Review of "Gradient Descent Finds Global Minima of Deep Neural Networks"

David Kewei Lin
Stanford University
linkewei@stanford.edu

Jensen Jinhui Wang
Stanford University
wangjh97@stanford.edu

May 3, 2022

1 Introduction

In this paper, we will review Du et al. [2019a]: a generalization of the previous approach to deep neural networks [Du et al., 2019b]). Once again, the ultimate goal is to explain why deep neural networks can attain global training minima via gradient descent despite loss functions non-convex in their parameters. Even though the same analysis holds for more complex architectures such as ResNet and Convolutional ResNet, we will focus on the case of a feed-forward network with H layers. Given an input $\mathbf{x} = \mathbf{x}^{(0)} \in \mathbb{R}^d$, the output of the h-th layer is defined recursively as:

$$\mathbf{x}^{(h)} = \sqrt{\frac{c_\sigma}{m}}\sigma\left(\mathbf{W}^{(h)}\mathbf{x}^{(h-1)}\right), \quad 1 \le h \le H \tag{1}$$

where $c_\sigma = \left(\mathbb{E}_{x\sim N(0,1)}\left[\sigma(x)^2\right]\right)^{-1}$ is a normalization factor during initialization. Here, $\mathbf{W}^{(1)} \in \mathbb{R}^{m\times d}$, $\mathbf{x}^{(h)} \in \mathbb{R}^m$ for $h \in [H]$ and $\mathbf{W}^{(h)} \in \mathbb{R}^{m\times m}$ for $2 \le h \le H$. The final prediction is:

$$f(\mathbf{x}, \theta) = \mathbf{a}^\top \mathbf{x}^{(H)} \tag{2}$$

where $\mathbf{a} \in \mathbb{R}^m$. $\mathbf{a}$ and the columns of $\mathbf{W}^{(h)}$ are drawn i.i.d from $\mathcal{N}(0, I)$, for which the largeness of m will again ensure concentration and convergence.

This network is optimized by gradient descent with the ℓ_2-loss (freezing the final layer):

$$L(\theta) = \frac{1}{2}\sum_{i=1}^{n}\left(f\left(\mathbf{x}_i, \theta\right) - y_i\right)^2 \tag{3}$$

$$\mathbf{W}^{(h)}(k) = \mathbf{W}^{(h)}(k-1) - \eta\frac{\partial L(\theta(k-1))}{\partial \mathbf{W}^{(h)}(k-1)} \tag{4}$$

where $\eta > 0$ is the step size. We will use $\theta(k) = \{\mathbf{W}^{(h)}(k), \mathbf{a}(k)\}_{h\in[H]}$ to denote the parameters of the network at iteration k. Given this set-up and assumptions on $\mathbf{x}_i$ and σ to be specified,the following theorem states the network **converges to a global minimum at a linear rate**.

Theorem 1. *Suppose*

$$m = \Omega\left(2^{O(H)}\max\left\{\frac{n^4}{\lambda_0^4}, \frac{n}{\delta}, \frac{n^2\log\left(\frac{Hn}{\delta}\right)}{\lambda_0^2}\right\}\right) \tag{5}$$

where $\mathbf{K}^{(H)}$ is a population Gram matrix to be defined and $\lambda_0 \triangleq \lambda_{\min}\left(\mathbf{K}^{(H)}\right)$. If the step size $\eta = O\left(\frac{\lambda_0}{n^2 2^{O(H)}}\right)$ then with probability at least $1-\delta$ over the random initialization, the loss at each iteration $k \in \mathbb{N}$ satisfies

$$L(\theta(k)) \le \left(1 - \frac{\eta\lambda_{\min}\left(\mathbf{K}^{(H)}\right)}{2}\right)^k L(\theta(0)) \tag{6}$$

Remark. Note that $\lambda_0 = \lambda_{\min}\left(\mathbf{K}^{(H)}\right)$ hides the dependency on H which is not ideal. A priori, one would expect it to decay exponentially in H due to the vanishing gradients phenomenon.

Remark. Despite having a polynomial dependence on n, m is still much larger than realistic settings

where the total number of parameters $O(m^2H)$ is a constant multiple of the dataset size n.

