Tópicos en Inteligencia Artificial
Deep Learning

Inicialización

Basado en el **ppt 07a – Initialization** de

Prof. Simon Prince

Adaptado por Prof. Fernando Crema García

Inicialización

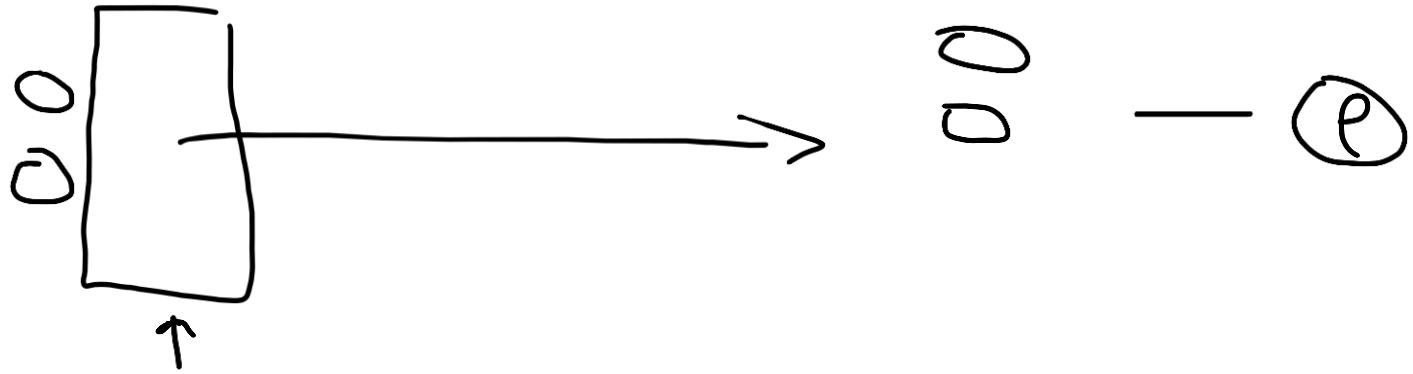
- Necesidad de inicialización
- Inicialización **He**
- Interludio: Valores esperados
- Demostrar que $\mathbb{E}[f_i'] = 0$
- Escribir la varianza de las preactivaciones *f'* en términos de las activaciones *h* en la capa anterior

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

- 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:

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

¿a?
- ReLU
- tanh
- N(,)

- ¿Cómo inicializamos los sesgos y los pesos?
- **<u>Equivalente</u>** 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*:

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

$$= \boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathrm{a}[\mathbf{f}_{k-1}]$$

- Fijar todos los sesgos a 0

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

- Pesos distribuidos normalmente
  - media 0
  - Varianza σ_Ω^2

$$\Omega_k \sim \mathcal{N}(0, \sigma_\Omega^2)$$

- ¿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:

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

Relu

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** → ReLU
- 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

