Machine Learning

CMPT 726

Mo Chen
SFU School of Computing Science
2022-11-01

Neural Networks

Backpropagation in MLPs: Practice

$\hat{\vec{y}} = W_L \vec{h}_L$, where $\vec{h}_L = \psi(W_{L-1}\vec{h}_{L-1})$ and $\vec{h}_{L-1} = \psi(W_{L-2}\vec{h}_{L-2}), \ldots,$ and $\vec{h}_1 = \psi(\vec{z}_1)$ and $\vec{z}_1 = W_0\vec{x}$.

- $\frac{\partial L}{\partial \vec{h}_L} = \frac{\partial \hat{\vec{y}}}{\partial \vec{h}_L}\frac{\partial L}{\partial \hat{\vec{y}}} = W_L^\top \frac{\partial L}{\partial \hat{\vec{y}}}$
- $\frac{\partial L}{\partial \vec{z}_l} = \frac{\partial \vec{h}_l}{\partial \vec{z}_l}\frac{\partial L}{\partial \vec{h}_l} = \begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix}\frac{\partial L}{\partial \vec{h}_l}$
- $\frac{\partial L}{\partial \vec{h}_l} = \frac{\partial \vec{z}_{l+1}}{\partial \vec{h}_l}\frac{\partial L}{\partial \vec{z}_{l+1}} = W_l^\top \frac{\partial L}{\partial \vec{z}_{l+1}}$
- $\frac{\partial L}{\partial \vec{w}_{l,j,\cdot}^\top} = \frac{\partial \vec{z}_{l+1}}{\partial \vec{w}_{l,j,\cdot}^\top}\frac{\partial L}{\partial \vec{z}_{l+1}} = \underbrace{(\vec{0} \quad \cdots \quad \vec{0} \quad \vec{h}_l \quad \vec{0} \quad \cdots \quad \vec{0})}_{j\text{th column}}\frac{\partial L}{\partial \vec{z}_{l+1}}$

Find $\frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^\top}$.

$$\frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^\top} = \underbrace{(\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0})}_{n_0 \times (n_1 - 1)}\frac{\partial L}{\partial \vec{z}_1}$$

$$= (\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0})\begin{pmatrix} g'(z_{1,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{1,n_1-1}) & 0 \end{pmatrix}\frac{\partial L}{\partial \vec{h}_1}$$

$$= (\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0})\begin{pmatrix} g'(z_{1,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{1,n_1-1}) & 0 \end{pmatrix} W_1^\top \frac{\partial L}{\partial \vec{z}_2}$$

Backpropagation in MLPs: Practice

$$\frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^{\top}} = (\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0}) \begin{pmatrix} g'(z_{1,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{1,n_1-1}) & 0 \end{pmatrix} W_1^{\top} \frac{\partial L}{\partial \vec{z}_2}$$

$$= (\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0}) \begin{pmatrix} g'(z_{1,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{1,n_1-1}) & 0 \end{pmatrix} W_1^{\top} \begin{pmatrix} g'(z_{2,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{2,n_2-1}) & 0 \end{pmatrix} \frac{\partial L}{\partial \vec{h}_2}$$

$$= (\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0}) \begin{pmatrix} g'(z_{1,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{1,n_1-1}) & 0 \end{pmatrix} W_1^{\top} \begin{pmatrix} g'(z_{2,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{2,n_2-1}) & 0 \end{pmatrix} W_2^{\top} \frac{\partial L}{\partial \vec{z}_3}$$

$\cdots$

$$= (\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0}) \prod_{l=1}^{L-1} \left[\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} W_l^{\top} \right] \frac{\partial L}{\partial \vec{z}_L}$$

$$= (\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0}) \prod_{l=1}^{L-1} \left[\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} W_l^{\top} \right] \begin{pmatrix} g'(z_{L,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{L,n_L-1}) & 0 \end{pmatrix} \frac{\partial L}{\partial \vec{h}_L}$$

$$= (\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0}) \prod_{l=1}^{L-1} \left[\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} W_l^{\top} \right] \begin{pmatrix} g'(z_{L,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{L,n_L-1}) & 0 \end{pmatrix} W_L^{\top} \frac{\partial L}{\partial \hat{\vec{y}}}$$

$$= (\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0}) \prod_{l=1}^{L} \left[\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} W_l^{\top} \right] \frac{\partial L}{\partial \hat{\vec{y}}}$$