Assumptions on inputs: The normalized inputs $\|\mathbf{x}_i\|_2 = 1$ for $1 \leq i \leq n$ are assumed to satisfy $\mathbf{x}_i \not\parallel \mathbf{x}_j$. Similar to the previous paper, this ensures the $\mathbf{K}^{(H)}$ has a minimum eigenvalue $\lambda_0 > 0$. To this end, by cleverly taking tensor products enough times, we can extend this pairwise independence to a linear independence of tensor products $\{\mathbf{x}_i^{\otimes(n-1)}\}_{i\in[n]}$.

Assumptions on activation: The activation function σ is assumed to satisfy

1. *(Lipschitz and smooth)* There exists a constant $c > 0$ such that $|\sigma(0)| \leq c$ and
$$|\sigma(z) - \sigma(z')| \leq c|z - z'| \quad \text{and} \quad |\sigma'(z) - \sigma'(z')| \leq c|z - z'| \quad \forall z, z' \in \mathbb{R} \tag{7}$$
Implicitly, we also get (by taking $z' \to z$) $|\sigma'(z)| \leq c$ for all $z \in \mathbb{R}$. The main use of this condition is the following technical but intuitive lemma.

   Lemma 2. *Let* $\mathbf{F}(\mathbf{A}) \triangleq \mathbb{E}_{\mathbf{U}\sim\mathcal{N}(0,\mathbf{A})}\left[\sigma(\mathbf{U})\sigma(\mathbf{U})^T\right]$. *There exists* $c_\sigma > 0$ *such that all positive definite* $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^m$,
$$\|\mathbf{F}(\mathbf{A}_1) - \mathbf{F}(\mathbf{A}_2)\|_{max} \leq c_\sigma \|\mathbf{A}_1 - \mathbf{A}_2\|_{max} \tag{8}$$

   Remark. The paper actually proved a weaker statement which required the diagonal values $\frac{1}{C} < (\mathbf{A}_1)_{ii}, (\mathbf{A}_2)_{ii} \leq C$ for a constant $C > 0$. This is undesirable since we will repeatedly use this lemma for different layers so we will again need a bootstrap argument to show that the same C applies to all layers. Our version bypasses this step by noting that Lemma G.3 in the paper can be improved to a universal constant by exploiting the boundedness of σ'.

   Later, $\mathbf{A}$ will represent the Gram matrix of inputs before the activation σ and $\mathbf{F}(\mathbf{A})$ will be the Gram matrix after σ. This inequality enables us to chain the Lipschitzness across layers.

2. *(Analyticity and non-polynomial)* σ is an analytic function that is not a polynomial.

   This main use of this assumption is to ensure the n-th derivative $\sigma^{(n)}(\cdot)$ is not trivially zero. This enables us to induct across layers to show that $\lambda_0 > 0$ via the following lemma:

   Lemma 3. *Consider data* $\mathbf{Z}^T = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$ *of* n *non-parallel points* $\{\mathbf{z}_i\}_{i\in[n]}$. *Define:*
$$\mathbf{G}(\mathbf{Z}) = \mathbb{E}_{\mathbf{w}\sim N(\mathbf{0},\mathbf{I})}\left[\sigma\left(\mathbf{Z}\mathbf{w}\right)\sigma\left(\mathbf{Z}\mathbf{w}\right)^\top\right] \tag{9}$$
$$\mathbf{H}(\mathbf{Z}) = \mathbb{E}_{\mathbf{w}\sim N(\mathbf{0},\mathbf{I})}\left[\sigma'\left(\mathbf{Z}\mathbf{w}\right)\sigma'\left(\mathbf{Z}\mathbf{w}\right)^\top \odot \mathbf{Z}\mathbf{Z}^\top\right] \tag{10}$$
   Then $\lambda_{\min}(\mathbf{G}(\mathbf{Z})), \lambda_{\min}(\mathbf{H}(\mathbf{Z})) > 0$.

