



Tópicos en Inteligencia Artificial Deep Learning

Inicialización

Basado en el ppt 07a – Initialization de

Prof. Simon Prince

Adaptado por Prof. Fernando Crema García



Inicialización

- Necesidad de inicialización

- Inicialización **He**

- Interludio: Valores esperados

- Demostrar que $\mathbb{E}[f'_i] = 0$

- Escribir la varianza de las preactivaciones f' en términos de las activaciones h en la capa anterior

$$\sigma_{f'}^2 = \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \mathbb{E}[h_j^2]$$

- Escribir la varianza de las preactivaciones f' en términos de las preactivaciones f de la capa anterior

$$\sigma_{f'}^2 = \frac{D_h \sigma_{\Omega}^2 \sigma_f^2}{2}$$

Inicialización

- Consideremos el bloque de construcción estándar de la red neuronal en términos de preactivaciones:

$$\begin{aligned}\mathbf{f}_k &= \boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathbf{h}_k \\ &= \boldsymbol{\beta}_k + \boldsymbol{\Omega}_k a[\mathbf{f}_{k-1}]\end{aligned}$$

- ¿Cómo inicializamos los sesgos y los pesos?
- **Equivalente** a elegir el punto de partida en los modelos de regresión Gabor/Lineal

Inicialización

- Considere el bloque de construcción estándar de red neuronal en términos de *preactivaciones*:

$$\begin{aligned}\mathbf{f}_k &= \boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathbf{h}_k \\ &= \boldsymbol{\beta}_k + \boldsymbol{\Omega}_k a[\mathbf{f}_{k-1}]\end{aligned}$$

- Fijar todos los sesgos a 0

$$\boldsymbol{\beta}_k = \mathbf{0}$$

- Pesos distribuidos normalmente

- media 0
- Varianza σ_{Ω}^2

- ¿Qué ocurrirá a medida que avancemos por la red si σ_{Ω}^2 es muy pequeña?
- ¿Qué ocurrirá a medida que avancemos por la red si σ_{Ω}^2 es muy grande?

Resumen del backprop

Backward pass: We start with the derivative $\partial\ell_i/\partial\mathbf{f}_K$ of the loss function ℓ_i with respect to the network output $\mathbf{f}_K$ and work backward through the network:

$$\begin{aligned}\frac{\partial\ell_i}{\partial\boldsymbol{\beta}_k} &= \frac{\partial\ell_i}{\partial\mathbf{f}_k} & k \in \{K, K-1, \dots, 1\} \\ \frac{\partial\ell_i}{\partial\boldsymbol{\Omega}_k} &= \frac{\partial\ell_i}{\partial\mathbf{f}_k} \mathbf{h}_k^T & k \in \{K, K-1, \dots, 1\} \\ \frac{\partial\ell_i}{\partial\mathbf{f}_{k-1}} &= \mathbb{I}[\mathbf{f}_{k-1} > 0] \odot \left(\boldsymbol{\Omega}_k^T \frac{\partial\ell_i}{\partial\mathbf{f}_k} \right), & k \in \{K, K-1, \dots, 1\}\end{aligned}\tag{7.13}$$

where $\odot$ denotes pointwise multiplication and $\mathbb{I}[\mathbf{f}_{k-1} > 0]$ is a vector containing ones where $\mathbf{f}_{k-1}$ is greater than zero and zeros elsewhere. Finally, we compute the derivatives with respect to the first set of biases and weights:

$$\begin{aligned}\frac{\partial\ell_i}{\partial\boldsymbol{\beta}_0} &= \frac{\partial\ell_i}{\partial\mathbf{f}_0} \\ \frac{\partial\ell_i}{\partial\boldsymbol{\Omega}_0} &= \frac{\partial\ell_i}{\partial\mathbf{f}_0} \mathbf{x}_i^T\end{aligned}$$

Inicialización

- Necesidad de inicialización
- Inicialización **He**
- Interludio: Valores esperados
- Demostrar que $\mathbb{E}[f'_i] = 0$
- Escribir la varianza de las preactivaciones f' en términos de las activaciones h en la capa anterior

$$\sigma_{f'}^2 = \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \mathbb{E}[h_j^2]$$

- Escribir la varianza de las preactivaciones f' en términos de las preactivaciones f de la capa anterior

$$\sigma_{f'}^2 = \frac{D_h \sigma_{\Omega}^2 \sigma_f^2}{2}$$

Inicialización de los pesos

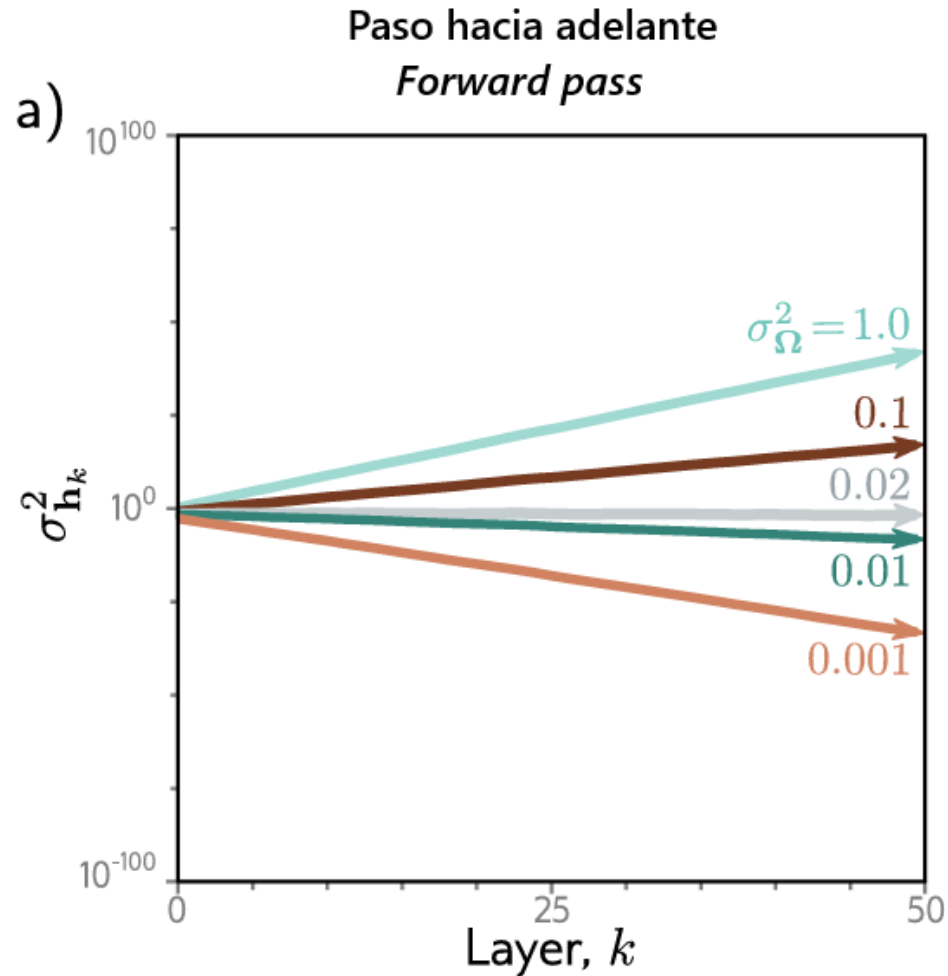


Figura 7.7

Setup

1) Número de capas $k = 50$

2) Número de neuronas por capa $D_h = 100$

3) El vector de entrada $X \sim N(0,1)^{100} = \begin{bmatrix} N(0,1) \\ \vdots \\ N(0,1) \end{bmatrix}$

4) El vector de salida es fijo con $Y = [0]^{100} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$

