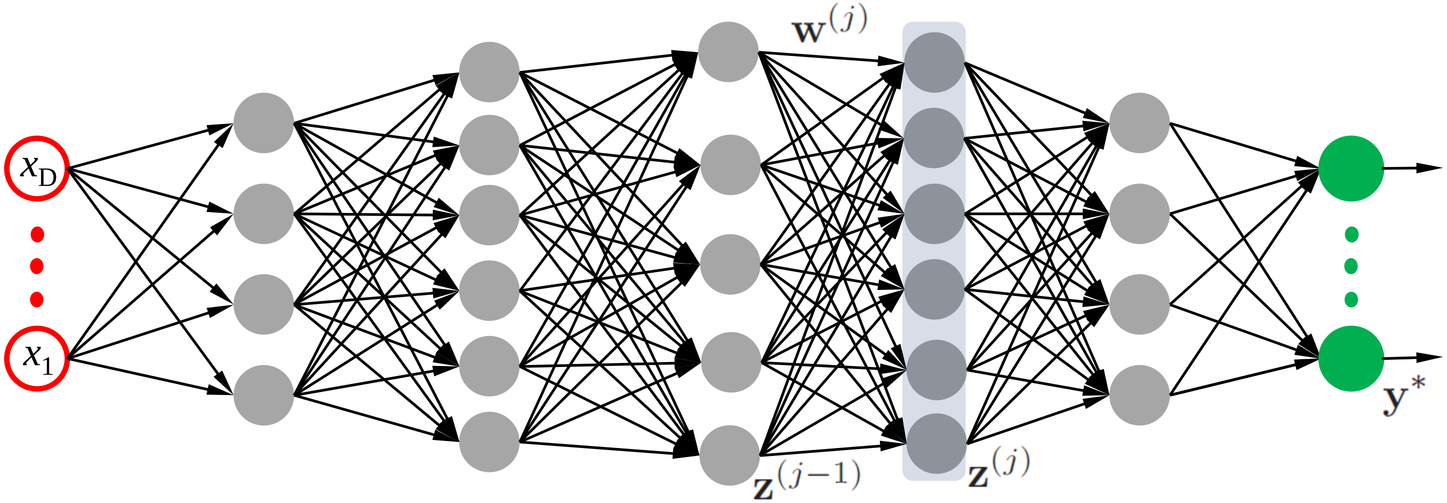
Training Deep Neural Networks: Exploding/vanishing gradients

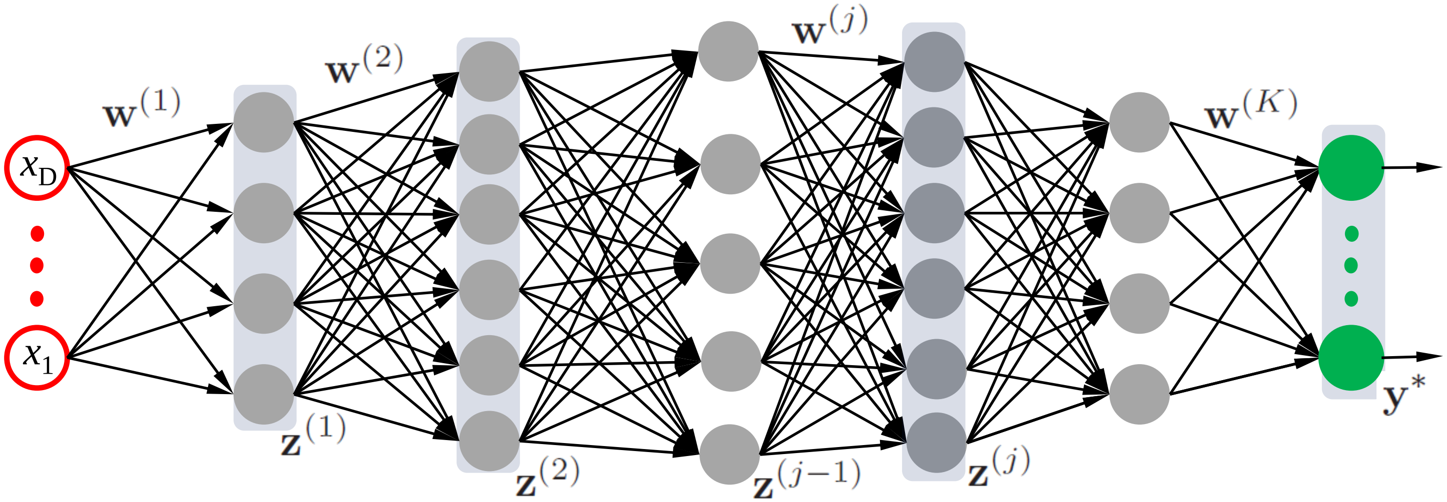
Dripta Mj
Department of Mathematics
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, India