   Proof. Firstly, $\mathbf{G}(\mathbf{Z})$ is evidently PSD. Now, suppose there exists $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{v}^\top\mathbf{G}(\mathbf{Z})\mathbf{v} = \mathbb{E}_{\mathbf{w}\sim N(\mathbf{0},\mathbf{I})}\left[\mathbf{v}^\top\sigma\left(\mathbf{Z}\mathbf{w}\right)\sigma\left(\mathbf{Z}\mathbf{w}\right)^\top\mathbf{v}\right] = 0$. Then, for all $\mathbf{w} \in \mathbb{R}^n$, we must have $\sigma(\mathbf{Z}\mathbf{w})\mathbf{v} = \sum_{i=1}^n v_i\sigma(\mathbf{z}_i^\top\mathbf{w}) = 0$. Taking the derivative with respect to $\mathbf{w}$ $(n-1)$-times, we obtain $\sum_{i=1}^n v_i\sigma^{(n-1)}(\mathbf{z}_i^\top\mathbf{w})\mathbf{z}_i^{\otimes(n-1)} = 0$ after which the linear independence of $\left\{\mathbf{z}_i^{\otimes(n-1)}\right\}$ implies $v_i\sigma^{(n-1)}(\mathbf{z}_i^\top\mathbf{w}) = 0$. Since $\sigma^{(n-1)}$ is analytic and non-zero, we can find $\mathbf{w}$ such that $\sigma^{(n-1)}(\mathbf{z}_i^\top\mathbf{w}) \neq 0$ for all $i \in [n]$, implying that $v_i = 0$ for all $i \in [n]$. Thus, $\mathbf{G}(\mathbf{Z})$ does not have a zero eigenvalue. The proof for $\lambda_{\min}(\mathbf{H}(\mathbf{Z}))$ is essentially the same. □

An example of σ that satisfies these assumptions is the softplus function $\sigma(z) = \log(1 + e^z)$. Henceforth, to simplify notation, we will use c_σ to denote constants dependent on the Lipschitz constant of σ and C for genuinely $O(1)$ constants, whose values may change from line to line.

2 General proof strategy

The overall proof strategy is the same as before. Defining the outputs $u_i = f(\mathbf{x}_i(k), \theta(k))$, we once again consider how $\|\mathbf{y} - \mathbf{u}(k+1)\|_2^2$ relates to $\|\mathbf{y} - \mathbf{u}(k)\|_2^2$. If we were still in a gradient flow set-up, we would have:
$$\frac{d}{dt}\left(\mathbf{u}(t) - \mathbf{y}\right) = -\mathbf{G}(t)\left(\mathbf{u}(t) - y\right) \tag{11}$$

where $\mathbf{G}(t) = \sum_{h=1}^{H} \mathbf{G}^{(h)}(t)$ is the sum of contributions from different layers and

$$\mathbf{G}^{(h)}(t) = \frac{c_\sigma}{m} \sum_{r=1}^{m} \frac{\partial \mathbf{u}}{\partial \mathbf{W}_r^{(h)}} \left(\frac{\partial \mathbf{u}}{\partial \mathbf{W}_r^{(h)}} \right)^\top \tag{12}$$

where $\frac{\partial \mathbf{u}}{\partial \mathbf{W}_r^{(h)}} \in \mathbb{R}^{n \times m}$ is the Jacobian with respect to the rth-row of $\mathbf{W}^{(h)}$. From this, one obtains:

$$\frac{d}{dt} \|\mathbf{y} - \mathbf{u}(t)\|_2^2 \leq -2\lambda_{\min}(\mathbf{G}(t)) \|\mathbf{y} - \mathbf{u}(t)\|_2^2 \leq -2\lambda_{\min}\left(\mathbf{G}^{(H)}(t)\right) \|\mathbf{y} - \mathbf{u}(t)\|_2^2 \tag{13}$$

Here, the last inequality stems from $\mathbf{G}^{(h)}(t)$ being manifestly PSD for all $h \in [H]$ so we can consider just $\mathbf{G}^{(H)}(t)$ at cost of the convergence rate. Returning to our case of discrete steps and recalling that the gradient flow is essentially a linear approximation, a Taylor expansion yields

$$\|y - \mathbf{u}(k+1)\|_2^2 \leq \left(1 - 2\eta\lambda_{\min}\left(\mathbf{G}^{(H)}(k)\right) + O(\eta^2)\right) \|y - \mathbf{u}(k)\|_2^2 \tag{14}$$

