

# Aprendizado Profundo 1

Inicialização de Pesos – Explosão e Dissipação de Gradientes

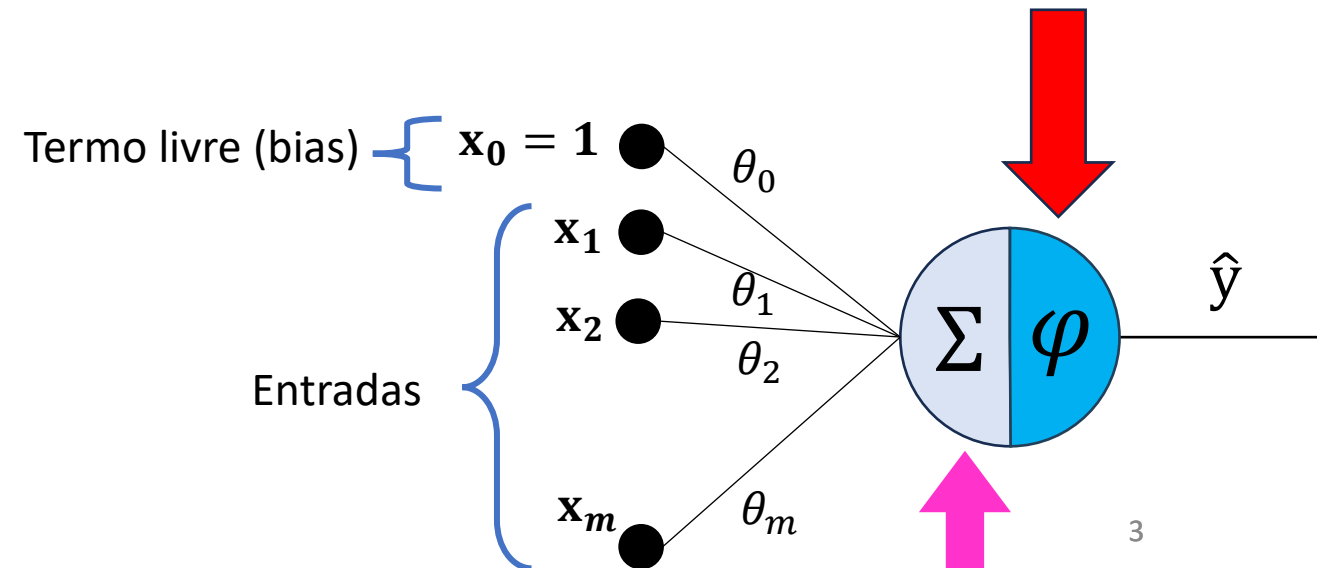
Professor: Lucas Silveira Kupssinskü

# Agenda

- Revisão sobre funções de ativação
- Breve Estudo de Caso
- Inicialização de Pesos
  - He
  - Xavier

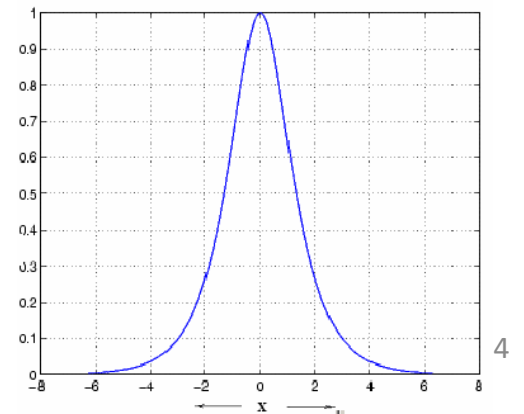
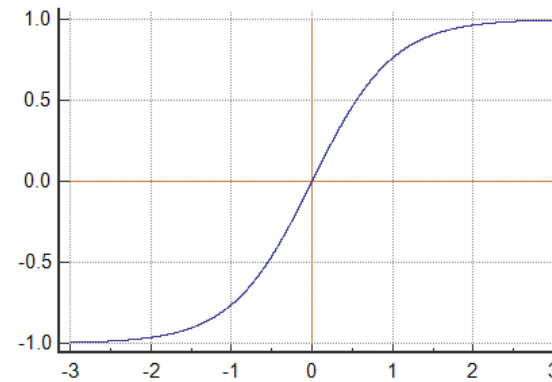
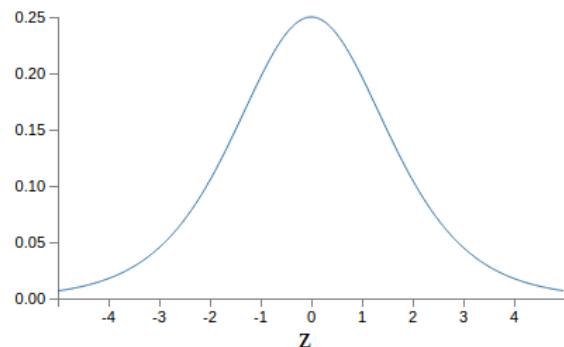
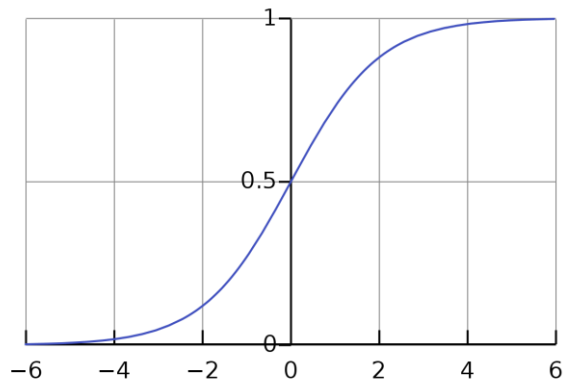
# Funções de Ativação

- São aplicadas na **pré-ativação**
- Nas camadas ocultas
  - são responsáveis por adicionar “não linearidades”
- Na camada de saída
  - são escolhidas conforme a tarefa
- Idealmente devem ser:
  - contínuas
  - diferenciáveis