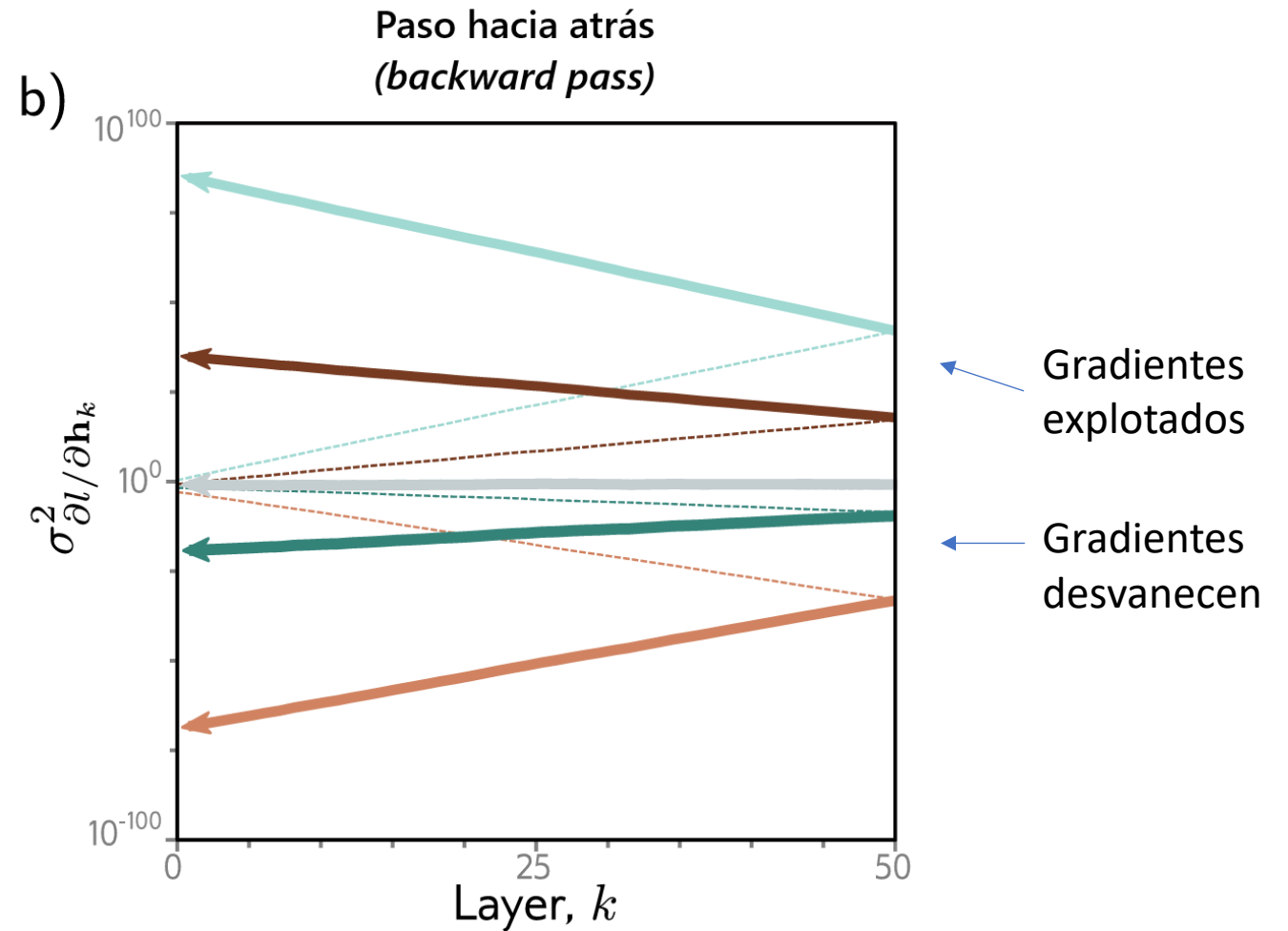
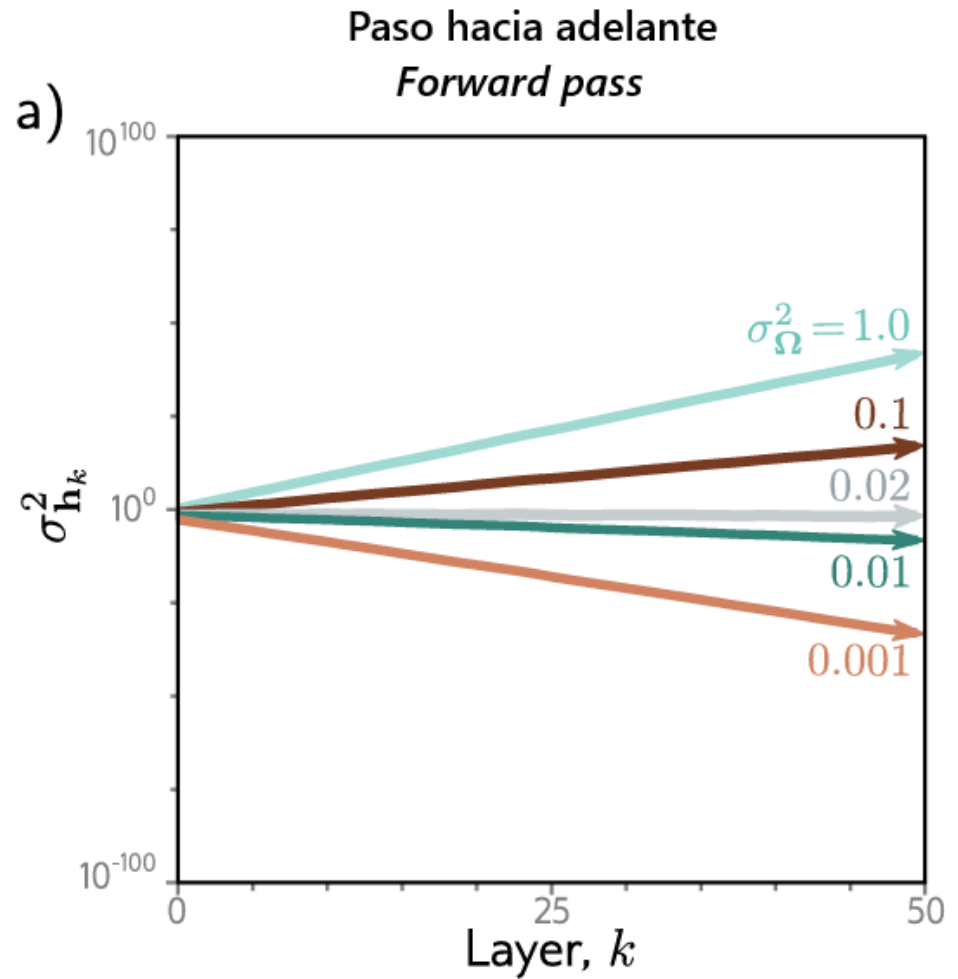
5) El vector de sesgo $\beta_k = [0]^{100} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$

6) La matriz de pesos

$$\Omega_k \sim N(0,1)^{100 \times 100} = \begin{bmatrix} N_{1,1}(0, \sigma_{\Omega}^2) & \cdots & N_{1,100}(0, \sigma_{\Omega}^2) \\ \vdots & \ddots & \vdots \\ N_{100,0}(0, \sigma_{\Omega}^2) & \cdots & N_{100,100}(0, \sigma_{\Omega}^2) \end{bmatrix}$$

7) Las varianzas $\sigma_{\Omega}^2 \in \{0.001, 0.01, 0.02, 0.1, 1\}$

Inicialización de los pesos



Inicialización He (asume ReLU como activación)

