Tight Upper Bound for Expectation of Lipschitz constant of Deep Random Neural Networks

Yash Jakhmola

In this article, we present a $\mathcal{O}(\sqrt{d})$ upper bound for lipschitz constant of deep random neural networks. This bound is tight (upto constants, and for large N, d) due to the lower bound shown in theorem 3.4, [1].

1 Setup

Consider a deep random neural network $\Phi : \mathbb{R}^d \to \mathbb{R}$ with $N \in \mathbb{N}$ hidden neurons in $L \in \mathbb{N}$ layers and ReLU activation ReLU:

$$\Phi(x) := \left(V^{(L)} \circ \mathrm{ReLU} \circ V^{(L-1)} \circ \cdots \circ \mathrm{ReLU} \circ V^{(0)}\right)(x)$$

for all $x \in \mathbb{R}^d$, where $V^{(l)}(x) := W^{(l)}x + b^{(l)}$ for all $l \in \{0, \ldots, L\}$, where $W^{(0)} \in \mathbb{R}^{N\times d}$, $W^{(l)} \in \mathbb{R}^{N\times N}$ for $l \in \{1, \ldots, L-1\}$, $W^{(L)} \in \mathbb{R}^{1\times N}$, $b^{(l)} \in \mathbb{R}^N$ for $l \in \{0, \ldots, L-1\}$, $b^{(L)} \in \mathbb{R}$ and

$$W_{ij}^{(l)} \overset{\text{iid}}{\sim} N\left(0, \frac{2}{N}\right)$$

$$W_j^{(L)} \overset{\text{iid}}{\sim} N(0, 1)$$

for all $l \in \{0, \cdots, L-1\}$ and the biases are distributed independently of the weights, such that their distribution is symmetric about zero.

Theorem 1.1 (Lipschitz constant of ReLU networks). *For a deep ReLU network Φ as defined above,*

$$Lip(\Phi) \leq \sup_{x\in\mathbb{R}^d} \|W^{(L)}D^{(L-1)}(x)W^{(L-1)} \cdots D^{(0)}(x)W^{(0)}\|_2 \tag{1}$$

where $D^{(l)}(x) := diag(\mathbb{I}(W^{(l)}x^{(l)} + b^{(l)} > 0)_i)$, $x^{(l+1)} := D^{(l)}(x)(W^{(l)}x^{(l)} + b^{(l)})$ and $x^{(0)} := x$.

Proof. See equation (4.2), [1]. □

2 The Proof

We make use of the fact that the weights matrices will be isometric in (square of) expectation of frobenius norm and that the variant of He-initialization used makes each layer isometric in expectation. We then use tower law for expectations in an induction argument to show that expectation of the squared norm in RHS of equation (1) is d. We end the proof by using Jensen's inequality to get the final bound.

Lemma 2.1. *Let $W \in \mathbb{R}^{N\times N}$ be a matrix such that all its entries are drawn iid from $N\big(0, \frac{2}{N}\big)$. Then, for any deterministic $X \in \mathbb{R}^{N\times d}$, $\mathbb{E}\,\|WX\|_F^2 = 2\|X\|_F^2$.*

Proof. We know $\|WX\|_F^2 = \sum_{i=1}^N \sum_{j=1}^d \langle W_{i,-}, X_{-,j}\rangle^2 \implies \mathbb{E}\,\|WX\|_F^2 = \sum_{i=1}^N \sum_{j=1}^d \mathbb{E}\left(\langle W_{i,-}, X_{-,j}\rangle^2\right)$. Since $\langle W_{i,-}, X_{-,j}\rangle$ are zero mean gaussians,

$$\mathbb{E}\left(\langle W_{i,-}, X_{-,j}\rangle^2\right) = \mathrm{Var}\left(\langle W_{i,-}, X_{-,j}\rangle\right) = \mathrm{Var}\left(\sum_{k=1}^N W_{ik}X_{kj}\right) = \frac{2}{N}\|X_{-,j}\|^2$$