# Funções de Ativação

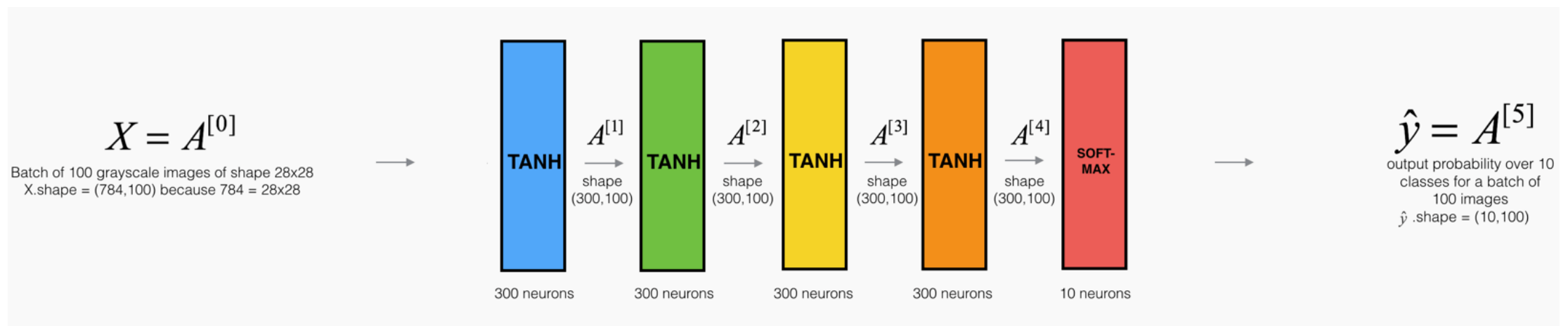
- Sigmoid e Tanh
  - Caíram em desuso para camadas ocultas
  - Facilitam o fenômeno de *Vanishing Gradient*
  - Você consegue identificar o motivo?



# Vamos fazer um estudo de caso



- Vamos usar um MLP para fazer a classificação dos dígitos escritos a mão (MNIST)
- Obviamente, não conseguimos acompanhar as ativações e os gradientes individualmente
  - Mas Podemos acompanhar os histogramas

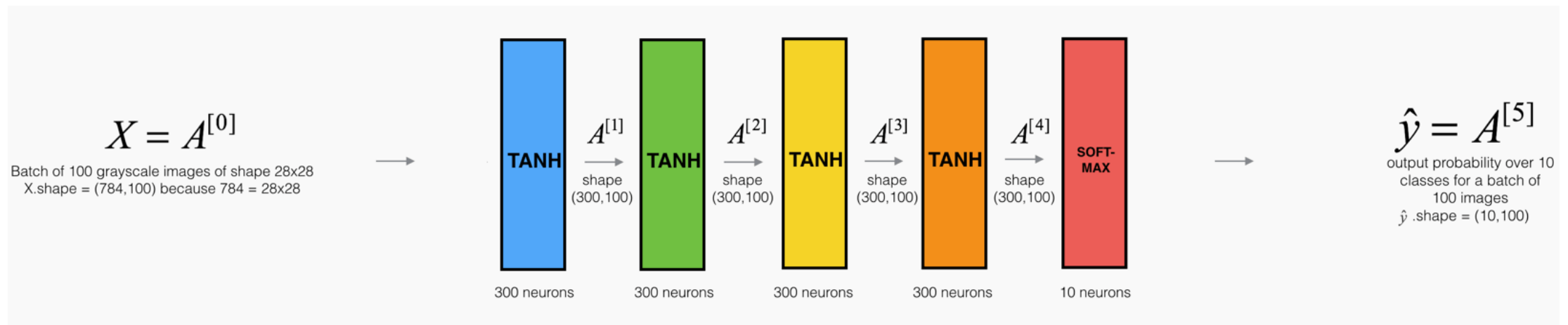


# Vamos fazer um estudo de caso



Inicializando  $\theta = 0$

O que será que vai acontecer?

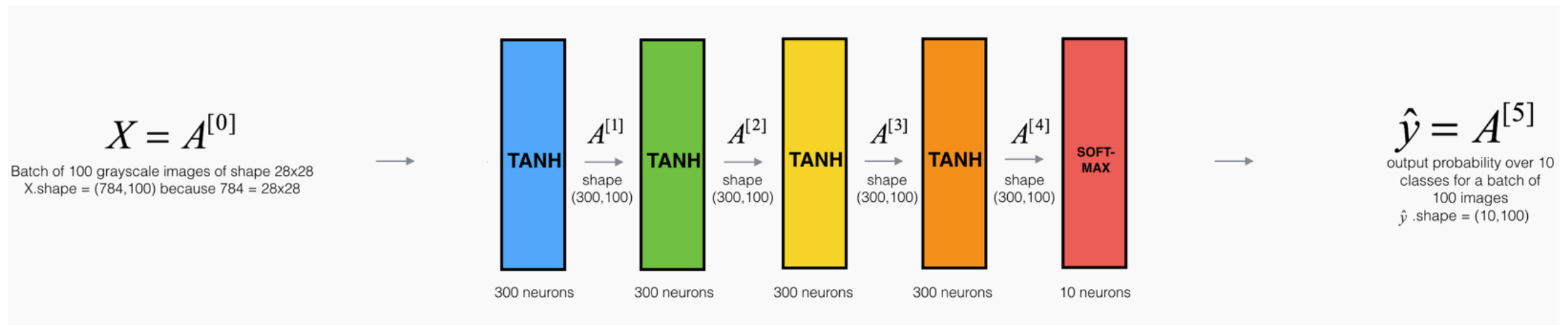
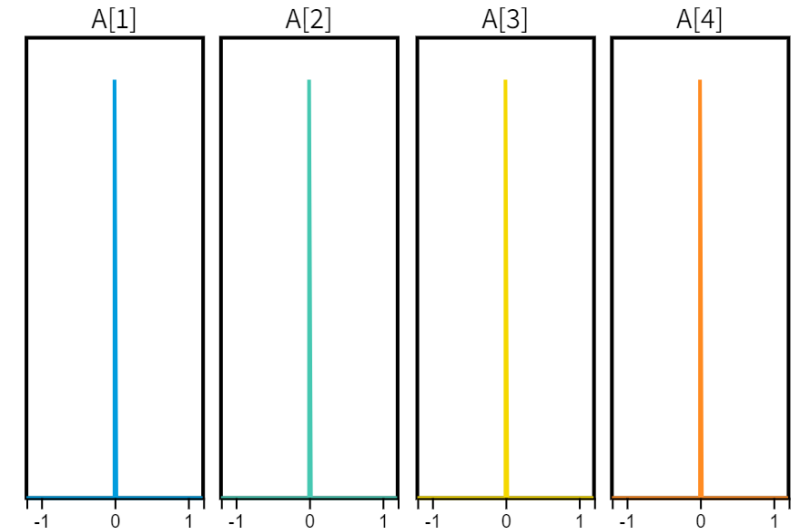


# Vamos fazer um estudo de caso



Inicializando  $\theta = 0$

O que será que vai acontecer?