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

- 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

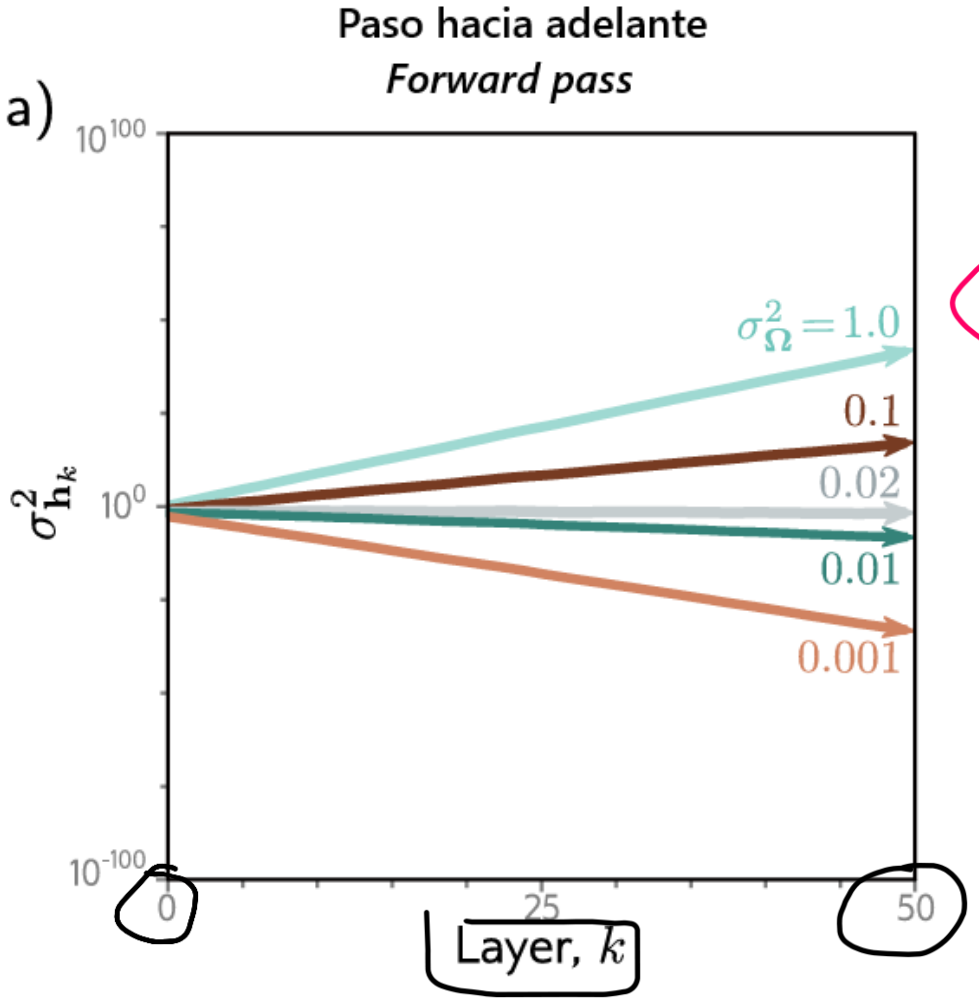


Figura 7.7

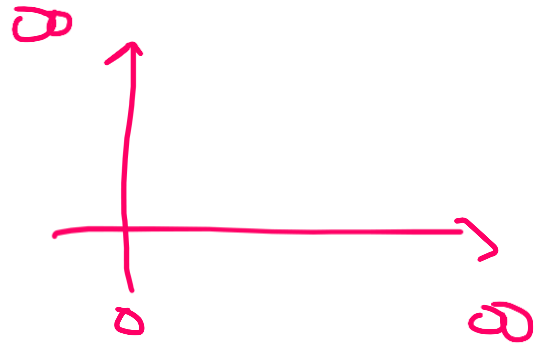
Setup

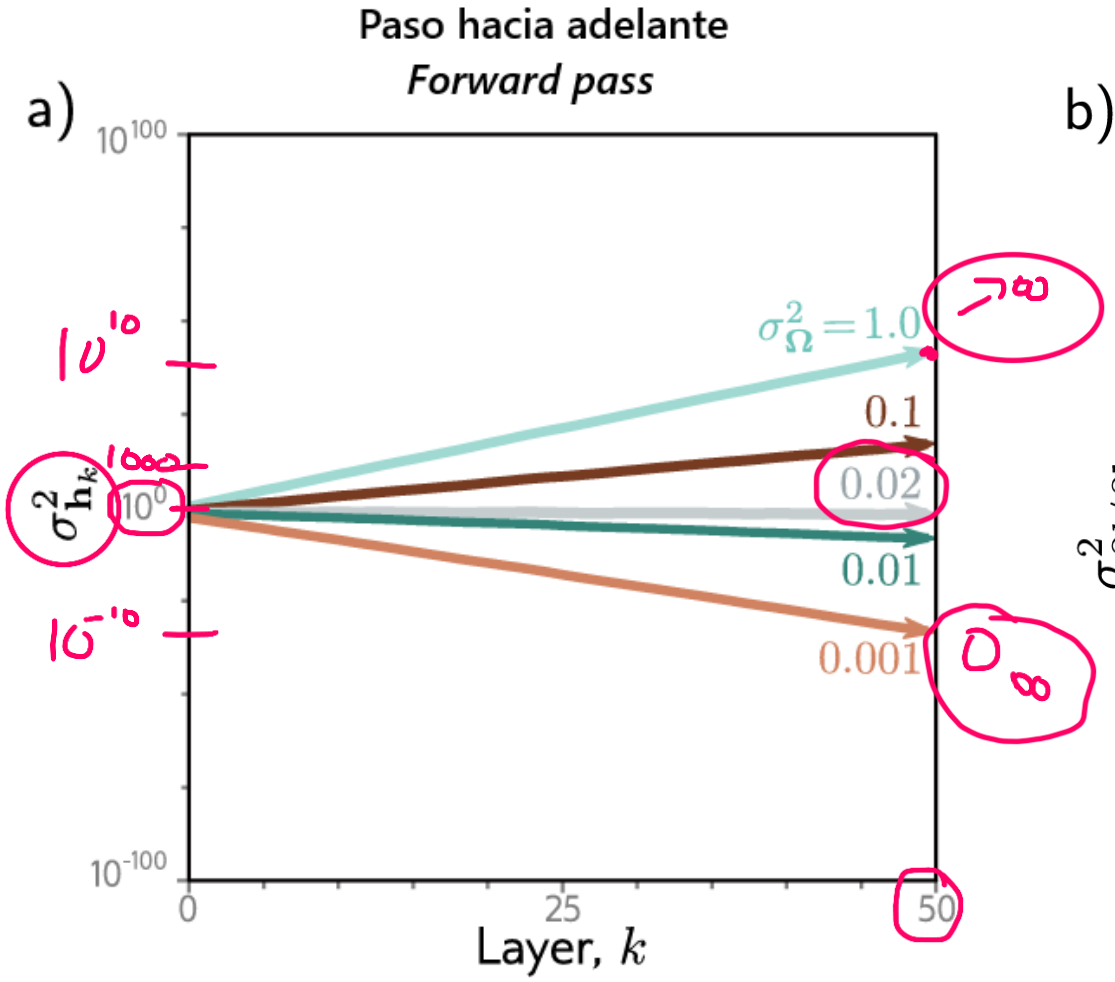
1) Número de capas $k = 50$
2) Número de neuronas por capa $D_h = 100$
3) El vector de entrada $X \sim N(0,1)^{100} = \begin{bmatrix} N(0,1) \\ \vdots \\ N(0,1) \end{bmatrix}$
4) El vector de salida es fijo con $Y = [0]^{100} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$
5) El vector de sesgo $\beta_k = [0]^{100} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$
6) La matriz de pesos