$n_0 \times (n_1 - 1)$ $\quad (n_l - 1) \times n_l$ $\quad n_{L+1} \times 1$

$n_l \times (n_{l+1} - 1)$, or $n_L \times n_{L+1}$

Vanishing Gradients

Recall: $\|A\|_2 = \sup_{\|\vec{x}\|_2=1} \|A\vec{x}\|_2 = \sigma_{1,1}(A)$, where $\sigma_{1,1}(A)$ denotes the largest singular value of A.

So, $\left\| A\left(\frac{\vec{y}}{\|\vec{y}\|_2}\right)\right\|_2 \leq \|A\|_2 = \sigma_{1,1}(A)$.

On the other hand, $\left\| A\left(\frac{\vec{y}}{\|\vec{y}\|_2}\right)\right\|_2 = \left\|\frac{1}{\|\vec{y}\|_2} A\vec{y}\right\|_2 = \frac{1}{\|\vec{y}\|_2}\|A\vec{y}\|_2$.

Hence, $\frac{1}{\|\vec{y}\|_2}\|A\vec{y}\| \leq \sigma_{1,1}(A) \Longrightarrow \|A\vec{y}\|_2 \leq \sigma_{1,1}(A)\|\vec{y}\|_2$.

We just showed that for any A and $\vec{y}$, $\|A\vec{y}\|_2 \leq \sigma_{1,1}(A)\|\vec{y}\|_2$.

Vanishing Gradients

$$\frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^{\top}} = \begin{pmatrix}\vec{0} & \cdots & \vec{0} & \vec{x} & \vec{0} & \cdots & \vec{0}\end{pmatrix} \prod_{l=1}^{L} \left[\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} W_l^{\top} \right] \frac{\partial L}{\partial \hat{\vec{y}}}$$

We just showed that for any A and $\vec{y}$, $\|A\vec{y}\|_2 \leq \sigma_{1,1}(A)\|\vec{y}\|_2$.

$$\left\| \frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^{\top}} \right\|_2 = \left\| \begin{pmatrix}\vec{0} & \cdots & \vec{0} & \vec{x} & \vec{0} & \cdots & \vec{0}\end{pmatrix} \prod_{l=1}^{L} \left[\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} W_l^{\top} \right] \frac{\partial L}{\partial \hat{\vec{y}}} \right\|_2$$

$$\leq \sigma_{1,1}\left(\begin{pmatrix}\vec{0} & \cdots & \vec{0} & \vec{x} & \vec{0} & \cdots & \vec{0}\end{pmatrix}\right) \left\| \prod_{l=1}^{L} \left[\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} W_l^{\top} \right] \frac{\partial L}{\partial \hat{\vec{y}}} \right\|_2$$

$$= \sigma_{1,1}\left(\begin{pmatrix}\vec{0} & \cdots & \vec{0} & \vec{x} & \vec{0} & \cdots & \vec{0}\end{pmatrix}\right) \left\| \begin{pmatrix} g'(z_{1,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{1,n_1-1}) & 0 \end{pmatrix} W_1^{\top} \prod_{l=2}^{L} \left[\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} W_l^{\top} \right] \frac{\partial L}{\partial \hat{\vec{y}}} \right\|_2$$