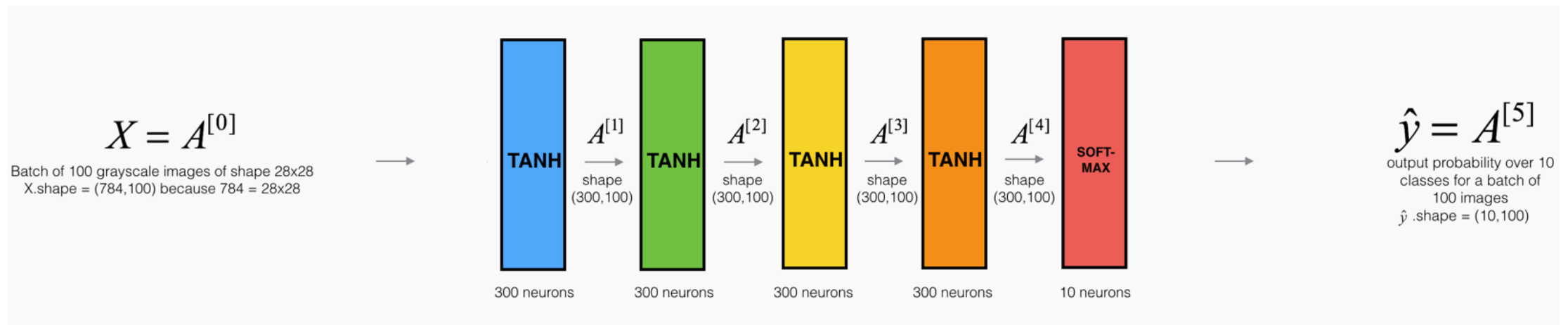


# Vamos fazer um estudo de caso



Inicializando  $\theta = \mathcal{N}(0,1)$

O que será que vai acontecer?



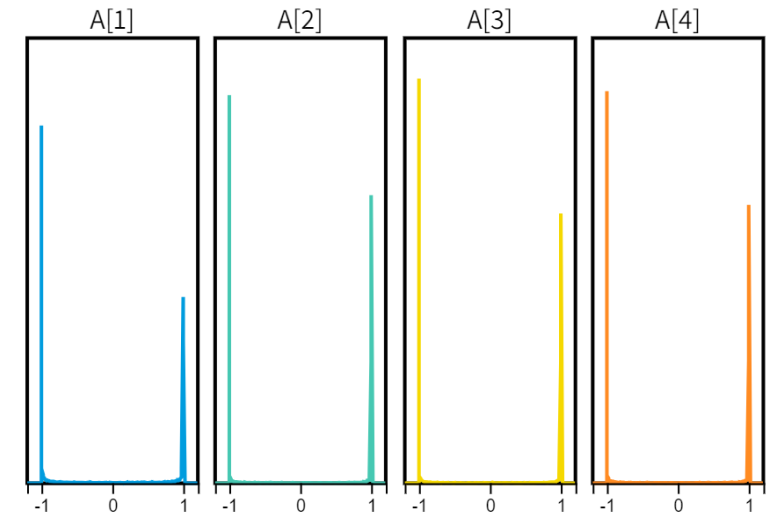


# Vamos fazer um estudo de caso



Inicializando  $\theta = \mathcal{N}(0,1)$

O que será que vai acontecer?



$X = A^{[0]}$   
Batch of 100 grayscale images of shape 28x28  
 $X.shape = (784, 100)$  because  $784 = 28 \times 28$



300 neurons

$A^{[1]}$   
shape  
(300, 100)



300 neurons

$A^{[2]}$   
shape  
(300, 100)



300 neurons

$A^{[3]}$   
shape  
(300, 100)



300 neurons

$A^{[4]}$   
shape  
(300, 100)



10 neurons



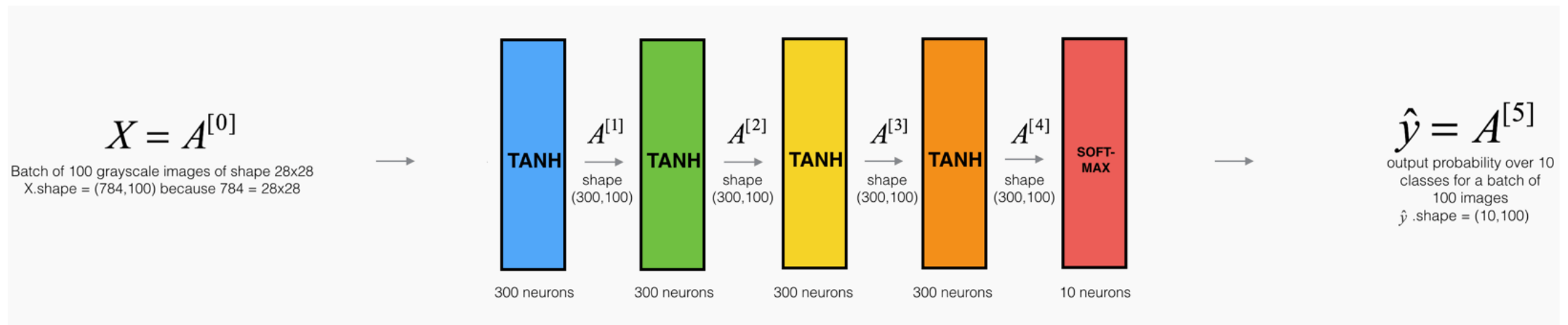
$\hat{y} = A^{[5]}$   
output probability over 10  
classes for a batch of  
100 images  
 $\hat{y}.shape = (10, 100)$

# Vamos fazer um estudo de caso



Inicializando  $\theta = \mathcal{U}(-1,1)$

O que será que vai acontecer?

