# Which Neural Net Architectures Give Rise to Exploding and Vanishing Gradients?

Boris Hanin

Texas A&M

Feb 6, 2018

# Problem Statement

# Problem Statement

- Fix $d \geq 1$ and $\mathbf{n} = (n_j)_{j=0}^{d}$ .
- $\mathfrak{N}(d, \mathbf{n})$ – depth $d$ ReLU nets with hidden layer widths $n_j$.
- $f_{\mathcal{N}}$ – function computed by $\mathcal{N} \in \mathfrak{N}(d, \mathbf{n})$

# Problem Statement

- Fix $d \geq 1$ and $\mathbf{n} = (n_j)_{j=0}^{d}$ .
- $\mathfrak{N}(d, \mathbf{n})$ − depth $d$ ReLU nets with hidden layer widths $n_j$.
- $f_{\mathcal{N}}$ − function computed by $\mathcal{N} \in \mathfrak{N}(d, \mathbf{n})$
- **Q.** How do $d, \mathbf{n}$ influence $Z = \|\nabla f_{\mathcal{N}}\|^2$ when weights and biases are random (i.e. at initialization)?

# Problem Statement

- Fix $d \geq 1$ and $\mathbf{n} = (n_j)_{j=0}^{d}$.
- $\mathfrak{N}(d, \mathbf{n})$ – depth $d$ ReLU nets with hidden layer widths $n_j$.
- $f_{\mathcal{N}}$ – function computed by $\mathcal{N} \in \mathfrak{N}(d, \mathbf{n})$
- **Q.** How do $d, \mathbf{n}$ influence $Z = \|\nabla f_{\mathcal{N}}\|^2$ when weights and biases are random (i.e. at initialization)?
- **A.** $\mathbb{E}\left[Z^K\right] = \exp\left(\Theta_K\left(\sum_j \frac{1}{n_j}\right)\right)$

# Motivation - Exploding and Vanishing Gradients

# Motivation - Exploding and Vanishing Gradients

- SGD fails if $f_{\mathcal{N}}$ has wild gradients: $\left|\partial f_{\mathcal{N}} / \partial w_{\alpha,\beta}^{(j)}\right| \in \{0, \infty\}$,

## Motivation - Exploding and Vanishing Gradients

- SGD fails if $f_{\mathcal{N}}$ has wild gradients: $\left|\partial f_{\mathcal{N}} / \partial w_{\alpha,\beta}^{(j)}\right| \in \{0, \infty\}$,
- For neural nets,

$$\frac{\partial f_{\mathcal{N}}}{\partial w_{\alpha,\beta}^{(j)}} = \frac{\partial f_{\mathcal{N}}}{\partial \,\mathrm{Act}_{\beta}^{(j)}} \; \frac{\partial \,\mathrm{Act}_{\beta}^{(j)}}{\partial w_{\alpha,\beta}^{(j)}}$$

- SGD fails if $f_{\mathcal{N}}$ has wild gradients: $\left|\partial f_{\mathcal{N}} \,/\partial w_{\alpha,\beta}^{(j)}\right| \in \{0, \infty\}$,
- For neural nets,

$$\frac{\partial f_{\mathcal{N}}}{\partial w_{\alpha,\beta}^{(j)}} = \frac{\partial f_{\mathcal{N}}}{\partial \,\mathrm{Act}_{\beta}^{(j)}} \; \frac{\partial \,\mathrm{Act}_{\beta}^{(j)}}{\partial w_{\alpha,\beta}^{(j)}}$$

- $f_{\mathcal{N}}(\mathrm{Act}^{(j)}) = (f_d \circ \cdots \circ f_{j+1})\,(\mathrm{Act}^{(j)})$

# Motivation - Exploding and Vanishing Gradients

- SGD fails if $f_{\mathcal{N}}$ has wild gradients: $\left|\partial f_{\mathcal{N}} / \partial w_{\alpha,\beta}^{(j)}\right| \in \{0, \infty\}$,
- For neural nets,

$$\frac{\partial f_{\mathcal{N}}}{\partial w_{\alpha,\beta}^{(j)}} = \frac{\partial f_{\mathcal{N}}}{\partial \operatorname{Act}_{\beta}^{(j)}} \frac{\partial \operatorname{Act}_{\beta}^{(j)}}{\partial w_{\alpha,\beta}^{(j)}}$$

- $f_{\mathcal{N}}(\operatorname{Act}^{(j)}) = (f_d \circ \cdots \circ f_{j+1})(\operatorname{Act}^{(j)})$
- Exploding and Vanishing gradients problem comes down to

$$\left|\frac{\partial f_{\mathcal{N}}}{\partial \operatorname{Act}_{\beta}^{(j)}}\right| \in \{0, \infty\}$$