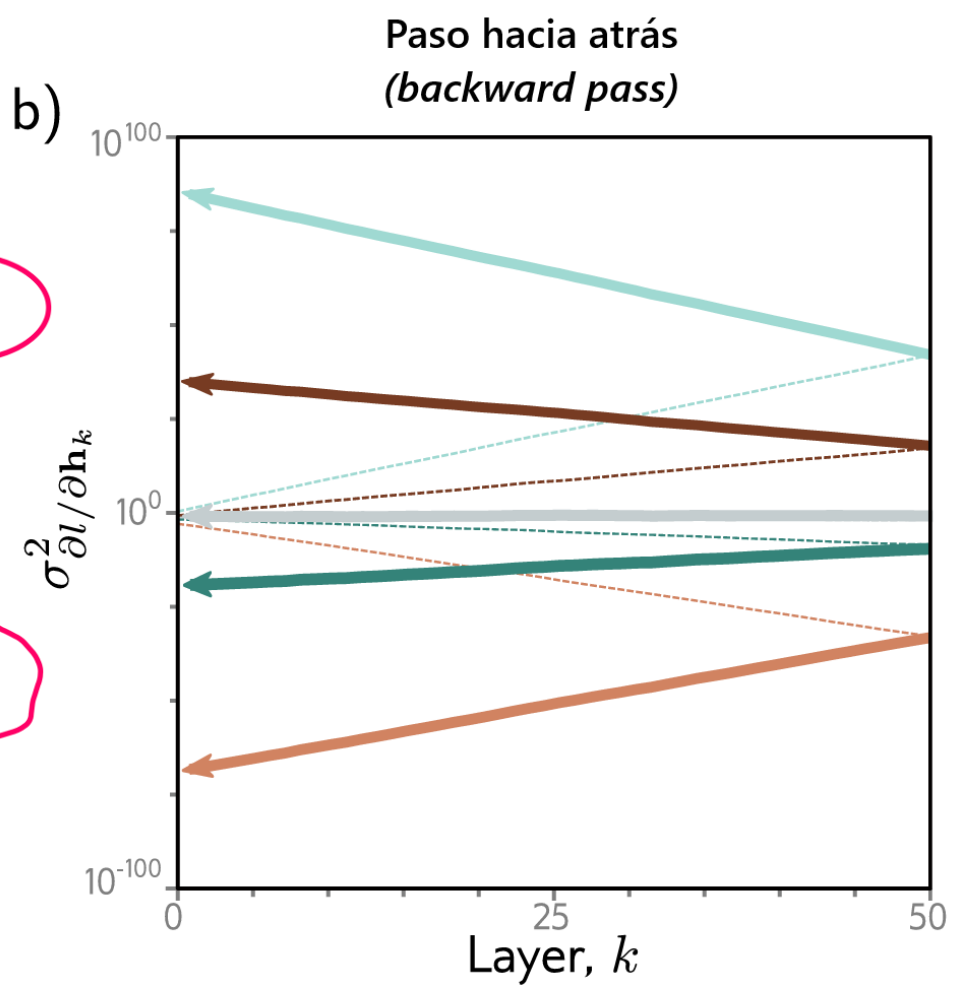
$$\Omega_k \sim N(0,1)^{100\times 100} = \begin{bmatrix} N_{1,1}(0,\sigma^2_\Omega) & \cdots & N_{1,100}(0,\sigma^2_\Omega) \\ \vdots & \ddots & \vdots \\ N_{100,0}(0,\sigma^2_\Omega) & \cdots & N_{100,100}(0,\sigma^2_\Omega) \end{bmatrix}$$

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

Inicialización de los pesos







Gradientes explotados

Gradientes desvanecen

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:

$$\frac{2}{100} = 0.02$$

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

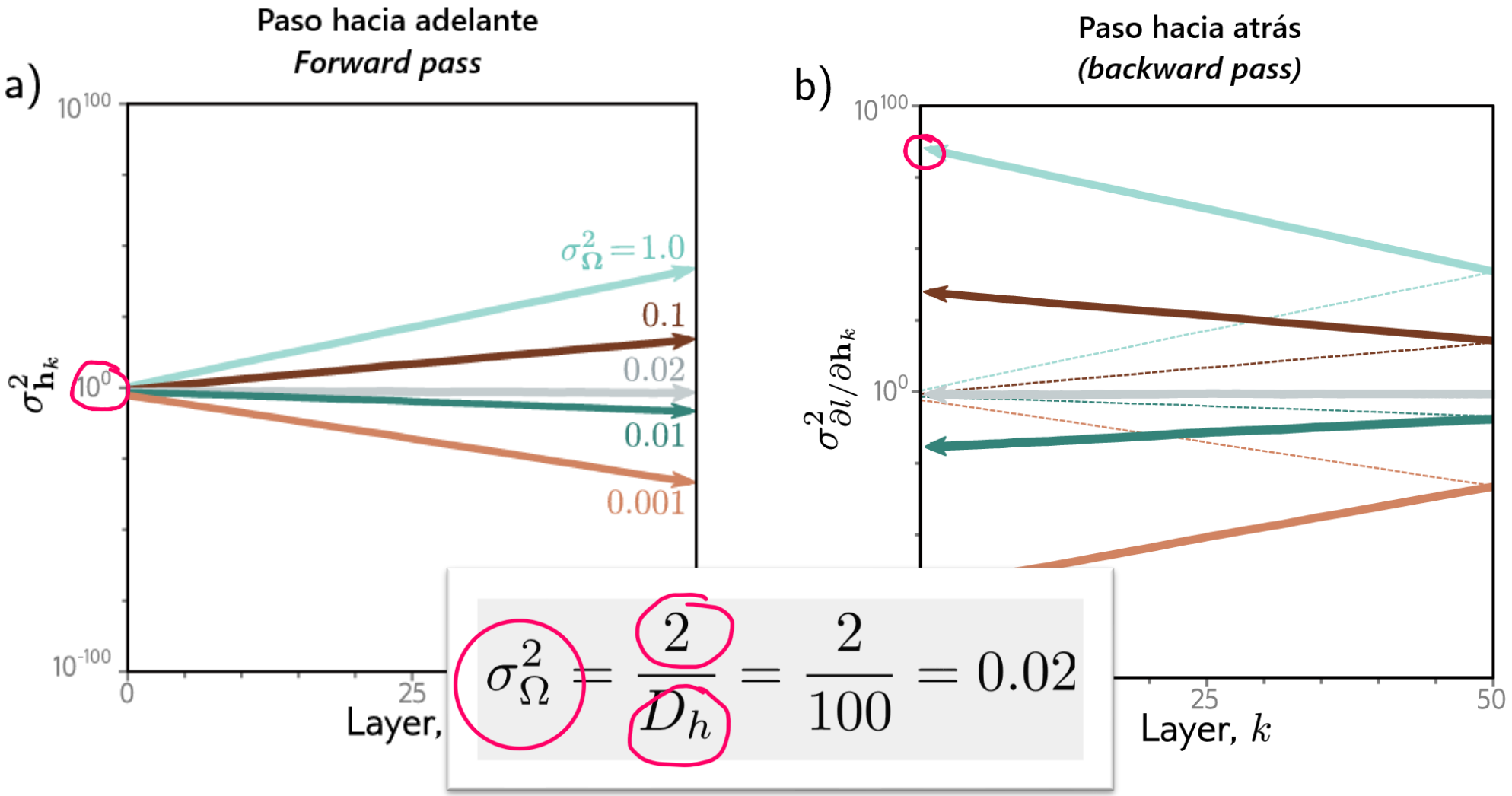
Número de unidades en la capa k

- 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:

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

Número de unidades en la capa k+1

Inicialización de los pesos



$$\sigma_\Omega^2 = \frac{2}{D_h} = \frac{2}{100} = 0.02$$

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

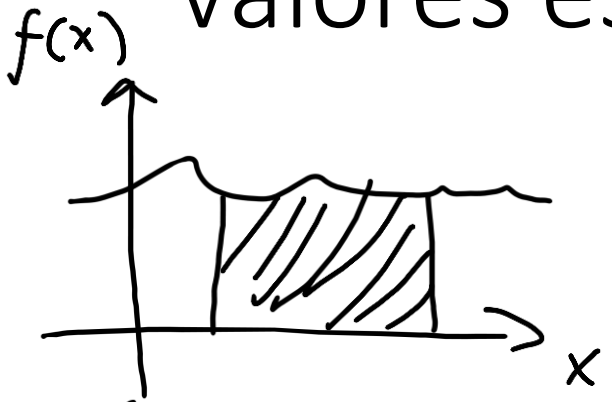
$$\sigma_{f'}^2 = \sigma_\Omega^2 \sum_{j=1}^{D_h} \mathbb{E}\left[h_j^2\right]$$

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

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

$\int_{-\infty}^{\infty} f(x)\,dx = 1$

Valores esperados



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

$X : \{1, 2, \dots, 6\} \times \{1, \dots, 6\} \to \mathbb{R}$

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:

$\int_a^b f(x) = F(b) - F(a) \Rightarrow$ F.D.A

$\int_a^a f(x)\,dx = 0$

$$\mathbb{E}\Big[\mathrm{g}[x]\Big] \approx \frac{1}{N} \sum_{n=1}^{N} \mathrm{g}[x_n^*]$$

where $x_n^* \sim Pr(x)$

$P(a < x < b) = F(b) - F(a)$

$$E(x) = \begin{cases} \sum_{\forall x \in \mathbb{R}_x} x \cdot P(x = x) \\ \int_{-\infty}^{\infty} x \cdot f(x) \cdot dx \end{cases}$$

Valores esperados

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

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:

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

ECDF

Valores esperados

$\mu = E[x]$

$x^k = E[x^k] = E[x^k - 0]$

$V(x) = \int (x-\mu)^2 f(x)\,dx$

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

Propiedades para manipular Valores esperados

$$\mathbb{E}\Big[\underset{\cdot}{k}\Big] = \underset{\cdot}{k}$$

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

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

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

Regla 1

$$\sum_{i=1}^{N} a F(i) = a \sum f(i)$$

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

$$\mathbb{E}\Big[\underline{\kappa}\Big] = \int \kappa Pr(x) dx$$

$$= \kappa \int_{-\infty}^{\infty} Pr(x) dx \quad (1)$$

$$= \kappa.$$

Regla 2

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

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

Regla 3

$g(x) = f(x) + t(x)$

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

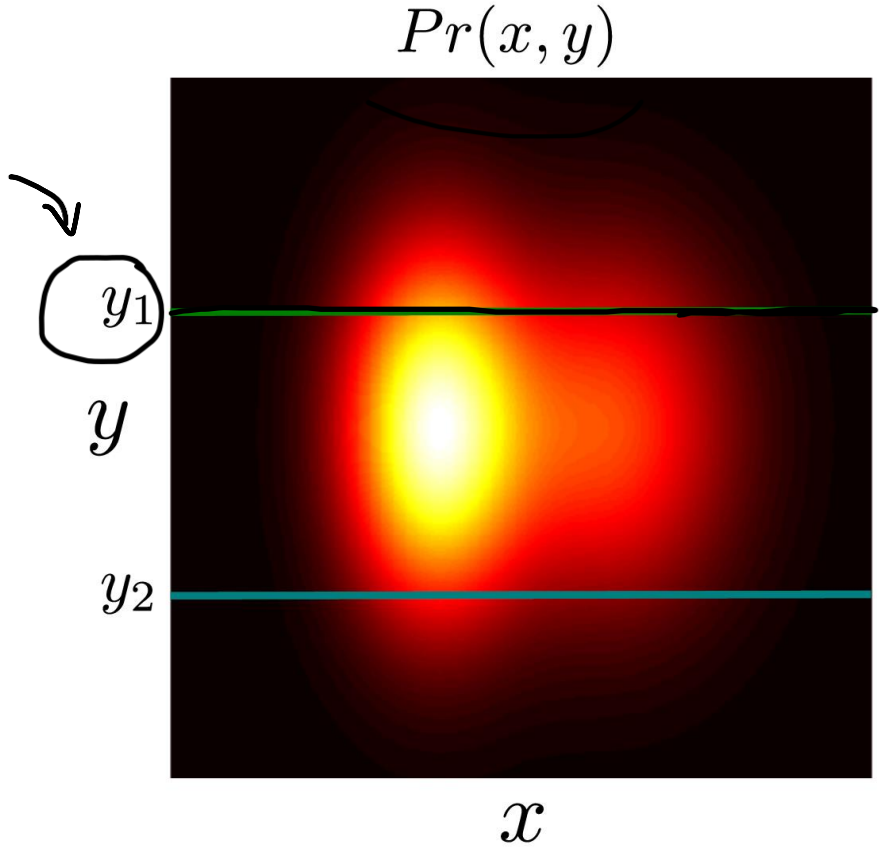
$$\mathbb{E}\Big[\mathrm{f}[x] + \mathrm{g}[x]\Big] = \int (\mathrm{f}[x] + \mathrm{g}[x]) Pr(x) dx \qquad f(x)$$

