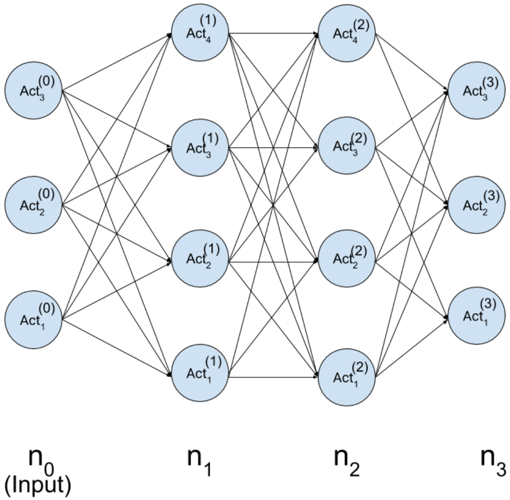
Gradients of Neural Nets and Products of Random Matrices

Boris Hanin* and Mihai Nica**



*Texas A&M University and **University of Toronto

Abstract

We prove that, on initialization, the **norm of the gradient** of a fully connected neural net with d layers of widths $n_0, n_1, \ldots n_d$, is approximately **log-normal** with variance:

$$\beta = \sum_{i=1}^{d} \frac{5}{n_i}$$

The result holds when the network is initialization with a large class of **symmetric random weights** and a **ReLu non-linearity**. This provides an explanation why very deep networks can suffer from the vanishing and exploding gradient problem. The proof goes by a connection to certain products of random matrices.

Definition: Fully Connected Neural Nets with ReLU

- Fix **depth** $d \in \mathbb{N}$ and **layer widths** $\{n_0, n_1, \ldots, n_d\} \in \mathbb{N}^d$.
- For each $1 \leq i \leq d$:
  - Let $W^{(i)} \in \mathbb{R}^{n_{i-1} \times n_i}$ be a **matrix of random weights** with
    - **iid entries**, $\mathbf{E}\left[W^{(i)}_{a,b}\right] = 0$, $\mathbf{Var}\left[W^{(i)}_{a,b}\right] = 1$, finite higher moments.
  - Let $B^{(i)} \in \mathbb{R}^{n_i}$ be a **random bias vector** with
    - **iid entries**, $\mathbf{E}\left[B^{(i)}_a\right] = 0$, $\mathbf{Var}\left[B^{(i)}_a\right] = 1$, finite higher moments.
  - Assume also the bias and weights are **symmetric** and have no atoms
- Let $Act^{(0)} \in \mathbb{R}^{n_0}$ be some fixed **input vector**.
- For each $1 \leq i \leq d$, recursively define the **activations** $Act^{(i)} \in \mathbb{R}^{n_i}$ by:

$$Act^{(i)} := \sigma_{ReLU}\left(\sqrt{\frac{2}{n_i}} W^{(i)} Act^{(i-1)} + \sqrt{\frac{2}{n_i}} B^{(i)}\right)$$

  - Note $\sigma_{ReLU}(x) = x 1_{\{x \geq 0\}}$. It is applied **entry-wise**

Definition: Gradient

- The Jacobian matrix $J^{(i)} \in \mathbb{R}^{n_i \times n_0}$ is the gradient of the i-th layer with respect to the input:

$$J^{(i)}_{a,b} := \frac{\mathrm{d}Act^{(i)}_a}{\mathrm{d}Act^{(0)}_b}$$

- It can be recursivly defined by:

$$J^{(i)} = Diag\left(\sigma'_{ReLU}\left(W^{(i)}Act^{(i-1)} + B^{(i)}\right)\right)\left(\sqrt{\frac{2}{n_i}}W^{(i)}\right)J^{(i-1)}$$

  - Note $\sigma'_{ReLU}(x) = 1_{\{x \geq 0\}}$.It is applied **entry-wise**

Connection To Products of Random Matrices

Proposition

Let $p \in (0,1)$. Let $D_p^{(i)} \in \mathbb{R}^{n_i \times n_i}$ be the diagonal random matrix whose entries are iid $\{0,1\}$ valued Bernoulli(p) random variables:

$$D_p^{(i)} = Diag\left(\xi_1^{(i)}, \ldots, \xi_{n_i}^{(i)}\right)$$

If the weights $W^{(i)}$ and biases $B^{(i)}$ are symmetricaly distributed, then, for any vector $\vec{u} \in \mathbb{R}^{n_0}$ we have:

$$\left\| J^{(i)} \vec{u} \right\|^2 \stackrel{d}{=} \left\| \left(\prod_{j=1}^{i} D_{\frac{1}{2}}^{(i)} \sqrt{\frac{2}{n_i}} W^{(i)} \right) \vec{u} \right\|^2$$

Limit Theorem - Moments

Theorem

For any vector $\vec{u} \in \mathbb{R}^{n_0}$ $\|\vec{u}\|^2 = 1$, *the moments of the* $\left\|J^{(i)}\vec{u}\right\|^2$ *are approximatly log-normal:*

$$\mathbf{E}\left[\left\|J^{(i)}\vec{u}\right\|^{2k}\right] = \mathbf{E}\left[\exp\left\{\sqrt{\beta}G - \frac{1}{2}\beta\right\}^k\right](1+\epsilon)^k$$

where $G \sim \mathcal{N}(0,1)$ *is a standard Gaussian and:*

$$\beta := \sum_{i=1}^{d} \frac{5}{n_i}$$

$$\epsilon = O\left(\frac{\sum_{i=1}^{d} n_i^{-2}}{\left(\sum_{i=1}^{d} n_i^{-1}\right)^2}\right)$$

Limit Theorem: Kolmogorov-Smirnov Distance

Theorem

For any vector $\vec{u} \in \mathbb{R}^{n_0}$ $\|\vec{u}\|^2 = 1$*, the distribution of* $\ln \left\| J^{(i)} \vec{u} \right\|^2$ *is approximatly normal:*

$$\sup_{t \in \mathbb{R}} \left| \mathbf{P}\left(\frac{\ln\left(\left\| J^{(i)} \vec{u} \right\|^2 \right) - \frac{1}{2}\beta}{\sqrt{\beta}} \leq t \right) - \mathbf{P}\left(G \leq t\right) \right| \leq \epsilon^{1/5}$$

where $G \sim \mathcal{N}(0,1)$ *is a standard Gaussian and:*

$$\beta := \sum_{i=1}^{d} \frac{5}{n_i}$$

$$\epsilon = O\left(\frac{\sum_{i=1}^{d} n_i^{-2}}{\left(\sum_{i=1}^{d} n_i^{-1} \right)^2} \right)$$

Random Matrix Limits

Theorem

Let $(p_1, \dots, p_d) \in (0,1)^d$. The same log-normal results hold for the norm of the random matrix product:

$$\left\| \left(\prod_{j=1}^{i} D_{p_i}^{(i)} \sqrt{\frac{1}{p_i n_i}} W^{(i)} \right) \vec{u} \right\|^2$$

The parameter β is now:

$$\beta := \sum_{i=1}^{d} \left(\frac{3}{p_i} - 1 \right) n_i^{-1}$$

Proof Ideas

Path counting along the net to get moments. Martingale CLT on the moment formulas to bound the KS distance.