Thus, $\mathbb{E}\,\|WX\|_F^2 = \sum_{i=1}^N \sum_{j=1}^d \frac{2}{N}\|X_{-,j}\|^2 = 2\|X\|_F^2$. □

Remark. Using a similar proof, one can show that if $W \in \mathbb{R}^{1\times N}$ is such that its entries are iid $N(0,1)$, then for any deterministic $X \in \mathbb{R}^{N\times d}$, $\mathbb{E}\,\|WX\|_2^2 = \|X\|_F^2$.

Fix $x \in \mathbb{R}^d$ and drop it for simpler notation. Define $A_0 := D^{(0)}W^{(0)}$ and $A_k := D^{(k)}W^{(k)}A_{k-1}$ for all $k \in \{1,\dots,L-1\}$.

Lemma 2.2. *For all $k \in \{1,\dots,L\}$, $\mathbb{E}\left[\|A_k\|_F^2 \mid A_{k-1}\right] = \frac{1}{2}\,\mathbb{E}\left[\|W^{(k)}A_{k-1}\|_F^2 \mid A_{k-1}\right]$.*

Proof. Let $\tilde{W} := W^{(k)}A_{k-1}$. Then,

$$\mathbb{E}\left[\|A_k\|_F^2 \mid A_{k-1}\right] = \sum_{i=1}^N \sum_{j=1}^d \mathbb{E}\left[D_{ii}^{(k)}\tilde{W}_{ij}^{(0)^2} \mid A_{k-1}\right]$$

Conditioned on A_{k-1}, entries of $\tilde{W}$ are all normally distributed. Now, let $z_i := \sum_{p=1}^d W_{ip}^{(k)}x_p^{(k)}$. Thus, conditioned on A_{k-1}, $(z_i, \tilde{W}_{ij}^{(k)})$ is jointly normally distributed with zero mean.

Thus, we can write $\tilde{W}_{ij}^{(k)} = \rho z_i + \varepsilon$, where ρ is a constant and ε is centered gaussian random variable, independent of z_i. Now,

$$\begin{aligned}
\mathbb{E}\left[D_{ii}^{(k)}\tilde{W}_{ij}^{(0)^2} \mid A_{k-1}\right] &= \mathbb{E}\left[\mathbb{I}(z_i + b_i^{(k)} > 0)(\rho z_i + \varepsilon)^2 \mid A_{k-1}\right] \\
&= \mathbb{E}\left[\mathbb{I}(z_i + b_i^{(k)} > 0)(\rho^2 z_i^2 + \varepsilon^2) \mid A_{k-1}\right] \\
&= \rho^2\,\mathbb{E}\left[\mathbb{I}(z_i + b_i^{(k)} > 0)z_i^2 \mid A_{k-1}\right] + \mathbb{E}\left[\mathbb{I}(z_i + b_i^{(k)} > 0)\varepsilon^2 \mid A_{k-1}\right] \\
&= \frac{1}{2}\rho^2\,\mathbb{E}\left[z_i^2 \mid A_{k-1}\right] + \frac{1}{2}\,\mathbb{E}\left[\varepsilon^2 \mid A_{k-1}\right] \\
&= \frac{1}{2}\,\mathbb{E}\left[\rho^2 z_i^2 + \varepsilon^2 \mid A_{k-1}\right] \\
&= \frac{1}{2}\,\mathbb{E}\left[(\rho z_i + \varepsilon)^2 \mid A_{k-1}\right] \\
&= \frac{1}{2}\,\mathbb{E}\left[\tilde{W}_{ij}^{(k)^2} \mid A_{k-1}\right]
\end{aligned}$$

where $1/2$ occurs since expectation of z_i^2 over $z_i + b_i^{(k)} > 0$ is half of expectation of z_i^2, because expectation of z_i^2 given either $z_i + b_i^{(k)} > 0$ or $z_i + b_i^{(k)} < 0$ is the same, and the probability that $z_i + b_i^{(k)} > 0$ is $1/2$. Summing over all i, j, we get the needed equality. □

Remark. Similar proof as above can be used to show that $\mathbb{E}\left[\|A_0\|_F^2\right] = \frac{1}{2}\,\mathbb{E}\left[\|W^{(0)}\|_F^2\right]$.