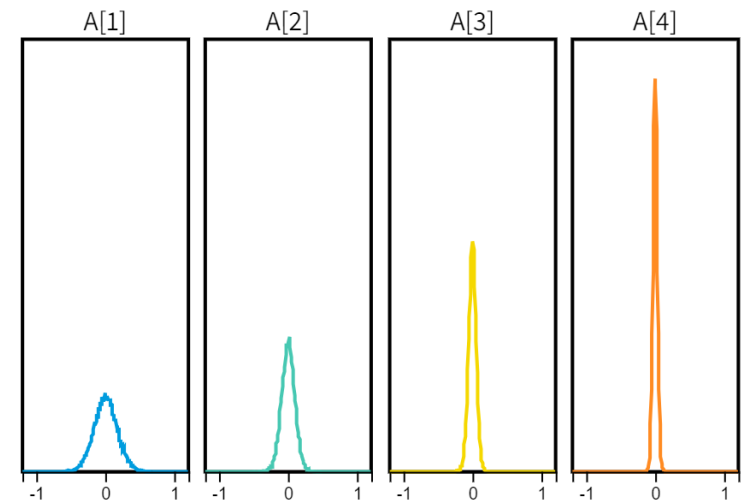


# Vamos fazer um estudo de caso



Inicializando  $\theta = \mathcal{U}(-1,1)$

O que será que vai acontecer?



$X = A^{[0]}$   
Batch of 100 grayscale images of shape 28x28  
 $X.shape = (784, 100)$  because  $784 = 28 \times 28$



300 neurons

$A^{[1]}$   
shape  
(300, 100)



300 neurons

$A^{[2]}$   
shape  
(300, 100)



300 neurons

$A^{[3]}$   
shape  
(300, 100)



300 neurons

$A^{[4]}$   
shape  
(300, 100)



10 neurons

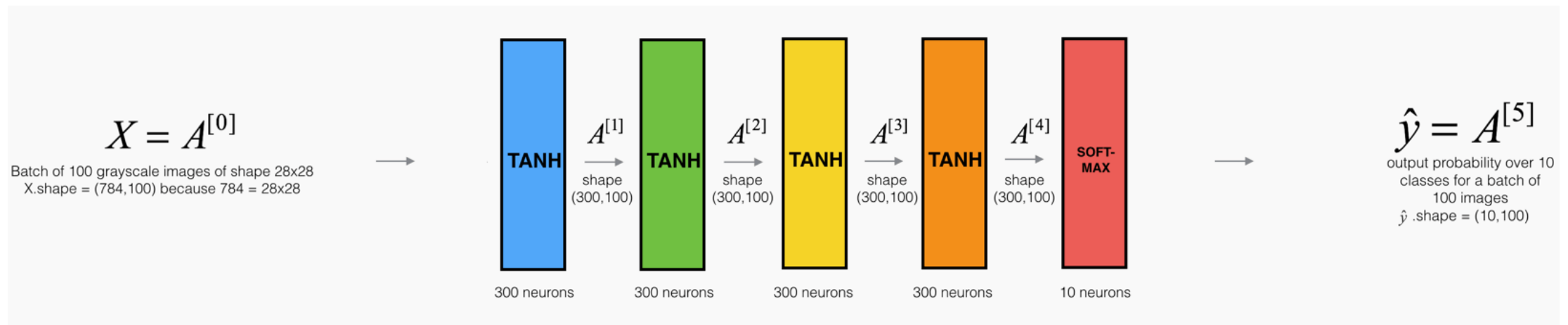


$\hat{y} = A^{[5]}$   
output probability over 10  
classes for a batch of  
100 images  
 $\hat{y}.shape = (10, 100)$

# Vamos fazer um estudo de caso



Conclusão: precisamos fazer melhor 😊



# Inicialização de Parâmetros

- Considere que:

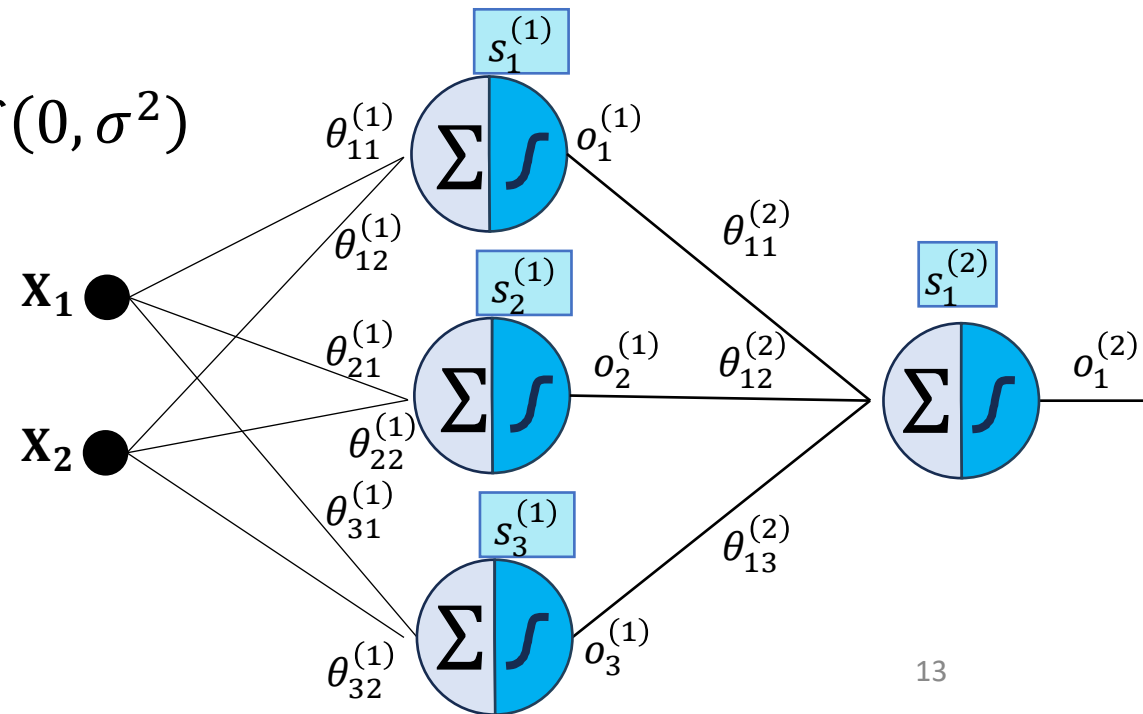
$$s^{(l)} = b^{(l)} + \theta^{(l)} \varphi(s^{(l-1)})$$