Thus, as long as $\eta = c_0 \lambda_{\min}\left(\mathbf{G}^{(H)}(k)\right)$ for all $k \in \mathbb{N}$ for a sufficiently small $c_0 > 0$, we will enjoy a linear convergence rate. The strategy is again to show that given large enough m, $\mathbf{G}^{(H)}(0) \approx \mathbf{K}^{(H)}$ for some fixed Gram matrix $\mathbf{K}^{(H)}$ initially such that $\lambda_{\min}\left(\mathbf{G}^{(H)}(0)\right) \geq \frac{3\lambda_0}{4}$. Then, we show that $\mathbf{G}^{(H)}(k)$ remains in a neighborhood around its initialization such that $\lambda_{\min}\left(\mathbf{G}^{(H)}(0)\right) \geq \frac{\lambda_0}{2}$. Since $\mathbf{G}^{(H)}$ depends on all layers, this is accomplished by proving $\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)$ is small for all $h \in [H]$. Finally, choosing small enough η such that the $O(\eta^2)$ terms is $\leq \frac{\eta\lambda_0}{2}$ yields Theorem 1.

The additional difficulty in this proof is the error propagation across layers in showing $\mathbf{G}^{(H)}(0) \approx \mathbf{K}^{(H)}$ and $\mathbf{W}^{(h)}(k) \approx \mathbf{W}^{(h)}(0)$ for all $h \in [H]$. Hence, we will focus on reviewing how this is accomplished by chaining Lipschitzness in matrix perturbations. This approach will be explicated for just the initialization stage which primarily determines how large m needs to be—the dynamical analysis is largely the same as the previous paper except one now has to consider all $h \in [H]$ simultaneously.

2.1 Definition and positive-definiteness of $\mathbf{K}^{(H)}$

In this section, we will motivate the rather convoluted form of $\mathbf{K}^{(H)}$ in the paper. Let $\mathbf{x}_i^{(h)}$ denote the output of the hth-layer of the network on example $\mathbf{x}_i$ and $\left(\mathbf{X}^{(h)}\right)^\top = \left(\mathbf{x}_1^{(h)}, \ldots, \mathbf{x}_n^{(h)}\right)$. Let's inductively assume that $\mathbf{X}^{(h)}\left(\mathbf{X}^{(h)}\right)^\top$ is fairly concentrated about its mean. At the $(h+1)$-th layer, the neurons compute $\mathbf{X}^{(h+1)} = \sqrt{\frac{c_\sigma}{m}}\sigma\left(\mathbf{U}^{(h)}\right)$ where $\mathbf{U}^{(h+1)} \in \mathbb{R}^{n \times m}$ is defined as:

$$\mathbf{U}^{(h+1)} = \mathbf{X}^{(h)}\left(\mathbf{W}^{(h+1)}\right)^\top \tag{15}$$

Noting that the columns of $\mathbf{U}^{(h+1)}$ have the same distribution, we obtain:

$$\mathbf{X}^{(h+1)}(\mathbf{X}^{(h+1)})^\top = \sum_{i=1}^{m} \mathbf{X}^{(h+1)} e_i e_i^\top \left(\mathbf{X}^{(h+1)}\right)^\top \tag{16}$$

$$= \frac{c_\sigma}{m} \sum_{i=1}^{m} \sigma\left(\mathbf{U}^{(h+1)} e_i\right) \sigma\left(\mathbf{U}^{(h+1)} e_i\right)^\top \tag{17}$$

$$\approx c_\sigma \mathbb{E}_{\mathbf{W}^{(h+1)}}\left[\sigma(\mathbf{U}^{(h+1)} e_1)\sigma(\mathbf{U}^{(h+1)} e_1)^\top\right] \tag{18}$$

$$= c_\sigma \mathbb{E}_{\mathbf{W}^{(h+1)}}\left[\sigma\left(\mathbf{X}^{(h)}(\mathbf{W}^{(h+1)})^\top e_1\right)\sigma\left(\mathbf{X}^{(h)}(\mathbf{W}^{(h+1)})^\top e_1\right)^\top\right] \tag{19}$$

$$= c_\sigma \mathbb{E}_{\mathbf{w} \sim \mathcal{N}\left(0, X^{(h)}(X^{(h)})^\top\right)}\left[\sigma(\mathbf{w})\sigma(\mathbf{w})^\top\right] \tag{20}$$