Network setup



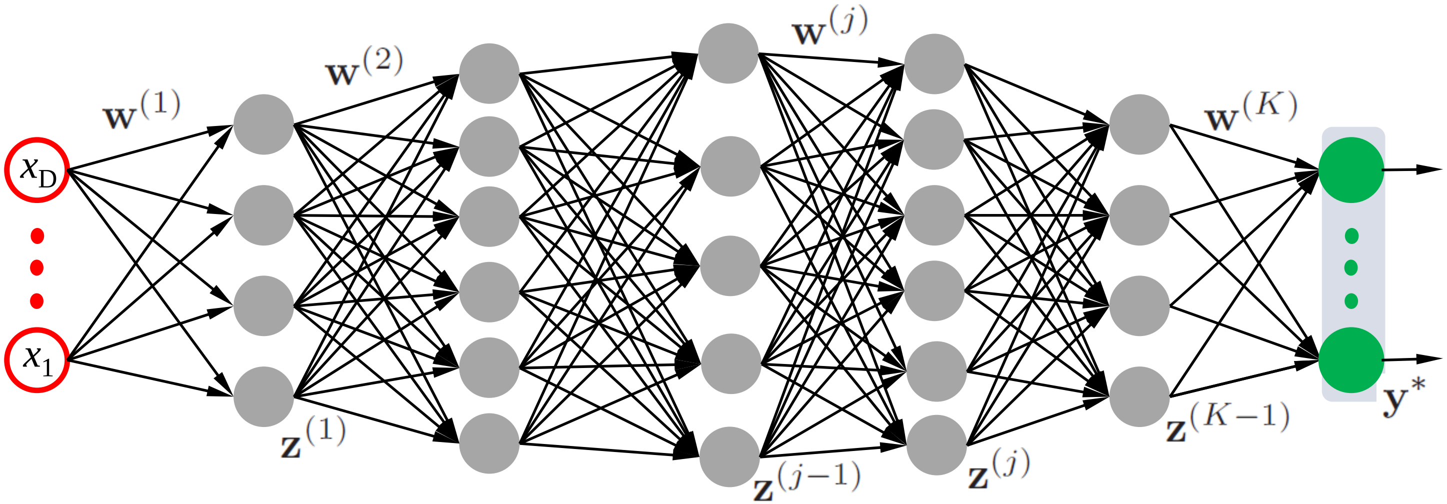
- Total number of layer: K.
- Input (vector) to the jth layer: $\mathbf{z}^{(j-1)}$.
- Output (vector) of the jth layer: $\mathbf{z}^{(j)}$.
- Weight matrix associated with the jth layer: $\mathbf{w}^{(j)}$
- Number of neurons in the jth layer: H_j.

Layer outputs



- Output from layer 1: $\mathbf{z}^{(1)} = \mathcal{A}^{(1)}\left(\mathbf{w}^{(1)\mathrm{T}}\mathbf{x}\right)$
- Output from layer 2: $\mathbf{z}^{(2)} = \mathcal{A}^{(2)}\left(\mathbf{w}^{(2)\mathrm{T}}\mathbf{z}^{(1)}\right)$
- Output from layer j: $\mathbf{z}^{(j)} = \mathcal{A}^{(j)}\left(\mathbf{w}^{(j)\mathrm{T}}\mathbf{z}^{(j-1)}\right)$
- Output from layer K (lastlayer): $\mathbf{y}^* = \mathcal{A}^{(K)}\left(\mathbf{w}^{(K)\mathrm{T}}\mathbf{z}^{(K-1)}\right)$

