to train them. This property allows neural nets to fit to any labels even those that are randomly generated [1]. Despite many empirical evidence of this capability the conditions under which this occurs is far from clear. In particular, due to this overparameterization, it is natural to expect the training loss to have numerous global optima that perfectly interpolate the training data. However, given the highly nonconvex nature of the training landscape it is far less clear why (stochastic) gradient descent can converge to such a globally optimal model without getting stock in subpar local optima or stationary points. Furthermore, what is the exact amount and kind of overpametrization that enables such global convergence? Yet another challenge is that due to overparameterization, the training loss may have infinitely many global minima and it is critical to understand the properties of the solutions found by first-order optimization schemes such as (stochastic) gradient descent starting from different initializations.

Recently there has been interesting progress aimed at demystifying the global convergence of gradient descent for overparameterized networks. However, most existing results focus on either quadratric activations [2], [3] or apply to very specialized forms of overparameterization [4]–[9] involving unrealistically wide neural networks where the number of hidden nodes are polynomially large in the size of the dataset. In contrast to this theoretical literature popular neural networks require much more modest amounts of overparameterization and do not typically involve extremely wide architectures. In particular (stochastic) gradient descent starting from a random initialization seems to find globally optimal network parameters that perfectly interpolate the training data as soon as the number of parameters exceed the size of the training data by a constant factor. See Section IV for some numerical experiments corroborating this claim. Also in such overparameterized regimes gradient descent seems to converge much faster than existing results suggest.

In this paper we take a step towards closing the significant gap between the theory and practice of overparameterized neural network training. We show that for training neural networks with one hidden layer, (stochastic) gradient descent starting from a random initialization finds globally

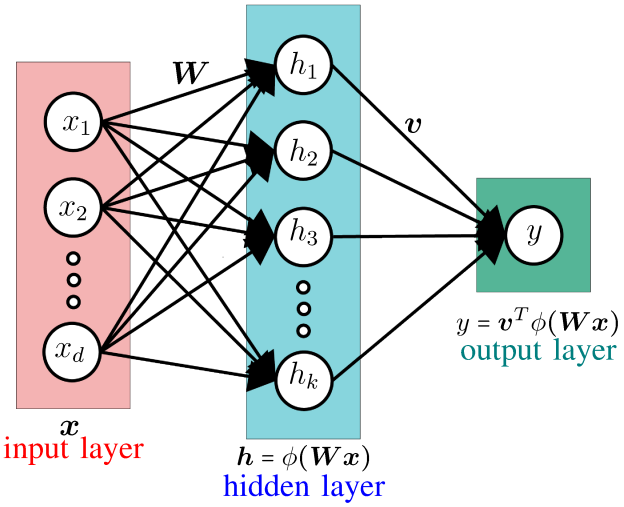


Fig. 1: Illustration of a one-hidden layer neural net with d inputs, k hidden units and a single output.

optimal weights that perfectly fit any labels as soon as the number of parameters in the model exceed the square of the size of the training data by numerical constants only depending on the input training data. This result holds for networks with differentiable activations. We also develop results of a similar flavor, albeit with slightly worse levels of overparameterization, for neural networks involving Rectified Linear Units (ReLU) activations. Our results also show that gradient descent converges at a much faster rate than existing gurantees. Our theory is based on combining recent results on overparameterized nonlinear learning [10] with more intricate tools from random matrix theory and bounds on the spectrum of Hadamard matrices. While in this paper we have focused on shallow neural networks with a quadratic loss, the mathematical techniques we develop are quite general and may apply more broadly. For instance, our techniques may help improve the existing guarantees for overparameterized deep networks ([6], [9]) or allow guarantees for other loss functions. We leave a detailed study of these cases to future work.

B. Model

We shall focus on neural networks with only one hidden layer with d inputs, k hidden neurons and a single output as depicted in Figure 1. The overall input-output relationship of the neural network in this case is a function $f(\cdot; \boldsymbol{W}) : \mathbb{R}^d \to \mathbb{R}$ that maps the input vector $\boldsymbol{x} \in \mathbb{R}^d$ into a scalar output via the following equation

$$\boldsymbol{x} \mapsto f(\boldsymbol{x}; \boldsymbol{W}) = \sum_{\ell=1}^{k} \boldsymbol{v}_\ell \phi\left(\langle \boldsymbol{w}_\ell, \boldsymbol{x} \rangle\right).$$

In the above the vectors $\boldsymbol{w}_\ell \in \mathbb{R}^d$ contains the weights of the edges connecting the input to the ℓth hidden node and $\boldsymbol{v}_\ell \in \mathbb{R}$ is the weight of the edge connecting the ℓth hidden node to the output. Finally, $\phi : \mathbb{R} \to \mathbb{R}$ denotes the activation function applied to each hidden node. For more compact notation we gather the weights $\boldsymbol{w}_\ell / \boldsymbol{v}_\ell$ into larger matrices $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ and $\boldsymbol{v} \in \mathbb{R}^k$ of the form

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \vdots \\ \boldsymbol{w}_k^T \end{bmatrix} \quad \text{and} \quad \boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix}.$$

We can now rewrite our input-output model in the more succinct form

$$\boldsymbol{x} \mapsto f(\boldsymbol{x}; \boldsymbol{W}) := \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}). \tag{I.1}$$

Here, we have used the convention that when ϕ is applied to a vector it corresponds to applying ϕ to each entry of that vector.

C. Notations

Before we begin discussing our main results we discuss some notation used throughout the paper. For a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ we use $\sigma_{\min}(\boldsymbol{X})$ and $\sigma_{\max}(\boldsymbol{X}) = \|\boldsymbol{X}\|$ to denote the minimum and maximum singular value of $\boldsymbol{X}$. For two matrices

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \\ \vdots \\ \boldsymbol{A}_p \end{bmatrix} \in \mathbb{R}^{p \times m} \quad \text{and} \quad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \\ \vdots \\ \boldsymbol{B}_p \end{bmatrix} \in \mathbb{R}^{p \times n},$$

we define their Khatri-Rao product as $\boldsymbol{A} * \boldsymbol{B} = \left[\boldsymbol{A}_1 \otimes \boldsymbol{B}_1, \ldots, \boldsymbol{A}_p \otimes \boldsymbol{B}_p\right] \in \mathbb{R}^{p \times mn}$, where $\otimes$ denotes the Kronecher product. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we denote their Hadamard (entrywise) product by $\boldsymbol{A} \odot \boldsymbol{B}$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, $\boldsymbol{A}^{\odot r} \in \mathbb{R}^{n \times n}$ is defined inductively via $\boldsymbol{A}^{\odot r} = \boldsymbol{A} \odot \left(\boldsymbol{A}^{\odot (r-1)}\right)$ with $\boldsymbol{A}^{\odot 0} = \boldsymbol{1}\boldsymbol{1}^T$. Similarly, for a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ with rows given by $\boldsymbol{x}_i \in \mathbb{R}^d$ we define the r-way Khatrio-Rao matrix $\boldsymbol{X}^{*r} \in \mathbb{R}^{n \times d^r}$ as a matrix with rows given by

$$\left[\boldsymbol{X}^{*r}\right]_i = \left(\underbrace{\boldsymbol{x}_i \otimes \boldsymbol{x}_i \otimes \ldots \otimes \boldsymbol{x}_i}_{r} \right)^T .$$

For a matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ we use $\text{vect}(\boldsymbol{W}) \in \mathbb{R}^{kd}$ to denote a column vector obtained by concatenating the rows $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k \in \mathbb{R}^d$ of $\boldsymbol{W}$. That is, $\text{vect}(\boldsymbol{W}) = \begin{bmatrix} \boldsymbol{w}_1^T & \boldsymbol{w}_2^T & \ldots & \boldsymbol{w}_k^T \end{bmatrix}^T$. Similarly, we use $\text{mat}\,(\boldsymbol{w}) \in \mathbb{R}^{k \times d}$ to denote a $k \times d$ matrix obtained by reshaping the vector $\boldsymbol{w} \in \mathbb{R}^{kd}$ across its rows. Throughout, for a differentiable function $\phi : \mathbb{R} \mapsto \mathbb{R}$ we use ϕ' and ϕ'' to denote the first and second derivative. If the function is not differentiable but only has isolated non-differentiable points we use ϕ' to denote a generalized derivative [11]. For instance, for $\phi(z) = ReLU(z) = \max(0, z)$ we have $\phi'(z) = \mathbb{1}_{\{z \geq 0\}}$ with $\mathbb{1}$ denoting the indicator mapping. We use c and C to denote numerical constants whose values may change from line to line. We also use the notation c_z to denote a numerical constant only depending on the variable or function z.

II. Main results

When training a neural network, one typically has access to a data set consisting of n feature/label pairs $(\boldsymbol{x}_i, y_i)$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ representing the features and y_i the associated labels. We wish to infer the best weights $\boldsymbol{v}, \boldsymbol{W}$ such that the mapping $f(\boldsymbol{x}; \boldsymbol{W}) := \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x})$ best fits the training data. In this paper we assume $\boldsymbol{v} \in \mathbb{R}^k$ is fixed and we train for the input-to-hidden weights $\boldsymbol{W}$ via a quadratic loss. The training optimization problem then takes the form

$$\min_{\boldsymbol{W} \in \mathbb{R}^{k \times d}} \mathcal{L}(\boldsymbol{W}) := \frac{1}{2} \sum_{i=1}^{n} \left(\boldsymbol{v}^T \phi\left(\boldsymbol{W}\boldsymbol{x}_i\right) - y_i\right)^2 . \tag{II.1}$$

To optimize this loss we run (stochastic) gradient descent starting from a random initialization $\boldsymbol{W}_0$. We wish to understand: (1) when such iterative updates lead to a globally optimal solution that perfectly interpolates the training data, (2) what are the properties of the solutions these algorithms converge to, and (3) what is the required amount of overparameterization necessary

for such events to occur. We begin by stating results for training via gradient descent for smooth activations in Section II-A followed by ReLU activations in Section II-B. Finally, we discuss results for training via Stochastic Gradient Descent (SGD) in Section II-C.

A. *Training networks with smooth activations via gradient descent*

In our first result we consider a one-hidden layer neural network with smooth activations and study the behavior of gradient descent in an over-parameterized regime where the number of parameters is sufficiently large.

Theorem 2.1: Consider a data set of input/label pairs $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$ aggregated as rows/entries of a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and a label vector $\boldsymbol{y} \in \mathbb{R}^n$. Without loss of generality we assume the dataset is normalized so that $\|\boldsymbol{x}_i\|_{\ell_2} = 1$. Also consider a one-hidden layer neural network with k hidden units and one output of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x})$ with $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ and $\boldsymbol{v} \in \mathbb{R}^k$ the input-to-hidden and hidden-to-output weights. We assume the activation ϕ has bounded derivatives i.e. $|\phi'(z)| \le B$ and $|\phi''(z)| \le B$ for all z and set $\mu_\phi = \mathbb{E}_{g\sim\mathcal{N}(0,1)}[g\phi'(g)]$. Furthermore, we set half of the entries of $\boldsymbol{v}$ to $\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{kn}}$ and the other half to $-\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{kn}}$[1] and train only over $\boldsymbol{W}$. Starting from an initial weight matrix $\boldsymbol{W}_0$ selected at random with i.i.d. $\mathcal{N}(0,1)$ entries we run Gradient Descent (GD) updates of the form $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{W}_\tau)$ on the loss (II.1) with step size $\eta = \frac{n\bar{\eta}}{2B^2\|\boldsymbol{y}\|_{\ell_2}^2\|\boldsymbol{X}\|^2}$ where $\bar{\eta} \le 1$. Then, as long as

$$\sqrt{kd} \ge c\frac{B^2}{\mu_\phi^2}(1+\delta)\kappa(\boldsymbol{X})n \quad \text{holds with} \quad \kappa(\boldsymbol{X}) := \frac{\sqrt{\frac{d}{n}}\,\|\boldsymbol{X}\|}{\sigma_{\min}^2(\boldsymbol{X}*\boldsymbol{X})}, \tag{II.2}$$

and c is a fixed numerical constant, then with probability at least $1 - \frac{1}{n} - e^{-\delta^2\frac{n}{2\|\boldsymbol{X}\|^2}}$ all GD iterates obey

$$\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le \left(1 - \frac{\bar{\eta}}{32}\frac{\mu_\phi^2}{B^2}\frac{\sigma_{\min}^2(\boldsymbol{X}*\boldsymbol{X})}{\|\boldsymbol{X}\|^2}\right)^\tau \|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2},$$

$$\frac{\mu_\phi}{\sqrt{32}}\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{n}}\sigma_{\min}(\boldsymbol{X}*\boldsymbol{X})\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F + \|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le \|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2}.$$

[1] If k is odd we set one entry to zero $\lfloor\frac{k-1}{2}\rfloor$ to $\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{kn}}$ and $\lfloor\frac{k-1}{2}\rfloor$ entries to $-\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{kn}}$.

Furthermore, the total gradient path obeys

$$\sum_{\tau=0}^{\infty} \|\boldsymbol{W}_{\tau+1} - \boldsymbol{W}_{\tau}\|_F \leq \frac{\sqrt{32}}{\mu_\phi} \frac{\sqrt{n}}{\|\boldsymbol{y}\|_{\ell_2}} \frac{\|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2}}{\sigma_{\min}(\boldsymbol{X} * \boldsymbol{X})}.$$

We would like to note that we have chosen to state our results based on easy to calculate quantities such as $\sigma^2_{\min}(\boldsymbol{X} * \boldsymbol{X})$ and μ_ϕ. As it becomes clear in the proofs a more general result holds where the theorem above and its conclusions can be stated with $\mu_\phi \sigma_{\min}(\boldsymbol{X} * \boldsymbol{X})$ replaced with a quantity that only depends on the expected minimum singular value of the Jacobian of the neural network mapping at the random initialization (See Theorem 6.2 in the proofs for details). Using this more general result combined with well known calculations involving Hermite polynomials one can develop other interpretable results. For instant we can show that $\sigma_{\min}(\boldsymbol{X} * \boldsymbol{X})$ can be replaced with higher order Khatrio-Rao products (i.e. $\sigma_{\min}(\boldsymbol{X}^{*r})$).

Before we start discussing the conclusions of this theorem let us briefly discuss the scaling of various quantities. When $n \gtrsim d$, in many cases we expect $\|\boldsymbol{X}\|$ to grow with $\sqrt{n/d}$ and $\sigma_{\min}(\boldsymbol{X} * \boldsymbol{X})$ to be roughly a constant so that $\kappa(\boldsymbol{X})$ is typically a constant (see Corollary 2.2 below for a precise statement). Thus based on (II.2) the typical scaling required in our results is $kd \gtrsim n^2$. That is, the conclusions of Theorem 2.1 holds with high probability as soon as the square-root of the number of parameters of the model exceed the number of training data by a fixed numerical constant. To the extent of our knowledge this result is the first of its kind only requiring the number of parameters to be sufficiently large w.r.t. the training data rather than the number of hidden units w.r.t. the size of the training data. That said, as we demonstrate in Section IV neural networks seem to work with even more modest amounts of overparameterization and when the number of parameters exceed the size of the training data by a numerical constant i.e. $kd \gtrsim n$. We hope to close this remaining gap in future work. We also note that based on this typical scaling the convergence rate is on the order of $\left(1 - c\frac{d}{n}\right)$.

We briefly pause to also discuss the case where one assumes $n \lesssim d$ (although this is not a typical regime of operation in neural networks). In this case both $\|\boldsymbol{X}\|$ and $\sigma_{\min}(\boldsymbol{X} * \boldsymbol{X})$ are of the order of one and thus κ scales as $\sqrt{d/n}$. Thus, the overparmeterization requirement (II.2) reduces to $k \gtrsim n$. Thus, in this regime we can perfectly fit any labels as soon as the number of hidden units exceeds the size of the training data. We also note that in this regime the convergence rate is a fixed numerical constant independent of any of the dimensions.

Before we start discussing the conclusions of this theorem let us state a simple corollary that clearly illustrates the scaling discussed above for randomly generated input data. The proof of this simple corollary is deferred to Appendix E.

Corollary 2.2: Consider the setting of Theorem 2.1 above with $\eta = \frac{n}{2B^2\|\boldsymbol{y}\|_{\ell_2}^2\|\boldsymbol{X}\|^2}$. Furthermore, assume the we use the softplus activation $\phi(z) = \log(1+e^z)$ and the input data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are generated i.i.d. uniformly at random from the unit sphere of $\mathbb{R}^d$ where $d \le n \le cd^2$. Then, as long as

$$\sqrt{kd} \ge Cn, \tag{II.3}$$

with probability at least $1 - \frac{2}{n} - e^{-\frac{d}{4}} - ne^{-\gamma_1\sqrt{n}} - (2n+1)e^{-\gamma_2 d}$ all GD iterates obey

$$\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le 3\left(1 - c_1\frac{d}{n}\right)^\tau \|\boldsymbol{y}\|_{\ell_2}, \tag{II.4}$$

$$\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F + c_2\frac{\sqrt{n}}{\|\boldsymbol{y}\|_{\ell_2}}\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le c_3\sqrt{n}. \tag{II.5}$$

Here, $\gamma_1, \gamma_2, c, C, c_1, c_2$, and c_3 are fixed numerical constants.

We would like to note that while for simplicity this corollary is stated for data points that are uniform on the unit sphere, as it becomes clear in the proof, this result continues to hold for a variety of other generic[2] data models with the same scaling. The corollary above clarifies that the typical scaling required in our results is indeed $kd \gtrsim n^2$. That is, the conclusions of Theorem 2.1 holds with high probability as soon as the square-root of the number of parameters of the model exceed the number of training data by a fixed numerical constant.

The theorem and corollary above show that under $kd \gtrsim n^2$ overparameterization Gradient Descent (GD) iterates have a few interesting properties properties:

Zero traning error: The first property demonstrated by Theorem 2.1 above is that the iterates converge to a global optima. This holds despite the fact that the fitting problem may be highly nonconvex in general. Indeed, based on (II.4) the fitting/training error $\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2}$ achieved by Gradient Descent (GD) iterates converges to zero. Therefore, GD can perfectly interpolate the data and achieve zero training error. Furthermore, the algorithm enjoys a fast geometric rate of convergence to this global optima. In particular to achieve a relative accuracy of ϵ

[2]Informally, we call a set of points generic as long as no subset of them belong to an algebraic manifold.

(i.e. $\|f(\boldsymbol{W}_\tau)-\boldsymbol{y}\|_{\ell_2}/\|\boldsymbol{y}\|_{\ell_2} \le \epsilon$) the required number of iterations τ is of the order of $\tau \gtrsim \frac{n}{d}\log(1/\epsilon)$.

Gradient descent iterates remain close to the initialization: The second interesting aspect of our result is that we guarantee the GD iterates never leave a neighborhood of radius of the order of $\sqrt{n}$ around the initial point. That is the GD iterates remain rather close to the initialization.[3] Furthermore, (II.5) shows that for all iterates the weighted sum of the distance to the initialization and the misfit error remains bounded so that as the loss decreases the distance to the initialization only moderately increases.

Gradient descent follows a short path: Another interesting aspect of the above results is that the total length of the path taken by gradient descent remains bounded and is of the order of $\sqrt{n}$.

B. Training ReLU networks via gradient descent

The results in the previous section focused on smooth activations and therefore does not apply to non-differentiable activations and in particular the widely popular ReLU activations. In the next theorem we show that a similar result continues to hold when ReLU activations are used.

Theorem 2.3: Consider the setting of Theorem 2.1 with the activations equal to $\phi(z) = ReLU(z) := \max(0,z)$ and the step size $\eta = \frac{n}{3\|\boldsymbol{y}\|_{\ell_2}^2\|\boldsymbol{X}\|^2}\bar{\eta}$ with $\bar{\eta} \le 1$. Then, as long as

$$\sqrt{kd} \ge C(1+\delta)\frac{n^2}{d}\kappa^3(\boldsymbol{X})\,\sigma_{\min}^2(\boldsymbol{X}*\boldsymbol{X}) \quad \text{holds with} \quad \kappa(\boldsymbol{X}) := \frac{\sqrt{\frac{d}{n}}\|\boldsymbol{X}\|}{\sigma_{\min}^2(\boldsymbol{X}*\boldsymbol{X})}, \tag{II.6}$$

and γ and c fixed numerical constants, then with probability at least $1-\frac{1}{n}-e^{-\delta^2\frac{n}{\|\boldsymbol{X}\|^2}}-ne^{-n}$ all GD iterates obey

$$\|f(\boldsymbol{W}_\tau)-\boldsymbol{y}\|_{\ell_2} \le \left(1-\frac{\bar{\eta}}{48\pi}\frac{\sigma_{\min}^2(\boldsymbol{X}*\boldsymbol{X})}{\|\boldsymbol{X}\|^2}\right)^\tau \|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2},$$

$$\frac{1}{12\sqrt{\pi}}\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{n}}\sigma_{\min}(\boldsymbol{X}*\boldsymbol{X})\|\boldsymbol{W}_\tau-\boldsymbol{W}_0\|_F + \|f(\boldsymbol{W}_\tau)-\boldsymbol{y}\|_{\ell_2} \le \|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2}.$$

Also similar to Corollary 2.2 we can state the following simple corollary to better understand the requirement in typical instances.