$s^{(l)}$ : pré-ativações camada  $l$

$b^{(l)}$ : bias da camada  $l$ , inicializados com 0

$\theta^{(l)}$ : parâmetros camada  $l$ , inicializados com  $\mathcal{N}(0, \sigma^2)$

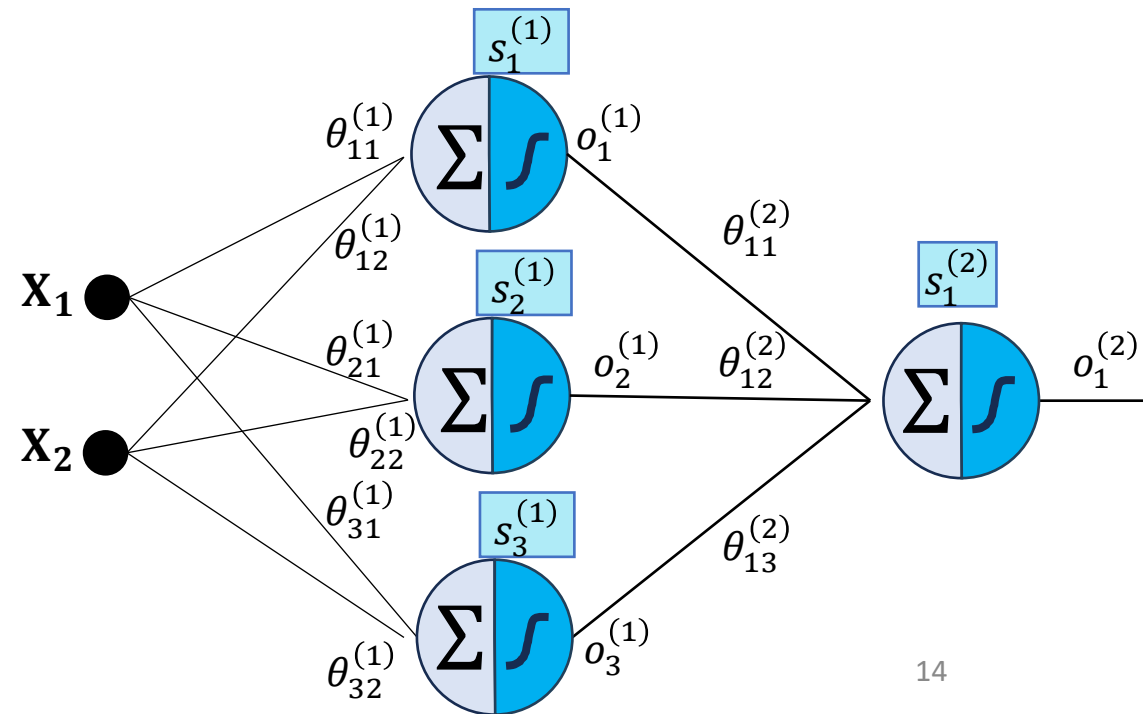
$\varphi$ : função de ativação (nesse caso ReLU)



# Inicialização de Parâmetros

$$s^{(l)} = b^{(l)} + \theta^{(l)} \varphi(s^{(l-1)})$$

Se  $\sigma^2$  for muito pequeno, o que vai acontecer com as pré-ativações de camadas mais internas da rede?

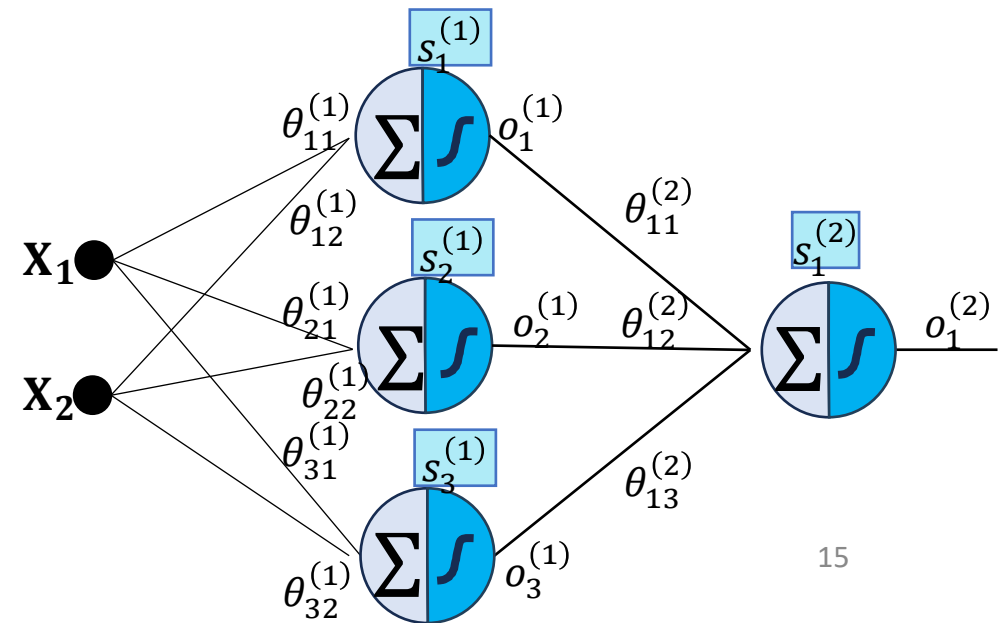


# Inicialização de Parâmetros

$$s^{(l)} = b^{(l)} + \theta^{(l)} \varphi(s^{(l-1)})$$

Se  $\sigma^2$  for muito pequeno, o que vai acontecer com as pré-ativações de camadas mais internas da rede?