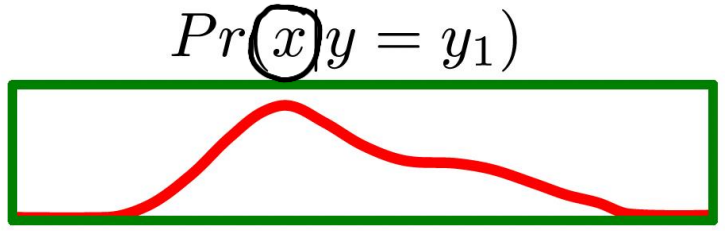
$$= \int (\mathrm{f}[x] Pr(x) + \mathrm{g}[x] Pr(x))\, dx$$

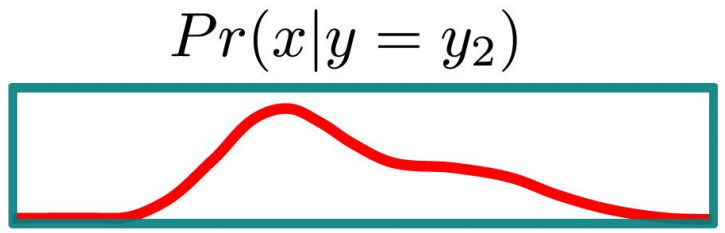
$$= \int \mathrm{f}[x] Pr(x) dx + \int \mathrm{g}[x] Pr(x) dx$$