[3]Note that $\|\boldsymbol{W}_0\|_F \approx \sqrt{kd} >> \sqrt{n}$ so that this radius is indeed small.

Corollary 2.4: Consider the setting of Theorem 2.3 above with $\eta = \frac{n}{\|\boldsymbol{y}\|_{\ell_2}^2 \|\boldsymbol{X}\|^2}$. Furthermore, assume the input data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are generated i.i.d. uniformly at random from the unit sphere of $\mathbb{R}^d$ where $d \le n \le cd^2$. Then, as long as

$$\sqrt{kd} \ge C\frac{n^2}{d}, \tag{II.7}$$

with probability at least $1 - \frac{2}{n} - e^{-d} - ne^{-\gamma_1\sqrt{n}} - (2n+1)e^{-\gamma_2 d}$ all GD iterates obey

$$\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le 3\left(1 - c_1\frac{d}{n}\right)^\tau \|\boldsymbol{y}\|_{\ell_2},$$

$$\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F + c_2\frac{\sqrt{n}}{\|\boldsymbol{y}\|_{\ell_2}}\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le c_3\sqrt{n}.$$

Here, $\gamma_1, \gamma_2, c, C, c_1, c_2$, and c_3 are fixed numerical constants.

The theorem and corollary above show that all the nice properties of GD with smooth activations continue to hold for ReLU activations. The only difference is that the required overparameterization is now of the form $\sqrt{kd} \ge C\frac{n^2}{d}$ which is suboptimal compared to the smooth case by a factor of n/d (factor of n^2/d^2 in terms of number of parameters kd).

Our discussion so far focused on results based on the minimum singular value of the second order Khatrio-Rao product $\boldsymbol{X} * \boldsymbol{X}$ or higher order products $\boldsymbol{X}^{*r}$. The reason we require these minimum singular values to be positive is to ensure diversity in the data set. Indeed, if two data points are the same but have different output labels there is no way of achieving zero training error. However, assuming these minimum singular values are positive is not the only way to ensure diversity and our results apply more generally (see Theorem 6.3 in the proofs). Another related and intuitive criteria for ensuring diversity is assuming the input samples are sufficiently separated as defined below.

Assumption 1 (δ-separable data): Let $\delta > 0$ be a scalar. Consider a data set consisting of n samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ all with unit Euclidian norm. We assume that any pair of points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ obey

$$\min(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\ell_2}, \|\boldsymbol{x}_i + \boldsymbol{x}_j\|_{\ell_2}) \ge \delta.$$

We now state a result based on this minimum separation assumption. This result is a corollary of our meta theorem (Theorem 6.3) discussed in the proofs.

Theorem 2.5: Consider the setting of Theorem 2.1 with the activations equal to $\phi(z) = ReLU(z) := \max(0, z)$ and the step size $\eta = \frac{n}{3\|\boldsymbol{y}\|_{\ell_2}^2 \|\boldsymbol{X}\|^2}\bar{\eta}$ with $\bar{\eta} \leq 1$. Suppose Assumption 1 holds for some $\delta > 0$ and let $c, C > 0$ be two numerical constants. Suppose number of hidden nodes satisfy

$$k \geq C(1+\nu)^2 \frac{n^9 \|\boldsymbol{X}\|^6}{\delta^4}, \tag{II.8}$$

Then with probability at least $1 - \frac{2}{n} - e^{-\nu^2 \frac{n}{\|\boldsymbol{X}\|^2}}$ all GD iterates obey

$$\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \leq \left(1 - c\frac{\bar{\eta}\delta}{n^2 \|\boldsymbol{X}\|^2}\right)^\tau \|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2},$$

We would like to note that related works [4], [6], [8] consider slight variations of this assumption for training ReLU networks to give overparameterized learning guarantees where the number of hidden nodes grow polynomially in n. Our results seem to have much better dependencies on n compared to these works. Furthermore, we do not require the number of hidden nodes to scale with the desired training accuracy ($\mathcal{L}(\boldsymbol{W}) \leq \epsilon$) as required by [4]. Finally, we would like to note that after the first version of this paper appeared on arxiv, the paper [12] improves the dependence of the width to $k \gtrsim n^8$. However, we note that the main results of our paper is based on minimum eigenvalue of Khatri-Rao product (or alternatively minimum eigenvalue of neural tangent kernels ($\lambda(X)$) per Definition 6.1 in the proofs). The theorem above is obtained by replacing these quantities with a crude lower bound in terms of the separation assumption (specifically $\lambda(\boldsymbol{X}) \gtrsim \frac{\delta}{n^2}$). Thus the main results of the two papers are not directly comparable (excluding theorem above of course). In fact, we believe a more accurate lower bound connects minimum eigenvalue of neural tangent kernels to the separation assumption (specifically $\lambda(\boldsymbol{X}) \gtrsim \frac{\delta}{n}$) which in turn would further improve our result above to $k \gtrsim n^5 \|\boldsymbol{X}\|^6$.

C. Training using SGD

The most widely used algorithm for training neural networks is Stochastic Gradient Descent (SGD) and its variants. A natural implementation of SGD is to sample a data point at random

and use that data point for the gradient updates. Specifically, let $\{\gamma_\tau\}_{\tau=0}^{\infty}$ be an i.i.d. sequence of integers chosen uniformly from $\{1, 2, \ldots, n\}$, the SGD iterates take the form

$$\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau + \eta G(\boldsymbol{\theta}_\tau; \gamma_\tau) \quad \text{where} \quad G(\boldsymbol{\theta}_\tau; \gamma_\tau) := (y_{\gamma_\tau} - f(\boldsymbol{x}_{\gamma_\tau}; \boldsymbol{W}_\tau))\nabla f(\boldsymbol{x}_{\gamma_\tau}; \boldsymbol{W}_\tau). \tag{II.9}$$

Here, $G(\boldsymbol{\theta}_\tau; \gamma_\tau)$ is the gradient on the γ_τth training sample. We are interested in understanding the trajectory of SGD for neural network training e.g. the required overparameterization and the associated rate of convergence. We state our result for smooth activations. An analogous result also holds for ReLU activations but we omit the statement to avoid repetition.

Theorem 2.6: Fix scalars $\bar{\eta} \le 1$ and $\nu \ge 4$ and consider the setting and assumptions of Theorem 2.1. Suppose the number of network parameters obey $\sqrt{kd} \ge c\frac{\nu B^2}{\mu_\phi^2}(1+\delta)\kappa(\boldsymbol{X})n$ and we use the SGD updates (II.9) in lieu of GD updates with a step size $\eta = \frac{\mu^2(\phi)}{9\nu B^4}\frac{n}{\|\boldsymbol{y}\|_{\ell_2}^2}\frac{\sigma_{\min}^2(\boldsymbol{X}*\boldsymbol{X})}{\|\boldsymbol{X}\|^2}\bar{\eta}$ with c a fixed numerical constant. Set initial weights $\boldsymbol{W}_0$ with i.i.d. $\mathcal{N}(0,1)$ entries. Then, with probability at least $1 - \frac{1}{n} - e^{-\delta^2\frac{n}{2\|\boldsymbol{X}\|^2}}$ over $\boldsymbol{W}_0$, there exists an event E[4] which holds with probability at least $\mathbb{P}(E) \ge 1 - \frac{4}{\nu}\left(\frac{3B\|\boldsymbol{X}\|}{\mu_\phi\sigma_{\min}(\boldsymbol{X}*\boldsymbol{X})}\right)^{\frac{1}{kd}}$ such that, starting from $\boldsymbol{W}_0$ and running stochastic gradient descent updates of the form (II.9), all iterates obey

$$\mathbb{E}\Big[\,\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2}^2\,\mathbb{1}_E\Big] \le \left(1 - \frac{\bar{\eta}}{144n}\frac{\mu^4(\phi)}{\nu B^4}\frac{\sigma_{\min}^4(\boldsymbol{X}*\boldsymbol{X})}{\|\boldsymbol{X}\|^2}\right)^{\tau}\|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2}^2\,, \tag{II.10}$$

Here, E is the event that the infinite SGD sequence never leads the neural network parameters outside of a certain neighborhood of the initialization $\boldsymbol{W}_0$. We prove that this event has high probability and all SGD iterations stay close despite the random nature of the iterates. Recall that the distance to initialization is critical for our gradient descent analysis (also see Sec. II-D). Thus, the event E naturally arises in the SGD analysis because outside of a nearby neighborhood of the initialization we have essentially no control over the optimization process and the SGD sequences in the complement of E might diverge.

This result shows that SGD converges to a global optima that is close to the initialization. Furthermore, SGD always remains in close proximity to the initialization with high probability. To assess the rate of convergence, let us assume generic data and $n \ge d$, so that we have $\|\boldsymbol{X}\| \sim \sqrt{n/d}$ and $\sigma_{\min}(\boldsymbol{X} * \boldsymbol{X})$ scales as a constant. Then, the result above shows that to achieve a relative

[4]This event is over the randomness introduced by the infinite sequence of random SGD updates given fixed $\boldsymbol{W}_0$.

accuracy of ε the number of SGD iterates required is of the order of $\tau \gtrsim \frac{n^2}{d}\log(\frac{1}{\varepsilon})$. This is essentially on par with our earlier result on gradient descent by noting that n SGD iterations require similar computational effort to one full gradient with both approaches requiring $\frac{n}{d}\log(1/\epsilon)$ passes through the data.

D. Key proof ideas

In this section, we briefly discuss the critical ingredients of our proof strategy. At a high level, our proof relies on two facts surrounding the Jacobian map of a neural network.

Importance of minimum singular value: The first observation is that, despite non-convexity, gradient descent iterations can decrease the loss function substantally. Specifically, the amount of decrease is tied to the minimum singular value of the Jacobian map of the neural network. Thus, as long as the Jacobian map remains well-conditioned, the loss function will keep decreasing (exponentially fast). Our core technical novelties along this direction are two-fold. First, in Section VI-D, we develop new and tighter bounds for the minimum singular value via intricate eigenvalue bounds for the Hadamard product of two matrices and random matrix theory. Second, in Section H, we further relate the Jacobian of neural networks to high-order Khatri-Rao products associated to the input matrix $\boldsymbol{X}$ to obtain more interpretable bounds.

Utilizing Lipschitzness and conditioning of the Jacobian via a Lyapunov analysis: The second critical ingredient of the proof is ensuring that the Jacobian map does not degrade over time i.e. it has a reasonable condition number throughout the optimization process. One way to achieve this is making the neural network extremely wide as it can be shown that (this work as well as others [7], [8]) the wider the neural network gets, the less the Jacobian deviates throughout the iterative updates. In our case, we argue that small width is sufficient which requires a tighter control on the deviation of the Jacobian from its initial state in a rather large neighborhood. We accomplish this by first bounding the Lipschitzness of the Jacobian map with respect to the parameters (Section C) and then showing that the condition number of the Jacobian is maintained in a relatively large neighborhood. Our tighter bounds on the Lipschitzness arise from eigenvalue inequalities of Section VI-D as well as ReLU specific analysis in Section C2. We also show that the minimum eigenvalue of the Jacobian is bounded away from zero in a

much larger neighborhood of the initialization compared to other related works [6]–[8]. This is based on a novel Lypunov analysis developed in our previous work [10].

To summarize, we achieve smaller over-parameterization by tightly controlling the critical properties of the neural network Jacobian (both for smooth and ReLU activations) which in turn allow us to accurately assess optimization dynamics and show global convergence.

III. The need for overparameterization beyond width

In this section we would like to further clarify why understanding overparameterization beyond width is particularly important. To see this, we shall set the input-to-hidden weights at random (as used for initialization) and consider the optimization over the output layer weights $\boldsymbol{v} \in \mathbb{R}^k$. This optimization problem has the form

$$\mathcal{L}(\boldsymbol{v}) := \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{v}^T\phi\left(\boldsymbol{W}\boldsymbol{x}_i\right) - \boldsymbol{y}_i\right)^2 = \frac{1}{2}\left\|\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v} - \boldsymbol{y}\right\|_{\ell_2}^2, \tag{III.1}$$

which is a simple least-squares problem with a globally optimal solution given by

$$\hat{\boldsymbol{v}} := \boldsymbol{\Phi}^T\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T\right)^{-1}\boldsymbol{y} \quad \text{where} \quad \boldsymbol{\Phi} := \phi\left(\boldsymbol{X}\boldsymbol{W}^T\right).$$

This simple observation shows that the simple least-squares optimization over the output weights achieves zero training as soon as $\boldsymbol{\Phi}$ has full column rank. Thus, in such a setting a simple kernel regression using the random features $\phi(\boldsymbol{W}\boldsymbol{x}_1), \phi(\boldsymbol{W}\boldsymbol{x}_2), \ldots, \phi(\boldsymbol{W}\boldsymbol{x}_n)$ suffices to perfectly interpolate the data. In this section we wish to understand the amount and kind of overparameterization where such a simple strategy suffices. We thus need to understand the conditions under which the matrix $\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)$ has full row rank. To make things quantitative we need the following definition.

Definition 3.1 (Output feature covariance and eigenvalue): We define the output feature covariance matrix as

$$\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{w}}\left[\phi\left(\boldsymbol{X}\boldsymbol{w}\right)\phi\left(\boldsymbol{X}\boldsymbol{w}\right)^T\right],$$

where $\boldsymbol{w} \in \mathbb{R}^d$ has a $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ distribution. We use $\widetilde{\lambda}(\boldsymbol{X})$ to denote the corresponding minimum eigenvalue i.e. $\widetilde{\lambda}(\boldsymbol{X}) = \lambda_{\min}\left(\widetilde{\Sigma}(\boldsymbol{X})\right)$.

With this definition in place we are now ready to state the main result of this section.

Theorem 3.2: Consider a data set of input/label pairs $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$ aggregated as rows/entries of a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and a label vector $\boldsymbol{y} \in \mathbb{R}^n$. Without loss of generality we assume the dataset is normalized so that $\|\boldsymbol{x}_i\|_{\ell_2} = 1$. Also consider a one-hidden layer neural network with k hidden units and one output of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x})$ with $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ and $\boldsymbol{v} \in \mathbb{R}^k$ the input-to-hidden and hidden-to-output weights. We assume the activation ϕ is bounded at zero i.e. $|\phi(0)| \leq B$ and has a bounded derivative i.e. $|\phi'(z)| \leq B$ for all z. We set $\boldsymbol{W}$ to be a random matrix with i.i.d. $\mathcal{N}(0, 1)$ entires. Also assume

$$k \geq C \log^2(n) \frac{n}{\widetilde{\lambda}(\boldsymbol{X})}.$$

Then, the matrix $\boldsymbol{\Phi} := \phi(\boldsymbol{X}\boldsymbol{W}^T)$ has full row rank with the minimum eigenvalue obeying

$$\lambda_{\min}\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T\right) \geq \frac{1}{2} k \left(\widetilde{\lambda}(\boldsymbol{X}) - \frac{6B}{n^{100}}\right).$$

Thus, the global optima of (III.1) achieves zero training error as long as $\widetilde{\lambda}(\boldsymbol{X}) \geq \frac{6B}{n^{100}}$.

We note that one can develop interpretable lower bounds for $\widetilde{\lambda}$ (see Appendix H). For instance, in Appendix H we show that

$$\widetilde{\lambda}(\boldsymbol{X}) \geq \gamma_\phi^2 \sigma_{\min}^2(\boldsymbol{X} * \boldsymbol{X}) \quad \text{with} \quad \gamma_\phi = \frac{1}{\sqrt{2}} \mathbb{E}[\phi(g)(g^2 - 1)].$$

As we discussed in the previous sections for generic or random data $\sigma_{\min}^2(\boldsymbol{X} * \boldsymbol{X})$ often scales like a constant. In turn, based on the above inequality $\widetilde{\lambda}(\boldsymbol{X})$ also scales like a constant. Thus, the above theorem shows that as long as the neural network is wide enough in the sense that $k \gtrsim n$, with high probability on can achieve perfect interpolation and the global optima by simply fitting the last layer with the input-to-hidden weights set randomly. Of course the optimization problem over $\boldsymbol{W}$ is significantly more challenging to analyze (the setting in this paper and other publications [4], [6], [8], [9], [13]). However, this simple baseline result suggests that there is no fundamental barrier to understanding perfect interpolation for $k \gtrsim n$ wide networks. In particular, as discussed earlier the result above can be thought of as kernel learning with random features. Indeed, in this settings one can also show the solutions found by (stochastic) gradient descent converges to the least-norm solution and does indeed generalize. Furthermore, neural networks are often trained with the number of hidden nodes at each intermediate layer significantly smaller than the size of the training dataset. Thus to truly understand the behavior of neural network

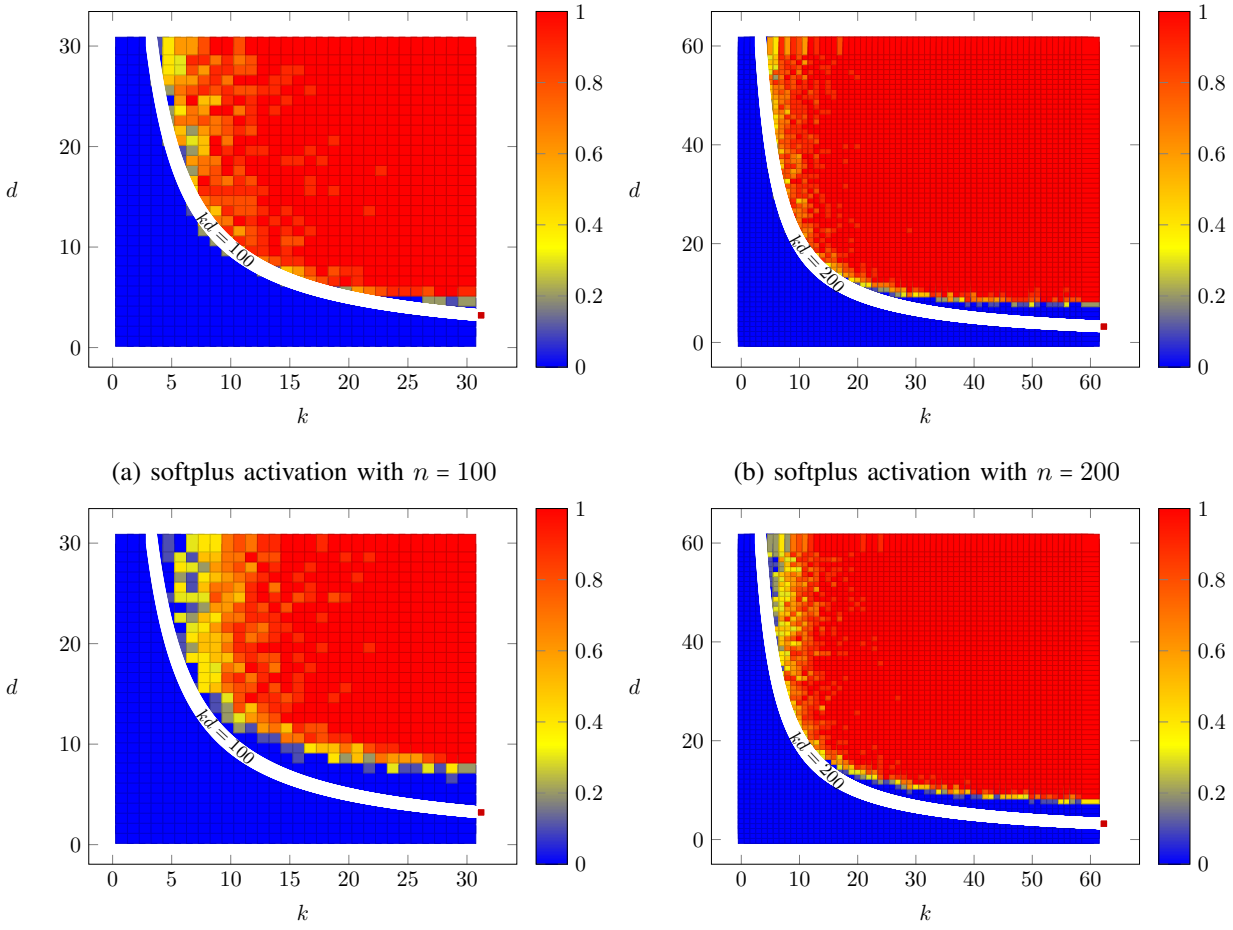
training and demystify their success beyond kernel learning it is crucially important to focus on *moderately* overparameterized networks where the number of data points is only moderately larger than the number of parameters used for training. We hope the discussion above can help focus future theoretical investigations to this moderately overparameterized regime.