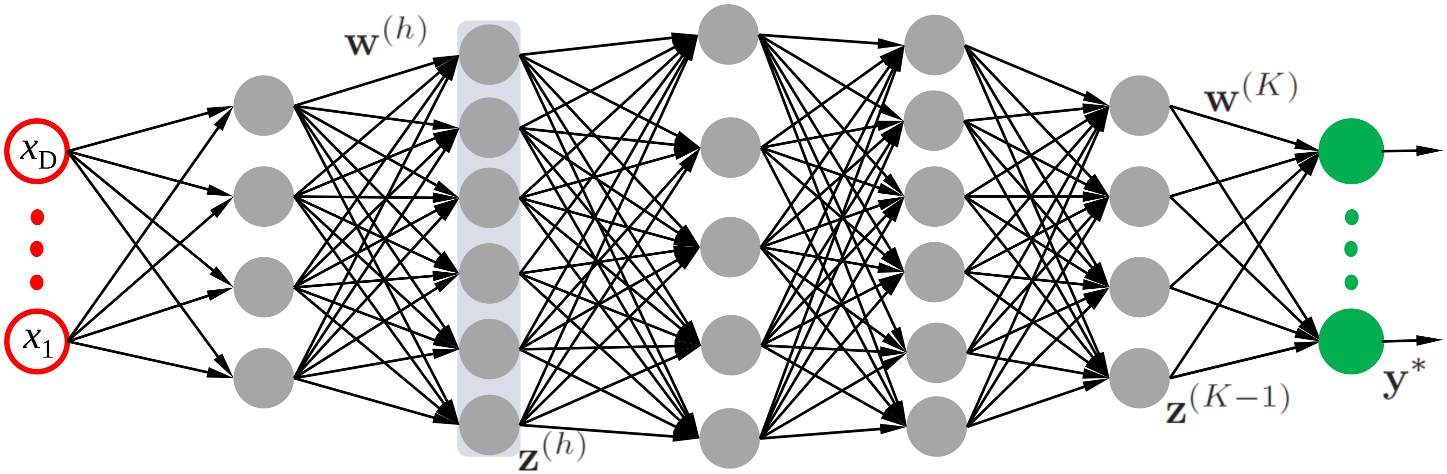
Final output



- Final output can also be written as

$$
\begin{aligned}
\mathbf{y}^* &= \mathcal{A}^{(K)}\left(\mathbf{w}^{(K)\mathrm{T}}\mathbf{z}^{(K-1)}\right) \\
&= \mathcal{A}^{(K)}\left(\mathbf{w}^{(K)\mathrm{T}}\mathcal{A}^{(K-1)}\left(\mathbf{w}^{(K-1)\mathrm{T}}\mathbf{z}^{(K-2)}\right)\right) \\
&= \mathcal{A}^{(K)}\left(\mathbf{w}^{(K)\mathrm{T}}\mathcal{A}^{(K-1)}\left(\mathbf{w}^{(K-1)\mathrm{T}}\ldots\ldots\mathcal{A}^{(j)}\left(\mathbf{w}^{(j)\mathrm{T}}\mathbf{z}^{(j-1)}\right)\ldots.\right)\right) \\
&= \mathcal{A}^{(K)}\left(\mathbf{w}^{(K)\mathrm{T}}\mathcal{A}^{(K-1)}\left(\mathbf{w}^{(K-1)\mathrm{T}}\ldots\ldots\ldots.\mathcal{A}^{(1)}\left(\mathbf{w}^{(1)\mathrm{T}}\mathbf{x}\right)\ldots\ldots\ldots\right)\right)
\end{aligned}
$$

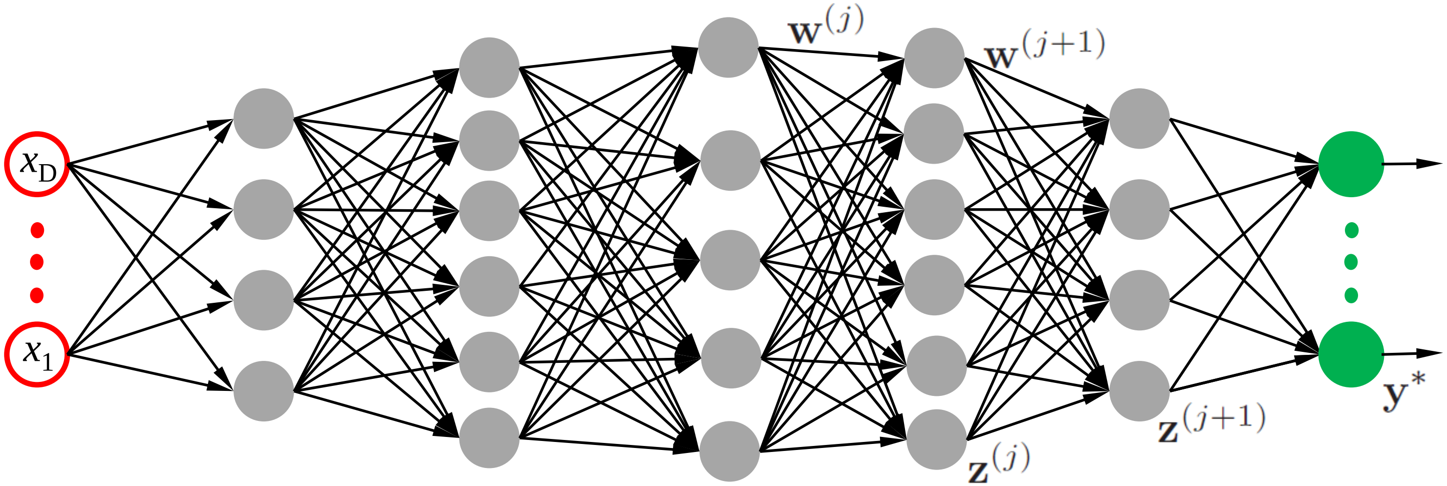
Partial derivative of the loss function



- Partial derivatives of the loss function $\mathcal{L}$ w.r.t. the weights in the hth layer:

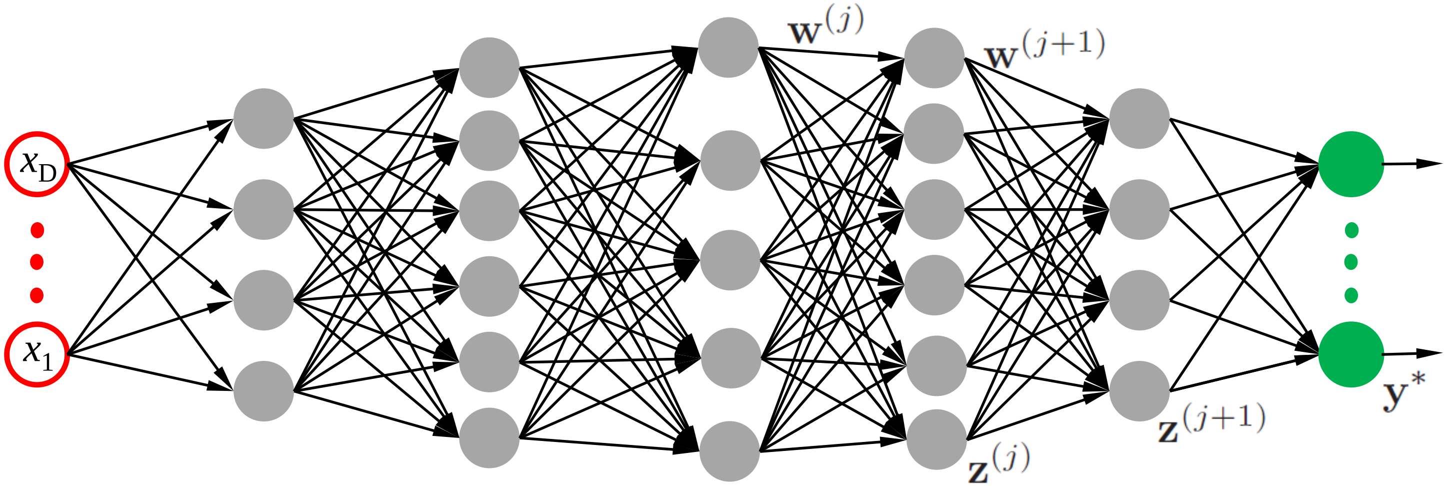
$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{w}^{(h)}} &= \frac{\partial \mathcal{L}}{\partial \mathbf{y}^*}\frac{\partial \mathbf{y}^*}{\partial \mathbf{w}^{(h)}} \\
&= \frac{\partial \mathcal{L}}{\partial \mathbf{y}^*}\frac{\partial \mathbf{y}^*}{\partial \mathbf{z}^{(K-1)}}\frac{\partial \mathbf{z}^{(K-1)}}{\partial \mathbf{w}^{(h)}} \\
&= \frac{\partial \mathcal{L}}{\partial \mathbf{y}^*}\frac{\partial \mathbf{y}^*}{\partial \mathbf{z}^{(K-1)}}\frac{\partial \mathbf{z}^{(K-1)}}{\partial \mathbf{z}^{(K-2)}}\frac{\partial \mathbf{z}^{(K-2)}}{\partial \mathbf{w}^{(h)}} \\
&\;\;. \\
&\;\;. \\
&= \frac{\partial \mathcal{L}}{\partial \mathbf{y}^*}\frac{\partial \mathbf{y}^*}{\partial \mathbf{z}^{(K-1)}}\frac{\partial \mathbf{z}^{(K-1)}}{\partial \mathbf{z}^{(K-2)}}\frac{\partial \mathbf{z}^{(K-2)}}{\partial \mathbf{z}^{(K-3)}}\cdots\frac{\partial \mathbf{z}^{(h+1)}}{\partial \mathbf{z}^{(h)}}\frac{\partial \mathbf{z}^{(h)}}{\partial \mathbf{w}^{(h)}}
\end{aligned}$$

Jacobian matrix



$$\frac{\partial \mathbf{z}^{(j+1)}}{\partial \mathbf{z}^{(j)}} = \begin{bmatrix} \frac{\partial z_1^{(j+1)}}{\partial z_1^{(j)}} & \frac{\partial z_1^{(j+1)}}{\partial z_2^{(j)}} & \cdots & \frac{\partial z_1^{(j+1)}}{\partial z_{H_j}^{(j)}} \\ \frac{\partial z_2^{(j+1)}}{\partial z_1^{(j)}} & \frac{\partial z_2^{(j+1)}}{\partial z_2^{(j)}} & \cdots & \frac{\partial z_2^{(j+1)}}{\partial z_{H_j}^{(j)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_{H_{j+1}}^{(j+1)}}{\partial z_1^{(j)}} & \frac{\partial z_{H_{j+1}}^{(j+1)}}{\partial z_2^{(j)}} & \cdots & \frac{\partial z_{H_{j+1}}^{(j+1)}}{\partial z_{H_j}^{(j)}} \end{bmatrix}$$

Jacobian matrix



- Consider one of the Jacobians $\dfrac{\partial \mathbf{z}^{(j+1)}}{\partial \mathbf{z}^{(j)}}$:

$$\frac{\partial \mathbf{z}^{(j+1)}}{\partial \mathbf{z}^{(j)}} = \frac{\partial\left(\mathcal{A}^{(j+1)}\left(\mathbf{w}^{(j+1)\mathrm{T}}\mathbf{z}^{(j)}\right)\right)}{\partial \mathbf{z}^{(j)}}$$

$$= \operatorname{diag}\left[\boldsymbol{\mathcal{A}}^{(j+1)'}\left(\mathbf{w}^{(j+1)\mathrm{T}}\mathbf{z}^{(j)}\right)\right]\mathbf{w}^{(j+1)\mathrm{T}}$$