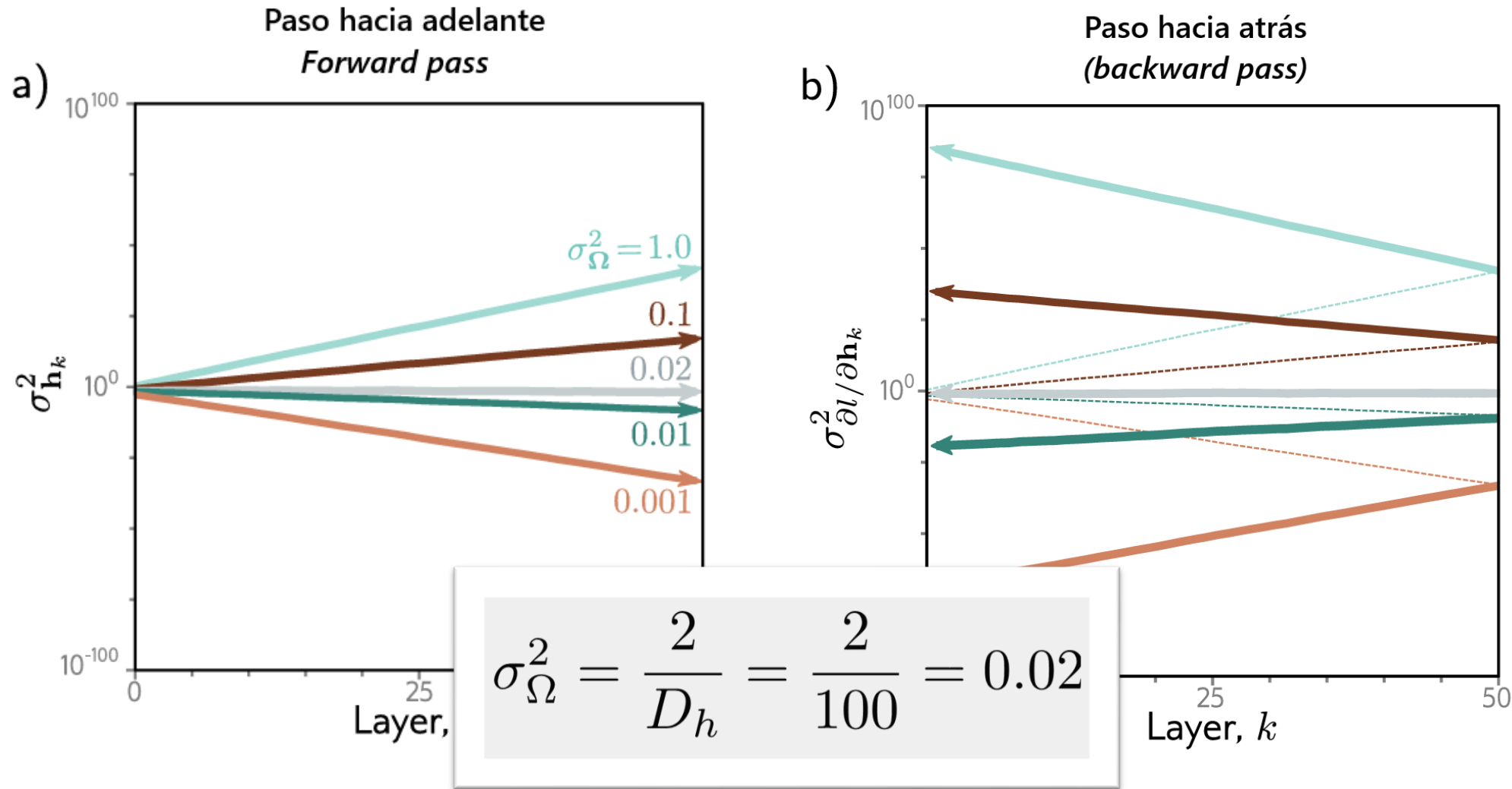
- Paso hacia adelante: desea que la varianza de las activaciones de la unidad oculta en la capa k+1 sea la misma que la varianza de las activaciones en la capa k:

$$\sigma_{\Omega}^2 = \frac{2}{D_h} \quad \leftarrow \text{Número de unidades en la capa k}$$

- Paso hacia atrás: se desea que la varianza de los gradientes en la capa k sea la misma que la varianza del gradiente en la capa k+1:

$$\sigma_{\Omega}^2 = \frac{2}{D_{h'}} \quad \leftarrow \text{Número de unidades en la capa k+1}$$

Inicialización de los pesos



Fíjense cómo
para $\sigma_{\Omega}^2 = 0.02$ en la
línea gris
para ambos
pases los
valores son
estables

Inicialización

- Necesidad de inicialización
- Inicialización **He**
- Interludio: Valores esperados
- Demostrar que $\mathbb{E}[f'_i] = 0$
- Escribir la varianza de las preactivaciones f' en términos de las activaciones h en la capa anterior

$$\sigma_{f'}^2 = \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \mathbb{E}[h_j^2]$$

- Escribir la varianza de las preactivaciones f' en términos de las preactivaciones f de la capa anterior

$$\sigma_{f'}^2 = \frac{D_h \sigma_{\Omega}^2 \sigma_f^2}{2}$$

Valores esperados

$$\mathbb{E}[g[x]] = \int g[x] Pr(x) dx,$$

Interpretación: ¿Cuál es el valor medio de $g[x]$ teniendo en cuenta la probabilidad de x ?

Podría aproximarse, muestreando muchos valores de x de la distribución, calculando $g[x]$, y tomando la media:

$$\mathbb{E}[g[x]] \approx \frac{1}{N} \sum_{n=1}^N g[x_n^*] \quad \text{where} \quad x_n^* \sim Pr(x)$$

Valores esperados

$$\mathbb{E}[g[x]] = \int g[x] Pr(x) dx,$$

Interpretación: ¿Cuál es el valor medio de $g[x]$ teniendo en cuenta la probabilidad de x ?

Podría aproximarse, muestreando muchos valores de x de la distribución, calculando $g[x]$, y tomando la media:

$$\mathbb{E}[g[x]] \approx \frac{1}{N} \sum_{n=1}^N g[x_n^*] \quad \text{where} \quad x_n^* \sim Pr(x)$$

“Integrales de Riemann”

Valores esperados

$$\mathbb{E}[g[x]] = \int g[x] Pr(x) dx,$$

Interpretación: ¿Cuál es el valor medio de $g[x]$ teniendo en cuenta la probabilidad de x ?

Podría aproximarse, muestreando muchos valores de x de la distribución, calculando $g[x]$, y tomando la media:

$$\mathbb{E}[g[x]] \approx \frac{1}{N} \sum_{n=1}^N g[x_n^*] \quad \text{where} \quad x_n^* \sim Pr(x)$$

Valores esperados

Function $g[\bullet]$	Valor esperado
x	media, μ
x^k	k – <i>ésimo</i> momento alrededor de cero
$(x - \mu)^k$	k – <i>ésimo</i> momento alrededor de la media
$(x - \mu)^2$	varianza
$(x - \mu)^3$	Asimetría (skew)
$(x - \mu)^4$	Curtosis (kurtosis)

Propiedades para manipular Valores esperados

$$\mathbb{E}[k] = k$$

$$\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$$

$$\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$$

$$\mathbb{E}[f[x]g[y]] = \mathbb{E}[f[x]]\mathbb{E}[g[y]] \quad \text{if } x, y \text{ independent}$$

Regla 1

$$\mathbb{E}[g[x]] = \int g[x]Pr(x)dx,$$

$$\begin{aligned}\mathbb{E}[\kappa] &= \int \kappa Pr(x)dx \\ &= \kappa \int Pr(x)dx \\ &= \kappa.\end{aligned}$$

Regla 2

$$\mathbb{E}[g[x]] = \int g[x]Pr(x)dx,$$

$$\begin{aligned}\mathbb{E}[\kappa \cdot g[x]] &= \int \kappa \cdot g[x]Pr(x)dx \\ &= \kappa \cdot \int g[x]Pr(x)dx \\ &= \kappa \cdot \mathbb{E}[g[x]]\end{aligned}$$