# Motivation - Exploding and Vanishing Gradients

- SGD fails if $f_{\mathcal{N}}$ has wild gradients: $\left|\partial f_{\mathcal{N}} \,/ \partial w_{\alpha,\beta}^{(j)}\right| \in \{0, \infty\}$,
- For neural nets,

$$\frac{\partial f_{\mathcal{N}}}{\partial w_{\alpha,\beta}^{(j)}} = \frac{\partial f_{\mathcal{N}}}{\partial \operatorname{Act}_{\beta}^{(j)}} \frac{\partial \operatorname{Act}_{\beta}^{(j)}}{\partial w_{\alpha,\beta}^{(j)}}$$

- $f_{\mathcal{N}}(\operatorname{Act}^{(j)}) = (f_d \circ \cdots \circ f_{j+1})\,(\operatorname{Act}^{(j)})$
- Exploding and Vanishing gradients problem comes down to

$$\left|\frac{\partial f_{\mathcal{N}}}{\partial \operatorname{Act}_{\beta}^{(j)}}\right| \in \{0, \infty\} \quad \Longleftrightarrow \quad \operatorname{Var}[Z] = \operatorname{Var}[\|\nabla f_{\mathcal{N}}\|^2] \gg 1.$$

# The Init

# The Init

- Weight and biases for neurons at layer $j = 1, \ldots, d$ are drawn i.i.d. from measures $\mu^{(j)}, \nu^{(j)}$

# The Init

- Weight and biases for neurons at layer $j = 1, \ldots, d$ are drawn i.i.d. from measures $\mu^{(j)}, \nu^{(j)}$ satisfying

1. $\mu^{(j)}, \nu^{(j)}$ are symmetric around 0
2. $\mathrm{Var}[\mu^{(j)}] = 2/n_{j-1}$
3. $\nu^{(j)}$ has not atoms

# Phase Transition for $Z$

# Phase Transition for $Z$

**Theorem (H)**

*Let* $\mathcal{N} \in \mathfrak{N}_{\mu,\nu}(d, \mathbf{n})$.

# Phase Transition for $Z$

## Theorem (H)

*Let* $\mathcal{N} \in \mathfrak{N}_{\mu,\nu}(d, \mathbf{n})$. *Then, with* $Z = \|\nabla f_{\mathcal{N}}\|^2$,

1. *For every* $d, \mathbf{n}$
$$\mathbb{E}[Z] = 1.$$

# Phase Transition for $Z$

## Theorem (H)

*Let* $\mathcal{N} \in \mathfrak{N}_{\mu,\nu}(d, \mathbf{n})$. *Then, with* $Z = \|\nabla f_{\mathcal{N}}\|^2$,

1. *For every* $d, \mathbf{n}$
$$\mathbb{E}[Z] = 1.$$

2. *There exists* $C > 0$
$$2 \exp\left(\frac{1}{2}\sum_{j=1}^{d-1}\frac{1}{n_j}\right) \leq \mathbb{E}\left[Z^2\right] \leq \exp\left(C\sum_{j=1}^{d-1}\frac{1}{n_j}\right).$$

# Phase Transition for $Z$

## Theorem (H)

*Let* $\mathcal{N} \in \mathfrak{N}_{\mu,\nu}(d, \mathbf{n})$. *Then, with* $Z = \|\nabla f_{\mathcal{N}}\|^2$,

1. *For every* $d, \mathbf{n}$
$$\mathbb{E}[Z] = 1.$$
2. *There exists* $C > 0$
$$2 \exp\left(\frac{1}{2}\sum_{j=1}^{d-1}\frac{1}{n_j}\right) \leq \mathbb{E}\left[Z^2\right] \leq \exp\left(C\sum_{j=1}^{d-1}\frac{1}{n_j}\right).$$
3. *For* $K < \min\{n_j\}$, *there exists* $c_K, C_K > 0$ *so that*
$$\exp\left(c_K\sum_{j=1}^{d-1}\frac{1}{n_j}\right) \leq \mathbb{E}\left[Z^K\right] \leq \exp\left(C_K\sum_{j=1}^{d-1}\frac{1}{n_j}\right).$$

# Implications for Architecture Selection

# Implications for Architecture Selection

- If $\sum_{j \leq d} \frac{1}{n_j}$ large, then will have exploding and vanishing gradient at initialization.

# Implications for Architecture Selection

- If $\sum_{j\leq d}\frac{1}{n_j}$ large, then will have exploding and vanishing gradient at initialization.