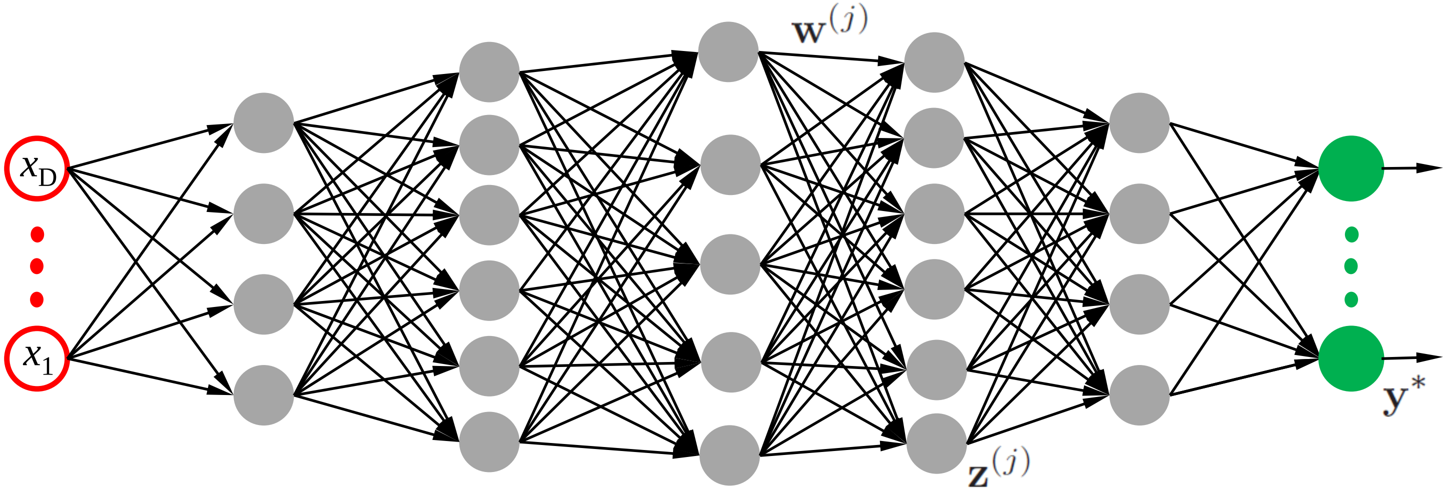
Jacobian matrix: components

$$\mathbf{w}^{(j+1)} = \begin{bmatrix} w_{11}^{(j+1)} & w_{12}^{(j+1)} & \cdot & \cdot & w_{1H_{j+1}}^{(j+1)} \\ w_{21}^{(j+1)} & w_{22}^{(j+1)} & \cdot & \cdot & w_{2H_{j+1}}^{(j+1)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ w_{H_j1}^{(j+1)} & w_{H_j2}^{(j+1)} & \cdot & \cdot & w_{H_jH_{j+1}}^{(j+1)} \end{bmatrix}$$

$$\operatorname{diag}\left[\boldsymbol{\mathcal{A}}^{(j+1)'}\left(\mathbf{w}^{(j+1)\mathrm{T}}\mathbf{z}^{(j)}\right)\right] =$$

$$\begin{bmatrix} \mathcal{A}_1^{(j+1)'}(\mathbf{w}_{\cdot 1}^{(j+1)\mathrm{T}}\mathbf{z}^{(j)}) & 0 & \cdot & \cdot & 0 \\ 0 & \mathcal{A}_2^{(j+1)'}(\mathbf{w}_{\cdot 2}^{(j+1)\mathrm{T}}\mathbf{z}^{(j)}) & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \mathcal{A}_{H_{j+1}}^{(j+1)'}(\mathbf{w}_{\cdot H_{j+1}}^{(j+1)\mathrm{T}}\mathbf{z}^{(j)}) \end{bmatrix}$$