Vanishing Gradients

$$\left\|\frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^{\top}}\right\|_2 \leq \sigma_{1,1}\left(\begin{pmatrix}\vec{0} & \cdots & \vec{0} & \vec{x} & \vec{0} & \cdots & \vec{0}\end{pmatrix}\right)\left\|\begin{pmatrix} g'(z_{1,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{1,n_1-1}) & 0 \end{pmatrix} W_1^{\top} \prod_{l=2}^{L}\left[\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} W_l^{\top}\right]\left|\frac{\partial L}{\partial \hat{\vec{y}}}\right\|_2$$

$$\leq \sigma_{1,1}\left(\begin{pmatrix}\vec{0} & \cdots & \vec{0} & \vec{x} & \vec{0} & \cdots & \vec{0}\end{pmatrix}\right)\sigma_{1,1}\left(\begin{pmatrix} g'(z_{1,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{1,n_1-1}) & 0 \end{pmatrix}\right)\left\|W_1^{\top} \prod_{l=2}^{L}\left[\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} W_l^{\top}\right]\left|\frac{\partial L}{\partial \hat{\vec{y}}}\right\|_2$$

$$\leq \sigma_{1,1}\left(\begin{pmatrix}\vec{0} & \cdots & \vec{0} & \vec{x} & \vec{0} & \cdots & \vec{0}\end{pmatrix}\right)\sigma_{1,1}\left(\begin{pmatrix} g'(z_{1,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{1,n_1-1}) & 0 \end{pmatrix}\right)\sigma_{1,1}(W_1^{\top})\left\|\prod_{l=2}^{L}\left[\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} W_l^{\top}\right]\left|\frac{\partial L}{\partial \hat{\vec{y}}}\right\|_2$$

$$\leq \sigma_{1,1}\left(\begin{pmatrix}\vec{0} & \cdots & \vec{0} & \vec{x} & \vec{0} & \cdots & \vec{0}\end{pmatrix}\right)\prod_{l=1}^{2}\left[\sigma_{1,1}\left(\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix}\right)\sigma_{1,1}(W_l^{\top})\right]\left\|\prod_{l=3}^{L}\left[\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} W_l^{\top}\right]\left|\frac{\partial L}{\partial \hat{\vec{y}}}\right\|_2$$

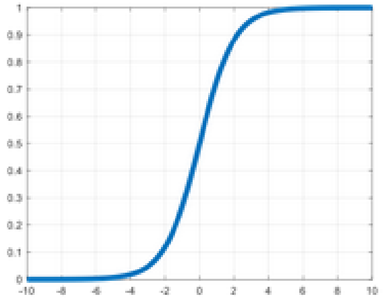
$$\vdots$$

$$\leq \sigma_{1,1}\left(\begin{pmatrix}\vec{0} & \cdots & \vec{0} & \vec{x} & \vec{0} & \cdots & \vec{0}\end{pmatrix}\right)\prod_{l=1}^{L}\left[\sigma_{1,1}\left(\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix}\right)\sigma_{1,1}(W_l^{\top})\right]\left\|\frac{\partial L}{\partial \hat{\vec{y}}}\right\|_2$$

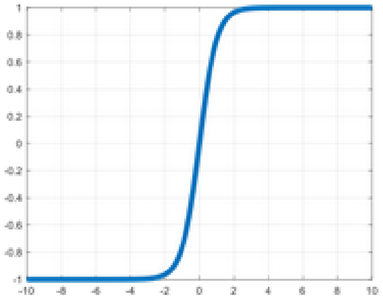
Vanishing Gradients

$$\left\| \frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^{\top}} \right\|_2 \leq \sigma_{1,1}\left(\left(\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0}\right)\right) \prod_{l=1}^{L} \left[\sigma_{1,1}\left(\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} \right) \sigma_{1,1}\left(W_l^{\top}\right) \right] \left\| \frac{\partial L}{\partial \hat{\vec{y}}} \right\|_2$$