- Power-mean inequality:

$$\left(\frac{1}{d}\sum_{j=1}^{d-1}\frac{1}{n_j}\right)^{-1} \leq \frac{1}{d}\sum_{j=1}^{d-1} n_j \leq \left(\frac{1}{d}\sum_{j=1}^{d-1} n_j^2\right)^{1/2},$$

with equality iff $n_j$ are all equal.

# Implications for Architecture Selection

- If $\sum_{j\leq d} \frac{1}{n_j}$ large, then will have exploding and vanishing gradient at initialization.

- Power-mean inequality:

$$\left(\frac{1}{d}\sum_{j=1}^{d-1}\frac{1}{n_j}\right)^{-1} \leq \frac{1}{d}\sum_{j=1}^{d-1} n_j \leq \left(\frac{1}{d}\sum_{j=1}^{d-1} n_j^2\right)^{1/2},$$

  with equality iff $n_j$ are all equal.

- $\sum_j n_j$ – total number of neurons

# Implications for Architecture Selection

- If $\sum_{j\leq d} \frac{1}{n_j}$ large, then will have exploding and vanishing gradient at initialization.

- Power-mean inequality:

$$\left(\frac{1}{d}\sum_{j=1}^{d-1}\frac{1}{n_j}\right)^{-1} \leq \frac{1}{d}\sum_{j=1}^{d-1} n_j \leq \left(\frac{1}{d}\sum_{j=1}^{d-1} n_j^2\right)^{1/2},$$

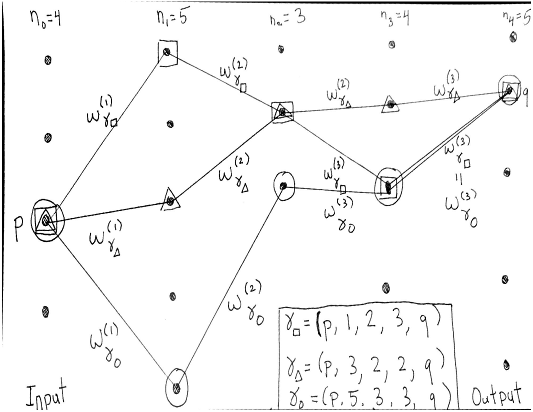
  with equality iff $n_j$ are all equal.

- $\sum_j n_j$ – total number of neurons

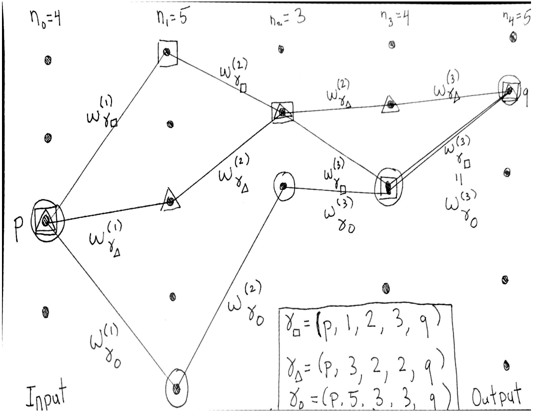
- $\sum_j n_j^2$ – total number of parameters