2-norm of the Jacobians

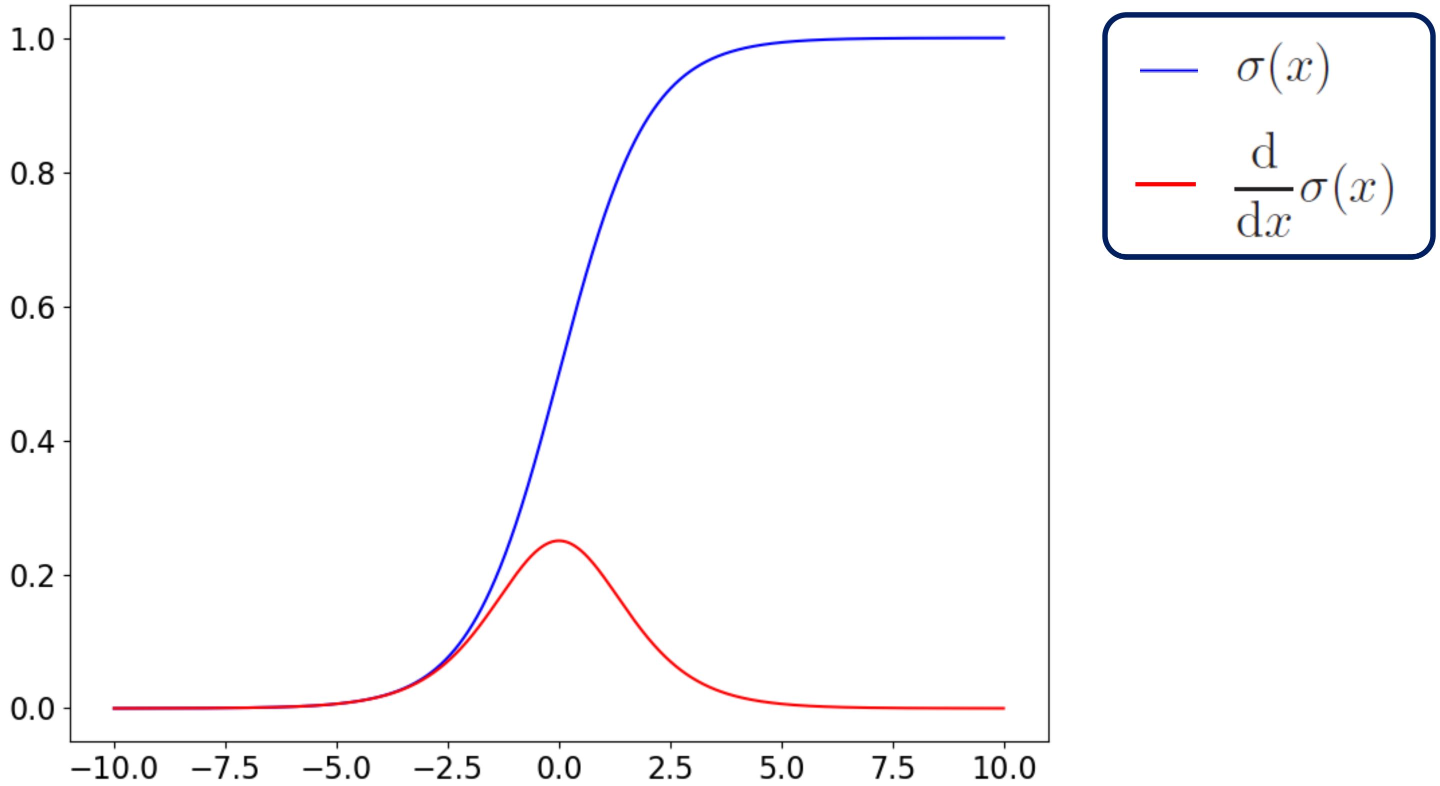


- 2-norm of the Jacobian:

$$\left|\left|\frac{\partial \mathbf{z}^{(j+1)}}{\partial \mathbf{z}^{(j)}}\right|\right| = \left|\left|\mathrm{diag}\left[\boldsymbol{\mathcal{A}}^{(j+1)'}\left(\mathbf{w}^{(j+1)\mathrm{T}}\mathbf{z}^{(j)}\right)\right]\mathbf{w}^{(j+1)\mathrm{T}}\right|\right|$$
$$\leq \left|\left|\mathrm{diag}\left[\boldsymbol{\mathcal{A}}^{(j+1)'}\left(\mathbf{w}^{(j+1)\mathrm{T}}\mathbf{z}^{(j)}\right)\right]\right|\right| \; \left|\left|\mathbf{w}^{(j+1)\mathrm{T}}\right|\right|$$