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

Independencia





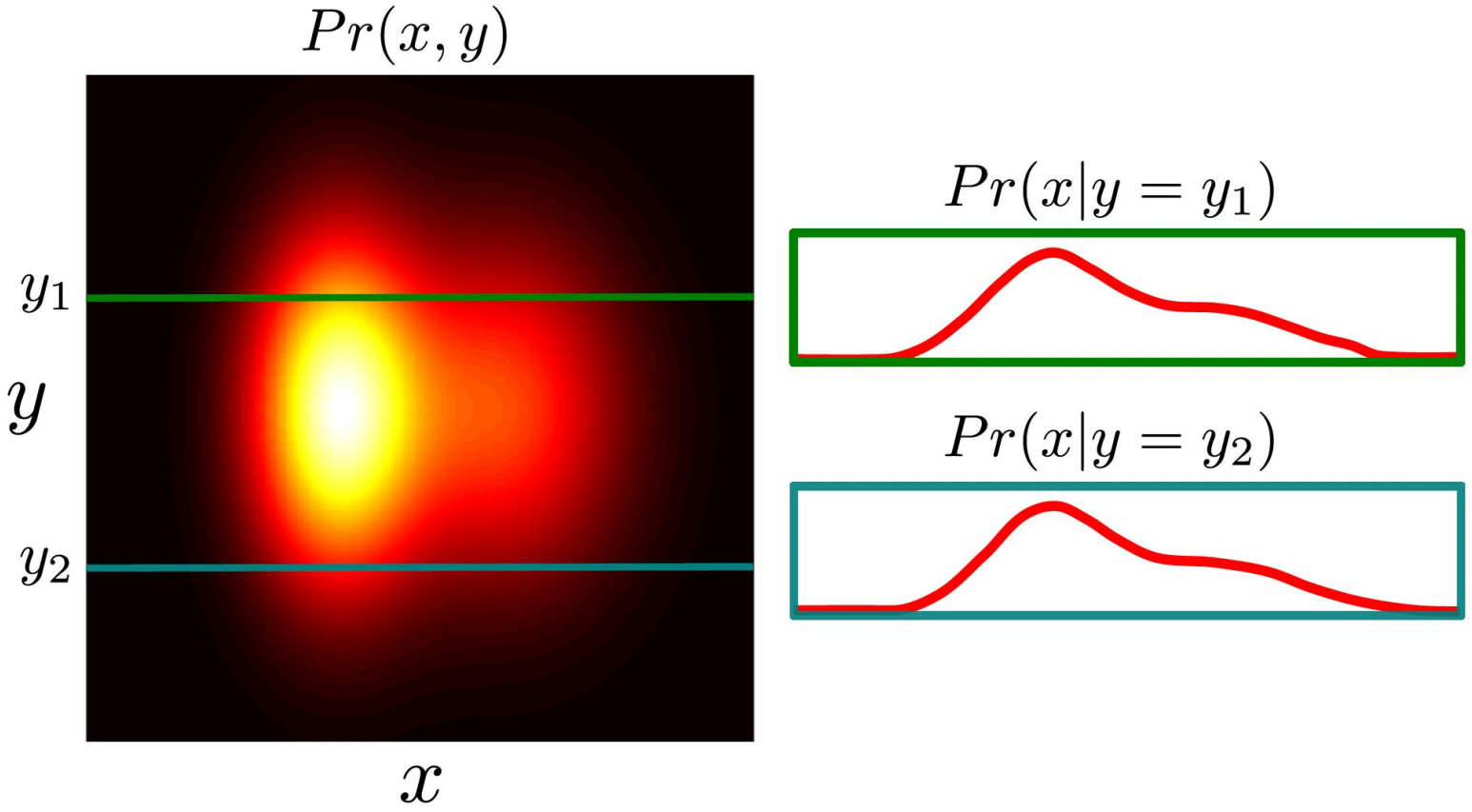


Probabilidad de x e y

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

Independencia

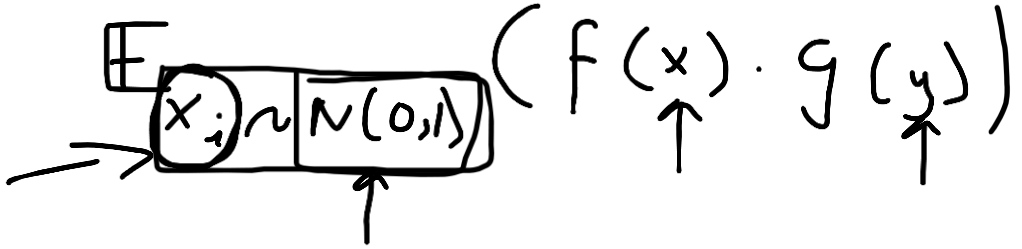
$$N(0,1)^{100} = \prod_{i=1}^{100} N(0,1)$$



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

Probabilidad de x e y

Regla 4



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

$$\begin{aligned}
\mathbb{E}\Big[\mathrm{f}[x] \cdot \mathrm{g}[y]\Big] &= \int\int \mathrm{f}[x] \cdot \mathrm{g}[y] Pr(x,y) dx dy \\
&= \int\int \mathrm{f}[x] \cdot \mathrm{g}[y] Pr(x) Pr(y) dx dy \\
&= \int \mathrm{f}[x] Pr(x) dx \int \mathrm{g}[y] Pr(y) dy \\
&= \mathbb{E}_x\Big[\mathrm{f}[x]\Big] \mathbb{E}_y\Big[\mathrm{g}[y]\Big] \qquad \text{if} \quad x, y \quad \text{independent}
\end{aligned}$$

Porque independientes

Ahora probemos:

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

Teniendo en cuenta:

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

Ahora probemos:

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

Teniendo en cuenta:

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

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

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

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

Definición $\mathbb{E}[x] = \mu$

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

$g(x) = (x-\mu)^2$

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

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

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

Definición $\mathbb{E}[x] = \mu$

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

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

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

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

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

Definición $\mathbb{E}[x] = \mu$

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

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

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

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

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

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

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

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

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

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

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

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

Definición $\mathbb{E}[x] = \mu$

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

1) V.A
$X: \Omega \to \mathbb{R}$
2) $\int f(x) = 1$
$\int x f(x) = E(x)$
$\int g(x) f(x) = E(g(x))$
3) Propi.

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

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

- 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*:

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

- Fijar todos los sesgos a 0

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

- Pesos distribuidos normalmente
  - media 0
  - varianza σ_Ω^2

$$\mathcal{N}(0, \sigma_\Omega^2)$$

- ¿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: $$\mathbb{E}\Big[k\Big] = k$$

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

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

Regla 4: $$\mathbb{E}\Big[\mathrm{f}[x]g[y]\Big] = \mathbb{E}\Big[\mathrm{f}[x]\Big]\mathbb{E}\Big[\mathrm{g}[y]\Big] \quad \text{if} \quad x, y \quad \text{independent}$$

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

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

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

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

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

Regla 4: $$\mathbb{E}\Big[\mathrm{f}[x]g[y]\Big] = \mathbb{E}\Big[\mathrm{f}[x]\Big]\mathbb{E}\Big[\mathrm{g}[y]\Big] \quad \text{if} \quad x, y \quad \text{independent}$$

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

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

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

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

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

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

Regla 4: $$\mathbb{E}\Big[\mathrm{f}[x]g[y]\Big] = \mathbb{E}\Big[\mathrm{f}[x]\Big]\mathbb{E}\Big[\mathrm{g}[y]\Big] \quad \text{if} \quad x, y \quad \text{independent}$$

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

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

Poner todos los sesgos a 0

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

Ponderaciones distribuidas normalmente

- media 0
- varianza σ_Ω^2

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

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

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

- 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

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

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

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

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

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

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

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

Regla 4: $$\mathbb{E}\Big[\mathrm{f}[x]g[y]\Big] = \mathbb{E}\Big[\mathrm{f}[x]\Big]\mathbb{E}\Big[\mathrm{g}[y]\Big] \quad \text{if} \quad x, y \quad \text{independent}$$

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

$$= \mathbb{E}\left[\left(\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j\right)^2\right] - 0$$

Poner todos los sesgos a 0

Ponderaciones distribuidas normalmente

media 0

varianza σ_Ω^2

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

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

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

Regla 4: $\mathbb{E}\Big[\mathrm{f}[x]g[y]\Big] = \mathbb{E}\Big[\mathrm{f}[x]\Big]\mathbb{E}\Big[\mathrm{g}[y]\Big] \quad \text{if} \quad x, y \quad \text{independent}$

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

$$= \mathbb{E}\left[\left(\cancel{\beta_i} + \sum_{j=1}^{D_h} \Omega_{ij} h_j\right)^2\right] - 0$$

$$= \mathbb{E}\left[\left(\sum_{j=1}^{D_h} \Omega_{ij} h_j\right)^2\right]$$

Poner todos los sesgos a 0

Ponderaciones distribuidas normalmente

media 0

varianza σ_Ω^2

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

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

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

Regla 4: $$\mathbb{E}\Big[\mathrm{f}[x]g[y]\Big] = \mathbb{E}\Big[\mathrm{f}[x]\Big]\mathbb{E}\Big[\mathrm{g}[y]\Big] \quad \text{if} \quad x, y \quad \text{independent}$$