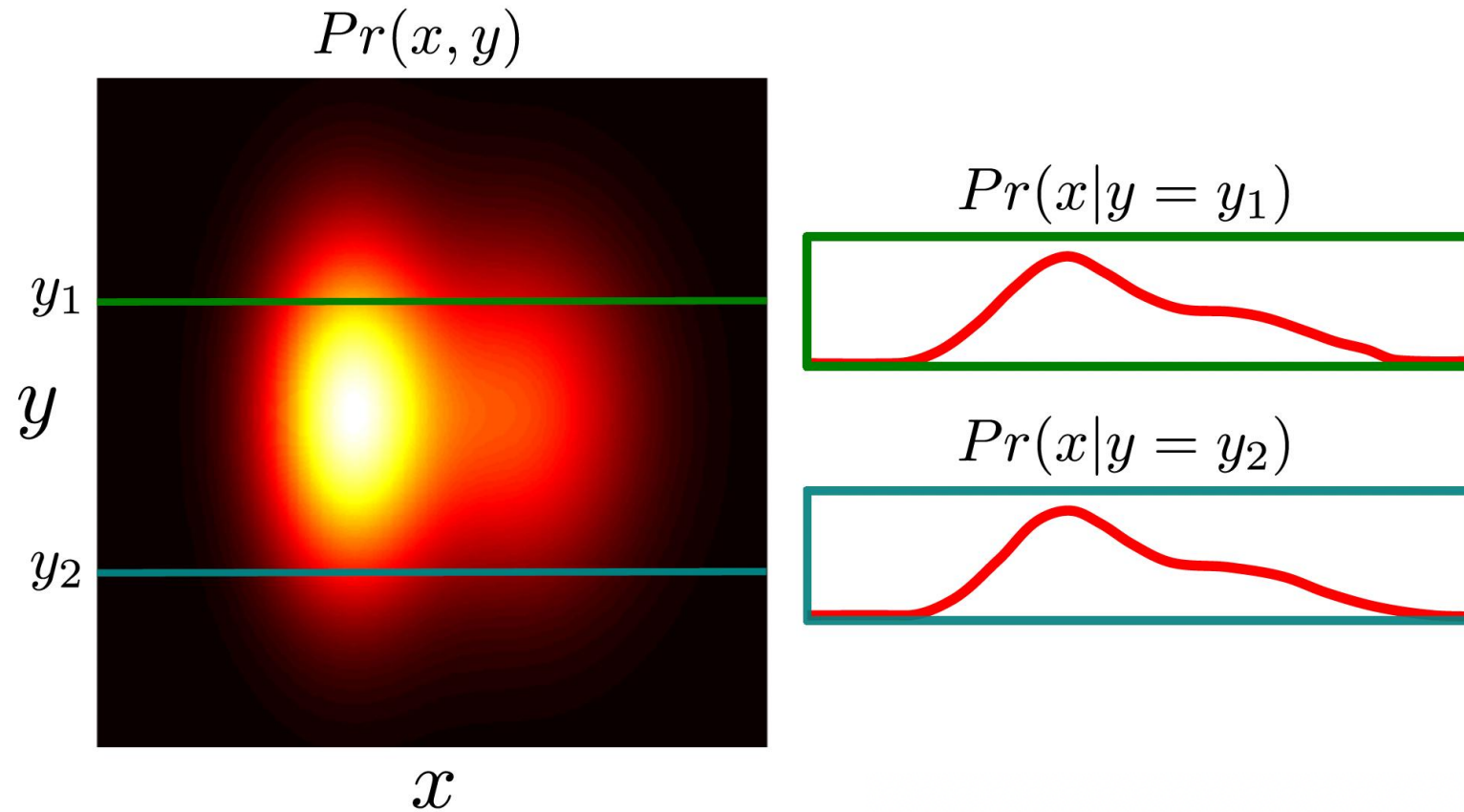
Regla 3

$$\mathbb{E}[g[x]] = \int g[x]Pr(x)dx,$$

$$\begin{aligned}\mathbb{E}[f[x] + g[x]] &= \int (f[x] + g[x])Pr(x)dx \\ &= \int (f[x]Pr(x) + g[x]Pr(x)) dx \\ &= \int f[x]Pr(x)dx + \int g[x]Pr(x)dx \\ &= \mathbb{E}[f[x]] + \mathbb{E}[g[x]]\end{aligned}$$



Independencia

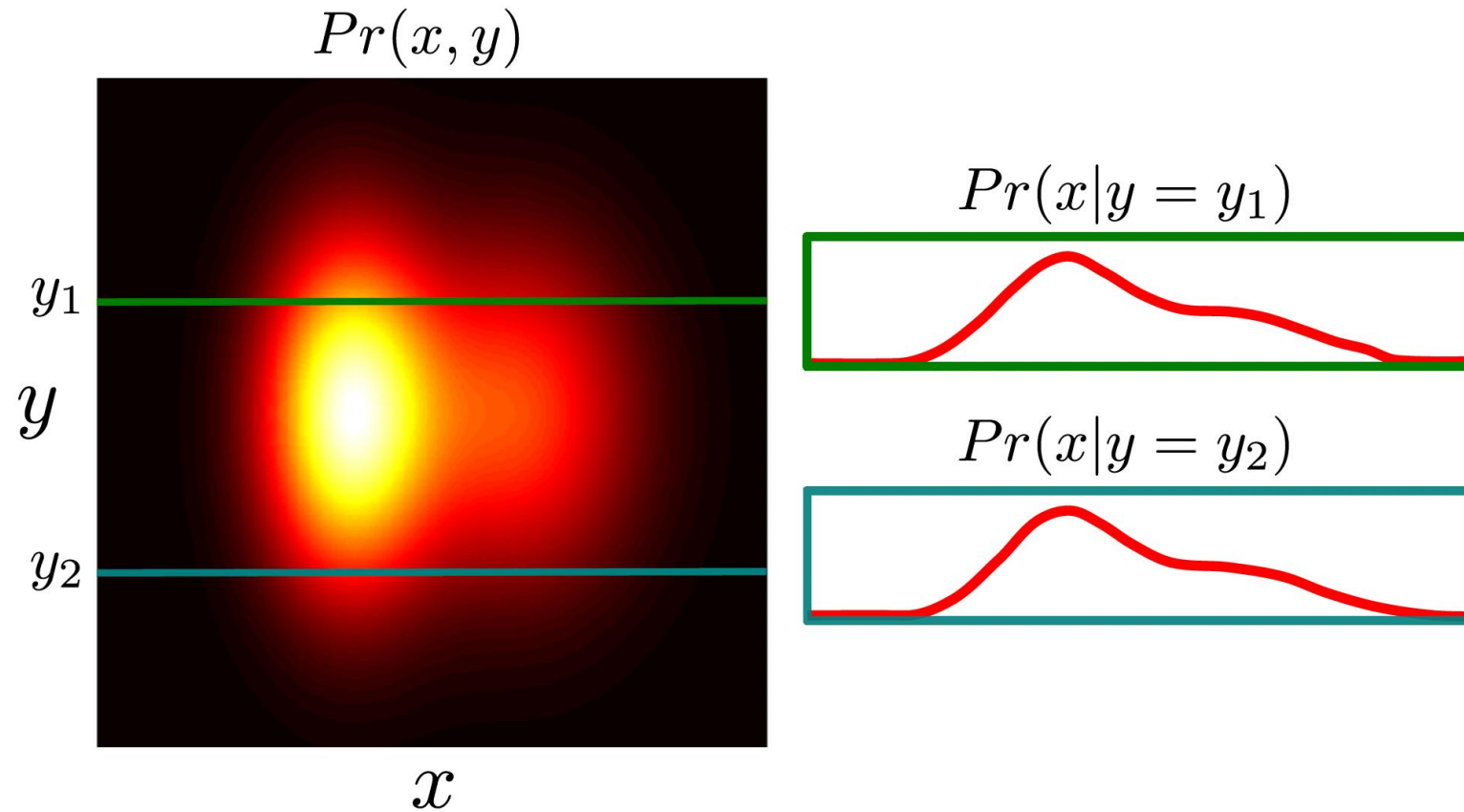


Probabilidad de x e y

$$Pr(x|y) = Pr(x)$$

$$Pr(y|x) = Pr(y)$$

Independencia



$$Pr(x, y) = Pr(x)Pr(y)$$

Probabilidad de x e y

Regla 4

$$\mathbb{E}[g[x]] = \int g[x] Pr(x) dx,$$

$$\begin{aligned}\mathbb{E}[f[x] \cdot g[y]] &= \int \int f[x] \cdot g[y] Pr(x, y) dx dy \\ &= \int \int f[x] \cdot g[y] Pr(x) Pr(y) dx dy \quad \leftarrow \text{Porque independientes} \\ &= \int f[x] Pr(x) dx \int g[y] Pr(y) dy \\ &= \mathbb{E}[f[x]] \mathbb{E}[g[y]] \quad \text{if } x, y \text{ independent}\end{aligned}$$