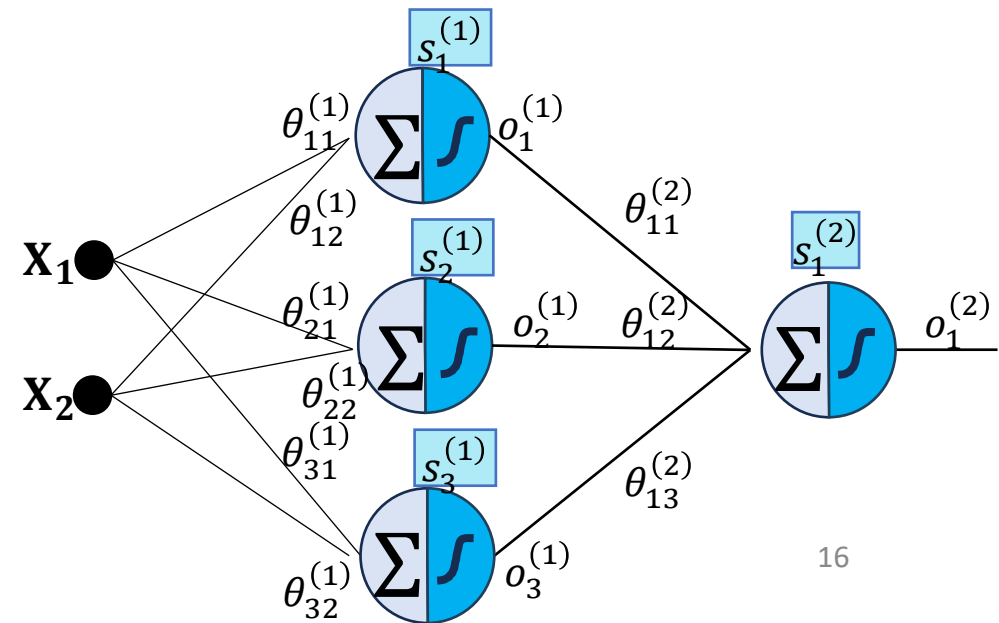
- Como  $s^{(l)}$  é uma soma ponderada que vai considerar pesos  $\theta^{(l)}$  pequenos, a tendência é que a saída tenha uma magnitude menor que a entrada



# Inicialização de Parâmetros

$$s^{(l)} = b^{(l)} + \theta^{(l)} \varphi(s^{(l-1)})$$

Se  $\sigma^2$  for muito grande, o que vai acontecer com as pré-ativações de camadas mais internas da rede?



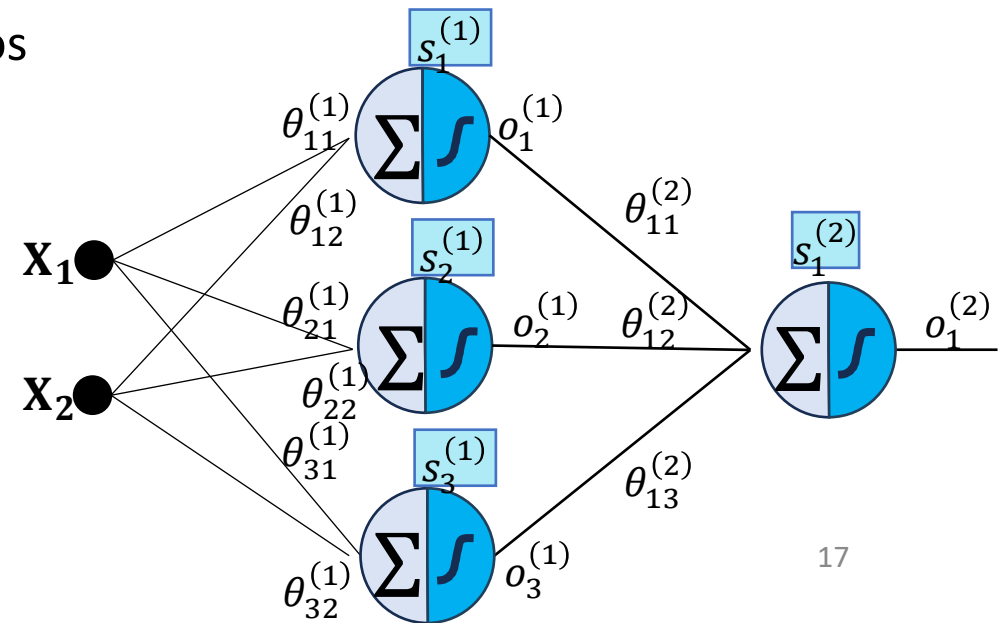


# Inicialização de Parâmetros

$$s^{(l)} = b^{(l)} + \theta^{(l)} \varphi(s^{(l-1)})$$

Se  $\sigma^2$  for muito grande, o que vai acontecer com as pré-ativações de camadas mais internas da rede?

- Como  $s^{(l)}$  é uma soma ponderada que vai considerar pesos  $\theta^{(l)}$  grande, a tendência é que a saída tenha uma magnitude maior que a entrada
  - Ocorre mesmo com ReLU cortando os valores negativos



# Inicialização de Parâmetros

- Nas duas situações vistas anteriormente as pré-ativações se tornam muito pequenas ou muito grandes
  - Intratável para aritmética de ponto flutuante
- A mesma lógica se aplica ao gradiente (lembra das expressões do *Backpropagation*?)
- Esses casos são chamados de *Vanishing* e *Exploding Gradient* respectivamente

# Inicialização de Parâmetros

- Vamos tentar fazer melhor
- Como será que as pré-ativações de uma camada se comportam em relação as pré-ativações da camada anterior?
- Considere que:

$$s^{(l)} = b^{(l)} + \theta^{(l)} \varphi(s^{(l-1)})$$

- $s^{(l)}$ : pré-ativações camada  $l$
- $b^{(l)}$ : bias da camada  $l$ , inicializados com 0
- $\theta^{(l)}$ : parâmetros camada  $l$ , inicializados com  $\mathcal{N}(0, \sigma_\theta^2)$
- $\varphi$ : função de ativação (nesse caso ReLU)