IV. Numerical experiments

In this section, we provide numerical evidence that neural networks trained with first order methods can fit to random data as long as the number of parameters exceed the size of the training dataset. In particular, we explore the fitting ability of a shallow neural network by fixing a dataset size n and scanning over the different values of hidden nodes k and input dimension d. The input samples are drawn i.i.d. from the unit sphere, the labels i.i.d. standard normal variables and the input/output weights of the network are initialized according to our theorems. We consider two activations softplus ($\phi(z) = softplus(z) = \log(1+e^z)$) and Rectified Linear Units ($\phi(z) = ReLU(z) = max(0,z)$). We pick a constant learning rate of $\eta = 0.15$ for softplus and $\eta = 0.1$ for ReLU activations. We run the updates for 15000 iterations or when the relative Euclidean error ($\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} / \|\boldsymbol{y}\|_{\ell_2}$) falls below 2.5×10^{-3}. Success is declared if the relative loss is less than 2.5×10^{-3}. To obtain an empirical probability, we average 10 independent realizations for each (k,d) pair.

Figure 2a plots the success probability where $n = 100$ and k and d are varied between 0 to 25. The solid white line represents the $n = kd$. There is a visible phase transition from failure to success as k and d grows. Perhaps more surprisingly, the success region is tightly surrounded by the $n = kd$ curve indicating that neural nets can overfit as soon as the problem is slightly overparameterized. Figure 2b repeats the same experiment with a larger dataset ($n = 200$). Phase transitions are more visible in higher dimensions due to concentration of measure phenomena. Indeed, $n = kd$ curve matches the success region even tighter indicating that $kd > (1+\varepsilon)n$ amount of overparametrization may suffice for fitting random data.

A related set of experiments are based on assigning random labels in classification problems [1]. These experiments shuffle the labels of real datasets (e.g. CIFAR10) and demonstrate that standard deep architectures can still fit them (even if the training takes a bit longer). While these experiments provide very interesting and useful insights the do not address the fundamental



(a) softplus activation with $n = 100$

(b) softplus activation with $n = 200$

(c) ReLU activation with $n = 100$

(d) ReLU activation with $n = 200$

Fig. 2: **Phase transitions for overparameterization.** These diagrams show the empirical probability that gradient descent from a random initialization successfully fits n random labels $\boldsymbol{y} \in \mathbb{R}^n$ when a one-hidden layer neural network is used. Here, d is the input dimension, k the number of hidden units, and n the size of the training data. The colormap tapers between red and blue where red represents certain success, while blue represents certain failure. The solid white line highlights $n = kd$ i.e. when the size of the training data is equal to the number of parameters.

tradeoffs surrounding problem parameters such as $n, k,$ and d. Finally, we emphasize that the dataset in our experiment is randomly generated. It is possible that worst case datasets exhibit different phase transitions. For instance, if two identical inputs receive different outputs a significantly higher amounts of overparameterization may be required.

V. Prior art

Optimization of neural networks is a challenging problem and it has been the topic of many recent works [1]. A large body of work focuses on understanding the optimization landscape of the simple nonlinearities or neural networks [14]–[22] when the labels are created according to a planted model. These works establish local convergence guarantees and use techniques such as tensor methods to initialize the network in the proper local neighborhood. Ideally, one would not need specialized initialization if loss surface has no spurious local minima. However, a few publications [23], [24] demonstrate that the loss surface of nonlinear networks do indeed contains spurious local minima even when the input data are random and the labels are created according to a planted model.

Over-parameterization seems to provide a way to bypass the challenging optimization landscape by relaxing the problem. Several works [1]–[3], [10], [25]–[35] study the benefits of overparameterization for training neural networks and related optimization problems. Very recent works [4], [6], [8], [9], [13] show that overparameterized neural networks can fit the data with random initialization if the number of hidden nodes are polynomially large in the size of the dataset. While these results are based on assuming the networks are sufficiently wide with respect to the size of the data set we only require the total number of parameters to be sufficiently large. Since our conclusions and assumptions are more closely related to [9], [13] we focus precise comparisons to these two publications. In particular, for smooth activations we show that neural networks can fit the data as soon as $kd \gtrsim n^2$ where as [9] requires $k \gtrsim n^4$. Thus, in terms of the hidden units our results are sharper by a factor on the order of $n^2 d$.[5] Focusing on ReLU networks we require $k \gtrsim \frac{n^4}{d^3}$ compared to $k \gtrsim n^6$ assumed in [13] so that our results are sharper

[5]Our results are also sharper in terms of dependence on the quantity λ defined in the proofs. In more detail, we require $kd \gtrsim \frac{n^2}{\lambda^2}$ where as [9] requires $k \gtrsim \frac{n^4}{\lambda^4}$.

by a factor n^2d^3. Our convergence rate for gradient descent also seems to be faster by a factor on the order of n compared to these results. In addition our results extend to SGD. We would like to note however that our results focus on one-hidden layer networks where as some of the publications above such as [6], [9] apply to deep architectures. That said, our results and proof strategy can be extended to deeper architectures and we hope to study such networks in our future work. Finally, these recent papers as well as our work is inherently based on connecting neural networks to kernel methods. We would like to note that the relationship between kernel methods and deep learning has been emphasized by a few interesting publications [36]–[39].

We would also like to note that a few interesting recent papers [31], [40]–[42] relate the empirical distribution of the network parameters to Wasserstein gradient flows using ideas from mean field analysis. However, this literature is focused on asymptotic characterizations rather than finite-size networks.

An equally important question to understanding the convergence behavior of optimization algorithms for overparameterized models is understanding their generalization capabilities. This is the subject of a few interesting recent papers [5], [38], [43]–[49]. While this work do not directly address generalization, techniques developed here (e.g. characterizing how far the global minima is) may help demystify the generalization capabilities of overparametrized networks trained via first order methods. Rigorous understanding of the relationship between optimization and generalization is an interesting and important subject for future research.

VI. Proofs

A. Preliminaries

We begin by noting that for a one-hidden layer neural network of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x})$, the Jacobian matrix with respect to $\mathrm{vect}(\boldsymbol{W}) \in \mathbb{R}^{kd}$ takes the form

$$\mathcal{J}(\boldsymbol{W}) = \begin{bmatrix} \mathcal{J}(\boldsymbol{w}_1) & \dots & \mathcal{J}(\boldsymbol{w}_k) \end{bmatrix} \in \mathbb{R}^{n\times kd} \quad \text{with} \quad \mathcal{J}(\boldsymbol{w}_\ell) := \boldsymbol{v}_\ell \mathrm{diag}(\phi'(\boldsymbol{X}\boldsymbol{w}_\ell))\boldsymbol{X}.$$

Alternatively this can be rewritten in the form

$$\mathcal{J}^T(\boldsymbol{W}) = \left(\mathrm{diag}(\boldsymbol{v})\phi'\left(\boldsymbol{W}\boldsymbol{X}^T\right)\right) * \boldsymbol{X}^T \tag{VI.1}$$

An alternative characterization of the Jacobian is

$$\operatorname{mat}\left(\mathcal{J}^T(\boldsymbol{W})\boldsymbol{u}\right) = \operatorname{diag}(\boldsymbol{v})\phi'\left(\boldsymbol{W}\boldsymbol{X}^T\right)\operatorname{diag}(\boldsymbol{u})\boldsymbol{X}$$

In particular, given a residual misfit $\boldsymbol{r} := \boldsymbol{r}(\boldsymbol{W}) := \phi\left(\boldsymbol{W}\boldsymbol{X}^T\right)^T\boldsymbol{v} - \boldsymbol{y} \in \mathbb{R}^n$ the gradient can be rewritten in the form

$$\nabla\mathcal{L}(\boldsymbol{W}) = \operatorname{mat}\left(\mathcal{J}^T(\boldsymbol{W})\boldsymbol{r}\right) = \operatorname{diag}(\boldsymbol{v})\phi'\left(\boldsymbol{W}\boldsymbol{X}^T\right)\operatorname{diag}(\boldsymbol{r})\boldsymbol{X}$$

We also note that

$$\mathcal{J}(\boldsymbol{W})\mathcal{J}^T(\boldsymbol{W}) = \sum_{\ell=1}^{k}\boldsymbol{v}_\ell^2\operatorname{diag}\left(\phi'\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\right)\boldsymbol{X}\boldsymbol{X}^T\operatorname{diag}\left(\phi'\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\right).$$

The latter can also be rewritten in the more compact form

$$\mathcal{J}(\boldsymbol{W})\mathcal{J}^T(\boldsymbol{W}) = \left(\phi'\left(\boldsymbol{X}\boldsymbol{W}^T\right)\operatorname{diag}\left(\boldsymbol{v}\right)\operatorname{diag}\left(\boldsymbol{v}\right)\phi'\left(\boldsymbol{W}\boldsymbol{X}^T\right)\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right).$$

B. Meta-theorems

In this section we will state two meta-theorems and discuss how the two main theorems stated in the main text follow from these results. Our results require defining the notion of a covariance matrix associated to a neural network.

Definition 6.1 (Neural network covariance matrix and eigenvalue): Let $\boldsymbol{w} \in \mathbb{R}^d$ be a random vector with a $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ distribution. Also consider a set of n input data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ aggregated into the rows of a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$. Associated to a network $\boldsymbol{x} \mapsto \boldsymbol{v}^T\phi\left(\boldsymbol{W}\boldsymbol{x}\right)$ and the input data matrix $\boldsymbol{X}$ we define the neural net covariance matrix as

$$\boldsymbol{\Sigma}(\boldsymbol{X}) := \mathbb{E}\Big[\left(\phi'\left(\boldsymbol{X}\boldsymbol{w}\right)\phi'\left(\boldsymbol{X}\boldsymbol{w}\right)^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\Big].$$

We also define the eigenvalue $\lambda(\boldsymbol{X})$ based on $\boldsymbol{\Sigma}(\boldsymbol{X})$ as

$$\lambda(\boldsymbol{X}) := \lambda_{\min}\left(\boldsymbol{\Sigma}(\boldsymbol{X})\right).$$

We note that the neural network covariance matrix is intimately related to the expected value of the Jacobian mapping of the neural network at the random initialization. In particular when the output weights have unit absolute value (i.e. $|\boldsymbol{v}_\ell| = 1$), then

$$\boldsymbol{\Sigma}(\boldsymbol{X}) = \frac{1}{k}\mathbb{E}_{\boldsymbol{W}_0}\Big[\mathcal{J}(\boldsymbol{W}_0)\mathcal{J}^T(\boldsymbol{W}_0)\Big],$$

where $\boldsymbol{W}_0 \in \mathbb{R}^{k\times d}$ is a matrix with i.i.d. $\mathcal{N}(0,1)$ entires.

As mentioned earlier we prove a more general version of Theorem 2.1 which we now state. The proof is deferred to Section 6.2.

Theorem 6.2 (Meta-theorem for smooth activations): Consider a data set of input/label pairs $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$ aggregated as rows/entries of a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and a label vector $\boldsymbol{y} \in \mathbb{R}^n$. Without loss of generality we assume the dataset is normalized so that $\|\boldsymbol{x}_i\|_{\ell_2} = 1$. Also consider a one-hidden layer neural network with k hidden units and one output of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x})$ with $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ and $\boldsymbol{v} \in \mathbb{R}^k$ the input-to-hidden and hidden-to-output weights. We assume the activation ϕ has bounded derivatives i.e. $|\phi'(z)| \le B$ and $|\phi''(z)| \le B$ for all z. Let $\lambda(\boldsymbol{X})$ be the minimum neural net eigenvalue per Definition 6.1. Furthermore, we set half of the entries of $\boldsymbol{v}$ to $\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{kn}}$ and the other half to $-\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{kn}}$ and train only over $\boldsymbol{W}$. Starting from an initial weight matrix $\boldsymbol{W}_0$ selected at random with i.i.d. $\mathcal{N}(0,1)$ entries we run Gradient Descent (GD) updates of the form $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{W}_\tau)$ on the loss (II.1) with step size $\eta = \frac{n\bar{\eta}}{2B^2\|\boldsymbol{y}\|_{\ell_2}^2\|\boldsymbol{X}\|^2}$ where $\bar{\eta} \le 1$. Then, as long as

$$\sqrt{kd} \ge cB^2(1+\delta)\widetilde{\kappa}(\boldsymbol{X})n \quad \text{holds with} \quad \widetilde{\kappa}(\boldsymbol{X}) := \frac{\sqrt{\frac{d}{n}}\,\|\boldsymbol{X}\|}{\lambda(\boldsymbol{X})}, \tag{VI.2}$$

and c a fixed numerical constant, then with probability at least $1 - \frac{1}{n} - e^{-\delta^2\frac{n}{2\|\boldsymbol{X}\|^2}}$ all GD iterates obey

$$\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le \left(1 - \frac{\bar{\eta}}{32}\frac{1}{B^2}\frac{\lambda(\boldsymbol{X})}{\|\boldsymbol{X}\|^2}\right)^\tau \|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2},$$

$$\frac{\sqrt{\lambda(\boldsymbol{X})}}{\sqrt{32}}\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{n}}\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F + \|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le \|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2}.$$

Furthermore, the total gradient path obeys

$$\sum_{\tau=0}^{\infty}\|\boldsymbol{W}_{\tau+1} - \boldsymbol{W}_\tau\|_F \le \sqrt{32}\frac{\sqrt{n}}{\|\boldsymbol{y}\|_{\ell_2}}\frac{\|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2}}{\sqrt{\lambda(\boldsymbol{X})}}.$$

Next we state our meta-theorem for ReLU activations. The proof is deferred to Section 6.2.

Theorem 6.3 (Meta-theorem for ReLU activations): Consider the setting of Theorem 6.2 with the activations equal to $\phi(z) = ReLU(z) := \max(0, z)$ and the $\eta = \frac{n}{3\|\boldsymbol{y}\|_{\ell_2}^2 \|\boldsymbol{X}\|^2}\bar{\eta}$ with $\bar{\eta} \le 1$. Then, as long as

$$k \ge C(1+\delta)^2 \frac{n\|\boldsymbol{X}\|^6}{\lambda^4(\boldsymbol{X})}, \tag{VI.3}$$

holds with C a fixed numerical constant, then with probability at least $1 - \frac{2}{n} - e^{-\delta^2 \frac{n}{\|\boldsymbol{X}\|^2}}$ all GD iterates obey

$$\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le \left(1 - \frac{\bar{\eta}}{24}\frac{\lambda(\boldsymbol{X})}{\|\boldsymbol{X}\|^2}\right)^\tau \|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2},$$

$$\frac{1}{6\sqrt{2}}\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{n}}\sqrt{\lambda(\boldsymbol{X})}\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F + \|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le \|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2}.$$

Our main theorems in Section II can be obtained by substituting the appropriate value of $\lambda(\boldsymbol{X})$ into the two meta theorems above.

C. Reduction to quadratic activations and proofs for Theorems 2.1 and 2.3

Theorems 2.1 and 2.3 are corollaries of the meta-Theorems 6.2 and 6.3. To see this connection we will focus on lower bounding the the quantity $\lambda(\boldsymbol{X})$ which is not very interpretable and also not easily computable based on data. In the next lemma we provide a lower bound on $\lambda(\boldsymbol{X})$ based on the minimum eigenvalue of the Khatri-Rao product of $\boldsymbol{X}$ with itself. This key lemma relates the neural network covariance (from Definition 6.1) for any activation ϕ to the case of where the activation is a quadratic of the form $\phi(z) = \frac{1}{2}z^2$. We defer the proof of this lemma to Appendix B. We also note that this lemma is a special case of a more general result containing higher order interactions between the data points. Please see Appendix H for more details.

Lemma 6.4 (Reduction to quadratic activations): For an activation $\phi : \mathbb{R} \mapsto \mathbb{R}$ define the quantities

$$\widetilde{\mu}_\phi = \mathbb{E}_{g\sim\mathcal{N}(0,1)}[\phi'(g)] \quad \text{and} \quad \mu_\phi = \mathbb{E}_{g\sim\mathcal{N}(0,1)}[g\phi'(g)].$$

Then, the neural network covariance matrix and eigenvalue obey

$$\boldsymbol{\Sigma}(\boldsymbol{X}) \ge \left(\widetilde{\mu}_\phi^2 \mathbf{1}\mathbf{1}^T + \mu_\phi^2 \boldsymbol{X}\boldsymbol{X}^T\right) \odot \left(\boldsymbol{X}\boldsymbol{X}^T\right) \ge \mu_\phi^2 \left(\boldsymbol{X}\boldsymbol{X}^T\right) \odot \left(\boldsymbol{X}\boldsymbol{X}^T\right), \tag{VI.4}$$

$$\lambda(\boldsymbol{X}) \ge \mu_\phi^2 \sigma_{\min}^2(\boldsymbol{X} * \boldsymbol{X}). \tag{VI.5}$$

To see the relationship with the quadratic activation note that for this activation

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{X}) &:= \mathbb{E}\left[\left(\phi'(\boldsymbol{X}\boldsymbol{w})\,\phi'(\boldsymbol{X}\boldsymbol{w})^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\right]\\ &= \mathbb{E}\left[\left(\boldsymbol{X}\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{W}^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\right]\\ &= \left(\mathbb{E}[\boldsymbol{X}\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{W}^T]\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\\ &= (\boldsymbol{X}\boldsymbol{X}^T)\odot(\boldsymbol{X}\boldsymbol{X}^T)\\ &= (\boldsymbol{X}*\boldsymbol{X})(\boldsymbol{X}*\boldsymbol{X})^T.\end{aligned}$$

Thus the right-hand side of (VI.4) is μ_ϕ^2 multiplied by the covariance matrix of a neural network with a quadratic activation $\phi(z) = \frac{1}{2}z^2$.

With this lemma in place we can now prove Theorem 2.1 as simple corollaries of Theorem 6.2 by noting that $\lambda(\boldsymbol{X}) \geq \mu_\phi^2\sigma_{\min}^2(\boldsymbol{X}*\boldsymbol{X})$ per (VI.5) from Lemma 6.4. Similarly, to prove Theorem 2.3 from Theorem 6.3 we again use the fact that $\lambda(\boldsymbol{X}) \geq \mu_\phi^2\sigma_{\min}^2(\boldsymbol{X}*\boldsymbol{X})$ where for the ReLU activation $\mu_\phi^2 = \frac{1}{2\pi}$.

D. Lower and upper bounds on the eigenvalues of the Jacobian

In this section we will state a few key lemmas that provide lower and upper bounds on the eigenvalues of Jacobian matrices. The results in this section apply to any one-hidden neural network with activations that have bounded generalized derivative. In particular, our results here do not require the activation to be differentiable or smooth and thus apply to both the softplus ($\phi(z) = \log(e^z + 1)$) and ReLU ($\phi(z) = \max(0, z)$) activations.

We begin this section by stating a key lemma regarding the spectrum of the Hadamard product of matrices due to Schur [50] which plays a crucial role in both the upper and lower bounds on the eigenvalues of the Jacobian discussed in this section as well as our results on the perturbation of eigenvalues of the Jacobian discussed in the next section.

Lemma 6.5 ([50]): Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ be two Positive Semi-Definite (PSD) matrices. Then,

$$\begin{aligned}\lambda_{\min}(\boldsymbol{A}\odot\boldsymbol{B}) &\geq \left(\min_i \boldsymbol{B}_{ii}\right)\lambda_{\min}(\boldsymbol{A}),\\ \lambda_{\max}(\boldsymbol{A}\odot\boldsymbol{B}) &\leq \left(\max_i \boldsymbol{B}_{ii}\right)\lambda_{\max}(\boldsymbol{A}).\end{aligned}$$

The next lemma focuses on upper bounding the spectral norm of the Jacobian. The proof is deferred to Appendix A1.