Ahora probemos:

$$\mathbb{E} [(x - \mu)^2] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

Teniendo en cuenta:

$$\mathbb{E}[x] = \mu$$

Ahora probemos:

$$\mathbb{E} [(x - \mu)^2] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

Teniendo en cuenta:

$$\mathbb{E}[x] = \mu$$

Regla 1: $\mathbb{E}[k] = k$

Regla 2: $\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$

Regla 3: $\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$

Definición $\mathbb{E}[x] = \mu$

$$\mathbb{E}[(x - \mu)^2] = \mathbb{E}[x^2 - 2x\mu + \mu^2]$$

Regla 1: $\mathbb{E}[k] = k$

Regla 2: $\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$

Regla 3: $\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$

Definición $\mathbb{E}[x] = \mu$



$$\begin{aligned}\mathbb{E}[(x - \mu)^2] &= \mathbb{E}[x^2 - 2x\mu + \mu^2] \\ &= \mathbb{E}[x^2] - \mathbb{E}[2x\mu] + \mathbb{E}[\mu^2]\end{aligned}$$

Regla 1: $\mathbb{E}[k] = k$

Regla 2: $\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$

Regla 3: $\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$

Definición $\mathbb{E}[x] = \mu$

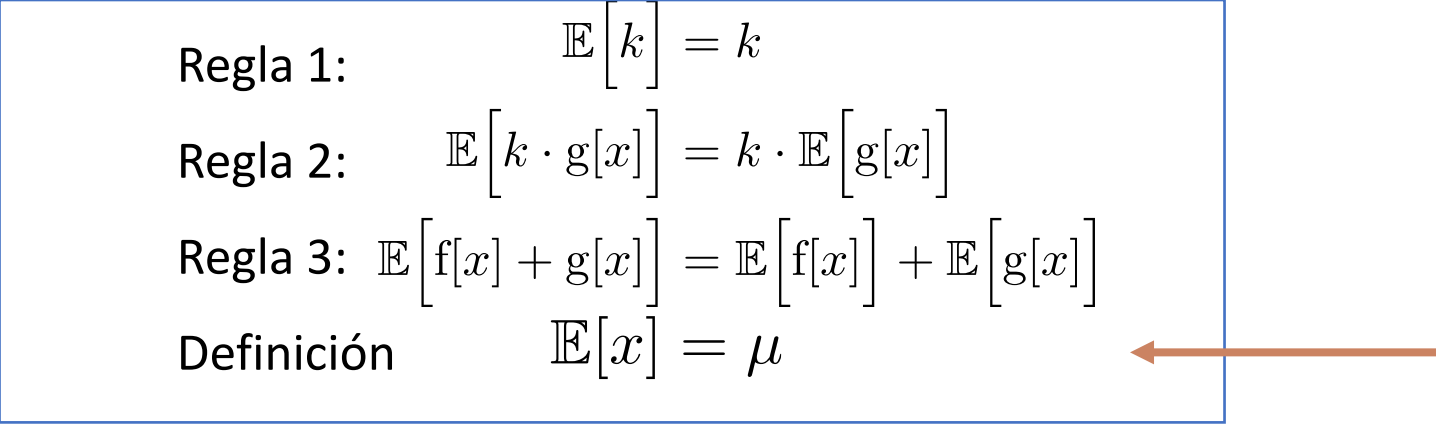
$$\begin{aligned}\mathbb{E}[(x - \mu)^2] &= \mathbb{E}[x^2 - 2x\mu + \mu^2] \\ &= \mathbb{E}[x^2] - \mathbb{E}[2x\mu] + \mathbb{E}[\mu^2] \\ &= \mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2\end{aligned}$$

Regla 1: $\mathbb{E}[k] = k$

Regla 2: $\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$

Regla 3: $\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$

Definición $\mathbb{E}[x] = \mu$



$$\begin{aligned}\mathbb{E}[(x - \mu)^2] &= \mathbb{E}[x^2 - 2x\mu + \mu^2] \\ &= \mathbb{E}[x^2] - \mathbb{E}[2x\mu] + \mathbb{E}[\mu^2] \\ &= \mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 \\ &= \mathbb{E}[x^2] - 2\mu^2 + \mu^2\end{aligned}$$

Regla 1: $\mathbb{E}[k] = k$

Regla 2: $\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$

Regla 3: $\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$

Definición $\mathbb{E}[x] = \mu$

$$\begin{aligned}\mathbb{E}[(x - \mu)^2] &= \mathbb{E}[x^2 - 2x\mu + \mu^2] \\ &= \mathbb{E}[x^2] - \mathbb{E}[2x\mu] + \mathbb{E}[\mu^2] \\ &= \mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 \\ &= \mathbb{E}[x^2] - 2\mu^2 + \mu^2 \\ &= \mathbb{E}[x^2] - \mu^2\end{aligned}$$

Regla 1: $\mathbb{E}[k] = k$

Regla 2: $\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$

Regla 3: $\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$

Definición $\mathbb{E}[x] = \mu$



$$\begin{aligned}\mathbb{E}[(x - \mu)^2] &= \mathbb{E}[x^2 - 2x\mu + \mu^2] \\ &= \mathbb{E}[x^2] - \mathbb{E}[2x\mu] + \mathbb{E}[\mu^2] \\ &= \mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 \\ &= \mathbb{E}[x^2] - 2\mu^2 + \mu^2 \\ &= \mathbb{E}[x^2] - \mu^2 \\ &= \mathbb{E}[x^2] - E[x]^2\end{aligned}$$

Inicialización

- Necesidad de inicialización
- Inicialización **He**
- Interludio: Valores esperados
- **Demostrar que** $\mathbb{E}[f'_i] = 0$
- Escribir la varianza de las preactivaciones f' en términos de las activaciones h en la capa anterior

$$\sigma_{f'}^2 = \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \mathbb{E}[h_j^2]$$

- Escribir la varianza de las preactivaciones f' en términos de las preactivaciones f de la capa anterior

$$\sigma_{f'}^2 = \frac{D_h \sigma_{\Omega}^2 \sigma_f^2}{2}$$

Inicialización

- Considere el bloque de construcción estándar de red neuronal en términos de *preactivaciones*:

$$\begin{aligned}\mathbf{f}_k &= \boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathbf{h}_k \\ &= \boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathbf{a}[\mathbf{f}_{k-1}]\end{aligned}$$

- Fijar todos los sesgos a 0

$$\boldsymbol{\beta}_k = \mathbf{0}$$

- Pesos distribuidos normalmente

- media 0
- varianza σ_{Ω}^2

- ¿Qué ocurrirá a medida que avancemos por la red si es muy pequeño?
- ¿Qué ocurrirá a medida que avancemos por la red si es muy grande?

Objetivo: mantener la misma varianza entre dos capas

$$\mathbf{f}' = \boldsymbol{\beta} + \boldsymbol{\Omega}\mathbf{h}$$

Considerar la media de las preactivaciones:

$$\mathbb{E}[f'_i] = \mathbb{E} \left[\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j \right]$$

- Regla 1: $\mathbb{E}[k] = k$
- Regla 2: $\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$
- Regla 3: $\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$
- Regla 4: $\mathbb{E}[f[x]g[y]] = \mathbb{E}[f[x]]\mathbb{E}[g[y]]$ if x, y independent
- 

$$\begin{aligned}\mathbb{E}[f'_i] &= \mathbb{E} \left[\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j \right] \\ &= \mathbb{E} [\beta_i] + \sum_{j=1}^{D_h} \mathbb{E} [\Omega_{ij} h_j]\end{aligned}$$

Regla 1:	$\mathbb{E}[k] = k$
Regla 2:	$\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$
Regla 3:	$\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$
Regla 4:	$\mathbb{E}[f[x]g[y]] = \mathbb{E}[f[x]]\mathbb{E}[g[y]]$ if x, y independent



$$\begin{aligned}
 \mathbb{E}[f'_i] &= \mathbb{E} \left[\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j \right] \\
 &= \mathbb{E} [\beta_i] + \sum_{j=1}^{D_h} \mathbb{E} [\Omega_{ij} h_j] \\
 &= \mathbb{E} [\beta_i] + \sum_{j=1}^{D_h} \mathbb{E} [\Omega_{ij}] \mathbb{E} [h_j]
 \end{aligned}$$