Hence, the concentration of $\mathbf{X}^{(h)}(\mathbf{X}^{(h)})^\top$ plus the Lipschitz property in Lemma (2) implies that $\mathbf{X}^{(h+1)}(\mathbf{X}^{(h+1)})^\top$ should also be concentrated around its mean, completing our "induction". This motivates us to define a sequence of population Gram matrices $\mathbf{K}^{(h)}$, to approximate $\mathbf{X}^{(h)}(\mathbf{X}^{(h)})^\top$ for $h \in [H-1]$, by the following recursive formula:

$$\mathbf{K}^{(0)} = \mathbf{X}^{(0)}(\mathbf{X}^{(0)})^\top \tag{21}$$

$$\mathbf{K}^{(h+1)} = c_\sigma \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, \mathbf{K}^{(h)})}[\sigma(\mathbf{w})\sigma(\mathbf{w})^\top], \qquad h = 0, 1, \cdots, (H-2) \tag{22}$$

A priori, this definition looks different from that of the paper (Definition 5.1), but they are in fact equivalent. Taking the (i,j)-th entry,

$$\mathbf{K}_{ij}^{(h+1)} = c_\sigma \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{K}^{(h)})}[e_i^\top \sigma(\mathbf{w})\sigma(\mathbf{w})^\top e_j] \tag{23}$$

$$= c_\sigma \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{K}^{(h)})}[\sigma(e_i^\top \mathbf{w})\sigma(e_j^\top \mathbf{w})] \tag{24}$$

where the pair $(e_i^\top \mathbf{w}, e_j^\top \mathbf{w})$ is jointly Gaussian with covariance $\mathbf{A}_{ij}^{(h)} = \begin{pmatrix} \mathbf{K}_{ii}^{(h)} & \mathbf{K}_{ij}^{(h)} \\ \mathbf{K}_{ji}^{(h)} & \mathbf{K}_{jj}^{(h)} \end{pmatrix}$. This is precisely the definition in Du et al. [2019a] for $\mathbf{K}^{(h+1)}$.

Given that $\mathbf{X}^{(H-1)}\left(\mathbf{X}^{(H-1)}\right)^\top$ concentrates around $\mathbf{K}^{(H-1)}$, to deduce the final form of $\mathbf{K}^{(H)}$, we compute via the chain rule

$$\begin{aligned}
\mathbf{G}^{(H)} &= \left(\mathbf{X}^{(H-1)}\right)\left(\mathbf{X}^{(H-1)}\right)^T \odot \frac{c_\sigma}{m}\sum_{i=1}^m a_i^2 \sigma'\left(\mathbf{X}^{(H-1)}\left(\mathbf{W}^{(H)}\right)^\top e_i\right)\sigma'\left(\mathbf{X}^{(H-1)}\left(\mathbf{W}^{(H)}\right)^\top e_i\right)^\top \\
&\approx \mathbf{K}^{(H-1)} \odot c_\sigma \mathbb{E}_{a\sim\mathcal{N}(0,1)}[a^2]\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{X}^{(H-1)}\left(\mathbf{X}^{(H-1)}\right)^\top}\left[\sigma'(\mathbf{w})\sigma'(\mathbf{w})^\top\right] \\
&\approx c_\sigma \mathbf{K}^{(H-1)} \odot \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{K}^{(H-1)})}\left[\sigma'(\mathbf{w})\sigma'(\mathbf{w})^\top\right]
\end{aligned}$$

Hence, we obtain

$$\mathbf{K}^{(H)} = c_\sigma \mathbf{K}^{(H-1)} \odot \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{K}^{(H-1)})}\left[\sigma'(\mathbf{w})\sigma'(\mathbf{w})^\top\right] \tag{25}$$