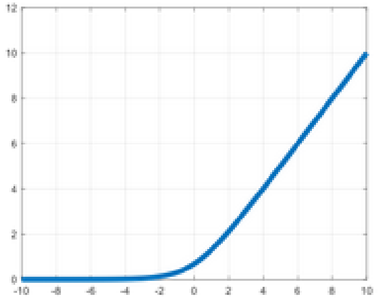
Recall: Activation functions $g(z)$



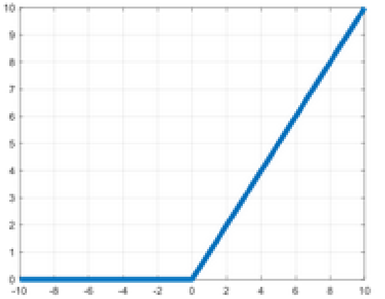
Sigmoid



Tanh



Softplus



ReLU

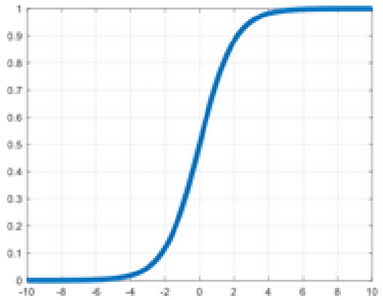
Sigmoids and tanh saturate on both ends, so if the magnitudes of the pre-activations $\vec{z}_l$ of all hidden units in some layer are moderately large (i.e.: >5),

$\sigma_{1,1}\left(\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} \right)$ would be small. As a result, $\left\| \frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^{\top}} \right\|_2$ would be small

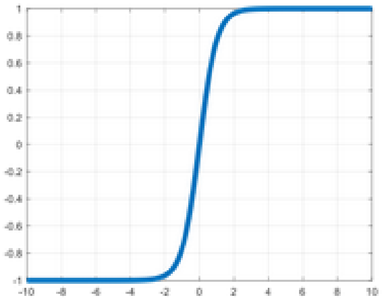
Vanishing Gradients

$$\left\| \frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^{\top}} \right\|_2 \leq \sigma_{1,1}\left(\left(\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0}\right)\right) \prod_{l=1}^{L} \left[\sigma_{1,1}\left(\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} \right) \sigma_{1,1}(W_l^{\top}) \right] \left\| \frac{\partial L}{\partial \hat{\vec{y}}} \right\|_2$$

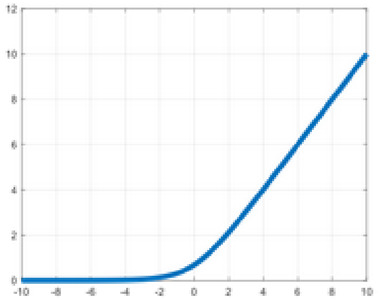
Recall: Activation functions $g(z)$



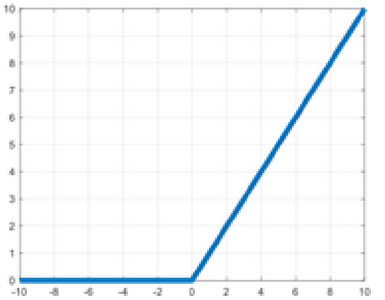
Sigmoid



Tanh



Softplus



ReLU

If there are many layers with small pre-activations, the magnitude of the gradient decays exponentially in the number of such layers.

This is known as the **vanishing gradients** problem.

It is therefore difficult to train MLPs with many layers (an example of a **deep neural network**) whose activation functions are sigmoid or tanh using gradient-based iterative optimization algorithms.

Exploding Gradients

We derived an upper bound on the gradient magnitude:

$$\left\|\frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^{\top}}\right\|_2 \leq \sigma_{1,1}\left(\left(\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0}\right)\right) \prod_{l=1}^{L}\left[\sigma_{1,1}\left(\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix}\right)\sigma_{1,1}\left(W_l^{\top}\right)\right]\left\|\frac{\partial L}{\partial \hat{\vec{y}}}\right\|_2$$

Fact: $\inf_{\|\vec{x}\|_2=1}\{\|A\vec{x}\|_2\} = \sigma_{n,n}(A)$, where $\sigma_{n,n}(A)$ denotes the smallest singular value of A. Using this fact and similar logic as before, we can also derive a lower bound on the gradient magnitude:

$$\left\|\frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^{\top}}\right\|_2 \geq \sigma_{n,n}\left(\left(\vec{0} \quad \cdots \quad \vec{0} \quad \vec{x} \quad \vec{0} \quad \cdots \quad \vec{0}\right)\right) \prod_{l=1}^{L}\left[\sigma_{n,n}\left(\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix}\right)\sigma_{n,n}\left(W_l^{\top}\right)\right]\left\|\frac{\partial L}{\partial \hat{\vec{y}}}\right\|_2$$