# Inicialização de Parâmetros

$$s^{(l)} = b^{(l)} + \theta^{(l)} \varphi(s^{(l-1)})$$

- Vamos tentar descobrir como a média e o variância da camada subsequente se comportam

# Inicialização de Parâmetros

$$s_i^{(l)} = b_i^{(l)} + \sum_{j=1}^n \theta_{ij}^{(l)} \varphi \left( s_j^{(l-1)} \right)$$

$$\mathbb{E}[s^{(l)}] = \mathbb{E} \left[ b_i^{(l)} + \sum_{j=1}^n \theta_{ij}^{(l)} \varphi \left( s_j^{(l-1)} \right) \right]$$

# Inicialização de Parâmetros

$$s_i^{(l)} = b_i^{(l)} + \sum_{j=1}^n \theta_{ij}^{(l)} \varphi \left( s_j^{(l-1)} \right)$$

$$\mathbb{E}[s^{(l)}] = \mathbb{E} \left[ b_i^{(l)} + \sum_{j=1}^n \theta_{ij}^{(l)} \varphi \left( s_j^{(l-1)} \right) \right]$$

$$\mathbb{E}[s^{(l)}] = \mathbb{E} \left[ b_i^{(l)} \right] + \mathbb{E} \left[ \sum_{j=1}^n \theta_{ij}^{(l)} \varphi \left( s_j^{(l-1)} \right) \right]$$

# Inicialização de Parâmetros

$$s_i^{(l)} = b_i^{(l)} + \sum_{j=1}^n \theta_{ij}^{(l)} \varphi(s_j^{(l-1)})$$

$$\mathbb{E}[s^{(l)}] = \mathbb{E} \left[ \overset{0}{\cancel{b_i^{(l)}}} \right] + \mathbb{E} \left[ \sum_{j=1}^n \theta_{ij}^{(l)} \varphi(s_j^{(l-1)}) \right]$$

$$\mathbb{E}[s^{(l)}] = \sum_{j=1}^n \mathbb{E} \left[ \theta_{ij}^{(l)} \varphi(s_j^{(l-1)}) \right]$$

# Inicialização de Parâmetros

$$s_i^{(l)} = b_i^{(l)} + \sum_{j=1}^n \theta_{ij}^{(l)} \varphi(s_j^{(l-1)})$$

$$\mathbb{E}[s^{(l)}] = \mathbb{E}[\cancel{b_i^{(l)}}] + \mathbb{E}\left[\sum_{j=1}^n \theta_{ij}^{(l)} \varphi(s_j^{(l-1)})\right]$$
$$\mathbb{E}[s^{(l)}] = \sum_{j=1}^U \mathbb{E}[\theta_{ij}^{(l)}] \mathbb{E}[\varphi(s_j^{(l-1)})]$$

Assumindo que  $\theta_{ij}^{(l)}$  e  $\varphi(s_j^{(l-1)})$  são independentes



# Inicialização de Parâmetros

$$s_i^{(l)} = b_i^{(l)} + \sum_{j=1}^n \theta_{ij}^{(l)} \varphi(s_j^{(l-1)})$$

$$\mathbb{E}[s^{(l)}] = \sum_{j=1}^n \mathbb{E}[\theta_{ij}^{(l)}] \mathbb{E}[\varphi(s_j^{(l-1)})]$$

$$\mathbb{E}[s^{(l)}] = \sum_{j=1}^n 0 \cdot \mathbb{E}[\varphi(s_j^{(l-1)})]$$

# Inicialização de Parâmetros

$$\mathbb{E}[s^{(l)}] = 0$$

Vamos usar esse resultado para computar a variância

# Inicialização de Parâmetros

$$\mathbb{E}[s^{(l)}] = 0$$

$$\sigma_{s^{(l)}}^2 = \mathbb{E}[(s^{(l)})^2] - \cancel{\mathbb{E}[s^{(l)}]^2}^0$$

# Inicialização de Parâmetros


$$\sigma_{s^{(l)}}^2 = \mathbb{E} \left[ \left( s^{(l)} \right)^2 \right] - \cancel{\mathbb{E} \left[ s^{(l)} \right]^2}^0$$

$$\sigma_{s^{(l)}}^2 = \mathbb{E} \left[ \left( \cancel{b_i^{(l)}}^0 + \sum_{j=1}^n \theta_{ij}^{(l)} \varphi \left( s_j^{(l-1)} \right) \right)^2 \right]$$

# Inicialização de Parâmetros

$$\sigma_{s^{(l)}}^2 = \mathbb{E} \left[ \left( \sum_{j=1}^n \theta_{ij}^{(l)} \varphi \left( s_j^{(l-1)} \right) \right)^2 \right]$$

$$\sigma_{s^{(l)}}^2 = \sum_{j=1}^U \mathbb{E} \left[ \left( \theta_{ij}^{(l)} \right)^2 \right] \mathbb{E} \left[ \left( \varphi \left( s_j^{(l-1)} \right) \right)^2 \right]$$



Assumindo que  $\theta_{ij}^{(l)}$  e  $\varphi \left( s_j^{(l-1)} \right)$  são independentes

# Inicialização de Parâmetros

$$\sigma_{s^{(l)}}^2 = \sum_{j=1}^n \mathbb{E} \left[ \left( \theta_{ij}^{(l)} \right)^2 \right] \mathbb{E} \left[ \left( \varphi \left( s_j^{(l-1)} \right)^2 \right) \right]$$

$$\sigma_{s^{(l)}}^2 = \sum_{j=1}^n \sigma_{\theta}^2 \mathbb{E} \left[ \left( \varphi \left( s_j^{(l-1)} \right)^2 \right) \right]$$

$$\sigma_{s^{(l)}}^2 = \sigma_{\theta}^2 \sum_{j=1}^n \mathbb{E} \left[ \left( \varphi \left( s_j^{(l-1)} \right)^2 \right) \right]$$