2.2 Positive-definiteness of $\mathbf{K}^{(H)}$

To prove $\lambda_0 = \lambda_{\min}\left(K^{(H)}\right) > 0$, observe that our non-parallel assumption on the data $\mathbf{X}^{(0)}$ allows us to use Lemma 3 to conclude that $\mathbf{K}^{(1)} = \mathbf{G}\left(\mathbf{X}^{(0)}\right)$ is positive. Then, applying Lemma 3 to $\mathbf{K}^{(h)} = \mathbf{G}\left(\left(\mathbf{K}^{(h-1)}\right)^{\frac{1}{2}}\right)$ for $h \in [H-1]$ and $\mathbf{K}^{(H)} = \mathbf{H}\left(\left(\mathbf{K}^{(H-1)}\right)^{\frac{1}{2}}\right)$ completes the proof (where the square-root for the PD matrix $\mathbf{K}^{(h)}$ exists and has no parallel columns since it has full rank).

3 Empirical gram matrix is close to population gram matrix

For a pedagogical analysis of the error propagation across layers, we prove that $\mathbf{G}^{(H)}(0)$ is close to $\mathbf{K}^{(H)}$ initially in this section. To this end, we define empirical Gram matrices for $h \in [H-1]$ as $\hat{\mathbf{K}}^{(h)} \triangleq \mathbf{X}^{(h)}\left(\mathbf{X}^{(h)}\right)^\top$. Then, we obtain the following concentration bound.

Theorem 4. *With probability $1-\delta$ over $\{\mathbf{W}^{(h)}\}_{h\in[h-1]}$, for any $1 \le h \le H-1$,*

$$\left\|\hat{\mathbf{K}}^{(h)} - \mathbf{K}^{(h)}\right\|_{max} \le 2^{O(H)}\sqrt{\frac{\log\frac{Hn}{\delta}}{m}} \tag{26}$$

Proof. Letting $\mathbb{E}^{(h)}$ denote the expectation of the hth layer conditioned on the previous $h-1$ layers, by a standard concentration inequality, for any $h \in [H-1]$ that with error probability $\frac{\delta}{H}$:

$$\left\|\hat{\mathbf{K}}^{(h)} - \mathbb{E}^{(h)}\hat{\mathbf{K}}^{(h)}\right\|_{max} \le C\sqrt{\frac{\log\frac{Hn}{\delta}}{m}} \tag{27}$$

Hence, to prove Eq. (26), it remains to bound

$$\begin{aligned}
\left\|\mathbf{K}^{(h)} - \mathbb{E}^{(h)}\hat{\mathbf{K}}^{(h)}\right\|_{max} &= c_\sigma\left\|\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{K}^{(h-1)})}\left[\sigma(\mathbf{w})\sigma(\mathbf{w})^T\right] - \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\hat{\mathbf{K}}^{(h-1)})}\left[\sigma(\mathbf{w})\sigma(\mathbf{w})^T\right]\right\|_{max} \\
&\le c_\sigma\left\|\mathbf{K}^{(h-1)} - \hat{\mathbf{K}}^{(h-1)}\right\|_{max} \qquad \text{(Lemma 2)}
\end{aligned}$$

Then, by the triangle inequality,

$$\left\|\mathbf{K}^{(h)} - \hat{\mathbf{K}}^{(h)}\right\|_{max} \le c_\sigma\left\|\mathbf{K}^{(h-1)} - \hat{\mathbf{K}}^{(h-1)}\right\|_{max} + C\sqrt{\frac{\log\frac{Hn}{\delta}}{m}} \tag{28}$$

Solving this recursive relation while accruing an additional $\frac{\delta}{H}$ error probability in each layer yields Eq. (26) with probability $\ge 1-\delta$. □

Remark. Since the recursive definition of $K^{(H)}$ differs from $K^{(h)}$ for $h \in [H-1]$, our theorem only holds until $K^{(H-1)}$. However, it turns out that if $\left\|\mathbf{K}^{(H-1)} - \mathbb{E}^{(H-1)}\hat{\mathbf{K}}^{(H-1)}\right\|_{max} \leq \frac{C\lambda_0}{n}$ and $m = \Omega\left(\frac{n^2\log(n/\delta)}{\lambda_0^2}\right)$, we can use another concentration inequality plus an application of Lemma 2 to σ' to show $\left\|\mathbf{G}^{(H)}(0) - \mathbf{K}^{(H)}\right\|_{op} \leq \frac{1}{4}$. To satisfy the former, our bound in Theorem 4 requires $m = \Omega\left(2^{O(H)}\frac{n^2\log\left(\frac{Hn}{\delta}\right)}{\lambda_0^2}\right)$ which is unfortunately plagued by an exponential dependence on H.