Lemma 6.6 (Spectral norm of the Jacobian): Consider a one-hidden layer neural network model of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x})$ where the activation ϕ has bounded derivatives obeying $|\phi'(z)| \le B$. Also assume we have n data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ aggregated as the rows of a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$. Then the Jacobian matrix with respect to the input-to-hidden weights obeys

$$\|\mathcal{J}(\boldsymbol{W})\| \le \sqrt{k} B \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{X}\|.$$

Next we focus on lower bounding the minimum eigenvalue of the Jacobian matrix at initialization. The proof is deferred to Appendix A2.

Lemma 6.7 (Minimum eigenvalue of the Jacobian at initialization): Consider a one-hidden layer neural network model of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x})$ where the activation ϕ has bounded derivatives obeying $|\phi'(z)| \le B$. Also assume we have n data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ with unit euclidean norm ($\|\boldsymbol{x}_i\|_{\ell_2} = 1$) aggregated as the rows of a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$. Then, as long as

$$\frac{\|\boldsymbol{v}\|_{\ell_2}}{\|\boldsymbol{v}\|_{\ell_\infty}} \ge \sqrt{20 \log n} \frac{\|\boldsymbol{X}\|}{\sqrt{\lambda(\boldsymbol{X})}} B,$$

the Jacobian matrix at a random point $\boldsymbol{W}_0 \in \mathbb{R}^{k\times d}$ with i.i.d. $\mathcal{N}(0,1)$ entries obeys

$$\sigma_{\min}(\mathcal{J}(\boldsymbol{W}_0)) \ge \frac{1}{\sqrt{2}} \|\boldsymbol{v}\|_{\ell_2} \sqrt{\lambda(\boldsymbol{X})},$$

with probability at least $1 - 1/n$.

E. Jacobian perturbation

In this section we discuss results regarding the perturbation of the Jacobian matrix.

Our first result focuses on smooth activations. In particular, we show the Lipschitz property of the Jacobian with smooth activations. The proof is deferred to Appendix C1.

Lemma 6.8 (Jacobian Lipschitzness): Consider a one-hidden layer neural network model of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x})$ where the activation ϕ has bounded second order derivatives obeying $|\phi''(z)| \le M$. Also assume we have n data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ with unit euclidean norm ($\|\boldsymbol{x}_i\|_{\ell_2} = 1$) aggregated as the rows of a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$. Then the Jacobian mapping with respect to the input-to-hidden weights obeys

$$\left\|\mathcal{J}(\widetilde{\boldsymbol{W}}) - \mathcal{J}(\boldsymbol{W})\right\| \le M \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{X}\| \left\|\widetilde{\boldsymbol{W}} - \boldsymbol{W}\right\|_F \quad \text{for all} \quad \widetilde{\boldsymbol{W}}, \boldsymbol{W} \in \mathbb{R}^{k\times d}.$$

Our second result focuses on perturbation of the Jacobian from the random initialization with ReLU activations. This requires an intricate perturbation bound stated below and proven in Appendix C2.

Lemma 6.9 (Jacobian perturbation): Consider a one-hidden layer neural network model of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x})$ wwith the activation $\phi(z) = ReLU(z) := \max(0, z)$. Also assume we have n data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ with unit euclidean norm ($\|\boldsymbol{x}_i\|_{\ell_2} = 1$) aggregated as the rows of a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$. Also let $\boldsymbol{W}_0 \in \mathbb{R}^{k \times d}$ be a matrix with i.i.d. $\mathcal{N}(0, 1)$ entries and set $m_0 = \frac{\|\boldsymbol{v}\|_{\ell_2}}{\sqrt{200}\|\boldsymbol{v}\|_{\ell_\infty}} \frac{\sqrt{\lambda(\boldsymbol{X})}}{\|\boldsymbol{X}\|}$. Then, for all $\boldsymbol{W}$ obeying

$$\|\boldsymbol{W} - \boldsymbol{W}_0\| \leq \frac{m_0^3}{2k},$$

with probability at least $1 - ne^{-\frac{m_0^2}{6n}}$ the Jacobian matrix $\mathcal{J}$ associated with the neural network obeys

$$\|\mathcal{J}(\boldsymbol{W}) - \mathcal{J}(\boldsymbol{W}_0)\| \leq \frac{1}{6\sqrt{2}} \|\boldsymbol{v}\|_{\ell_2} \sqrt{\lambda(\boldsymbol{X})}. \tag{VI.6}$$

F. Proofs for meta-theorem with smooth activations (Proof of Theorem 6.2)

To prove this theorem we will utilize a result from [10] stated below.

Theorem 6.10: Consider a nonlinear least-squares optimization problem of the form

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2} \|f(\boldsymbol{\theta}) - \boldsymbol{y}\|_{\ell_2}^2,$$

with $f : \mathbb{R}^p \mapsto \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$. Suppose the Jacobian mapping associated with f obeys

$$\alpha \leq \sigma_{\min}(\mathcal{J}(\boldsymbol{\theta})) \leq \|\mathcal{J}(\boldsymbol{\theta})\| \leq \beta \tag{VI.7}$$

over a ball $\mathcal{D}$ of radius $R := \frac{4\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}}{\alpha}$ around a point $\boldsymbol{\theta}_0 \in \mathbb{R}^p$.[6] Furthermore, suppose

$$\|\mathcal{J}(\boldsymbol{\theta}_2) - \mathcal{J}(\boldsymbol{\theta}_1)\| \leq L \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_{\ell_2}, \tag{VI.8}$$

[6]That is, $\mathcal{D} = \mathcal{B}\left(\boldsymbol{\theta}_0, \frac{4\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}}{\alpha}\right)$ with $\mathcal{B}(\boldsymbol{c}, r) = \left\{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta} - \boldsymbol{c}\|_{\ell_2} \leq r\right\}$

holds for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{D}$ and set $\eta \leq \frac{1}{2\beta^2} \cdot \min\left(1, \frac{\alpha^2}{L\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}\right)$. Then, running gradient descent updates of the form $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)$ starting from $\boldsymbol{\theta}_0$, all iterates obey

$$\|f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\|_{\ell_2}^2 \leq \left(1 - \frac{\eta\alpha^2}{2}\right)^\tau \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}^2, \tag{VI.9}$$

$$\frac{1}{4}\alpha\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} + \|f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\|_{\ell_2} \leq \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}. \tag{VI.10}$$

Furthermore, the total gradient path is bounded. That is,

$$\sum_{\tau=0}^{\infty} \|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|_{\ell_2} \leq \frac{4\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}}{\alpha}. \tag{VI.11}$$

It is more convenient to work with a simpler variation of this theorem that only requires assumption (VI.7) to hold at the initialization point. We state this corollary below and defer its proof to Appendix D.

Corollary 6.11: Consider the setting and assumptions of Theorem 6.10 where

$$\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)) \geq 2\alpha, \tag{VI.12}$$

holds only at the initialization point $\boldsymbol{\theta}_0$ in lieu of the left-hand side of (VI.7). Furthermore, assume

$$\frac{\alpha^2}{4L} \geq \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}, \tag{VI.13}$$

holds. Then, the conclusions of Theorem 6.10 continue to hold.

To be able to use this corollary it thus suffices to prove the conditions (VI.8), $\|\mathcal{J}(\boldsymbol{\theta})\| \leq \beta$, (VI.12), and (VI.13) hold for proper choices of α, β, and L. First, by Lemma 6.8 and our choice of $\boldsymbol{v}$ we can use

$$L = B\|\boldsymbol{v}\|_{\ell_\infty}\|\boldsymbol{X}\| = \frac{B}{\sqrt{kn}}\|\boldsymbol{y}\|_{\ell_2}\|\boldsymbol{X}\|. \tag{VI.14}$$

Second, by Lemma 6.6 and our choice of $\boldsymbol{v}$ we can use

$$\beta = \sqrt{k}B\|\boldsymbol{v}\|_{\ell_\infty}\|\boldsymbol{X}\| = \frac{B}{\sqrt{n}}\|\boldsymbol{y}\|_{\ell_2}\|\boldsymbol{X}\|. \tag{VI.15}$$

Next note that

$$\lambda(\boldsymbol{X}) = \lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{X})) \leq \boldsymbol{e}_1^T\boldsymbol{\Sigma}(\boldsymbol{X})\boldsymbol{e}_1 = \mathbb{E}_{g\sim\mathcal{N}(0,1)}[(\phi'(g))^2] \leq B^2 \quad \Rightarrow \quad \sqrt{\lambda(\boldsymbol{X})} \leq B. \tag{VI.16}$$

Thus, as long as (VI.2) holds then

$$\begin{aligned}\sqrt{k} \geq & c\sqrt{n}B^2\frac{\|\boldsymbol{X}\|}{\lambda(\boldsymbol{X})}\\ \overset{(a)}{\geq} & \sqrt{20\log n}B^2\frac{\|\boldsymbol{X}\|}{\lambda(\boldsymbol{X})}\\ \overset{(b)}{\geq} & \sqrt{20\log n}\frac{\|\boldsymbol{X}\|}{\sqrt{\lambda(\boldsymbol{X})}}B.\end{aligned}$$

Here, (a) follows from the fact that $n \geq \log n$ for $n \geq 1$ and (b) from (VI.16). Thus by our choice of $\boldsymbol{v}$ we have

$$\frac{\|\boldsymbol{v}\|_{\ell_2}}{\|\boldsymbol{v}\|_{\ell_\infty}} = \sqrt{k} \geq \sqrt{20\log n}B\frac{\|\boldsymbol{X}\|}{\sqrt{\lambda(\boldsymbol{X})}},$$

so that Lemma 6.7 applies and we can use

$$\alpha = \frac{1}{2\sqrt{2}}\|\boldsymbol{v}\|_{\ell_2}\sqrt{\lambda(\boldsymbol{X})} = \frac{1}{2\sqrt{2}}\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{n}}\sqrt{\lambda(\boldsymbol{X})}. \tag{VI.17}$$

All that remains is to prove the theorem using Corollary (6.11) is to check that (VI.13) holds. To this aim we upper bound the initial misfit in the next lemma. The proof is deferred to Section VI-F1.

Lemma 6.12 (Upper bound on initial misfit): Consider a one-hidden layer neural network model of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x})$ where the activation ϕ has bounded derivatives obeying $|\phi'(z)| \leq B$. Also assume we have n data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ with unit euclidean norm ($\|\boldsymbol{x}_i\|_{\ell_2} = 1$) aggregated as rows of a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and the corresponding labels given by $\boldsymbol{y} \in \mathbb{R}^n$. Furthermore, assume we set half of the entries of $\boldsymbol{v} \in \mathbb{R}^k$ to $\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{kn}}$ and the other half to $-\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{kn}}$. Then for $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ with i.i.d. $\mathcal{N}(0,1)$ entries

$$\left\|\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v} - \boldsymbol{y}\right\|_{\ell_2} \leq \|\boldsymbol{y}\|_{\ell_2}\left(1 + (1+\delta)B\right),$$

holds with probability at least $1 - e^{-\delta^2\frac{n}{2\|\boldsymbol{X}\|^2}}$.

To do this we will use Lemma 6.12 to conclude that

$$\begin{aligned}\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2} &:= \left\|\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v} - \boldsymbol{y}\right\|_{\ell_2}\\ &\leq \|\boldsymbol{y}\|_{\ell_2}\left(1 + (1+\delta)B\right)\end{aligned} \tag{VI.18}$$

holds with probability at least $1-e^{-\delta^2\frac{n}{2\|\boldsymbol{X}\|^2}}$. Thus, as long as

$$\begin{aligned}\sqrt{kd} \geq & 32nB\left(1+(1+\delta)B\right)\frac{\sqrt{\frac{d}{n}}\|\boldsymbol{X}\|}{\lambda(\boldsymbol{X})}\\ :=& 32B\left(1+(1+\delta)B\right)\widetilde{\kappa}(\boldsymbol{X})n\end{aligned} \tag{VI.19}$$

then

$$\begin{aligned}\frac{\alpha^2}{4L} =& \frac{\frac{1}{8}\frac{\|\boldsymbol{y}\|_{\ell_2}^2}{n}\lambda(\boldsymbol{X})}{4\frac{B}{\sqrt{kn}}\|\boldsymbol{y}\|_{\ell_2}\|\boldsymbol{X}\|}\\ =& \frac{1}{32B}\frac{\sqrt{k}}{\sqrt{n}}\|\boldsymbol{y}\|_{\ell_2}\frac{\lambda(\boldsymbol{X})}{\|\boldsymbol{X}\|}\\ =& \frac{1}{32B}\frac{\sqrt{kd}}{\widetilde{\kappa}(\boldsymbol{X})n}\|\boldsymbol{y}\|_{\ell_2}\\ \geq & \|\boldsymbol{y}\|_{\ell_2}\left(1+(1+\delta)B\right).\end{aligned}$$

Thus, as long as (VI.2) (equivalent to (VI.19)) holds, then also (VI.13) holds and hence $\frac{\alpha^2}{L\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}} \geq 4$. Therefore, using a step size

$$\eta \leq \frac{1}{2kB^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2} = \frac{1}{2\beta^2} = \frac{1}{2\beta^2}\cdot\min(1,4) \leq \frac{1}{2\beta^2}\cdot\min\left(1,\frac{\alpha^2}{L\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}\right),$$

all the assumptions of Corollary 6.11 hold and so do its conclusions, completing the proof of Theorem 6.2.

1) Upper bounding the initial misfit (Proof of Lemma 6.12)

To begin first note that for any two matrices $\widetilde{\boldsymbol{W}},\boldsymbol{W}\in\mathbb{R}^{k\times d}$ we have

$$\begin{aligned}\left|\left\|\phi\left(\boldsymbol{X}\widetilde{\boldsymbol{W}}^T\right)\boldsymbol{v}\right\|_{\ell_2}-\left\|\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v}\right\|_{\ell_2}\right| \leq & \left\|\phi\left(\boldsymbol{X}\widetilde{\boldsymbol{W}}^T\right)\boldsymbol{v}-\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v}\right\|_{\ell_2}\\ \leq & \left\|\phi\left(\boldsymbol{X}\widetilde{\boldsymbol{W}}^T\right)-\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\right\|\|\boldsymbol{v}\|_{\ell_2}\\ \leq & \left\|\phi\left(\boldsymbol{X}\widetilde{\boldsymbol{W}}^T\right)-\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\right\|_F\|\boldsymbol{v}\|_{\ell_2}\\ \overset{(a)}{=} & \left\|\left(\phi'\left(\boldsymbol{S}\odot\boldsymbol{X}\widetilde{\boldsymbol{W}}^T+(1_{k\times n}-\boldsymbol{S})\odot\boldsymbol{X}\boldsymbol{W}^T\right)\right)\odot\left(\boldsymbol{X}(\widetilde{\boldsymbol{W}}-\boldsymbol{W})^T\right)\right\|_F\|\boldsymbol{v}\|_{\ell_2}\\ \leq & B\left\|\boldsymbol{X}(\widetilde{\boldsymbol{W}}-\boldsymbol{W})^T\right\|_F\|\boldsymbol{v}\|_{\ell_2}\\ \leq & B\|\boldsymbol{X}\|\|\boldsymbol{v}\|_{\ell_2}\left\|\widetilde{\boldsymbol{W}}-\boldsymbol{W}\right\|_F,\end{aligned}$$

where in (a) we used the mean value theorem with $\boldsymbol{S}$ a matrix with entries obeying $0 \le \boldsymbol{S}_{i,j} \le 1$ and $1_{n\times n}$ the matrix of all ones. Thus, $\|\phi(\boldsymbol{X}\boldsymbol{W}^T)\boldsymbol{v}\|_{\ell_2}$ is a $B\|\boldsymbol{X}\|\|\boldsymbol{v}\|_{\ell_2}$-Lipschitz function of $\boldsymbol{W}$. Thus for a matrix $\boldsymbol{W}$ with i.i.d. Gaussian entries

$$\left\|\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v}\right\|_{\ell_2} \le \mathbb{E}\left[\left\|\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v}\right\|_{\ell_2}\right] + t, \tag{VI.20}$$

holds with probability at least $1 - e^{-\frac{t^2}{2B^2\|\boldsymbol{v}\|_{\ell_2}^2\|\boldsymbol{X}\|^2}}$. We now upper bound the expectation via

$$\begin{aligned}
\mathbb{E}\left[\left\|\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v}\right\|_{\ell_2}\right] &\overset{(a)}{\le} \sqrt{\mathbb{E}\left[\|\phi(\boldsymbol{X}\boldsymbol{W}^T)\boldsymbol{v}\|_{\ell_2}^2\right]} \\
&= \sqrt{\sum_{i=1}^{n}\mathbb{E}\left[\left(\boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x}_i)\right)^2\right]} \\
&\overset{(b)}{=} \sqrt{n}\sqrt{\mathbb{E}_{\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_k)}\left[\left(\boldsymbol{v}^T\phi(\boldsymbol{g})\right)^2\right]} \\
&\overset{(c)}{=} \sqrt{n}\sqrt{\|\boldsymbol{v}\|_{\ell_2}^2\,\mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[\left(\phi(g)-\mathbb{E}[\phi(g)]\right)^2\right] + (\boldsymbol{1}^T\boldsymbol{v})^2(\mathbb{E}_{g\sim\mathcal{N}(0,1)}[\phi(g)])^2} \\
&\overset{(d)}{=} \sqrt{n}\,\|\boldsymbol{v}\|_{\ell_2}\sqrt{\mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[\left(\phi(g)-\mathbb{E}[\phi(g)]\right)^2\right]} \\
&\overset{(e)}{\le} \sqrt{n}B\,\|\boldsymbol{v}\|_{\ell_2}.
\end{aligned}$$

Here, (a) follows from Jensen's inequality, (b) from linearity of expectation and the fact that for $\boldsymbol{x}_i$ with unit Euclidean norm $\boldsymbol{W}\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$, (c) from simple algebraic manipulations, (d) from the fact that $\boldsymbol{1}^T\boldsymbol{v} = 0$, (e) from $|\phi'(z)| \le B$ a long with the fact that for a B-Lipschitz function ϕ and normal random variable we have $\mathrm{Var}(\phi(g)) \le B^2$ based on the Poincare inequality (e.g. see [51, p. 49]). Thus using $t = \delta B\sqrt{n}\,\|\boldsymbol{v}\|_{\ell_2}$ in (VI.20) we conclude that

$$\begin{aligned}
\left\|\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v}\right\|_{\ell_2} &\le \|\boldsymbol{v}\|_{\ell_2}\sqrt{n}\,(1+\delta)\,B, \\
&= \|\boldsymbol{y}\|_{\ell_2}\,(1+\delta)\,B,
\end{aligned}$$

holds with probability at least $1 - e^{-\delta^2\frac{n}{2\|\boldsymbol{X}\|^2}}$. Thus,

$$\left\|\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v} - \boldsymbol{y}\right\|_{\ell_2} \le \left\|\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v}\right\|_{\ell_2} + \|\boldsymbol{y}\|_{\ell_2} \le \|\boldsymbol{y}\|_{\ell_2}\left(1 + (1+\delta)B\right),$$

holds with probability at least $1 - e^{-\delta^2\frac{n}{2\|\boldsymbol{X}\|^2}}$ concluding the proof.

G. Proofs for meta-theorem with ReLU activations (Proof of Theorem 6.3)

To prove Theorem 6.3 we start by stating a general overparameterized fitting of non-smooth functions. This can be thought of a counter part to Theorem 6.10 for non-smooth mappings. We note that we do not require the mapping f to be differentiable rather here the Jacobian is defined based on a generalized derivative. Consider a nonlinear least-squares optimization problem of the form

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2}\left\|f(\boldsymbol{\theta})-\boldsymbol{y}\right\|_{\ell_2}^2,$$

with $f : \mathbb{R}^p \mapsto \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$. Suppose the Jacobian mapping associated with f obeys the following three assumptions.

Assumption 2: We assume $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)) \geq 2\alpha$ for a point $\boldsymbol{\theta}_0 \in \mathbb{R}^p$.

Assumption 3: We assume that for all $\boldsymbol{\theta} \in \mathbb{R}^d$ we have $\|\mathcal{J}(\boldsymbol{\theta})\| \leq \beta$.