# Sum Over Paths Formula for $Z$

# Sum Over Paths Formula for $Z$



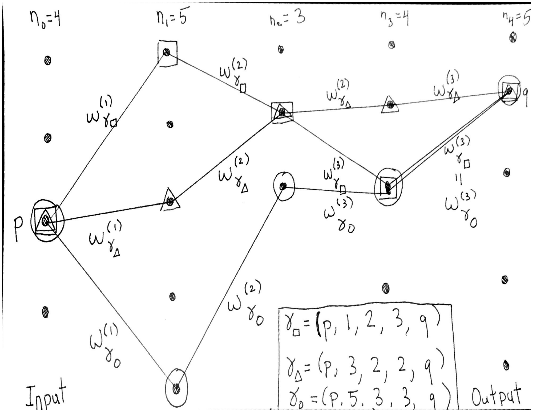
# Sum Over Paths Formula for $Z$




- We have $Z_q = \sum_{p=1}^{n_0} Z_{p,q}^2$ with

$$Z_{p,q} = \frac{\partial \left(f_{\mathcal{N}}\right)_q}{\partial x_p}$$

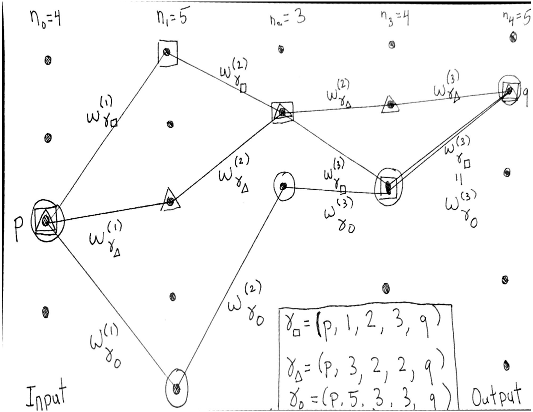
# Sum Over Paths Formula for $Z$




- We have $Z_q = \sum_{p=1}^{n_0} Z_{p,q}^2$ with

$$Z_{p,q} = \frac{\partial \left(f_{\mathcal{N}}\right)_q}{\partial x_p} = \sum_{\gamma : p \to q} \prod_{j=1}^{d} w_{\gamma}^{(j)} \mathbf{1}_{\left\{\mathrm{act}_{\gamma(j)}^{(j)} > 0\right\}}$$

# Sum Over Paths Formula for $Z$




- We have $Z_q = \sum_{p=1}^{n_0} Z_{p,q}^2$ with

$$Z_{p,q} = \frac{\partial\left(f_{\mathcal{N}}\right)_q}{\partial x_p} = \sum_{\gamma : p \to q} \prod_{j=1}^{d} w_\gamma^{(j)} \, \mathbf{1}_{\left\{\mathrm{act}_{\gamma(j)}^{(j)} > 0\right\}}$$

- $\Gamma = (\gamma_\square, \gamma_\Delta, \gamma_O)$ has $\Gamma(2) = \{2, 3\}$ and $|\Gamma_{3,q}(5)| = 2$.

# Sum Over Paths Formula for Moments of $Z_{p,q}$

# Sum Over Paths Formula for Moments of $Z_{p,q}$

## Theorem (H)

*Let* $\mathcal{N} \in \mathfrak{N}_{\mu,\nu}(d, \mathbf{n})$. *Write* $Z_{p,q} = \partial (f_{\mathcal{N}})_q / \partial x_p$. *For every* $K \geq 0$,

$$\mathbb{E}\left[Z_{p,q}^{2K}\right] = \sum_{\substack{\Gamma = (\gamma_k)_{k=1}^{2K} \\ \gamma_k : p \to q}} \prod_{j=1}^{d} \left(\frac{1}{2}\right)^{|\Gamma(j)|} \prod_{\substack{\alpha \in \Gamma(j-1) \\ \beta \in \Gamma(j)}} \mu^{(j)}_{\left|\Gamma_{\alpha,\beta}(j)\right|},$$

*where*

$$\mu_r^{(j)} = \int x^r \, d\mu^{(j)}(x).$$

# Sum Over Paths Formula for Moments of $Z_{p,q}$

## Theorem (H)

*Let* $\mathcal{N} \in \mathfrak{N}_{\mu,\nu}(d, \mathbf{n})$. *Write* $Z_{p,q} = \partial (f_{\mathcal{N}})_q / \partial x_p$. *For every* $K \geq 0$,

$$\mathbb{E}\left[Z_{p,q}^{2K}\right] = \sum_{\substack{\Gamma = (\gamma_k)_{k=1}^{2K} \\ \gamma_k : p \to q}} \prod_{j=1}^{d} \left(\frac{1}{2}\right)^{|\Gamma(j)|} \prod_{\substack{\alpha \in \Gamma(j-1) \\ \beta \in \Gamma(j)}} \mu^{(j)}_{|\Gamma_{\alpha,\beta}(j)|},$$

*where*

$$\mu_r^{(j)} = \int x^r \, d\mu^{(j)}(x).$$

## Remark

*The expression above is true for arbitrary connectivity and for convnets (when input is randomized).*

# Second Moment for $Z$

# Second Moment for $Z$

- Decompose $Z = \sum_{p=1}^{n_0} Z_{p,q}^2$

# Second Moment for $Z$

- Decompose $Z = \sum_{p=1}^{n_0} Z_{p,q}^2$
- Compute

$$\mathbb{E}\left[Z_{p,q}^2\right] = \sum_{\substack{\Gamma=(\gamma_1,\gamma_2)\\ \gamma_k : p \to q}} \prod_{j=1}^{d} \left(\frac{1}{2}\right)^{|\Gamma(j)|} \mu^{(j)}_{|\Gamma_{\alpha,\beta}(j)|}$$