Regla 1:	$\mathbb{E}[k] = k$
Regla 2:	$\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$
Regla 3:	$\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$
Regla 4:	$\mathbb{E}[f[x]g[y]] = \mathbb{E}[f[x]]\mathbb{E}[g[y]]$ if x, y independent

$$\mathbb{E}[f'_i] = \mathbb{E} \left[\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j \right]$$

$$= \mathbb{E} [\beta_i] + \sum_{j=1}^{D_h} \mathbb{E} [\Omega_{ij} h_j]$$

Poner todos los sesgos a 0

$$= \mathbb{E} [\beta_i] + \sum_{j=1}^{D_h} \mathbb{E} [\Omega_{ij}] \mathbb{E} [h_j]$$

Ponderaciones distribuidas normalmente

media 0

varianza σ_{Ω}^2

$$= 0 + \sum_{j=1}^{D_h} 0 \cdot \mathbb{E} [h_j] = 0$$

Inicialización

- Necesidad de inicialización
- Inicialización **He**
- Interludio: Valores esperados
- Demostrar que $\mathbb{E}[f'_i] = 0$
- Escribir la varianza de las preactivaciones f' en términos de las activaciones h en la capa anterior

$$\sigma_{f'}^2 = \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \mathbb{E}[h_j^2]$$

- Escribir la varianza de las preactivaciones f' en términos de las preactivaciones f de la capa anterior

$$\sigma_{f'}^2 = \frac{D_h \sigma_{\Omega}^2 \sigma_f^2}{2}$$

Objetivo: mantener la misma varianza entre dos capas

$$\mathbf{f}' = \boldsymbol{\beta} + \boldsymbol{\Omega}\mathbf{h}$$

$$\mathbf{h} = \mathbf{a}[\mathbf{f}],$$

$$\sigma_{f'}^2 = \mathbb{E}[f_i'^2] - \mathbb{E}[f_i']^2$$

$$\longrightarrow \mathbb{E}[(x - \mu)^2] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

Regla 1:	$\mathbb{E}[k] = k$
Regla 2:	$\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$
Regla 3:	$\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$
Regla 4:	$\mathbb{E}[f[x]g[y]] = \mathbb{E}[f[x]]\mathbb{E}[g[y]]$ if x, y independent

$$\begin{aligned}\sigma_{f'}^2 &= \mathbb{E}[f_i'^2] - \mathbb{E}[f_i']^2 \\ &= \mathbb{E} \left[\left(\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j \right)^2 \right] - 0\end{aligned}$$

Poner todos los sesgos a 0

Ponderaciones distribuidas normalmente

media 0

varianza σ_{Ω}^2

Regla 1:	$\mathbb{E}[k] = k$
Regla 2:	$\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$
Regla 3:	$\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$
Regla 4:	$\mathbb{E}[f[x]g[y]] = \mathbb{E}[f[x]]\mathbb{E}[g[y]]$ if x, y independent

$$\begin{aligned}\sigma_{f'}^2 &= \mathbb{E}[f_i'^2] - \mathbb{E}[f_i']^2 \\ &= \mathbb{E} \left[\left(\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j \right)^2 \right] - 0 \\ &= \mathbb{E} \left[\left(\sum_{j=1}^{D_h} \Omega_{ij} h_j \right)^2 \right]\end{aligned}$$

Poner todos los sesgos a 0



Ponderaciones distribuidas normalmente

media 0

varianza σ_{Ω}^2

Regla 1:	$\mathbb{E}[k] = k$
Regla 2:	$\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$
Regla 3:	$\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$
Regla 4:	$\mathbb{E}[f[x]g[y]] = \mathbb{E}[f[x]]\mathbb{E}[g[y]]$ if x, y independent



$$\begin{aligned}
 \sigma_{f'}^2 &= \mathbb{E}[f_i'^2] - \mathbb{E}[f_i']^2 \\
 &= \mathbb{E} \left[\left(\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j \right)^2 \right] - 0 \\
 &= \mathbb{E} \left[\left(\sum_{j=1}^{D_h} \Omega_{ij} h_j \right)^2 \right] \\
 &= \sum_{j=1}^{D_h} \mathbb{E} [\Omega_{ij}^2] \mathbb{E} [h_j^2]
 \end{aligned}$$

Poner todos los sesgos a 0

Ponderaciones distribuidas normalmente
media 0
varianza σ_{Ω}^2

Regla 1:	$\mathbb{E}[k] = k$
Regla 2:	$\mathbb{E}[k \cdot g[x]] = k \cdot \mathbb{E}[g[x]]$
Regla 3:	$\mathbb{E}[f[x] + g[x]] = \mathbb{E}[f[x]] + \mathbb{E}[g[x]]$
Regla 4:	$\mathbb{E}[f[x]g[y]] = \mathbb{E}[f[x]]\mathbb{E}[g[y]]$ if x, y independent

$$\begin{aligned}\sigma_{f'}^2 &= \mathbb{E}[f_i'^2] - \mathbb{E}[f_i']^2 \\ &= \mathbb{E} \left[\left(\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j \right)^2 \right] - 0 \\ &= \mathbb{E} \left[\left(\sum_{j=1}^{D_h} \Omega_{ij} h_j \right)^2 \right]\end{aligned}$$

Poner todos los sesgos a 0

Ponderaciones distribuidas normalmente
media 0
varianza σ_{Ω}^2

$$\begin{aligned}&= \sum_{j=1}^{D_h} \mathbb{E} [\Omega_{ij}^2] \mathbb{E} [h_j^2] \\ &= \sum_{j=1}^{D_h} \sigma_{\Omega}^2 \mathbb{E} [h_j^2] = \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \mathbb{E} [h_j^2]\end{aligned}$$

Inicialización

- Necesidad de inicialización
- Inicialización **He**
- Interludio: Valores esperados
- Demostrar que $\mathbb{E}[f'_i] = 0$
- Escribir la varianza de las preactivaciones f' en términos de las activaciones h en la capa anterior

$$\sigma_{f'}^2 = \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \mathbb{E}[h_j^2]$$

- Escribir la varianza de las preactivaciones f' en términos de las preactivaciones f de la capa anterior

$$\sigma_{f'}^2 = \frac{D_h \sigma_{\Omega}^2 \sigma_f^2}{2}$$

$$\begin{aligned}
\sigma_{f'}^2 &= \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \mathbb{E} [h_j^2] \\
&= \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \mathbb{E} [\text{ReLU}[f_j]^2] \\
&= \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \int_{-\infty}^{\infty} \text{ReLU}[f_j]^2 Pr(f_j) df_j \\
&= \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \int_{-\infty}^{\infty} (\mathbb{I}[f_j > 0] f_j)^2 Pr(f_j) df_j \\
&= \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \int_0^{\infty} f_j^2 Pr(f_j) df_j \\
&= \sigma_{\Omega}^2 \sum_{j=1}^{D_h} \frac{\sigma_f^2}{2} = \frac{D_h \sigma_{\Omega}^2 \sigma_f^2}{2}
\end{aligned}$$

Problema 7.14

Objetivo:

mantener la misma varianza entre dos capas

$$\sigma_{f'}^2 = \frac{D_h \sigma_{\Omega}^2 \sigma_f^2}{2}$$

Debería elegir:

$$\sigma_{\Omega}^2 = \frac{2}{D_h}$$

Esto se llama **inicialización He**.

Objetivo:

mantener la misma varianza entre dos capas

$$\sigma_{f'}^2 = \frac{D_h \sigma_{\Omega}^2 \sigma_f^2}{2}$$

Buscamos $\sigma_{f'}^2 = \sigma_f^2$ si asumimos que $\sigma_{f'}^2 = \sigma_f^2 > 0$

Basta con hacer $\sigma_{f'}^2 = \sigma_f^2 = a$ y tenemos

$$\sigma_{f'}^2 = \frac{\sigma_f^2 D_h \sigma_{\Omega}^2}{2} \equiv a = \frac{a D_h \sigma_{\Omega}^2}{2} \text{ que implica entonces que}$$

$$\sigma_{\Omega}^2 = \frac{2}{D_h}$$

A esto lo llamamos **inicialización He**.