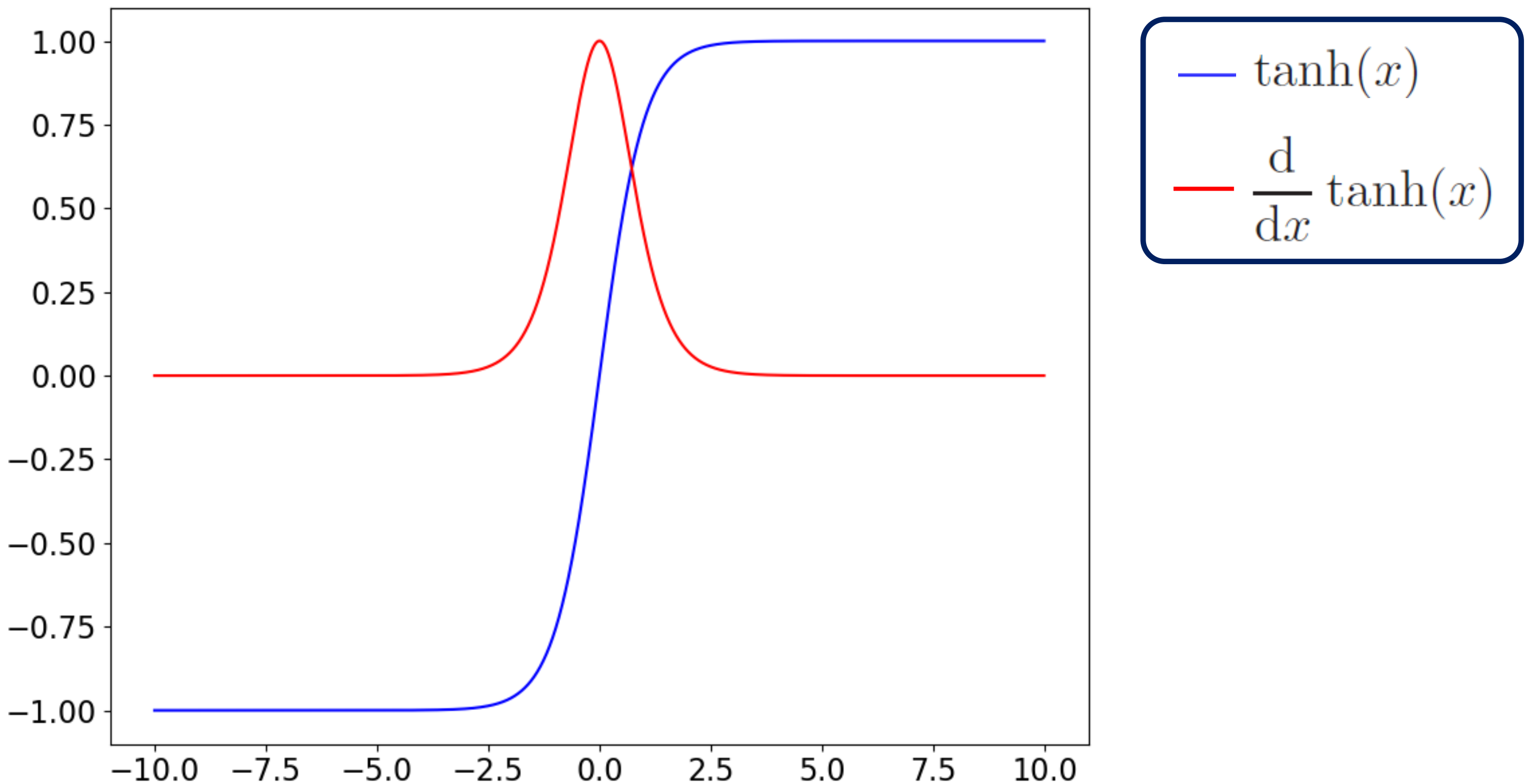
- Nonlinear activation functions $\mathcal{A}(x)$ for which $|\mathcal{A}'(x)|$ is bounded, $\mathrm{diag}\left[\boldsymbol{\mathcal{A}}^{(j+1)'}\left(\mathbf{w}^{(j+1)\mathrm{T}}\mathbf{z}^{(j)}\right)\right]$ is also bounded depending on the singular values.
- So there is a limit on how much a vector will be scaled by multiplying the Jacobian with it.

Activation function: sigmoid



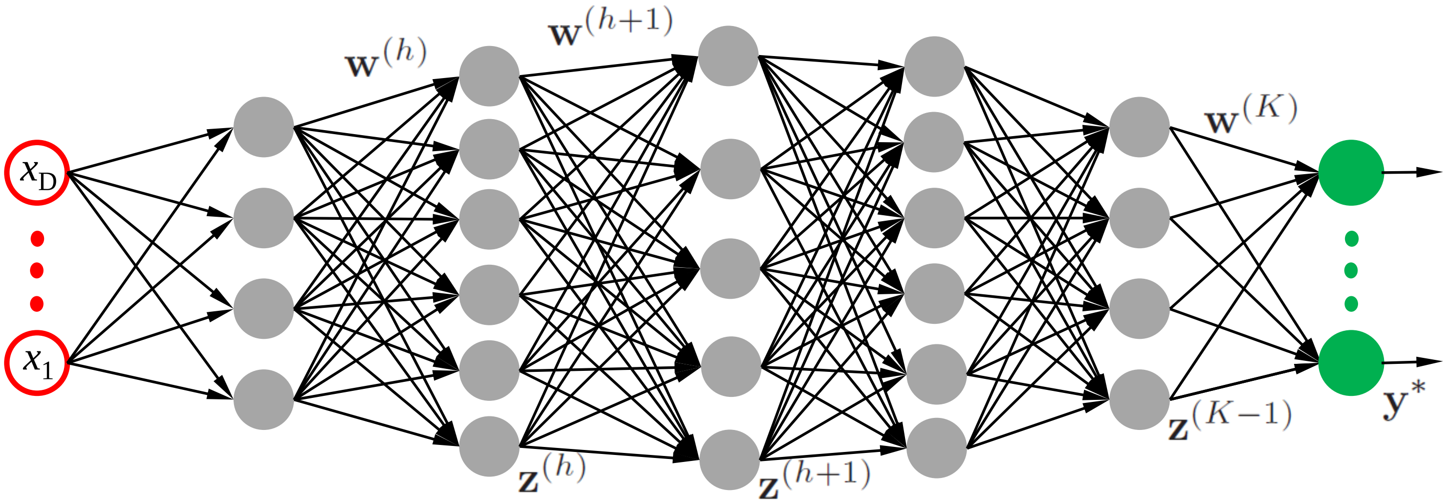
- If $\mathcal{A}^{(j+1)}(\cdot) = \sigma(\cdot)$, then $\left|\left|\mathrm{diag}\left[\boldsymbol{\mathcal{A}}^{(j+1)'}\left(\mathbf{w}^{(j+1)\mathrm{T}}\mathbf{z}^{(j)}\right)\right]\right|\right| \leq \frac{1}{4}$