# Second Moment for $Z$

- Decompose $Z = \sum_{p=1}^{n_0} Z_{p,q}^2$
- Compute

$$\mathbb{E}\left[Z_{p,q}^2\right] = \sum_{\substack{\Gamma=(\gamma_1,\gamma_2) \\ \gamma_k : p \to q}} \prod_{j=1}^{d} \left(\frac{1}{2}\right)^{|\Gamma(j)|} \mu_{|\Gamma_{\alpha,\beta}(j)|}^{(j)}$$

- $\mu_1 = 0$ so only $\gamma_1 = \gamma_2$ survives

# Second Moment for $Z$

- Decompose $Z = \sum_{p=1}^{n_0} Z_{p,q}^2$
- Compute

$$\mathbb{E}\left[Z_{p,q}^2\right] = \sum_{\substack{\Gamma=(\gamma_1,\gamma_2)\\ \gamma_k : p\to q}} \prod_{j=1}^{d} \left(\frac{1}{2}\right)^{|\Gamma(j)|} \mu_{|\Gamma_{\alpha,\beta}(j)|}^{(j)}$$

- $\mu_1 = 0$ so only $\gamma_1 = \gamma_2$ survives:

$$\mathbb{E}\left[Z_{p,q}^2\right] = \sum_{\gamma : p\to q} \prod_{j=1}^{d} \frac{1}{2} \cdot \frac{2}{n_{j-1}} = \prod_{j=1}^{d-1} n_j \cdot \prod_{j=1}^{d} \frac{1}{n_{j-1}} = \frac{1}{n_0}.$$

# Proof of Path Expression Moments

# Proof of Path Expression Moments

- Recall

$$Z_{p,q}^{2K} = \sum_{\gamma_k : p \to q} \prod_{k=1}^{2K} \prod_{j=1}^{d} w_{\gamma_k}^{(j)} \mathbf{1}_{\left\{ \mathrm{act}_{\gamma_k(j)}^{(j)} > 0 \right\}}$$

# Proof of Path Expression Moments

- Recall

$$Z_{p,q}^{2K} = \sum_{\gamma_k : p \to q} \prod_{k=1}^{2K} \prod_{j=1}^{d} w_{\gamma_k}^{(j)} \mathbf{1}_{\left\{\mathrm{act}_{\gamma_k(j)}^{(j)} > 0\right\}}$$

- Use that $f_{\mathcal{N}}$ is a Markov Chain:

$$\mathbb{E}\left[Z_{p,q}^{2K}\right] = \sum_{\gamma_k : p \to q} \mathbb{E}\left[\prod_{k=1}^{2K} \prod_{j=1}^{d} w_{\gamma_k}^{(j)} \mathbf{1}_{\left\{\mathrm{act}_{\gamma_k(j)}^{(j)} > 0\right\}}\right.$$

$$\left.\mathbb{E}\left[\prod_{k=1}^{2K} w_{\gamma_k}^{(d)} \mathbf{1}_{\left\{\mathrm{act}_{\gamma_k(d)}^{(d)} > 0\right\}} \;\middle|\; \mathrm{Act}^{(d-1)}\right]\right]$$

# Proof of Path Expression Moments

- Recall

$$Z_{p,q}^{2K} = \sum_{\gamma_k : p \to q} \prod_{k=1}^{2K} \prod_{j=1}^{d} w_{\gamma_k}^{(j)} \mathbf{1}_{\left\{\text{act}_{\gamma_k(j)}^{(j)} > 0\right\}}$$

- Use that $f_{\mathcal{N}}$ is a Markov Chain:

$$\mathbb{E}\left[Z_{p,q}^{2K}\right] = \sum_{\gamma_k : p \to q} \mathbb{E}\left[\prod_{k=1}^{2K} \prod_{j=1}^{d} w_{\gamma_k}^{(j)} \mathbf{1}_{\left\{\text{act}_{\gamma_k(j)}^{(j)} > 0\right\}} \right.$$

$$\left. \mathbb{E}\left[\prod_{k=1}^{2K} w_{\gamma_k}^{(d)} \mathbf{1}_{\left\{\text{act}_{\gamma_k(d)}^{(d)} > 0\right\}} \;\middle|\; \text{Act}^{(d-1)}\right]\right]$$

- Use independence of neurons and symmetrize:

$$\mathbb{E}\left[\prod_{k=1}^{2K} w_{\gamma_k}^{(d)} \mathbf{1}_{\left\{\text{act}_{\gamma_k(d)}^{(d)} > 0\right\}} \;\middle|\; \text{Act}^{(d-1)}\right] = \prod_{\beta \in \Gamma(d)} \frac{1}{2} \mathbb{E}\left[\prod_{k=1}^{2K} w_{\gamma_k}^{(d)}\right].$$