Leer más

Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. **Kaiming He** et al. Microsoft

Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun

Microsoft Research
{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

This product is the key to the initialization design. A proper initialization method should avoid reducing or magnifying the magnitudes of input signals exponentially. So we expect the above product to take a proper scalar (*e.g.*, 1). A sufficient condition is:

$$\frac{1}{2}n_l \text{Var}[w_l] = 1, \quad \forall l. \quad (10)$$

This leads to a zero-mean Gaussian distribution whose standard deviation (std) is $\sqrt{2/n_l}$. This is our way of initialization. We also initialize $\mathbf{b} = 0$.

For the first layer ($l = 1$), we should have $n_1 \text{Var}[w_1] = 1$ because there is no ReLU applied on the input signal. But the factor $1/2$ does not matter if it just exists on one layer. So we also adopt Eqn.(10) in the first layer for simplicity.

Objetivo:

mantener la misma varianza entre dos capas

$$\sigma_{f'}^2 = \frac{D_h \sigma_{\Omega}^2 \sigma_f^2}{2}$$

Cuando las dimensiones de las capas lineales no son cuadradas. Es decir

$$D_h \neq D_{h'}$$

Podemos hacer el compromiso con usando el promedio $\frac{D_h + D_{h'}}{2}$ por lo que

$$\sigma_{\Omega}^2 = \frac{4}{D_h + D_{h'}}$$

A esto lo llamamos **inicialización He**.

Inicialización Xavier

- Glorot & Bengio, "Understanding the difficulty of training deep feedforward neural networks", AISTATS 2010.

- Xavier Uniform

$$W \sim \mathcal{U}(-a, a), \quad a = \sqrt{\frac{6}{\text{fan_in} + \text{fan_out}}}$$

- Xavier Normal

$$W \sim \mathcal{N}\left(0, \sqrt{\frac{2}{\text{fan_in} + \text{fan_out}}}\right)$$

Inicialización Xavier

- Xavier Uniform $W \sim \mathcal{U}(-a, a), \quad a = \sqrt{\frac{6}{\text{fan_in} + \text{fan_out}}}$

Si hacemos $W \sim \text{Unif}(-a, a)$ entonces la varianza de los pesos es

$$\text{Var}(W) = \frac{(b - a)^2}{12} = \frac{(2a)^2}{12} = \frac{4a^2}{12} = \frac{a^2}{3}$$

Si entonces queremos igualar la varianza a $\frac{1}{\text{fan_in}}$

$$\frac{a^2}{3} = \frac{1}{\text{fan_in}} \Rightarrow a = \sqrt{3} \cdot \frac{1}{\sqrt{\text{fan_in}}}$$

Inicialización Xavier

- Xavier Uniform $W \sim \mathcal{U}(-a, a), \quad a = \sqrt{\frac{6}{\text{fan_in} + \text{fan_out}}}$

Si hacemos $W \sim \text{Unif}(-a, a)$ entonces la varianza de los pesos es

$$\text{Var}(W) = \frac{(b - a)^2}{12} = \frac{(2a)^2}{12} = \frac{4a^2}{12} = \frac{a^2}{3}$$

Si entonces queremos igualar la varianza a $\frac{2}{\text{fan_in} + \text{fan_out}}$

$$\frac{a^2}{3} = \frac{2}{\text{fan_in} + \text{fan_out}} = \sqrt{\left(\frac{6}{\text{fan_in} + \text{fan_out}}\right)}$$

Leer más

- “Understanding the difficulty of training deep feedforward neural networks.” Xavier Glorot and Yoshua Bengio. DIRO.

Understanding the difficulty of training deep feedforward neural networks

Xavier Glorot

DIRO, Université de Montréal, Montréal, Québec, Canada

Yoshua Bengio

We initialized the biases to be 0 and the weights W_{ij} at each layer with the following commonly used heuristic:

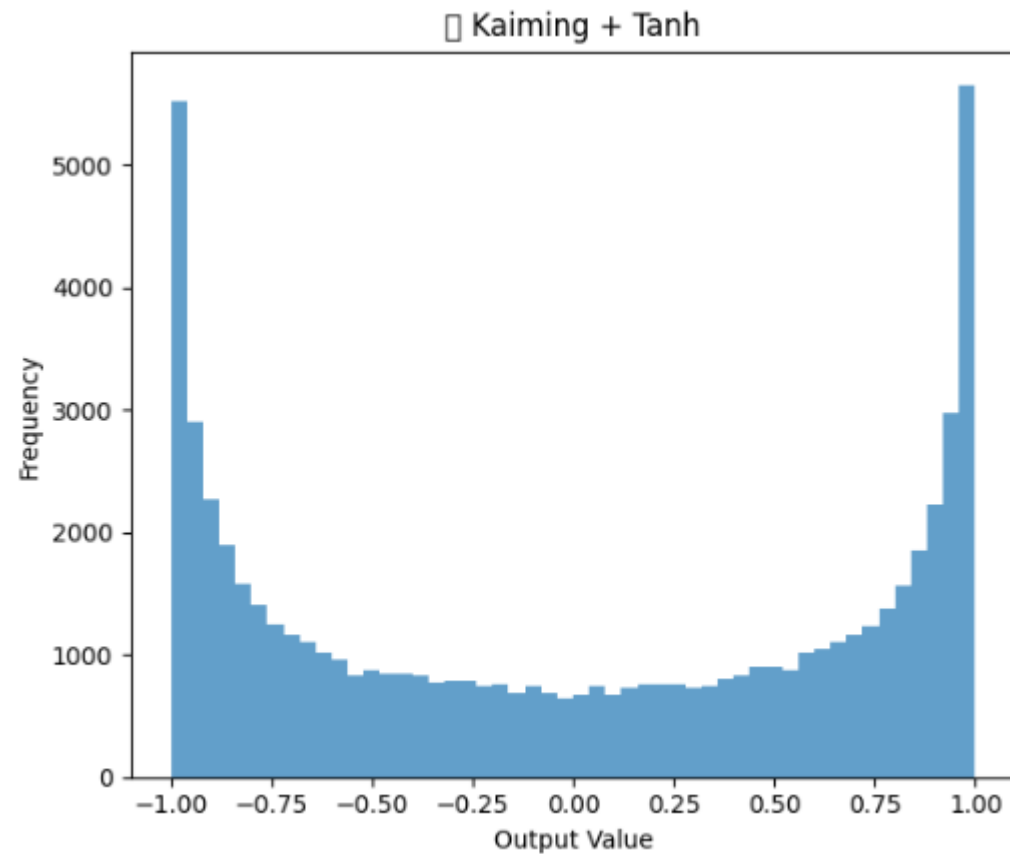
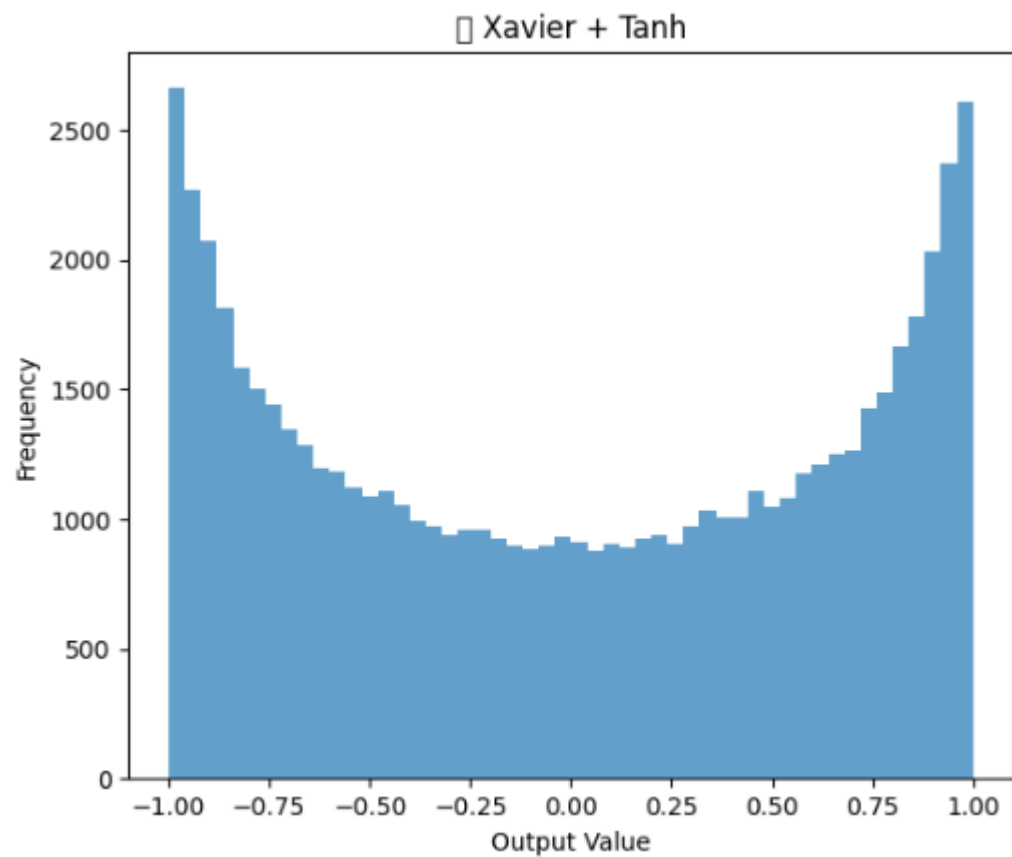
$$W_{ij} \sim U\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right], \quad (1)$$