Assumption 4: Let $\|\cdot\|$ denote a norm that is dominated by the Euclidean norm i.e. $\|\boldsymbol{\theta}\| \leq \|\boldsymbol{\theta}\|_{\ell_2}$ holds for all $\boldsymbol{\theta} \in \mathbb{R}^p$. Fix a point $\boldsymbol{\theta}_0$ and a number $R > 0$. For any $\boldsymbol{\theta}$ satisfying $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq R$, we have that $\|\mathcal{J}(\boldsymbol{\theta}_0)-\mathcal{J}(\boldsymbol{\theta})\| \leq \alpha/3$.

Under these assumptions we can state the following theorem. We defer the proof of this Theorem to Appendix G.

Theorem 6.13 (Non-smooth Overparameterized Optimization): Given $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, suppose Assumptions 2, 3, and 4 hold with

$$R = \frac{3\|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}}{\alpha}.$$

Then, using a learning rate $\eta \leq \frac{1}{3\beta^2}$, all gradient iterations obey

$$\|\boldsymbol{y}-f(\boldsymbol{\theta}_\tau)\|_{\ell_2} \leq \left(1-\eta\alpha^2\right)^\tau \|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}, \tag{VI.21}$$

$$\frac{\alpha}{3}\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\| + \|\boldsymbol{y}-f(\boldsymbol{\theta}_\tau)\|_{\ell_2} \leq \|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}. \tag{VI.22}$$

We shall apply this theorem to the case where the parameter is $\boldsymbol{W}$, the nonlinear mapping is given by $f(\boldsymbol{W}) = \boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{X}^T)$ with $\phi = ReLU$, and the norm $\|\cdot\|$ is the spectral norm of a matrix.

Completing the proof of Theorem 6.3. With this result in place we are now ready to complete the proof of Theorem 6.3. As in the smooth case (VI.3) guarantees the condition of Lemma 6.7

(i.e. $k \geq 20 \log n \frac{\|\boldsymbol{X}\|^2}{\lambda(\boldsymbol{X})}$) holds. Thus, using Lemma 6.7 with probability at least $1-1/n$, Assumption 2 holds with

$$\alpha = \frac{1}{2\sqrt{2n}} \|\boldsymbol{y}\|_{\ell_2} \sqrt{\lambda(\boldsymbol{X})}.$$

Furthermore, Lemma 6.6 allows us to conclude that Assumption 3 holds with

$$\beta = \frac{1}{\sqrt{n}} \|\boldsymbol{y}\|_{\ell_2} \|\boldsymbol{X}\|.$$

To be able to apply Theorem 6.13, all that remains is to prove Assumption 4 holds. To this aim note that using Lemma 6.12 with $B = 1$ and $\delta \leftarrow 2\delta$, to conclude that the initial misfit obeys

$$\|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2} \leq 2(1+\delta)\|\boldsymbol{y}\|_{\ell_2},$$

with probability at least $1 - e^{-\delta^2 \frac{n}{\|\boldsymbol{X}\|^2}}$. Therefore, with high probability

$$R := \frac{3\|\boldsymbol{y} - f(\boldsymbol{W}_0)\|_{\ell_2}}{\alpha} \leq 12(1+\delta)\sqrt{2n} \frac{1}{\sqrt{\lambda(\boldsymbol{X})}}.$$

Thus, when (VI.3) holds using the perturbation Lemma 6.9 with $m_0 = \sqrt{\frac{k}{200}} \frac{\sqrt{\lambda(\boldsymbol{X})}}{\|\boldsymbol{X}\|}$, with probability at least $1 - ne^{-\frac{k}{1200}\frac{\lambda(\boldsymbol{X})}{\|\boldsymbol{X}\|^2}} - e^{-\delta^2 \frac{n}{\|\boldsymbol{X}\|^2}}$, for all $\boldsymbol{W}$ obeying

$$\begin{aligned}
\|\boldsymbol{W} - \boldsymbol{W}_0\| \leq & R \\
\leq & 12(1+\delta)\sqrt{2n} \frac{1}{\sqrt{\lambda(\boldsymbol{X})}} \\
\overset{\text{(VI.3)}}{\leq} & \frac{\sqrt{k}\lambda^{\frac{3}{2}}(\boldsymbol{X})}{2(200)^{\frac{3}{2}}\|\boldsymbol{X}\|^3} \\
= & \frac{m_0^3}{2k}
\end{aligned}$$

we have

$$\|\mathcal{J}(\boldsymbol{W}) - \mathcal{J}(\boldsymbol{W}_0)\| \leq \frac{1}{6\sqrt{2n}} \|\boldsymbol{y}\|_{\ell_2} \sqrt{\lambda(\boldsymbol{X})} = \frac{\alpha}{3}.$$

This guarantees Assumption 4 also holds concluding the proof of Theorem 6.3 via Theorem 6.13.

ACKNOWLEDGEMENTS

M. Soltanolkotabi is supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship in Mathematics, an NSF-CAREER under award #1846369, the Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) under award #FA9550-18-1-0078, DARPA Learning with Less Labels (LwLL) and Fast Network Interface Cards (FastNICs) programs, an NSF-CIF award #1813877, and a Google faculty research award. S. Oymak is supported by the NSF award CNS-1932254.

REFERENCES

[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

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

[3] L. Venturi, A. Bandeira, and J. Bruna. Spurious valleys in two-layer neural network optimization landscapes. *arXiv preprint arXiv:1802.06384*, 2018.

[4] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *NeurIPS*, 2018.

[5] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

[6] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

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

[8] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

[9] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

[10] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? *arXiv preprint arXiv:1812.10004*, 2018.

[11] Frank H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–247, 1975.

[12] Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2053–2062, 2019.

[13] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. 10 2018.

[14] Pan Zhou and Jiashi Feng. The landscape of deep learning algorithms. *arXiv preprint arXiv:1705.07038*, 2017.

[15] Mahdi Soltanolkotabi. Learning ReLUs via gradient descent. *arXiv preprint arXiv:1705.04591*, 2017.

[16] Chi Jin, Lydia T Liu, Rong Ge, and Michael I Jordan. On the local minima of the empirical risk. In *Advances in Neural Information Processing Systems*, pages 4901–4910, 2018.

[17] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*, 2016.

[18] A. Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with Gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.

[19] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

[20] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.

[21] Samet Oymak. Stochastic gradient descent learns state equations with nonlinear activations. *arXiv preprint arXiv:1809.03019*, 2018.

[22] Haoyu Fu, Yuejie Chi, and Yingbin Liang. Local geometry of one-hidden-layer neural networks for logistic regression. *arXiv preprint arXiv:1802.06463*, 2018.

[23] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. A critical view of global optimality in deep learning. *arXiv preprint arXiv:1802.03487*, 2018.

[24] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *arXiv preprint arXiv:1712.08968*, 2017.

[25] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

[26] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[27] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

[28] Mei Song, A Montanari, and P Nguyen. A mean field view of the landscape of two-layers neural networks. In *Proceedings of the National Academy of Sciences*, volume 115, pages E7665–E7671, 2018.

[29] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[30] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*, 2016.

[31] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

[32] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. *arXiv preprint arXiv:1802.06509*, 2018.

[33] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. 10 2018.

[34] Zhihui Zhu, Daniel Soudry, Yonina C. Eldar, and Michael B. Wakin. The global optimization geometry of shallow linear neural networks. 05 2018.

[35] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. 05 2016.

[36] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *arXiv preprint arXiv:1802.01396*, 2018.

[37] Lenaic Chizat and Francis Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

[38] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? 06 2018.

[39] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. 08 2018.

[40] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layers neural networks. *arXiv preprint arXiv:1804.06561*, 2018.

[41] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. 08 2018.

[42] Grant M. Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. 05 2018.

[43] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

[44] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

[45] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. 02 2018.

[46] Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. 06 2017.

[47] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. 12 2017.

[48] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. 10 2017.

[49] Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. 06 2018.

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

[51] M. Ledoux. The concentration of measure phenomenon. *volume 89 of Mathematical Surveys and Monographs. American Matheamtical Society, Providence, RI*, 2001.

[52] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. 01 2019.

[53] Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv preprint arXiv:1906.05392*, 2019.

APPENDIX

A. *Proofs for bounding the eigenvalues of the Jacobian*

1) Proof for the spectral norm of the Jacobian (Proof of Lemma 6.6)

To bound the spectral norm note that as stated earlier

$$\mathcal{J}(\boldsymbol{W})\mathcal{J}^T(\boldsymbol{W}) = \left(\phi'\left(\boldsymbol{X}\boldsymbol{W}^T\right)\operatorname{diag}(\boldsymbol{v})\operatorname{diag}(\boldsymbol{v})\,\phi'\left(\boldsymbol{W}\boldsymbol{X}^T\right)\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right).$$

Thus using Lemma 6.5 we have

$$\begin{aligned}\|\mathcal{J}(\boldsymbol{W})\|^2 \leq& \left(\max_i \ \|\operatorname{diag}(\boldsymbol{v})\,\phi'(\boldsymbol{W}\boldsymbol{x}_i)\|_{\ell_2}^2\right)\lambda_{\max}\left(\boldsymbol{X}\boldsymbol{X}^T\right)\\ =& \left(\max_i \ \|\operatorname{diag}(\boldsymbol{v})\,\phi'(\boldsymbol{W}\boldsymbol{x}_i)\|_{\ell_2}^2\right)\|\boldsymbol{X}\|^2\\ \leq& \|\boldsymbol{v}\|_{\ell_\infty}^2\left(\max_i \ \|\phi'(\boldsymbol{W}\boldsymbol{x}_i)\|_{\ell_2}^2\right)\|\boldsymbol{X}\|^2\\ \leq& kB^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2,\end{aligned}$$

completing the proof.

2) Proofs for minimum eigenvalue of the Jacobian at initialization (Proof of Lemma 6.7)

To lower bound the minimum eigenvalue of $\mathcal{J}(\boldsymbol{W}_0)$, we focus on lower bounding the minimum eigenvalue of $\mathcal{J}(\boldsymbol{W}_0)\mathcal{J}(\boldsymbol{W}_0)^T$. To do this we first lower bound the minimum eigenvalue of the expected value $\mathbb{E}\left[\mathcal{J}(\boldsymbol{W}_0)\mathcal{J}(\boldsymbol{W}_0)^T\right]$ and then related the matrix $\mathcal{J}(\boldsymbol{W}_0)\mathcal{J}(\boldsymbol{W}_0)^T$ to its expected value. We proceed by simplifying the expected value. To this aim we use the identity

$$\begin{aligned}\mathcal{J}(\boldsymbol{W})\mathcal{J}^T(\boldsymbol{W}) =& \left(\phi'\left(\boldsymbol{X}\boldsymbol{W}^T\right)\operatorname{diag}(\boldsymbol{v})\operatorname{diag}(\boldsymbol{v})\,\phi'\left(\boldsymbol{W}\boldsymbol{X}^T\right)\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\\ =& \left(\sum_{\ell=1}^{k}\boldsymbol{v}_\ell^2\phi'(\boldsymbol{X}\boldsymbol{w}_\ell)\,\phi'(\boldsymbol{X}\boldsymbol{w}_\ell)^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right),\end{aligned}$$

mentioned earlier to conclude that

$$\begin{aligned}\mathbb{E}\left[\mathcal{J}(\boldsymbol{W}_0)\mathcal{J}(\boldsymbol{W}_0)^T\right] =& \|\boldsymbol{v}\|_{\ell_2}^2\left(\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\phi'(\boldsymbol{X}\boldsymbol{w})\,\phi'(\boldsymbol{X}\boldsymbol{w})^T\right]\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right),\\ :=& \|\boldsymbol{v}\|_{\ell_2}^2\,\Sigma(\boldsymbol{X}).\end{aligned}\tag{A.1}$$

Thus

$$\lambda_{\min}\left(\mathbb{E}\left[\mathcal{J}(\boldsymbol{W}_0)\mathcal{J}(\boldsymbol{W}_0)^T\right]\right) \geq \|\boldsymbol{v}\|_{\ell_2}^2\,\lambda(\boldsymbol{X}).\tag{A.2}$$

To relate the minimum eigenvalue of the expectation to that of $\mathcal{J}(\boldsymbol{W}_0)\mathcal{J}(\boldsymbol{W}_0)^T$ we utilize the matrix Chernoff identity stated below.

Theorem A.1 (Matrix Chernoff): Consider a finite sequence $\boldsymbol{A}_\ell \in \mathbb{R}^{n\times n}$ of independent, random, Hermitian matrices with common dimension n. Assume that $\boldsymbol{0} \preceq \boldsymbol{A}_\ell \preceq R\boldsymbol{I}$ for $\ell = 1, 2, \ldots, k$. Then

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_{\ell=1}^{k}\boldsymbol{A}_\ell\right) \le (1-\delta)\lambda_{\min}\left(\sum_{\ell=1}^{k}\mathbb{E}[\boldsymbol{A}_\ell]\right)\right\} \le n\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\frac{\lambda_{\min}\left(\sum_{\ell=1}^{k}\mathbb{E}[\boldsymbol{A}_\ell]\right)}{R}}$$

for $\delta \in [0,1)$.

We shall apply this theorem with $\boldsymbol{A}_\ell := \mathcal{J}(\boldsymbol{w}_\ell)\mathcal{J}^T(\boldsymbol{w}_\ell) = \boldsymbol{v}_\ell^2 \mathrm{diag}(\phi'(\boldsymbol{X}\boldsymbol{w}_\ell))\boldsymbol{X}\boldsymbol{X}^T\mathrm{diag}(\phi'(\boldsymbol{X}\boldsymbol{w}_\ell))$. To this aim note that

$$\boldsymbol{v}_\ell^2 \mathrm{diag}(\phi'(\boldsymbol{X}\boldsymbol{w}_\ell))\boldsymbol{X}\boldsymbol{X}^T\mathrm{diag}(\phi'(\boldsymbol{X}\boldsymbol{w}_\ell)) \preceq B^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2\boldsymbol{I},$$

so that we can use Chernoff Matrix with $R = B^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2$ to conclude that

$$\mathbb{P}\left\{\lambda_{\min}\left(\mathcal{J}(\boldsymbol{W}_0)\mathcal{J}^T(\boldsymbol{W}_0)\right) \le (1-\delta)\lambda_{\min}\left(\mathbb{E}\left[\mathcal{J}(\boldsymbol{W}_0)\mathcal{J}^T(\boldsymbol{W}_0)\right]\right)\right\}$$
$$\le n\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\frac{\lambda_{\min}\left(\mathbb{E}\left[\mathcal{J}(\boldsymbol{W}_0)\mathcal{J}^T(\boldsymbol{W}_0)\right]\right)}{B^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2}}.$$

Thus using (A.2) in the above with $\delta = \frac{1}{2}$ we have

$$\mathbb{P}\left\{\lambda_{\min}\left(\mathcal{J}(\boldsymbol{W}_0)\mathcal{J}^T(\boldsymbol{W}_0)\right) \le \frac{1}{2}\|\boldsymbol{v}\|_{\ell_2}^2\lambda(\boldsymbol{X})\right\} \le n\cdot e^{-\frac{1}{10}\frac{\|\boldsymbol{v}\|_{\ell_2}^2\lambda(\boldsymbol{X})}{B^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2}}.$$

Therefore, as long as

$$\frac{\|\boldsymbol{v}\|_{\ell_2}}{\|\boldsymbol{v}\|_{\ell_\infty}} \ge \sqrt{20\log n}\frac{\|\boldsymbol{X}\|}{\sqrt{\lambda(\boldsymbol{X})}}B,$$

then

$$\sigma_{\min}\left(\mathcal{J}(\boldsymbol{W}_0)\right) \ge \frac{1}{\sqrt{2}}\|\boldsymbol{v}\|_{\ell_2}\sqrt{\lambda(\boldsymbol{X})},$$

holds with probability at leat $1-\frac{1}{n}$.

B. Reduction to quadratic activations (Proof of Lemma 6.4)

First we note that (VI.5) simply follows from (VI.4) by noting that

$$(\boldsymbol{X}\boldsymbol{X}^T) \odot (\boldsymbol{X}\boldsymbol{X}^T) = (\boldsymbol{X} * \boldsymbol{X})(\boldsymbol{X} * \boldsymbol{X})^T.$$

Thus we focus on proving (VI.4). We begin the proof by noting two simple identities. First, using multivariate Stein identity we have

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{X}\boldsymbol{w}\phi'(\boldsymbol{X}\boldsymbol{w})^T\right] &= \sum_{i=1}^{n} \mathbb{E}\left[\boldsymbol{X}\boldsymbol{w} \cdot \phi'\left(\boldsymbol{e}_i^T \boldsymbol{X}\boldsymbol{w}\right)\right]\boldsymbol{e}_i^T \\
&= \sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{X}^T \mathbb{E}\left[\phi''\left(\boldsymbol{e}_i^T \boldsymbol{X}\boldsymbol{w}\right)\boldsymbol{e}_i\right]\boldsymbol{e}_i^T \\
&= \boldsymbol{X}\boldsymbol{X}^T \mathrm{diag}\left(\mathbb{E}[\phi''(\boldsymbol{X}\boldsymbol{w})]\right) \\
&= \mathbb{E}_{g\sim\mathcal{N}(0,1)}[\phi''(g)]\boldsymbol{X}\boldsymbol{X}^T \\
&= \mathbb{E}_{g\sim\mathcal{N}(0,1)}[g\phi'(g)]\boldsymbol{X}\boldsymbol{X}^T \\
&= \mu_\phi \boldsymbol{X}\boldsymbol{X}^T,
\end{aligned} \tag{A.3}$$

where in the last line we used the fact that $\|\boldsymbol{x}_i\|_{\ell_2} = 1$. We note that while for clarity of exposition we carried out the above proof using the fact that ϕ' is differentiable the identity above continues to hold without assuming ϕ' is differentiable with a simple modification to the above proof. Next we note that

$$\mathbb{E}[\phi'(\boldsymbol{X}\boldsymbol{w})] = \mathbb{E}_{g\sim\mathcal{N}(0,1)}[\phi'(g)]\mathbf{1} := \widetilde{\mu}_\phi. \tag{A.4}$$

We continue by noting that

$$\mathbb{E}\left[(\phi'(\boldsymbol{X}\boldsymbol{w}) - \eta\mathbf{1} - \gamma\boldsymbol{X}\boldsymbol{w})(\phi'(\boldsymbol{X}\boldsymbol{w}) - \eta\mathbf{1} - \gamma\boldsymbol{X}\boldsymbol{w})^T\right] \succeq \mathbf{0}. \tag{A.5}$$

Thus, using (A.3) and (A.4) we have

$$\begin{aligned}
&\mathbb{E}\left[(\phi'(\boldsymbol{X}\boldsymbol{w}) - \eta\mathbf{1} - \gamma\boldsymbol{X}\boldsymbol{w})(\phi'(\boldsymbol{X}\boldsymbol{w}) - \eta\mathbf{1} - \gamma\boldsymbol{X}\boldsymbol{w})^T\right] \\
&= \mathbb{E}\left[\phi'(\boldsymbol{X}\boldsymbol{w})\phi'(\boldsymbol{X}\boldsymbol{w})^T\right] - 2\eta\widetilde{\mu}_\phi\mathbf{1}\mathbf{1}^T - 2\gamma\mu_\phi\boldsymbol{X}\boldsymbol{X}^T \\
&\quad + \eta^2\mathbf{1}\mathbf{1}^T + \gamma^2\boldsymbol{X}\boldsymbol{X}^T \\
&= \mathbb{E}\left[\phi'(\boldsymbol{X}\boldsymbol{w})\phi'(\boldsymbol{X}\boldsymbol{w})^T\right] + \eta(\eta - 2\widetilde{\mu}_\phi)\mathbf{1}\mathbf{1}^T \\
&\quad + \gamma(\gamma - 2\mu_\phi)\boldsymbol{X}\boldsymbol{X}^T.
\end{aligned}$$

Combining the latter with (A.5) we arrive at

$$\mathbb{E}\left[\phi'\left(\boldsymbol{X}\boldsymbol{w}\right)\phi'\left(\boldsymbol{X}\boldsymbol{w}\right)^T\right] \succeq \eta\left(2\widetilde{\mu}_\phi - \eta\right)\mathbf{1}\mathbf{1}^T + \gamma\left(2\mu_\phi - \gamma\right)\boldsymbol{X}\boldsymbol{X}^T.$$

Hence, setting $\eta = \widetilde{\mu}_\phi$ and $\gamma = \mu_\phi$ we conclude that

$$\mathbb{E}\left[\phi'\left(\boldsymbol{X}\boldsymbol{w}\right)\phi'\left(\boldsymbol{X}\boldsymbol{w}\right)^T\right] \succeq \widetilde{\mu}_\phi^2\mathbf{1}\mathbf{1}^T + \mu_\phi^2\boldsymbol{X}\boldsymbol{X}^T.$$

Thus

$$\begin{aligned}\boldsymbol{\Sigma}\left(\boldsymbol{X}\right) &= \left(\mathbb{E}\left[\phi'\left(\boldsymbol{X}\boldsymbol{w}\right)\phi'\left(\boldsymbol{X}\boldsymbol{w}\right)^T\right]\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\\ &\succeq \left(\widetilde{\mu}_\phi^2\mathbf{1}\mathbf{1}^T + \mu_\phi^2\boldsymbol{X}\boldsymbol{X}^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\\ &\succeq \mu_\phi^2\left(\boldsymbol{X}\boldsymbol{X}^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\end{aligned}$$

completing the proof of (VI.4) and the lemma.