# Inicialização de Parâmetros

$$\sigma_{s^{(l)}}^2 = \sigma_{\theta}^2 \sum_{j=1}^n \mathbb{E} \left[ \left( \varphi \left( s_j^{(l-1)} \right)^2 \right) \right]$$

$$\sigma_{s^{(l)}}^2 = \sigma_{\theta}^2 \sum_{j=1}^n \frac{\sigma_{s^{(l-1)}}^2}{2}$$

Assumindo que  $s_j^{(l-1)}$  é simétrico em 0, metade das pré-ativações serão zeradas. Isso implica que

$$\mathbb{E} \left[ \left( \varphi \left( s_j^{(l-1)} \right)^2 \right) \right] = \frac{\sigma_{s^{(l-1)}}^2}{2}$$




# Inicialização de Parâmetros

$$\sigma_{s^{(l)}}^2 = \sigma_{\theta}^2 \sum_{j=1}^n \frac{\sigma_{s^{(l-1)}}^2}{2}$$

$$\sigma_{s^{(l)}}^2 = \frac{n}{2} \sigma_{\theta}^2 \sigma_{s^{(l-1)}}^2$$



# Inicialização de Parâmetros

$$\sigma_{s^{(l)}}^2 = \frac{n}{2} \sigma_{\theta}^2 \sigma_{s^{(l-1)}}^2$$


$$\sigma_{\theta}^2 = \frac{2}{n}$$

Como queremos manter a variância da camada subsequente igual a variância da camada anterior

$$\sigma_{s^{(l)}}^2 = \sigma_{s^{(l-1)}}^2$$

# Inicialização de Parâmetros

$$\theta \sim \mathcal{N}(0, 2/n)$$

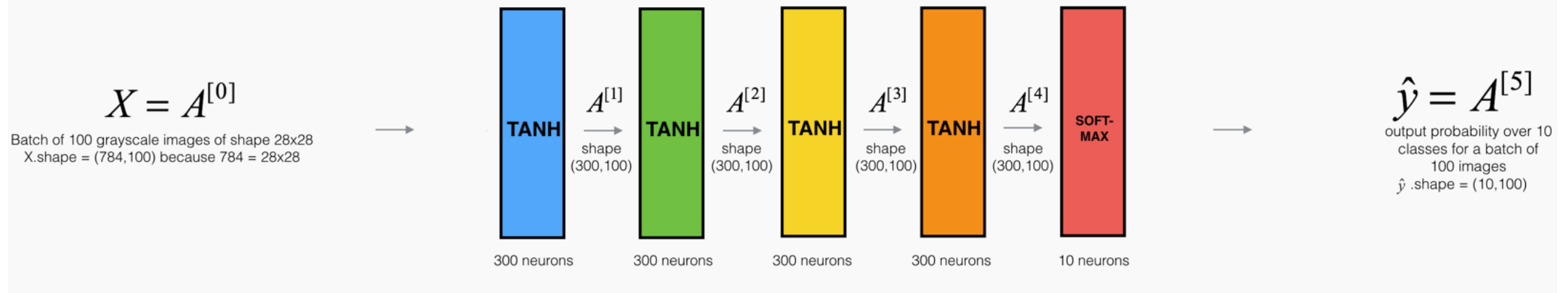
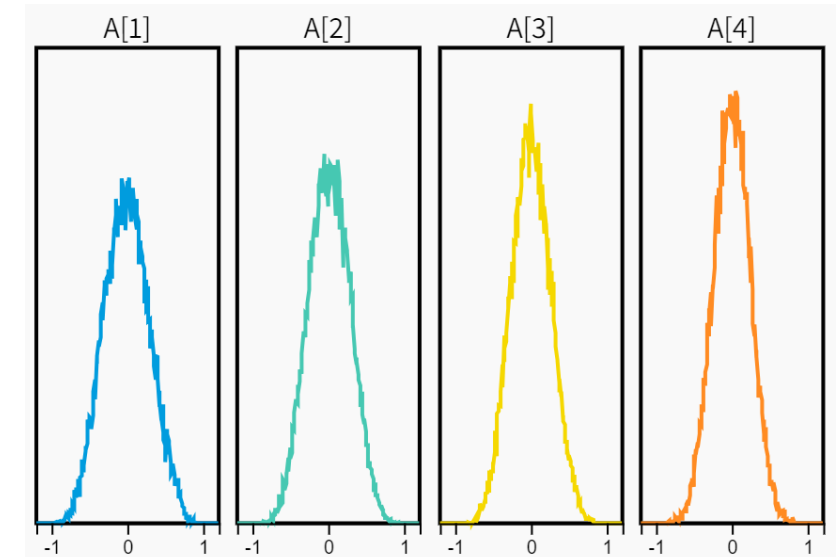
Essa inicialização é conhecida como *Kaiming* (He)

Um argumento similar pode ser feito usando a função de ativação tanh, levando a inicialização *Xavier*  $\theta \sim \mathcal{N}(0, 1/n)$

# Inicialização de Parâmetros



$$\theta \sim \mathcal{N}(0, 1/n)$$



# Inicialização de Parâmetros

- Tanto as ativações quanto os gradientes podem dissipar ou explodir
  - Nosso argumento levou em conta apenas o *forward pass*
- Para contabilizar tanto *forward* quanto *backward* podemos modificar a inicialização para considerar o número médio de unidades

$$\theta^l \sim \mathcal{N} \left( 0, \frac{4}{(n^{(l)} + n^{(l-1)})} \right)$$

# Exercício

- Considerando uma variável aleatória  $a$  cuja variância  $Var[a] = \sigma^2$  com uma distribuição simétrica em relação à média  $\mathbb{E}[a] = 0$
- Prove que se passarmos a variável por uma ReLU, o segundo momento da variável transformada  $\mathbb{E}[b^2] = \frac{\sigma^2}{2}$

$$b = ReLU(a) = \max(0, a)$$

Dicas:

$$\mathbb{E}[x] = \int_{-\infty}^{+\infty} xp(x)dx$$
$$Var[x] = \mathbb{E}[(x - \mathbb{E}[x])^2]$$

# Referências:

- Sugere-se ***fortemente*** a leitura de:
  - Capítulo 7 de Understanding Deep Learning
    - <https://udlbook.github.io/udlbook/>