Theorem 2.3. *For a deep ReLU network Φ as defined above,*

$$\mathbb{E}\,\|W^{(L)}D^{(L-1)}(x)W^{(L-1)}\cdots D^{(0)}(x)W^{(0)}\|_2^2 = d$$

for all $x \in \mathbb{R}^d$.

Proof. Using the remark above,

$$\mathbb{E}\,\|A_0\|_F^2 = \mathbb{E}\,\|D^{(0)}W^{(0)}\|_F^2 = \frac{1}{2}\,\mathbb{E}\,\|W^{(0)}\|_F^2 = \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^d \mathbb{E}\left(\left(W_{ij}^{(0)}\right)^2\right) = \frac{1}{2}Nd\left(\frac{2}{N}\right) = d$$

Using lemma 2.2, lemma 2.1 and assuming $\mathbb{E}\,\|A_{k-1}\|_F^2 = d$, we get

$$\begin{aligned}
\underset{A_k}{\mathbb{E}}\,\|A_k\|_F^2 &= \underset{D^{(k)},W^{(k)},A_{k-1}}{\mathbb{E}}\,\|D^{(k)}W^{(k)}A_{k-1}\|_F^2 \\
&= \underset{A_{k-1}}{\mathbb{E}}\left[\underset{D^{(k)},W_k}{\mathbb{E}}\left[\|D^{(k)}W^{(k)}A_{k-1}\|_F^2 \mid A_{k-1}\right]\right] \\
&= \frac{1}{2}\underset{A_{k-1}}{\mathbb{E}}\left[\underset{W_k}{\mathbb{E}}\left[\|W^{(k)}A_{k-1}\|_F^2 \mid A_{k-1}\right]\right] \\
&= \underset{A_{k-1}}{\mathbb{E}}\left[\|A_{k-1}\|_F^2\right] \\
&= d
\end{aligned}$$

Finally, to extend this to $W^{(L)}$, use tower law yet again.

$$\begin{aligned}
&\mathbb{E}\,\|W^{(L)}D^{(L-1)}W^{(L-1)}\cdots D^{(0)}W^{(0)}\|_2^2 \\
&\quad= \mathbb{E}\,\|W^{(L)}A_{L-1}\|_2^2 \\
&\quad= \underset{A_{L-1}}{\mathbb{E}}\left[\underset{W^{(L)}}{\mathbb{E}}\left[\|W^{(L)}A_{L-1}\|_2^2 \mid A_{L-1}\right]\right] \\
&\quad= \underset{A_{L-1}}{\mathbb{E}}\left[\|A_{L-1}\|_F^2\right] \\
&\quad= d
\end{aligned}$$

□

Theorem 2.4. *For a deep ReLU network Φ as defined above,*

$$\mathit{Lip}(\Phi) \leq \sqrt{d}$$

Proof. Using Jensen's inequality, we get

$$\mathbb{E}\,\|W^{(L)}D^{(L-1)}W^{(L-1)}\cdots D^{(0)}W^{(0)}\|_2 \leq \left(\mathbb{E}\,\|W^{(L)}D^{(L-1)}W^{(L-1)}\cdots D^{(0)}W^{(0)}\|_2^2\right)^{1/2} = \sqrt{d}$$

Since this holds for all $x \in \mathbb{R}^d$, we are done. □

References

[1] Paul Geuchen, Dominik Stöger, Thomas Telaar, and Felix Voigtlaender. “Upper and Lower Bounds for the Lipschitz constant of random neural networks”. In: (2025).