C. Proofs for Jacobian perturbation

1) Proof for Lipschitzness of the Jacobian with smooth activations (Proof of Lemma 6.8)

To prove this lemma first note that using the form (VI.1) we have

$$\mathcal{J}\left(\widetilde{\boldsymbol{W}}\right) - \mathcal{J}\left(\boldsymbol{W}\right) = \left(\mathrm{diag}(\boldsymbol{v})\left(\phi'\left(\boldsymbol{X}\widetilde{\boldsymbol{W}}^T\right) - \phi'\left(\boldsymbol{X}\boldsymbol{W}^T\right)\right)\right) * \boldsymbol{X}.$$

Now using the fact that $(\boldsymbol{A} * \boldsymbol{B})(\boldsymbol{A} * \boldsymbol{B})^T = (\boldsymbol{A}\boldsymbol{A}^T)\odot(\boldsymbol{B}\boldsymbol{B}^T)$ we conclude that

$$\begin{aligned}&\left(\mathcal{J}\left(\widetilde{\boldsymbol{W}}\right) - \mathcal{J}\left(\boldsymbol{W}\right)\right)\left(\mathcal{J}\left(\widetilde{\boldsymbol{W}}\right) - \mathcal{J}\left(\boldsymbol{W}\right)\right)^T\\ &\quad= \left(\left(\phi'\left(\boldsymbol{X}\widetilde{\boldsymbol{W}}^T\right) - \phi'\left(\boldsymbol{X}\boldsymbol{W}^T\right)\right)\mathrm{diag}(\boldsymbol{v})\mathrm{diag}(\boldsymbol{v})\left(\phi'\left(\widetilde{\boldsymbol{W}}\boldsymbol{X}^T\right) - \phi'\left(\boldsymbol{W}\boldsymbol{X}^T\right)\right)\right)\\ &\qquad\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right).\end{aligned} \tag{A.6}$$

To continue further we use Lemma 6.5 combined with (A.6) to conclude that

$$\begin{aligned}\left\|\mathcal{J}\left(\widetilde{\boldsymbol{W}}\right) - \mathcal{J}\left(\boldsymbol{W}\right)\right\|^2 &\le \left\|\mathrm{diag}(\boldsymbol{v})\left(\phi'\left(\widetilde{\boldsymbol{W}}\boldsymbol{X}^T\right) - \phi'\left(\boldsymbol{W}\boldsymbol{X}^T\right)\right)\right\|^2\left(\max_i\|\boldsymbol{x}_i\|_{\ell_2}^2\right)\\ &\le \|\boldsymbol{v}\|_{\ell_\infty}^2\left\|\phi'\left(\widetilde{\boldsymbol{W}}\boldsymbol{X}^T\right) - \phi'\left(\boldsymbol{W}\boldsymbol{X}^T\right)\right\|^2\\ &\overset{(a)}{=} \|\boldsymbol{v}\|_{\ell_\infty}^2\left\|\phi''\left((\boldsymbol{S}\odot\boldsymbol{W} + (1-\boldsymbol{S})\odot\widetilde{W})\boldsymbol{X}^T\right)\odot\left((\widetilde{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{X}^T\right)\right\|^2\\ &\le \|\boldsymbol{v}\|_{\ell_\infty}^2\left\|\phi''\left((\boldsymbol{S}\odot\boldsymbol{W} + (1-\boldsymbol{S})\odot\widetilde{W})\boldsymbol{X}^T\right)\odot\left((\widetilde{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{X}^T\right)\right\|_F^2\\ &\le \|\boldsymbol{v}\|_{\ell_\infty}^2 B^2\left\|(\widetilde{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{X}^T\right\|_F^2\\ &\le \|\boldsymbol{v}\|_{\ell_\infty}^2 B^2\|\boldsymbol{X}\|^2\left\|\widetilde{\boldsymbol{W}} - \boldsymbol{W}\right\|_F^2,\end{aligned}$$

completing the proof of this lemma. Here, (a) holds by the mean value theorem for some matrix $\boldsymbol{S} \in \mathbb{R}^{k\times d}$ with entries $0 \le \boldsymbol{S}_{ij} \le 1$.

2) Jacobian perturbation results for ReLU networks (Proof of Lemma 6.9)

To prove Lemma 6.9 we first relate the perturbation of the Jacobian to perturbation of the activation pattern $\phi'(\boldsymbol{X}\boldsymbol{W}^T)$ as follows.

Lemma A.2: Consider the matrices $\boldsymbol{W}, \widetilde{\boldsymbol{W}} \in \mathbb{R}^{k\times d}$ and a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with unit Euclidean norm rows. Then,

$$\|\mathcal{J}(\boldsymbol{W}) - \mathcal{J}(\widetilde{\boldsymbol{W}})\| \le \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{X}\| \cdot \max_{1\le i\le n} \left\|\phi'\left(\boldsymbol{W}\boldsymbol{x}_i\right) - \phi'\left(\widetilde{\boldsymbol{W}}\boldsymbol{x}_i\right)\right\|_{\ell_2}.$$

Proof Similar to the smooth case in the previous section, the Jacobian difference is given by

$$\mathcal{J}(\boldsymbol{W}) - \mathcal{J}(\widetilde{\boldsymbol{W}}) = \left(\operatorname{diag}(\boldsymbol{v})\left(\phi'(\boldsymbol{X}\boldsymbol{W}^T) - \phi'(\boldsymbol{X}\widetilde{\boldsymbol{W}}^T)\right)\right) * \boldsymbol{X}.$$

Consequently,

$$\begin{aligned}
\left\|\mathcal{J}(\boldsymbol{W}) - \mathcal{J}(\widetilde{\boldsymbol{W}})\right\|^2 &= \left\|\left(\mathcal{J}(\boldsymbol{W}) - \mathcal{J}(\widetilde{\boldsymbol{W}})\right)\left(\mathcal{J}(\boldsymbol{W}) - \mathcal{J}(\widetilde{\boldsymbol{W}})\right)^T\right\| \\
&\le \Big\|\Big(\left(\phi'\left(\boldsymbol{X}\boldsymbol{W}^T\right) - \phi'\left(\boldsymbol{X}\widetilde{\boldsymbol{W}}^T\right)\right)\operatorname{diag}(\boldsymbol{v})\operatorname{diag}(\boldsymbol{v})\left(\phi'\left(\boldsymbol{W}\boldsymbol{X}^T\right) - \phi'\left(\widetilde{W}\boldsymbol{X}^T\right)\right)\Big) \\
&\quad \odot \left(\boldsymbol{X}\boldsymbol{X}^T\right)\Big\| \\
&\le \left(\max_{1\le i\le n}\left\|\operatorname{diag}(\boldsymbol{v})\left(\phi'\left(\boldsymbol{W}\boldsymbol{x}_i\right) - \phi'\left(\widetilde{W}\boldsymbol{x}_i\right)\right)\right\|_{\ell_2}^2\right) \cdot \|\boldsymbol{X}\|^2 \\
&\le \|\boldsymbol{v}\|_{\ell_\infty}^2 \|\boldsymbol{X}\|^2 \cdot \max_{1\le i\le n}\left\|\phi'\left(\boldsymbol{W}\boldsymbol{x}_i\right) - \phi'\left(\widetilde{\boldsymbol{W}}\boldsymbol{x}_i\right)\right\|_{\ell_2}^2
\end{aligned}$$

■

The lemma above implies that, we simply need to control $\phi'(\boldsymbol{W}\boldsymbol{x}_i)$ around a neighborhood of $\boldsymbol{W}_0$. To continue note that since ϕ' is the step function, we shall focus on the number of sign flips between the matrices $\boldsymbol{W}\boldsymbol{X}^T$ and $\boldsymbol{W}_0\boldsymbol{X}^T$. Let $\|\boldsymbol{v}\|_{m-}$ denote the mth smallest entry of $\boldsymbol{v}$ after sorting its entries in terms of absolute value. We first state a intermediate lemma.

Lemma A.3: Given an integer m, suppose

$$\|\boldsymbol{W} - \boldsymbol{W}_0\| \le \sqrt{m}\,|\boldsymbol{W}_0\boldsymbol{x}_i|_{m-},$$

holds for $i = 1, 2, \ldots, n$. Then

$$\max_{1\le i\le n}\|\phi'(\boldsymbol{W}\boldsymbol{x}_i) - \phi'(\boldsymbol{W}_0\boldsymbol{x}_i)\|_{\ell_2} \le \sqrt{2m}.$$

Proof We will prove this result by contradiction. Suppose there is an $\boldsymbol{x}_i$ such that $\phi'(\boldsymbol{W}\boldsymbol{x}_i)$ and $\phi'(\boldsymbol{W}_0\boldsymbol{x}_i)$ have (at least) $2m$ different entries. Let $\{(a_r, b_r)\}_{r=1}^{2m}$ be (a subset of) entries of $\boldsymbol{W}\boldsymbol{x}_i, \boldsymbol{W}_0\boldsymbol{x}_i$ at these differing locations respectively and suppose a_r's are sorted decreasingly in absolute value. By definition $|a_r| \geq |\boldsymbol{W}_0\boldsymbol{x}_i|_{m-}$ for $r \leq m$. Consequently, using $\text{sign}(a_r) \neq \text{sign}(b_r)$,

$$\begin{aligned}\|\boldsymbol{W} - \boldsymbol{W}_0\|^2 &\geq \|(\boldsymbol{W} - \boldsymbol{W}_0)\boldsymbol{x}_i\|_{\ell_2}^2 \\ &\geq \sum_{r=1}^{2m} |a_r - b_r|^2 \\ &\geq \sum_{r=1}^{2m} |a_r|^2 \\ &\geq m\,|\boldsymbol{W}_0\boldsymbol{x}_i|_{m-}^2 .\end{aligned}$$

This implies $\|\boldsymbol{W} - \boldsymbol{W}_0\| \geq \sqrt{m}\,|\boldsymbol{W}_0\boldsymbol{x}_i|_{m-}$ contradicting the assumption of the lemma and thus concluding the proof. ■

Now note that by setting $m = m_0^2$ in Lemma A.3 as long as

$$\|\boldsymbol{W} - \boldsymbol{W}_0\| \leq m_0\,|\boldsymbol{W}_0\boldsymbol{x}_i|_{m_0^2-}, \tag{A.7}$$

we have

$$\max_{1\leq i\leq n} \|\phi'(\boldsymbol{W}\boldsymbol{x}_i) - \phi'(\boldsymbol{W}_0\boldsymbol{x}_i)\|_{\ell_2} \leq \frac{\|\boldsymbol{v}\|_{\ell_2}}{10\|\boldsymbol{v}\|_{\ell_\infty}} \frac{\sqrt{\lambda(\boldsymbol{X})}}{\|\boldsymbol{X}\|} := \sqrt{2}m_0. \tag{A.8}$$

Using Lemma A.2, this in turn implies

$$\begin{aligned}\|\mathcal{J}(\boldsymbol{W}) - \mathcal{J}(\boldsymbol{W}_0)\| &\leq \|\boldsymbol{v}\|_{\ell_\infty}\,\|\boldsymbol{X}\| \cdot \max_{1\leq i\leq n} \|\phi'(\boldsymbol{W}\boldsymbol{x}_i) - \phi'(\boldsymbol{W}_0\boldsymbol{x}_i)\|_{\ell_2} \leq \frac{1}{10}\|\boldsymbol{v}\|_{\ell_2}\sqrt{\lambda(\boldsymbol{X})} \\ &\leq \frac{1}{6\sqrt{2}}\|\boldsymbol{v}\|_{\ell_2}\sqrt{\lambda(\boldsymbol{X})}.\end{aligned}$$

Thus to complete the proof of Lemma 6.9 all that remains is to prove (A.7). To this aim, we state the following lemma proven later in this section.

Lemma A.4: Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ be the input data point with unit Euclidean norm. Also let $\boldsymbol{W}_0 \in \mathbb{R}^{k\times d}$ be a matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Then, with probability at least $1 - ne^{-\frac{m}{6}}$,

$$|\boldsymbol{W}_0\boldsymbol{x}_i|_{m-} \geq \frac{m}{2k} \quad \text{for all} \quad i = 1, 2, \ldots, n.$$

Now applying Lemma A.4 we conclude that with probability at least $1 - ne^{-\frac{1}{1200}\frac{\|\boldsymbol{v}\|_{\ell_2}^2}{\|\boldsymbol{v}\|_{\ell_\infty}^2}\frac{\lambda(\boldsymbol{X})}{\|\boldsymbol{X}\|^2}}}$

$$m_0 \left|\boldsymbol{W}_0 \boldsymbol{x}_i\right|_{m_0^2-} \geq \frac{m_0^3}{2k},$$

holds for all $i = 1, 2, \ldots, n$. Hence, with same probability, all $\|\boldsymbol{W} - \boldsymbol{W}_0\| \leq \frac{m_0^3}{2k}$ obeys (A.7) concluding the proof of Lemma 6.9.

3) Proof of Lemma A.4

Observe that $\boldsymbol{W}_0\boldsymbol{x}_1, \boldsymbol{W}_0\boldsymbol{x}_2, \ldots, \boldsymbol{W}_0\boldsymbol{x}_n$ are all standard normal however they depend on each other. We begin by focusing on one such vector. We begin by proving that with probability at least $1 - e^{-\frac{m}{6}}$, at most m of the entries of $\boldsymbol{W}_0\boldsymbol{x}_i$ are less than $\frac{m}{2k}$. To this aim let γ_α be the number for which $\mathbb{P}\{|g| \leq \gamma_\alpha\} = \alpha$ where $g \sim \mathcal{N}(0,1)$ (i.e. the inverse cumulative density function of $|g|$). γ_α trivially obeys $\gamma_\alpha \geq \sqrt{\pi/2}\alpha$. To continue set $\boldsymbol{g} := \boldsymbol{W}_0\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_k)$ and the Bernouli random variables δ_ℓ given by

$$\delta_\ell = \begin{cases} 1 & \text{if } |g_\ell| \leq \gamma_\delta \\ 0 & \text{if } |g_\ell| > \gamma_\delta \end{cases}$$

with $\delta = \frac{m}{2k}$. Note that

$$\mathbb{E}\left[\sum_{\ell=1}^{k} \delta_\ell\right] = \sum_{\ell=1}^{k} \mathbb{E}[\delta_\ell] = \sum_{\ell=1}^{k} \mathbb{P}\left\{|g_\ell| \leq \gamma_\delta\right\} = \delta k = \frac{m}{2}.$$

Since the δ_ℓ's are i.i.d., applying a standard Chernoff bound we obtain

$$\mathbb{P}\left\{\sum_{\ell=1}^{k} \delta_\ell \geq m\right\} \leq e^{-\frac{m}{6}}.$$

The complementary event implies that at most m entries are less than $\frac{m}{2k}$. This together with the union bound completes the proof.

D. Proof of Corollary 6.11

First note that (VI.13) can be rewritten in the form

$$R := \frac{4}{\alpha}\left\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\right\|_{\ell_2} \leq \frac{\alpha}{L}.$$

Thus using the Lipschitzness of the Jacobian from (VI.8) for all $\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_0, R)$ we have

$$\left\|\mathcal{J}(\boldsymbol{\theta}) - \mathcal{J}(\boldsymbol{\theta}_0)\right\| \leq L\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|_{\ell_2} \leq LR \leq \alpha.$$

Combining the latter with the triangular inequality we conclude that

$$\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta})) \geq \sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)) - \|\mathcal{J}(\boldsymbol{\theta}) - \mathcal{J}(\boldsymbol{\theta}_0)\| \geq 2\alpha - \alpha = \alpha,$$

so that (VI.7) holds under the assumptions of the corollary. Therefore, all of the assumptions of Theorem 6.10 continue to hold and thus so do its conclusions.

E. Proof of Corollary 2.2

The proof follows from a simple application of Theorem 2.1. We just need to calculate the various constants involved in this result. First, we focus on the constants related to the activation.It is trivial to check that $B = M = 1$ and $\mu_\phi \approx 0.207$ so that (II.2) reduces to $\sqrt{kd} \geq \tilde{c}(1+\delta)\kappa(\boldsymbol{X})n$ with $\tilde{c}$ a fixed numerical constant. Next we focus on the constant $\kappa(\boldsymbol{X})$ that depends on the data matrix. To this aim we note that standard results regarding the concentration of spectral norm of random matrices with i.i.d. rows imply that

$$\|\boldsymbol{X}\| \leq 2\sqrt{\frac{n}{d}},$$

holds with probability at least $1 - e^{-\gamma_2 d}$. Furthermore, based on a simple modification of [2, Corollary 6.5]

$$\sigma_{\min}\left(\boldsymbol{X} * \boldsymbol{X}\right) \geq c, \tag{A.9}$$

holds with probability at least $1 - ne^{-\gamma_1\sqrt{n}} - \frac{1}{n} - 2ne^{-\gamma_2 d}$ where c, γ_1, and γ_2 are fixed numerical constants.

F. Proof of Theorem 2.6

The proof of this result follows from [10, Theorem 3.1] similar to how Theorem 2.1 follows from Theorem 6.10 from the same paper. There are two differences in the proof. First [10, Theorem 3.1] requires Jacobian regularity condition (VI.7) to hold over the larger region $R_{sgd} := \frac{\nu\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$ compared to $R_{gd} := \frac{4\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$ (recall $\nu \geq 4$).

This requires us to use a wider network as follows. Recalling Theorem 6.10, the minimum and maximum singular values α and β of a wider network obeys the exact same guarantees with better or equal probabilities. However, in light of Corollary 6.11, we have to ensure that

$R_{sgd}L_{sgd} \le \alpha$ where L_{sgd} is the Lipschitz constant of the Jacobian in Theorem 2.6 and R_{sgd} is the larger SGD region defined above i.e. instead of (VI.13), we need to use $\frac{\alpha^2}{\nu L} \ge \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}$. This will lead to a smaller Lipschitz constant compared to gradient descent analysis via the relation $L_{sgd} = \alpha/R_{sgd} = \frac{R_{gd}}{R_{sgd}} L_{gd}$ where L_{gd} is the Lipschitz constant required for the gradient descent analysis. As we show next, this leads to the width requirement of $k_{sgd} = (\nu/4)^2 k_{gd}$. To be rigorous, using the α, L_{sgd}, R_{sgd} estimates of (VI.14), Lemma 6.12, (VI.17) we need

$$R_{sgd}L_{sgd} \le \alpha \iff \frac{1}{\frac{B}{\sqrt{kn}} \|\boldsymbol{y}\|_{\ell_2} \|\boldsymbol{X}\|} \ge \frac{8\nu \|\boldsymbol{y}\|_{\ell_2} (1 + (1+\delta)B)}{\frac{\|\boldsymbol{y}\|_{\ell_2}^2}{n} \lambda(\boldsymbol{X})}$$

which implies

$$\sqrt{k} \ge \frac{8\nu\sqrt{n} B \|\boldsymbol{X}\| (1 + (1+\delta)B)}{\lambda(\boldsymbol{X})} \iff \sqrt{kd} \ge 8\nu B n \widetilde{\kappa}(\boldsymbol{X}) (1 + (1+\delta)B).$$

Thus our bound on the SGD network requires exactly $(\nu/4)^2$ times as many parameters as the gradient descent bound (i.e. compare to (VI.19)).