Activation function: tanh



- If $\mathcal{A}^{(j+1)}(\cdot) = \tanh(\cdot)$, $\left|\left|\mathrm{diag}\left[\boldsymbol{\mathcal{A}}^{(j+1)'}\left(\mathbf{w}^{(j+1)\mathrm{T}}\mathbf{z}^{(j)}\right)\right]\right|\right| \leq 1$

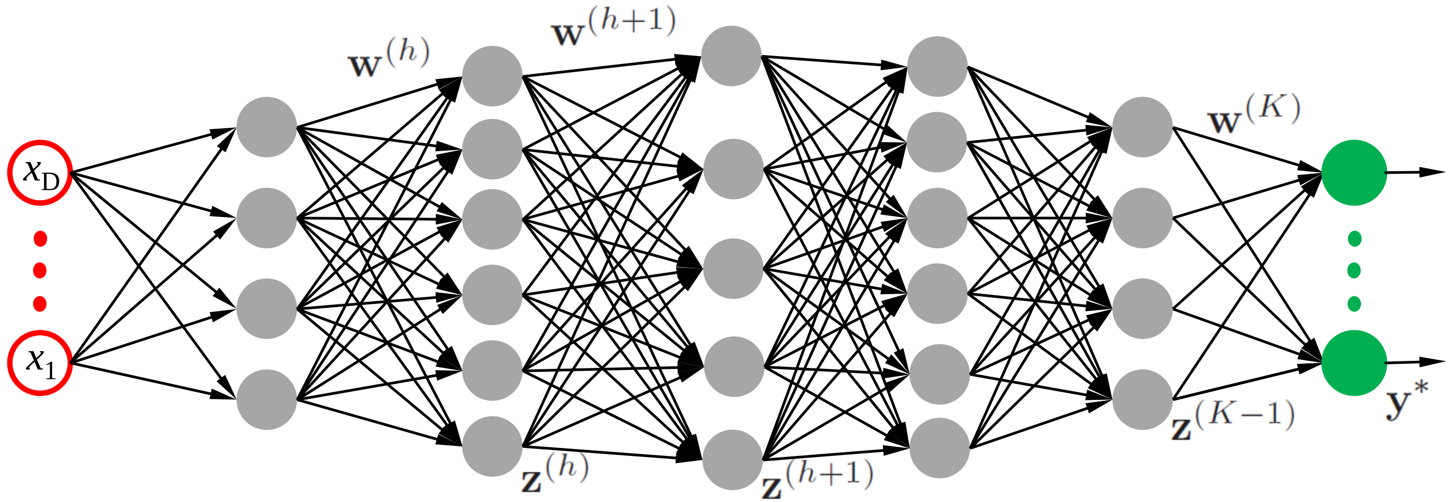
Effect of the weight matrix



$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}^{(h)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{y}^*} \frac{\partial \mathbf{y}^*}{\partial \mathbf{z}^{(K-1)}} \frac{\partial \mathbf{z}^{(K-1)}}{\partial \mathbf{z}^{(K-2)}} \frac{\partial \mathbf{z}^{(K-2)}}{\partial \mathbf{z}^{(K-3)}} \cdots \cdot \frac{\partial \mathbf{z}^{(h+1)}}{\partial \mathbf{z}^{(h)}} \frac{\partial \mathbf{z}^{(h)}}{\partial \mathbf{w}^{(h)}}$$

- Effect of the weight matrices on the chain product:
  - Expansion in directions where the singular values of the weight matrices are greater than one.
  - Shrink along directions where the singular values of the weight matrices are less than one.
- Multiplications by the weight matrices can lead to **exploding** or **vanishing** gradients.

Overall effect



$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}^{(h)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{y}^*} \frac{\partial \mathbf{y}^*}{\partial \mathbf{z}^{(K-1)}} \frac{\partial \mathbf{z}^{(K-1)}}{\partial \mathbf{z}^{(K-2)}} \frac{\partial \mathbf{z}^{(K-2)}}{\partial \mathbf{z}^{(K-3)}} \cdots \frac{\partial \mathbf{z}^{(h+1)}}{\partial \mathbf{z}^{(h)}} \frac{\partial \mathbf{z}^{(h)}}{\partial \mathbf{w}^{(h)}}$$

- Overall effect of the Jacobian matrices on the chain product:
  - Expansion in directions where the Jacobian matrices have singular values greater than one.
  - Shrinkage in directions where the Jacobian matrices have singular values less than one.