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

$$= \mathbb{E}\left[\left(\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j\right)^2\right] - 0$$

$$= \mathbb{E}\left[\left(\sum_{j=1}^{D_h} \Omega_{ij} h_j\right)^2\right]$$

$$= \sum_{j=1}^{D_h} \mathbb{E}\left[\Omega_{ij}^2\right] \mathbb{E}\left[h_j^2\right]$$

Poner todos los sesgos a 0

Ponderaciones distribuidas normalmente
- media 0
- varianza σ_Ω^2

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

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

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

Regla 4: $\mathbb{E}\left[\mathrm{f}[x]g[y]\right] = \mathbb{E}\left[\mathrm{f}[x]\right]\mathbb{E}\left[\mathrm{g}[y]\right]$ if x, y independent

$V(x) = E(x^2) - E(x)^2$

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

$$= \mathbb{E}\left[\left(\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j\right)^2\right] - 0$$

$$= \mathbb{E}\left[\left(\sum_{j=1}^{D_h} \Omega_{ij} h_j\right)^2\right]$$

Poner todos los sesgos a 0

$$= \sum_{j=1}^{D_h} \mathbb{E}\left[\Omega_{ij}^2\right] \mathbb{E}\left[h_j^2\right]$$

Ponderaciones distribuidas normalmente
- media 0
- varianza σ_Ω^2

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

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

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

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

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

$$\sigma^2_{f'} = \sigma^2_\Omega \sum_{j=1}^{D_h} \mathbb{E}\left[h_j^2\right]$$

$$= \sigma^2_\Omega \sum_{j=1}^{D_h} \mathbb{E}\left[\mathrm{ReLU}[f_j]^2\right]$$

$$= \sigma^2_\Omega \sum_{j=1}^{D_h} \int_{-\infty}^{\infty} \mathrm{ReLU}[f_j]^2 Pr(f_j) df_j$$

$$= \sigma^2_\Omega \sum_{j=1}^{D_h} \int_{-\infty}^{\infty} (\mathbb{I}[f_j \geq 0] f_j)^2 Pr(f_j) df_j$$

$$= \sigma^2_\Omega \sum_{j=1}^{D_h} \int_{0}^{\infty} f_j^2 Pr(f_j) df_j$$

$$= \sigma^2_\Omega \sum_{j=1}^{D_h} \frac{\sigma_f^2}{2} = \frac{D_h \sigma^2_\Omega \sigma^2_f}{2}$$

$h_j^{f_j} = a(f_j)$

$g(f_j) = \mathrm{ReLU}$

$\int_{-\infty}^{\infty} \mathbb{I}(f_j > 0) f_j^2 1$

$V(f_j) = \sigma_f^2$

$\sigma^2_{f_j} \sum_{j=1}^{D_h} \mathbb{E}(f_j^2)$

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

$$\sigma^2_{f'} = \frac{D_h \sigma^2_\Omega \sigma^2_f}{2}$$

Buscamos $\sigma^2_{f'} = \sigma^2_f$ si asumimos que $\sigma^2_{f'} = \sigma^2_f > 0$

Basta con hacer $\sigma^2_{f'} = \sigma^2_f = a$ y tenemos

$\sigma^2_{f'} = \frac{\sigma^2_f D_h \sigma^2_\Omega}{2} \equiv a = \frac{a D_h \sigma^2_\Omega}{2}$ que implica entonces que

$$\sigma^2_\Omega = \frac{2}{D_h}$$

$N\left(0, \theta^2_\Omega = \frac{2}{D_h}\right)$

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:

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

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 $n_1 Var[w_1] = 1$ 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

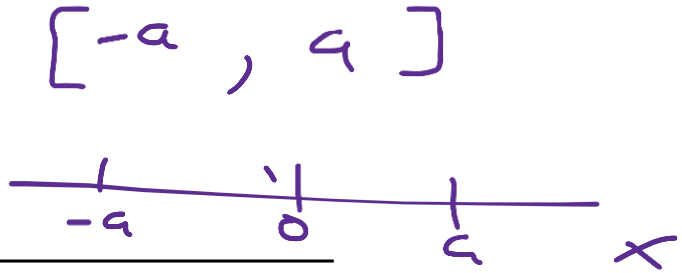
- Glorot & Bengio, "Understanding the difficulty of training deep feedforward neural networks", AISTATS 2010.
- Xavier Uniform

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

- Xavier Normal

$$W \sim \mathcal{N}\left(0, \sqrt{\frac{2}{\text{fan\_in} + \text{fan\_out}}}\right)$$

Inicialización Xavier



- Xavier Uniform $W \sim \mathcal{U}(-a, a), \quad a = \sqrt{\dfrac{6}{\text{fan\_in} + \text{fan\_out}}}$

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

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

$[-\sqrt{3}\frac{1}{\sqrt{fan\_in}}, \sqrt{3}\frac{1}{\sqrt{fan\_in}}]$

Si entonces queremos igualar la varianza a $\frac{1}{fan\_in}$

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

Inicialización Xavier

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

Si hacemos $W \sim Unif(-a, a)$ 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 $\frac{2}{fan\_in + fan\_out}$

$$\frac{a^2}{3} = \frac{2}{fan\_in + fan\_out} = \sqrt{\left(\frac{6}{fan_{in} + fan_{out}}\right)}$$

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 **Yoshua Bengio**
DIRO, Université de Montréal, Montréal, Québec, Canada