Secondly, the learning rate choice in [10, Theorem 3.1] requires us to calculate is the maximum Euclidean norm of the rows of the Jacobian matrix. For neural networks this takes the following form

$$\begin{aligned}
\max_i \ \|\mathcal{J}_i(\boldsymbol{W})\|_{\ell_2} &= \left\|\mathrm{diag}(\boldsymbol{v})\phi'(\boldsymbol{W}\boldsymbol{x}_i)\,\boldsymbol{x}_i^T\right\|_F \\
&= \|\mathrm{diag}(\boldsymbol{v})\phi'(\boldsymbol{W}\boldsymbol{x}_i)\|_F \|\boldsymbol{x}_i\|_{\ell_2} \\
&= \|\mathrm{diag}(\boldsymbol{v})\phi'(\boldsymbol{W}\boldsymbol{x}_i)\|_F \\
&\le \|\phi'(\boldsymbol{W}\boldsymbol{x}_i)\|_{\ell_\infty} \|\boldsymbol{v}\|_{\ell_2} \\
&\le B\|\boldsymbol{v}\|_{\ell_2}.
\end{aligned}$$

G. Proofs for nonsmooth optimization (Proof of Theorem 6.13)

To prove this theorem we begin by stating a few preliminary results and definitions.

Lemma A.5 (Asymmetric PSD perturbation): Consider the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n\times p}$ obeying $\|\boldsymbol{B} - \boldsymbol{C}\| \le \varepsilon$ and $\|\boldsymbol{A} - \boldsymbol{C}\| \le \varepsilon$. Then, for all $\boldsymbol{r} \in \mathbb{R}^n$,

$$|\boldsymbol{r}^T \boldsymbol{B}\boldsymbol{A}^T \boldsymbol{r} - \|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2}^2| \le 2\varepsilon \|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2} \|\boldsymbol{r}\|_{\ell_2} + \varepsilon^2 \|\boldsymbol{r}\|_{\ell_2}^2.$$

Proof We have

$$\boldsymbol{r}^T\boldsymbol{B}\boldsymbol{A}^T\boldsymbol{r} - \|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2}^2 = \boldsymbol{r}^T(\boldsymbol{B}-\boldsymbol{C})(\boldsymbol{A}-\boldsymbol{C})^T\boldsymbol{r} + \boldsymbol{r}^T\boldsymbol{C}(\boldsymbol{A}-\boldsymbol{C})^T\boldsymbol{r} + \boldsymbol{r}^T(\boldsymbol{B}-\boldsymbol{C})\boldsymbol{C}^T\boldsymbol{r}.$$

This implies

$$\begin{aligned}|\boldsymbol{r}^T\boldsymbol{B}\boldsymbol{A}^T\boldsymbol{r} - \|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2}^2| &\le |\boldsymbol{r}^T(\boldsymbol{B}-\boldsymbol{C})(\boldsymbol{A}-\boldsymbol{C})^T\boldsymbol{r}| + \|(\boldsymbol{A}-\boldsymbol{C})^T\boldsymbol{r}\|_{\ell_2}\|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2} \\ &\quad + \|(\boldsymbol{B}-\boldsymbol{C})^T\boldsymbol{r}\|_{\ell_2}\|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2} \\ &\le \varepsilon^2\|\boldsymbol{r}\|_{\ell_2}^2 + 2\varepsilon\|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2}\|\boldsymbol{r}\|_{\ell_2},\end{aligned}$$

concluding the proof. ∎

Definition A.6 (Average Jacobian): We define the average Jacobian along the path connecting two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ as

$$\mathcal{J}(\boldsymbol{y}, \boldsymbol{x}) := \int_0^1 \mathcal{J}(\boldsymbol{x} + \alpha(\boldsymbol{y}-\boldsymbol{x}))d\alpha. \tag{A.10}$$

Lemma A.7: Suppose $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ satisfy $\|\boldsymbol{x}-\boldsymbol{\theta}_0\|, \|\boldsymbol{y}-\boldsymbol{\theta}_0\| \le R$. Then, under Assumptions 2 and 4, for any $\boldsymbol{r} \in \mathbb{R}^d$, we have

$$\begin{aligned}\boldsymbol{r}^T\mathcal{J}(\boldsymbol{y}, \boldsymbol{x})\mathcal{J}(\boldsymbol{x})^T\boldsymbol{r} &\ge \frac{\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}^2}{2}, \\ \|\mathcal{J}(\boldsymbol{x})^T\boldsymbol{r}\|_{\ell_2}^2 &\le 1.5\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}^2.\end{aligned}$$

Proof Under Assumptions 2 and 4, applying Lemma A.5 with $\boldsymbol{A} = \mathcal{J}(\boldsymbol{x})$, $\boldsymbol{B} = \mathcal{J}(\boldsymbol{y}, \boldsymbol{x})$, $\boldsymbol{C} = \mathcal{J}(\boldsymbol{\theta}_0)$, and $\varepsilon = \alpha/3$, we conclude that

$$\begin{aligned}\boldsymbol{r}^T\mathcal{J}(\boldsymbol{y}, \boldsymbol{x})\mathcal{J}(\boldsymbol{x})^T\boldsymbol{r} - \|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}^2 &\ge -\left(\frac{2\alpha}{3}\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}\|\boldsymbol{r}\|_{\ell_2} + \frac{\alpha^2}{9}\|\boldsymbol{r}\|_{\ell_2}^2\right) \\ &\ge -\left(\frac{2\alpha}{3}\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}\|\boldsymbol{r}\|_{\ell_2} + \frac{\alpha}{18}\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}\|\boldsymbol{r}\|_{\ell_2}\right) \\ &\ge -\alpha\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}\|\boldsymbol{r}\|_{\ell_2} \\ &\ge -\frac{\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}^2}{2}.\end{aligned}$$

This implies $\boldsymbol{r}^T\mathcal{J}(\boldsymbol{y}, \boldsymbol{x})\mathcal{J}(\boldsymbol{x})^T\boldsymbol{r} \ge \frac{\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}^2}{2}$. The upper bound similarly follows from Lemma A.5 by setting $\boldsymbol{A} = \boldsymbol{B} = \mathcal{J}(\boldsymbol{x})$ and observing that the deviation is again upper bounded by $\frac{\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}^2}{2}$. ∎

Lemma A.8: Suppose Assumptions 2 and 4 hold. Consider two consequent iterative updates $\boldsymbol{\theta}_\tau$ and $\boldsymbol{\theta}_{\tau+1}$ which by definition obey

$$\boldsymbol{\theta}_{\tau+1} := \boldsymbol{\theta}_\tau - \eta \mathcal{J}^T(\boldsymbol{\theta}_\tau)\left(f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\right),$$

with $\eta \leq \frac{1}{3\beta^2}$. Also, denote the corresponding residuals by $\boldsymbol{r}_{\tau+1} := f(\boldsymbol{\theta}_{\tau+1}) - \boldsymbol{y}$ and $\boldsymbol{r}_\tau := f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}$. Finally, assume $\boldsymbol{\theta}_\tau, \boldsymbol{\theta}_{\tau+1}$ satisfy $\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_0\|, \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\| \leq R$. Then

$$\|\boldsymbol{r}_{\tau+1}\|_{\ell_2} \leq \|\boldsymbol{r}_\tau\|_{\ell_2} - \frac{\eta}{4}\frac{\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}^2}{\|\boldsymbol{r}\|_{\ell_2}}.$$

Proof For this proof we use the short-hand $\mathcal{J}_{\tau+1,\tau} := \mathcal{J}(\boldsymbol{\theta}_\tau, \boldsymbol{\theta}_\tau)$ and $\mathcal{J}_\tau := \mathcal{J}(\boldsymbol{\theta}_\tau)$. We expand the residual at $\boldsymbol{\theta}_{\tau+1}$ using Lemma A.7 as follows

$$\begin{aligned}
\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}^2 &= \|(\boldsymbol{I} - \eta\mathcal{J}_{\tau+1,\tau}\mathcal{J}_\tau^T)\boldsymbol{r}_\tau\|_{\ell_2}^2 \\
&= \|\boldsymbol{r}_\tau\|_{\ell_2}^2 - 2\eta\boldsymbol{r}_\tau^T\mathcal{J}_{\tau+1,\tau}\mathcal{J}_\tau^T\boldsymbol{r}_\tau + \eta^2\|\mathcal{J}_{\tau+1,\tau}\mathcal{J}_\tau^T\boldsymbol{r}_\tau\|_{\ell_2}^2 \\
&\leq \|\boldsymbol{r}_\tau\|_{\ell_2}^2 - \eta\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}^2 + \eta^2\beta^2\|\mathcal{J}_\tau^T\boldsymbol{r}_\tau\|_{\ell_2}^2 \\
&\leq \|\boldsymbol{r}_\tau\|_{\ell_2}^2 - \eta\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}^2 + \frac{3}{2}\eta^2\beta^2\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}_\tau\|_{\ell_2}^2
\end{aligned}$$

Using the fact that $\eta \leq \frac{1}{3\beta^2}$, we conclude that

$$\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}^2 \leq \|\boldsymbol{r}_\tau\|_{\ell_2}^2 - \frac{\eta}{2}\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}^2 \implies \|\boldsymbol{r}_{\tau+1}\|_{\ell_2} \leq \|\boldsymbol{r}_\tau\|_{\ell_2} - \frac{\eta}{4}\frac{\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}^2}{\|\boldsymbol{r}\|_{\ell_2}}.$$

■

1) Completing the proof of Theorem 6.13

With these lemmas in place we are now ready to complete the proof of Theorem 6.13. To this aim suppose the conclusions hold until iteration $\tau > 0$. We shall show the result for iteration $\tau + 1$. We first prove that iterates still stays inside the region $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq R$. To this aim first note that by the induction hypothesis we know that

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\| \leq R - \frac{3\|\boldsymbol{y} - f(\boldsymbol{\theta}_\tau)\|_{\ell_2}}{\alpha}.$$

Combining this with the gradient update rule, $\eta \leq 1/\beta^2$ and $\|\mathcal{J}\| \leq \beta$ yields

$$\begin{aligned}
\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_0\| &\leq \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\| + \eta\|\mathcal{J}(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau\| \\
&\leq \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\| + \eta\,\|\mathcal{J}(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau\|_{\ell_2} \\
&\leq R - \frac{3\|\boldsymbol{y} - f(\boldsymbol{\theta}_\tau)\|_{\ell_2}}{\alpha} + \eta\|\mathcal{J}(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau\|_{\ell_2} \\
&\leq R - \frac{3\|\boldsymbol{y} - f(\boldsymbol{\theta}_\tau)\|_{\ell_2}}{\alpha} + \frac{1}{\beta}\|\boldsymbol{r}_\tau\|_{\ell_2} \\
&\leq R.
\end{aligned}$$

Now that we have shown $\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_0\| \leq R$, we can apply Lemma A.8 to conclude that

$$\|\boldsymbol{r}_{\tau+1}\|_{\ell_2} \leq \|\boldsymbol{r}_\tau\|_{\ell_2} - \frac{\eta}{4}\frac{\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}^2}{\|\boldsymbol{r}\|_{\ell_2}} \leq \|\boldsymbol{r}_\tau\|_{\ell_2} - \frac{\eta\alpha}{2}\|\mathcal{J}(\boldsymbol{\theta}_0)^T\boldsymbol{r}\|_{\ell_2}. \tag{A.11}$$

Next, we complement this by using Lemma A.7 to control the increase in the distance of the iterates to the initial point. This allows us to conclude that

$$\begin{aligned}
\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_0\| &\leq \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\| + \eta\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|, \\
\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_0\| &\leq \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\| + \eta\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}, \\
&\leq \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\| + \eta\|\mathcal{J}^T(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau\|_{\ell_2}, \\
&\leq \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\| + 1.25\eta\|\mathcal{J}^T(\boldsymbol{\theta}_0)\boldsymbol{r}_\tau\|_{\ell_2}.
\end{aligned}$$

Adding the latter two identities, we obtain

$$\|\boldsymbol{r}_{\tau+1}\|_{\ell_2} + \frac{\alpha}{3}\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_0\| \leq \|\boldsymbol{r}_\tau\|_{\ell_2} + \frac{\alpha}{3}\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\| \leq \|\boldsymbol{r}_0\|_{\ell_2},$$

completing the proof of (VI.22). Finally, the convergence rate guarantee (VI.21) follows from (A.11) can be upper bounded by $(1 - \eta\alpha^2)\|\boldsymbol{r}_\tau\|_{\ell_2}$.

H. Lower bounds on the minimum eigenvalue of covariance matrices

In this section we discuss lower bounds on the minimum eigenvalue of the neural network and output feature covariance matrices which involve higher order Khatri-Rao products. This results involve the Hermite expansion of the activation and its derivatives. For any ϕ with bounded

Gaussian meaure i.e. $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\phi^2(g)e^{-\frac{g^2}{2}}dg<\infty$ the Hermite coefficients $\{\mu_r(\phi)\}_{r=0}^{+\infty}$ associated to ϕ are defined as

$$\mu_r(\phi) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\phi(g)h_r(g)e^{-\frac{g^2}{2}}dg,$$

where $h_r(g)$ is the normalized probabilists' Hermite polynomial defined by

$$h_r(x) := \frac{1}{\sqrt{r!}}(-1)^r e^{\frac{x^2}{2}}\frac{d^r}{dx^r}e^{-\frac{x^2}{2}}.$$

Using these expansions we prove the following simple lemma. The first one is a generalization of the reduction to quadratic activation Lemma (Lemma 6.4). See also [52] for related calculations. We note that Lemma 6.4 is a special case as $\widetilde{\mu}_\phi = \mu_0(\phi)$ and $\mu_\phi = \mu_1(\phi)$.

Lemma A.9: For an activation $\phi : \mathbb{R} \mapsto \mathbb{R}$ and a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with unit Euclidean norm rows the neural network covariance matrix and eigenvalue obey

$$\boldsymbol{\Sigma}(\boldsymbol{X}) = \left(\mu_0^2(\phi')\mathbf{1}\mathbf{1}^T + \sum_{r=1}^{+\infty}\mu_r^2(\phi)\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{\odot r}\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right) \succeq \mu_r^2(\phi')\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{\odot(r+1)}, \tag{A.12}$$

$$\lambda(\boldsymbol{X}) \geq \mu_r^2(\phi')\sigma_{\min}^2\left(\boldsymbol{X}^{*(r+1)}\right) \quad \text{for any } r = 0, 1, 2, \ldots. \tag{A.13}$$

As a reminder, for a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{A}^{\odot r} \in \mathbb{R}^{n\times n}$ is defined inductively via $\boldsymbol{A}^{\odot r} = \boldsymbol{A}\odot(\boldsymbol{A}^{\odot(r-1)})$ with $\boldsymbol{A}^{\odot 0} = \mathbf{1}\mathbf{1}^T$. Similarly, for a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with rows given by $\boldsymbol{x}_i \in \mathbb{R}^d$ we define the matrix $\boldsymbol{X}^{*r} \in \mathbb{R}^{n\times d^r}$ as

$$\left[\boldsymbol{X}^{*r}\right]_i = \left(\underbrace{\boldsymbol{x}_i \otimes \boldsymbol{x}_i \otimes \ldots \otimes \boldsymbol{x}_i}_{r}\right)^T$$

Proof To prove this result note that by the properties of Hermite expansions we have

$$\begin{aligned}\left[\mathbb{E}[\phi'(\boldsymbol{X}\boldsymbol{w})\phi'(\boldsymbol{X}\boldsymbol{w})^T]\right]_{ij} &= \mathbb{E}[\phi'(\boldsymbol{x}_i^T\boldsymbol{w})\phi'(\boldsymbol{x}_j^T\boldsymbol{w})] \\ &= \sum_{r=0}^{\infty}\mu_r^2(\phi')(\boldsymbol{x}_i^T\boldsymbol{x}_j)^r\end{aligned}$$

Thus

$$\boldsymbol{\Sigma}(\boldsymbol{X}) = \left(\sum_{r=0}^{\infty}\mu_r^2(\phi')\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{\odot r}\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right).$$

Furthermore,

$$\sum_{r=0}^{\infty}\mu_r^2(\phi')\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{\odot r} = \sum_{r=0}^{\infty}\left(\mu_r(\phi')\boldsymbol{X}^{*r}\right)\left(\mu_r(\phi')\boldsymbol{X}^{*r}\right)^T \succeq \mu_r^2(\phi')\left(\boldsymbol{X}^{*r}\right)\left(\boldsymbol{X}^{*r}\right)^T = \mu_r^2(\phi')\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{\odot r}.$$

Using the latter combined with the fact that the Hadamard product of two PSD matrices are PSD we arrive at (A.12). The latter also implies (A.13). ■

Similarly, it is also easy to prove the following result about the output feature covariance.

Lemma A.10: For an activation $\phi : \mathbb{R} \mapsto \mathbb{R}$ and a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with unit Euclidean norm rows the output feature covariance matrix and eigenvalue obey

$$\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{X}) = \left(\mu_0^2(\phi)\mathbf{1}\mathbf{1}^T + \sum_{r=1}^{+\infty} \mu_r^2(\phi)\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{\odot r}\right) \geq \mu_r^2(\phi)\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{\odot(r)}, \tag{A.14}$$

$$\widetilde{\lambda}(\boldsymbol{X}) \geq \mu_r^2(\phi)\sigma_{\min}^2(\boldsymbol{X}^{*r}) \qquad \text{for any } r = 1, 2, \ldots. \tag{A.15}$$

Proof To prove this result note that by the properties of Hermite expansions we have

$$\begin{aligned}\left[\mathbb{E}[\phi(\boldsymbol{X}\boldsymbol{w})\,\phi(\boldsymbol{X}\boldsymbol{w})^T]\right]_{ij} &= \mathbb{E}[\phi(\boldsymbol{x}_i^T\boldsymbol{w})\phi(\boldsymbol{x}_j^T\boldsymbol{w})] \\ &= \sum_{r=0}^{\infty} \mu_r^2(\phi)(\boldsymbol{x}_i^T\boldsymbol{x}_j)^r\end{aligned}$$

Thus

$$\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{X}) = \sum_{r=0}^{\infty} \mu_r^2(\phi)\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{\odot r} \geq \mu_r^2(\phi)\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{\odot(r)},$$

concluding the proof of (A.14). This in turn also implies (A.15). ■

We begin by stating a result regarding the covariance of the indicator mapping. Below we use $\mathcal{I}$ to denote the step function i.e. $\mathcal{I}(z) = \mathbb{1}_{\{z\geq 0\}}$.