where $U[-a, a]$ is the uniform distribution in the interval $(-a, a)$ and n is the size of the previous layer (the number of columns of W).

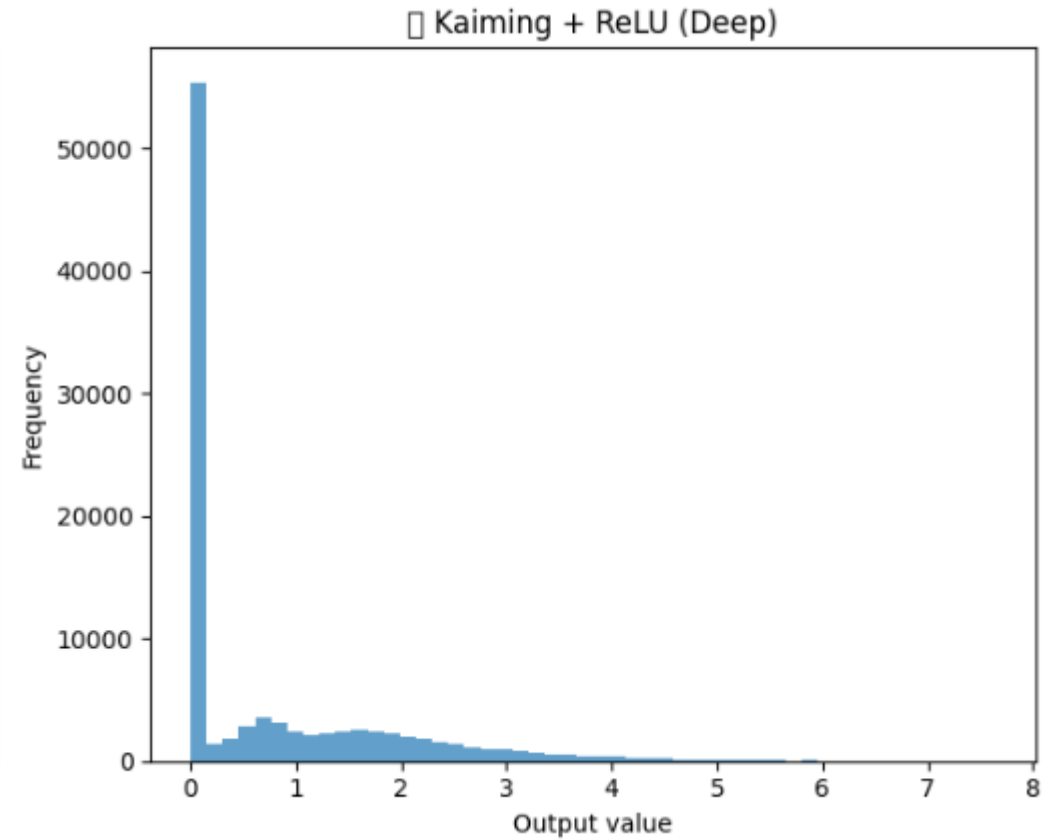
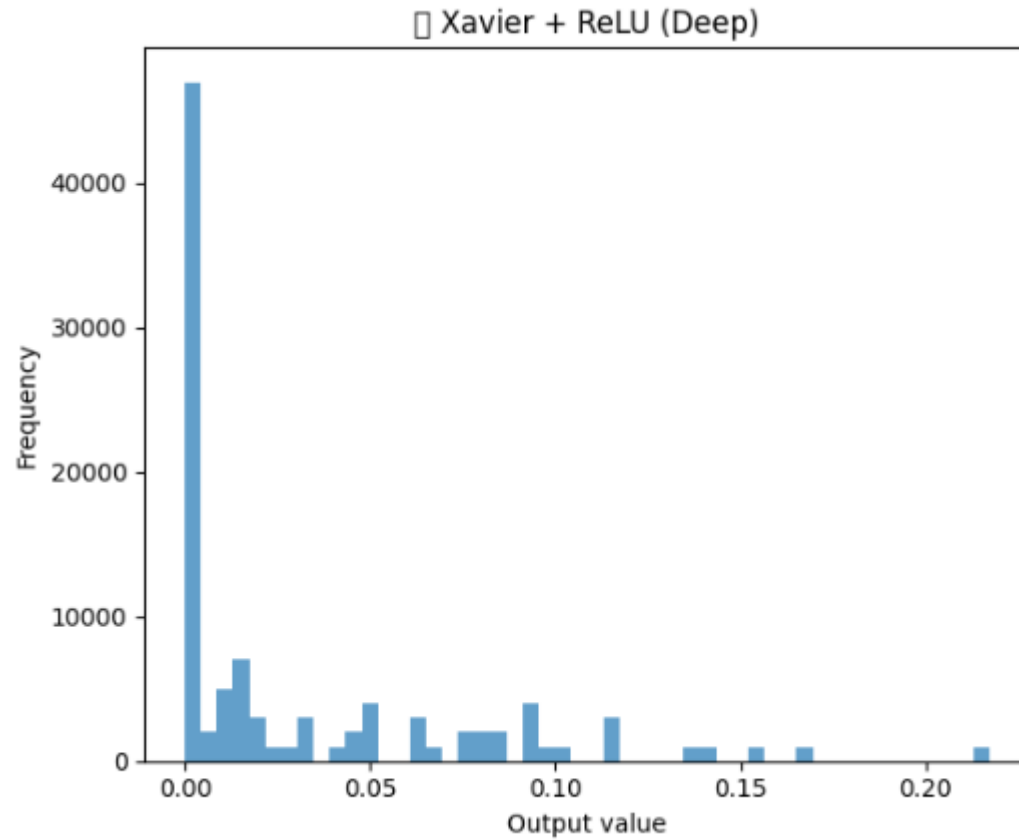
Resumen

Inicialización	Activación	Salida esperada
Xavier + tanh	Ideal	Estable. Valores balanceados
Kaiming + ReLU	Ideal	Buena varianza, pocas neuronas muertas
Xavier + ReLU	Riesgoso	Riesgo de neuronas muertas (varianza baja)
Kaiming + tanh	Riesgoso	Puede ocasionar explosión de los gradientes

Experimentos



Experimentos



En PyTorch

- Inicialización He
https://docs.pytorch.org/docs/stable/nn.init.html#torch.nn.init.kaiming_uniform
- Inicialización Xavier
https://docs.pytorch.org/docs/stable/nn.init.html#torch.nn.init.xavier_uniform