3.1 Efficiency of residual connections

A secondary result of paper is that for the ResNet architecture, the dependence of m on H is reduced to poly(H). The output of the h-th layer of ResNet is defined as

$$\mathbf{x}^{(h)} = \mathbf{x}^{(h-1)} + \frac{c_{res}}{H\sqrt{m}}\sigma(\mathbf{W}^{(h)}\mathbf{x}^{(h-1)}), \qquad 1 \leq h \leq H$$

with the final prediction similar to that of the fully-connected network. We may also define the Gram matrix $\mathbf{K}^{(h)}$ in a similar fashion to the fully-connected case. However, the analogous form of Eq. (28) is now

$$\left\|\mathbf{K}^{(h)} - \hat{\mathbf{K}}^{(h)}\right\|_{max} \leq \left(1 + \frac{c_\sigma}{H}\right)\left\|\mathbf{K}^{(h-1)} - \hat{\mathbf{K}}^{(h-1)}\right\|_{max} + C\sqrt{\frac{\log\frac{Hn}{\delta}}{m}} \tag{29}$$

where the 1 in $1 + \frac{c_\sigma}{H}$ originates from the residual connection and the $\frac{1}{H}$ factor stems from the rescaling. Since $\left(1 + \frac{c_\sigma}{H}\right)^H = O(1)$, m only needs to be $\Omega\left(\frac{n^2\log\left(\frac{Hn}{\delta}\right)}{\lambda_0^2}\right)$ during initialization. After accounting for the dynamics which require $G^{(H)}(0)$ to remain close to $G^{(H)}(k)$, the final lower bound for m is $\Omega\left(\frac{\text{poly}(n,1/H,1/\lambda_0)}{\delta}\right)$. Remarkably, for the ResNet architecture, one can also establish that $\lambda_0 = \Omega(\text{poly}(1/H))$ as opposed to the previous implicit dependence on H. Thus, the number of neurons required for ResNet to converge is truly polynomial in H.

4 Limitations and Extensions

1. In the case of fully-connected networks, λ_0 depends implicitly on H and is likely to be exponentially decaying. Ultimately, the $m = \Omega(2^{O(H)})$ bound seems too loose as compared to networks in practice. Now, one might think that adding a $\frac{1}{H}$ scaling to the fully-connected network would yield a smaller bound $m = O\left(\left(\frac{c_H}{H}\right)^H\right)$ but notice that for large H, the output of the network is essentially zero so the network is trivial. This isn't the case for ResNet due to the residual connections which are independent of H.

2. The improved bound on the number of nodes for ResNet can be entirely attributed to the rescaling of the update term (which depends on the number of layers). The authors' justification was that this rescaling is equivalent to changing the initialization, which isn't necessarily the case since it affects the dynamics too.

3. This paper doesn't develop a novel analysis of error propagation or Gram matrices and simply reapplies the same recursion and union bounds to complex architectures. As such, the proof is proportionately more involved. It may be desirable to have a more abstract, unifying perspective to expedite the analysis, especially for more complicated set-ups.

4. Once again, the network only demonstrates a trivial form of implicit regularization by converging in a local neighborhood of its initialization. In a similar vein, the paper only analyzes the training error and does not consider the network's generalizability.

5. **Potential extension**: In Eq. (13), we bounded $\lambda_{\min}(\mathbf{G}(t))$ by $\lambda_{\min}(\mathbf{G}^{(H)}(t))$, but we could have used any other index h other than H. This may be useful since a smaller h suffers from a smaller error propagation during initialization, suggesting a smaller bound for m. However, it is likely that h still needs to be sufficiently large for the dynamics to remain in a neighborhood of initialization (since the later layers directly affect the output). By analyzing this trade-off, one can determine a suitable index h. Another possibility is to consider all possible $h \in [H]$ though this might further complicate the already convoluted proof.

References

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks, 2019a.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks, 2019b.