Exploding Gradients

$$\left\|\frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^{\top}}\right\|_2 \geq \sigma_{n,n}\left(\begin{pmatrix}\vec{0} & \cdots & \vec{0} & \vec{x} & \vec{0} & \cdots & \vec{0}\end{pmatrix}\right) \prod_{l=1}^{L}\left[\sigma_{n,n}\left(\begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix}\right)\sigma_{n,n}\left(W_l^{\top}\right)\right]\left\|\frac{\partial L}{\partial \hat{\vec{y}}}\right\|_2$$

Large $\sigma_{n,n}\left(W_l^{\top}\right)$ will lead to large $\left\|\frac{\partial L}{\partial \vec{w}_{0,j,\cdot}^{\top}}\right\|_2$, all else being equal.

Assuming we use a non-saturating activation functions like ReLU or softplus, if there are many layers whose weight matrices' singular values are large, the magnitude of the gradient grows exponentially in the number of such layers.

This is known as the **exploding gradients** problem.

Therefore, the initialization of the weight matrices must be chosen carefully to make deep neural networks trainable using gradient-based iterative optimization algorithms.

Ideally, we would like to initialize the weight matrices so that their singular values are close to 1.

Initialization

However, the weight matrices are initialized randomly, it is tricky to design them such that the singular values lie within a range.

(The limiting distribution of singular values can be derived, however. Look up Marchenko–Pastur distribution.)

In practice, the following heuristics are popular and work well:

Xavier initialization (named after Xavier Glorot):

Initialize weight matrix W_l with $\mathcal{N}\left(\vec{0}, \frac{2}{n_l+n_{l+1}} I\right)$

He initialization (named after Kaiming He):

Initialize weight matrix W_l with $\mathcal{N}\left(\vec{0}, \frac{4}{n_l+n_{l+1}} I\right)$

Weight Decay

Add squared Frobenius norms of the weight matrices excluding the biases to the loss function.

E.g.: $L(\{W_l\}_{l=0}^{L}) = \sum_{i=1}^{N}\left[\left\|\vec{y}_i - f\left(\vec{x}_i; \{W_l\}_{l=0}^{L}\right)\right\|_2 + \lambda\left\|W_{\underbrace{l,\cdot,1\ldots n_l-1}}\right\|_F^2\right]$

Exclude biases

Two purposes:

Avoiding Exploding Gradients: Don't want the weight matrices to have large singular values, since they would cause exploding gradients.

Whereas $\|A\|_2 = \sigma_{1,1}(A)$, $\|A\|_F = \sqrt{\sum_i \sigma_{i,i}(A)^2}$

Avoiding Overfitting: Akin to L2 regularization in linear regression (i.e.: ridge regression), i.e.: don't want the prediction to be very sensitive to perturbations to the input.

Layer Normalization

Instead of directly scaling the weights, scale the pre-activations of each hidden layer.

If $\|A\vec{x}\|_2 = \|\vec{x}\|_2$ for all $\vec{x}$, the singular values of A must be all 1s. Therefore, it suffices to control the norms of the pre-activations.

Originally in each layer: $\vec{z}_{l+1} = W_l\vec{h}_l$,$\vec{h}_{l+1} = \psi(\vec{z}_{l+1})$

With layer normalization: $\vec{z}_{l+1} = W_l\vec{h}_l$, $\overline{z}_{l+1} = \frac{1}{n_{l+1}-1}\sum_{i=1}^{n_{l+1}-1} z_{l+1,i}$, $s_{l+1} = \| \vec{z}_{l+1} - \overline{z}_{l+1}\vec{1} \|_2$, $\vec{h}_{l+1}$ $= \psi\left(\frac{\sqrt{n_{l+1}-1}}{s_{l+1}}\mathrm{diag}(\vec{v}_{l+1})\left(\vec{z}_{l+1} - \overline{z}_{l+1}\vec{1}\right) + \vec{u}_{l+1}\right)$,where $\vec{u}_{l+1}$ and $\vec{v}_{l+1}$ are additional model parameters.

Other related normalization schemes: batch normalization, instance normalization, group normalization (all try to normalize pre-activations or post-activations in some way)