Theorem A.11: Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be points in $\mathbb{R}^d$ with unit Euclidian norm and $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Form the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d} = [\boldsymbol{x}_1 \ \ldots \ \boldsymbol{x}_n]^T$. Suppose there exists $\delta > 0$ such that for every $1 \leq i \neq j \leq n$ we have that

$$\min(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\ell_2}, \|\boldsymbol{x}_i + \boldsymbol{x}_j\|_{\ell_2}) \geq \delta.$$

Then, the covariance of the vector $\mathcal{I}(\boldsymbol{X}\boldsymbol{w})$ obeys

$$\mathbb{E}[\mathcal{I}(\boldsymbol{X}\boldsymbol{w})\mathcal{I}(\boldsymbol{X}\boldsymbol{w})^T] \geq \frac{\delta}{100n^2}. \tag{A.16}$$

Proof Fix a unit length vector $\boldsymbol{a} \in \mathbb{R}^n$. Suppose there exists constants c_1, c_2 such that

$$\mathbb{P}(|\boldsymbol{a}^T\mathcal{I}(\boldsymbol{X}\boldsymbol{w})| \geq c_1\|\boldsymbol{a}\|_{\ell_\infty}) \geq \frac{c_2\delta}{n}. \tag{A.17}$$

This would imply that

$$\mathbb{E}[(\boldsymbol{a}^T\mathcal{I}(\boldsymbol{X}\boldsymbol{w}))^2] \geq \mathbb{E}[|\boldsymbol{a}^T\mathcal{I}(\boldsymbol{X}\boldsymbol{w})|]^2 \geq c_1^2\|\boldsymbol{a}\|_{\ell_\infty}^2\frac{c_2\delta}{n} \geq c_1^2c_2\frac{\delta}{n^2}.$$

Since this is true for all $\boldsymbol{a}$, we find (A.19) with $c_1^2c_2 = \frac{1}{100}$ by choosing $c_1 = 1/2, c_2 = 1/25$ as described later. Hence, our goal is proving (A.17). For the most part, our argument is based on exploiting independence of orthogonal decomposition associated with Gaussian vectors and we will refine the argument of [8]. Without losing generality, assume $|a_1| = \|\boldsymbol{a}\|_{\ell_\infty}$ and construct an orthonormal basis $\boldsymbol{Q}$ in $\mathbb{R}^d$ where the first column is equal to $\boldsymbol{x}_1$ and $\boldsymbol{Q} = [\boldsymbol{x}_1\ \bar{\boldsymbol{Q}}]$. Note that $\boldsymbol{g} = \boldsymbol{Q}^T\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and we have

$$\boldsymbol{w} = \boldsymbol{Q}\boldsymbol{g} = g_1\boldsymbol{x}_1 + \bar{\boldsymbol{Q}}\bar{\boldsymbol{g}}.$$

For $0 \leq \gamma \leq 1/2$, Gaussian small ball guarantees

$$\mathbb{P}(|g_1| \leq \gamma) \geq \frac{7\gamma}{10}.$$

Next, we argue that $\boldsymbol{z}_i = \langle\bar{\boldsymbol{Q}}\bar{\boldsymbol{g}}, \boldsymbol{x}_i\rangle$ is small for all $i \neq 1$. For a fixed $i \geq 2$, observe that

$$\boldsymbol{z}_i \sim \mathcal{N}(0, 1 - (\boldsymbol{x}_1^T\boldsymbol{x}_i)^2).$$

Note that

$$1 - |\boldsymbol{x}_1^T\boldsymbol{x}_i| = \frac{\min(\|\boldsymbol{x}_1 - \boldsymbol{x}_i\|_{\ell_2}^2, \|\boldsymbol{x}_1 + \boldsymbol{x}_i\|_{\ell_2}^2)}{2} \geq \frac{\delta^2}{2}.$$

Hence $1 - (\boldsymbol{x}_1^T\boldsymbol{x}_i)^2 \geq \delta^2/2$. From Gaussian small ball and variance bound on $\boldsymbol{z}_i$, we have

$$\mathbb{P}(|\boldsymbol{z}_i| \leq \gamma) \leq \sqrt{\frac{2}{\pi}}\frac{\gamma}{\sqrt{1 - (\boldsymbol{x}_1^T\boldsymbol{x}_i)^2}} \leq \frac{2\gamma}{\delta\sqrt{\pi}}$$

Union bounding, we find that, with probability $1 - \frac{2n\gamma}{\sqrt{\pi}\delta}$, we have that, $|\boldsymbol{z}_i| > \gamma$ for all $i \geq 2$. Since $\bar{\boldsymbol{g}}$ is independent of g_1, setting $\gamma = \frac{\delta}{2\sqrt{2}n}$ (which is at most $1/2$ since $\delta \leq \sqrt{2}$),

$$\mathbb{P}(E) := \mathbb{P}(|g_1| \leq \gamma,\ |\boldsymbol{z}_i| > \gamma\ \forall\ i \geq 2) \geq (1 - \frac{2n\gamma}{\sqrt{\pi}\delta})\frac{2\gamma}{5} \geq \frac{\delta}{12n}.$$

To proceed, note that

$$f(\boldsymbol{g}) := \boldsymbol{a}^T\mathcal{I}(\boldsymbol{X}\boldsymbol{w}) = a_1\mathcal{I}(g_1) + \sum_{i=2}^{n}\left(a_i \times \mathcal{I}(\boldsymbol{x}_i^T\boldsymbol{x}_1g_1 + \boldsymbol{x}_i^T\bar{\boldsymbol{Q}}\bar{\boldsymbol{g}})\right)$$

On the event E, we have that $\mathcal{I}(\boldsymbol{x}_i^T\boldsymbol{x}_1 g_1 + \boldsymbol{x}_i^T\bar{\boldsymbol{Q}}\bar{\boldsymbol{g}}) = \mathcal{I}(\boldsymbol{x}_i^T\bar{\boldsymbol{Q}}\bar{\boldsymbol{g}})$ since $|g_1| \le \gamma \le |\boldsymbol{x}_i^T\bar{\boldsymbol{Q}}\bar{\boldsymbol{g}}|$. Hence, on E,

$$f(\boldsymbol{g}) = a_1\mathcal{I}(g_1) + \text{rest}(\bar{\boldsymbol{g}}),$$

where $\text{rest}(\bar{\boldsymbol{g}}) = \sum_{i=2}^{n}(a_i \times \mathcal{I}(\boldsymbol{x}_i^T\bar{\boldsymbol{Q}}\bar{\boldsymbol{g}}))$. Furthermore, conditioned on E, $g_1, \bar{\boldsymbol{g}}$ are independent as $\boldsymbol{z}_i$'s are function of $\bar{\boldsymbol{g}}$ alone hence, E can be split into two equally likely events that are symmetric with respect to g_1 i.e. $g_1 \ge 0$ and $g_1 < 0$. Consequently,

$$\mathbb{P}(|f(\boldsymbol{g})| \ge \max(|a_1\mathcal{I}(g_1) + \text{rest}(\bar{\boldsymbol{g}})|, |a_1\mathcal{I}(-g_1) + \text{rest}(\bar{\boldsymbol{g}})|) \;\big|\; E) \ge 1/2 \tag{A.18}$$

Now, using $\max(|a|,|b|) \ge |a-b|/2$, we find

$$\mathbb{P}(|f(\boldsymbol{g})| \ge |a_1||\mathcal{I}(g_1) - \mathcal{I}(-g_1)|/2 \;\big|\; E) = \mathbb{P}(|f(\boldsymbol{g})| \ge |a_1|/2 \;\big|\; E) = \mathbb{P}(|f(\boldsymbol{g})| \ge \|\boldsymbol{a}\|_{\ell_\infty}/2 \;\big|\; E) \ge 1/2.$$

This yields $\mathbb{P}(|f(\boldsymbol{g})| \ge \|\boldsymbol{a}\|_{\ell_\infty}/2) \ge \mathbb{P}(E)/2 \ge \delta/24n$, concluding the proof by using $c_1 = 1/2,\ c_2 = 1/25$. ■

Corollary A.12 (Covariance of ReLU Jacobian): Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be points in $\mathbb{R}^d$ with unit Euclidian norm and $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Form the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d} = [\boldsymbol{x}_1 \ \ldots \ \boldsymbol{x}_n]^T$. Suppose there exists $\delta > 0$ such that for every $1 \le i \ne j \le n$, the input sample pairs have δ distance i.e.

$$\min(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\ell_2}, \|\boldsymbol{x}_i + \boldsymbol{x}_j\|_{\ell_2}) \ge \delta.$$

Then, using Lemma 6.5 and Theorem A.11

$$\mathbb{E}[\mathcal{I}(\boldsymbol{X}\boldsymbol{w})\mathcal{I}(\boldsymbol{X}\boldsymbol{w})^T \odot \boldsymbol{X}\boldsymbol{X}^T] \ge \frac{\delta}{100n^2}. \tag{A.19}$$

Proof of Theorem 2.5

Proof For proof, we wish to apply the Meta-Theorem 6.3 with proper value of $\lambda(\boldsymbol{X})$. Under Assumption 1, using Corollary A.12, we have that

$$\lambda(\boldsymbol{X}) \ge \frac{\delta}{100n^2}.$$

Substituting this $\lambda(\boldsymbol{X})$ value results in the advertised result $k \ge \mathcal{O}((1+\nu)^2 n^9 \|\boldsymbol{X}\|^6/\delta^4)$ and the associated learning rate. ■

I. Proofs for training the output layer (Proof of Theorem 3.2)

To begin note that

$$\boldsymbol{\Phi}\boldsymbol{\Phi}^T = \phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\phi\left(\boldsymbol{W}\boldsymbol{X}^T\right) = \sum_{\ell=1}^{k}\phi\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\phi\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T \geq \sum_{\ell=1}^{k}\phi\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\phi\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T \mathbb{1}_{\{\|\phi(\boldsymbol{X}\boldsymbol{w}_\ell)\|_{\ell_2}\leq T_n\}}.$$

Here T_n a function of n whose value shall be determined later in the proofs. To continue we need the matrix Chernoff result stated below.

Theorem A.13 (Matrix Chernoff): Consider a finite sequence $\boldsymbol{A}_\ell \in \mathbb{R}^{n\times n}$ of independent, random, Hermitian matrices with common dimension n. Assume that $\boldsymbol{0} \preceq \boldsymbol{A}_\ell \preceq R\boldsymbol{I}$ for $\ell = 1, 2, \ldots, k$. Then

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_{\ell=1}^{k}\boldsymbol{A}_\ell\right) \leq (1-\delta)\lambda_{\min}\left(\sum_{\ell=1}^{k}\mathbb{E}[\boldsymbol{A}_\ell]\right)\right\} \leq n\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\frac{\lambda_{\min}\left(\sum_{\ell=1}^{k}\mathbb{E}[\boldsymbol{A}_\ell]\right)}{R}}$$

for $\delta \in [0, 1)$.

Applying this theorem with $\boldsymbol{A}_\ell = \phi\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\phi\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T \mathbb{1}_{\{\|\phi(\boldsymbol{X}\boldsymbol{w}_\ell)\|_{\ell_2}\leq T_n\}}$, $R = T_n^2$ and $\widetilde{\boldsymbol{A}}(\boldsymbol{w}) := \phi\left(\boldsymbol{X}\boldsymbol{w}\right)\phi\left(\boldsymbol{X}\boldsymbol{w}\right)^T \mathbb{1}_{\left\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2}\leq T_n\right\}}$

$$\lambda_{\min}\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T\right) \geq (1-\delta)k\lambda_{\min}\left(\mathbb{E}[\widetilde{\boldsymbol{A}}(\boldsymbol{w})]\right), \tag{A.20}$$

holds with probability at least $1 - n\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\frac{k\lambda_{\min}\left(\mathbb{E}[\widetilde{\boldsymbol{A}}(\boldsymbol{w})]\right)}{T_n^2}}$.

Next we shall connect the the expected value of the truncated matrix $\widetilde{\boldsymbol{A}}(\boldsymbol{w})$ to one that is not

truncated defined as $\boldsymbol{A}(\boldsymbol{w}) = \phi(\boldsymbol{X}\boldsymbol{w})\phi(\boldsymbol{X}\boldsymbol{w})^T$. To do this note that

$$\begin{aligned}
\left\|\mathbb{E}[\widetilde{\boldsymbol{A}}(\boldsymbol{w}) - \boldsymbol{A}(\boldsymbol{w})]\right\| &= \left\|\mathbb{E}\left[\phi(\boldsymbol{X}\boldsymbol{w})\phi(\boldsymbol{X}\boldsymbol{w})^T \mathbb{1}_{\left\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\right\}}\right]\right\| \\
&\overset{(a)}{\leq} \mathbb{E}\left[\left\|\phi(\boldsymbol{X}\boldsymbol{w})\phi(\boldsymbol{X}\boldsymbol{w})^T \mathbb{1}_{\left\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\right\}}\right\|\right] \\
&\leq \mathbb{E}\left[\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2}^2 \mathbb{1}_{\left\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\right\}}\right] \qquad \text{(A.21)} \\
&\overset{(b)}{\leq} 2\,\mathbb{E}\left[\|\phi(\boldsymbol{X}\boldsymbol{w}) - \phi(\boldsymbol{0})\|_{\ell_2}^2 \mathbb{1}_{\left\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\right\}}\right] + 2\,\mathbb{E}\left[\|\phi(\boldsymbol{0})\|_{\ell_2}^2 \mathbb{1}_{\left\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\right\}}\right] \\
&\overset{(c)}{\leq} 2B^2\,\mathbb{E}\left[\|\boldsymbol{X}\boldsymbol{w}\|_{\ell_2}^2 \mathbb{1}_{\left\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\right\}}\right] + 2nB^2\mathbb{P}\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\} \\
&\overset{(d)}{\leq} 2B^2\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\boldsymbol{w}\|_{\ell_2}^4\right]\mathbb{P}\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\}} + 2nB^2\mathbb{P}\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\} \\
&\overset{(e)}{\leq} 2\sqrt{n}B^2\sqrt{\left(\sum_{i=1}^n \mathbb{E}\left[|\boldsymbol{x}_i^T\boldsymbol{w}|^4\right]\right)\mathbb{P}\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\}} + 2nB^2\mathbb{P}\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\} \\
&\overset{(f)}{\leq} 2\sqrt{3}nB^2\sqrt{\mathbb{P}\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\}} + 2nB^2\mathbb{P}\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\} \\
&\leq 6nB^2\sqrt{\mathbb{P}\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\}}. \qquad \text{(A.22)}
\end{aligned}$$

Here, (a) follows from Jensen's inequality, (b) from the simple identity $(a+b)^2 \leq 2(a^2+b^2)$, (c) from $|\phi'(z)| \leq B$, (d) from the Cauchy-Schwarz inequality, (e) from Jensen's inequality, and (f) from the fact that for a standard moment random variable X we have $\mathbb{E}[X^4] = 3$.

To continue we need to show that $\mathbb{P}\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\}$ is small. To this aim note that for any activation ϕ with $|\phi'(z)| \leq B$ we have

$$\|\phi(\boldsymbol{X}\boldsymbol{w}_2) - \phi(\boldsymbol{X}\boldsymbol{w}_1)\|_{\ell_2} \leq B\,\|\boldsymbol{X}\|\,\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_{\ell_2}$$

Thus by Lipschitz concentration of Gaussian functions for a random vector $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ we have

$$\begin{aligned}
\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} &\leq \mathbb{E}[\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2}] + t, \\
&\leq \sqrt{\mathbb{E}[\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2}^2]} + t, \\
&= \sqrt{n}\sqrt{\mathbb{E}_{g\sim\mathcal{N}(0,1)}[\phi^2(g)]} + t, \\
&\leq B\sqrt{2n} + t,
\end{aligned}$$

holds with probability at least $1 - e^{-\frac{t^2}{2B^2\|\boldsymbol{X}\|^2}}$. Thus using $t = \Delta B\sqrt{n}$ we conclude that

$$\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} \leq (\Delta + \sqrt{2})B\sqrt{n},$$

holds with probability at least $1 - e^{-\frac{\Delta^2}{2}\frac{n}{\|\boldsymbol{X}\|^2}}$. Thus using $\Delta = c\sqrt{\log n}$ and $T_n = CB\sqrt{n\log n}$ we can conclude that

$$\mathbb{P}\{\|\phi(\boldsymbol{X}\boldsymbol{w})\|_{\ell_2} > T_n\} \leq \frac{1}{n^{202}}.$$

Thus, using (A.21) we can conclude that

$$\left\|\mathbb{E}[\widetilde{\boldsymbol{A}}(\boldsymbol{w}) - \boldsymbol{A}(\boldsymbol{w})]\right\| \leq \frac{6B}{n^{100}}.$$

Combining this with (A.20) with $\delta = 1/2$ we conclude that

$$\lambda_{\min}\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T\right) \geq \frac{1}{2}k\left(\lambda_{\min}\left(\mathbb{E}[\boldsymbol{A}(\boldsymbol{w})]\right) - \frac{6B}{n^{100}}\right) = \frac{1}{2}k\left(\widetilde{\lambda}(\boldsymbol{X}) - \frac{6B}{n^{100}}\right),$$

holds with probability at least $1 - ne^{-\gamma\frac{k\widetilde{\lambda}(\boldsymbol{X})}{T_n^2}}$. The latter probability is larger than $1 - \frac{1}{n^{100}}$ as long as

$$k \geq C\log^2(n)\frac{n}{\widetilde{\lambda}(\boldsymbol{X})},$$

concluding the proof.

J. Utilizing Jacobian structure for even smaller networks

The results we have stated so far study the required over-parameterization to fit to any training data including those involving adversarial corruption or pure noise. On practical data sets however neural networks require significantly less number of parameters to perfectly fit to the training data. Intuitively for semantically meaningful data where the labels are related to the input data we expect it to be easier to perfectly fit to the training data compared to the case where we wish to fit to pure noise. In this section we discuss our results (based on the companion paper [53]) that can harness the low-rank representation of semantically meaningful datasets via the Jacobian of the neural net to approximately fit to the training data as soon as the network width is larger than a fixed numerical constant. We remark that [53] appeared after the initial submission of the present manuscript and in fact utilizes some of the techniques presented here. Our goal in

presenting these result here is to provide an alternative insight into gradient descent dynamics by contrasting the network width required to approximately fit the training data vs that of finding a perfect fit.

Global convergence over a restricted subspace: To guide the discussion, consider the eigenvalue decomposition of the neural network covariance matrix (Neural Tangent Kernel) $\boldsymbol{\Sigma}(\boldsymbol{X}) = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = \sum_{i=1}^n \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$. Let $\lambda_{\text{cut}} > 0$ be a scalar of our choice and suppose r is the index of eigenvalue satisfying $\lambda_r \geq \lambda_{\text{cut}} \geq \lambda_{r+1}$ so that the top r eigenvalues are larger than λ_{cut}. Define the top and bottom eigenspaces as

$$\mathcal{T} = \text{span}((\boldsymbol{u}_i)_{i=1}^r) \quad \text{and} \quad \mathcal{B} = \text{span}((\boldsymbol{u}_i)_{i=r+1}^n). \tag{A.23}$$

Also let Π_S be the projection operator on a subspace S. The result below is obtained as a corollary to Theorem 6.22 of [53]. Specifically, the result of [53] is stated for multiclass problems and below we state it for a network with single output. This shows that one can interpolate over the top subspace $\mathcal{T}$ and the network capacity depends only on λ_{cut} rather than $\lambda_{\min}$. However, this comes at the expense of not obtaining a perfect fit.

Theorem A.14: Let $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^n$ be a dataset where input samples have unit Euclidian norm and the concatenated label vector obeys $\|\boldsymbol{y}\|_{\ell_2} = \sqrt{n}$. Suppose $|\phi'|, |\phi''| \leq B$. Fix tolerance level ζ and output variance σ as $10B\sigma = \zeta \leq c/2$. Set output layer $\boldsymbol{v}$ with half $\sigma/\sqrt{k}$ and half $-\sigma/\sqrt{k}$ entries. Let $B^2\|\boldsymbol{X}\|^2 \geq \lambda_{\text{cut}} > 0$ be the spectrum cutoff level and set the top/bottom subspaces $\mathcal{T}$ and $\mathcal{B}$ as described in (A.23). Choose λ_{cut} to ensure $\|\Pi_{\mathcal{T}}(\boldsymbol{y})\|_{\ell_2} \geq c\|\boldsymbol{y}\|_{\ell_2}$ for some constant $c > 0$. Pick $\eta \leq \frac{1}{B^2\sigma^2\|\boldsymbol{X}\|^2}$. Set

$$\bar{\lambda}_{\text{cut}} = \frac{\lambda_{\text{cut}}}{B^2\|\boldsymbol{X}\|^{3/2} n^{1/4}} \quad \text{and} \quad k \gtrsim \frac{\Gamma^4 \log(n)}{\zeta^4 \bar{\lambda}_{\text{cut}}^4}. \tag{A.24}$$

Fix $\Gamma \geq 1$. With probability at least $1 - 2^{-50n/\|\boldsymbol{X}\|^2}$, after $T = \frac{\Gamma}{\eta\sigma^2\lambda_{\text{cut}}}$, iterations, we have that

$$\|f(\boldsymbol{W}_T) - \boldsymbol{y}\|_{\ell_2} \leq \|\Pi_{\mathcal{B}}(\boldsymbol{y})\|_{\ell_2} + \mathrm{e}^{-\Gamma}\|\Pi_{\mathcal{T}}(\boldsymbol{y})\|_{\ell_2} + 4\zeta\sqrt{n}. \tag{A.25}$$

This theorem achieves near-global convergence over the top subspace and the bias over the bottom subspace is not affected and stays around $\|\Pi_{\mathcal{B}}(\boldsymbol{y})\|_{\ell_2}$ in the worst case. Most notably, setting ζ, Γ to be constant and cutoff to be $\lambda_{\text{cut}} \sim \mathcal{O}(n)$ we observe that k grows logarithmic in

the size of the dataset which improves over our requirement for global convergence which is quadratic in sample size.

We note that the result above is not informative if the label vector lies on the bottom subspace $\mathcal{B}$. In contrast, Theorem 2.1 applies regardless of choice of labels i.e. it is guaranteed to work for any labels(noisy or random labels) regardless of any semantic link to the input data. We also remark that in (A.24) number of hidden nodes grow with the fourth power of $\bar{\lambda}_{\text{cut}}^{-1}$ which is a weaker dependence than quadratic growth obtained by Theorem 2.1. Closing this gap is an interesting direction for future research.