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

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

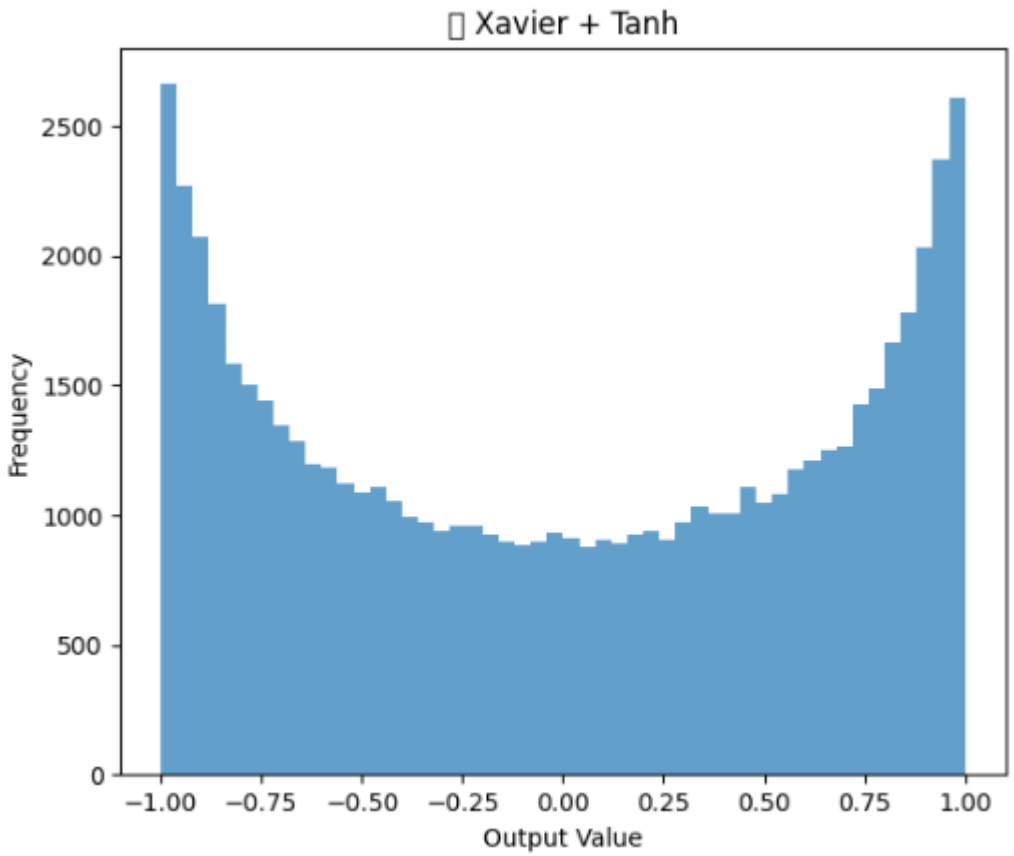
where $U[-a, a]$ is the uniform distribution in the interval $(-a, a)$ and n is the size of the previous layer (the number of columns of W).

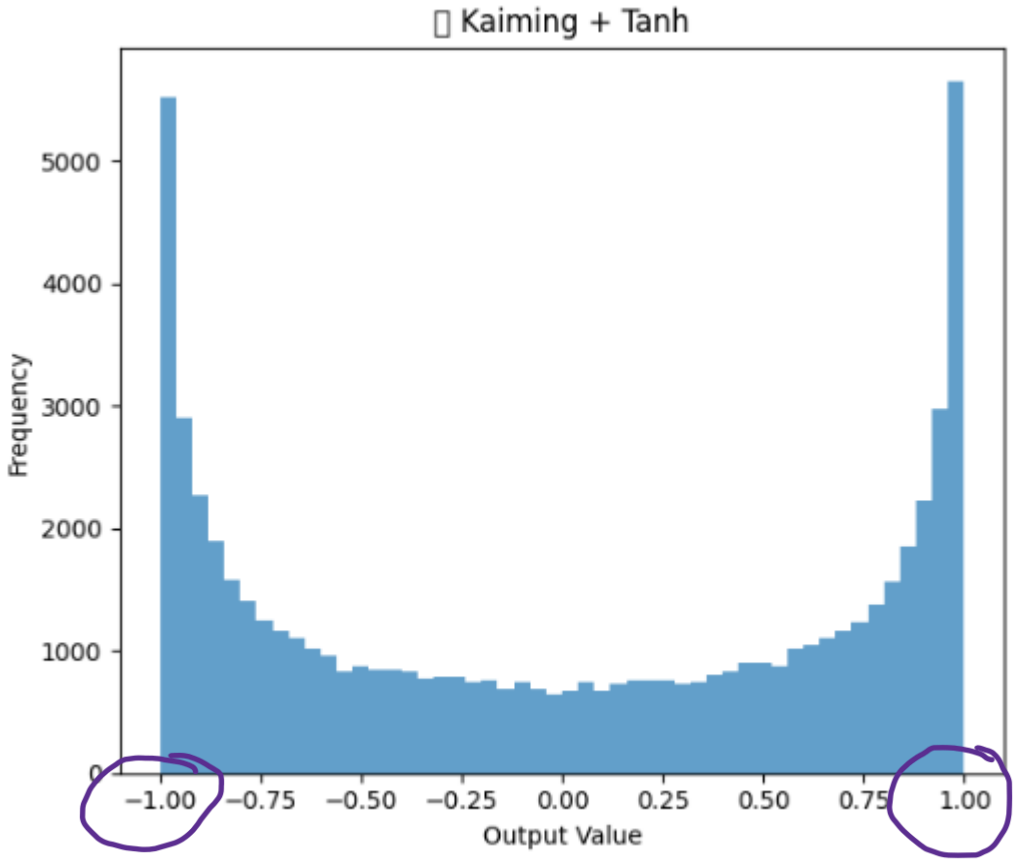
Resumen

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

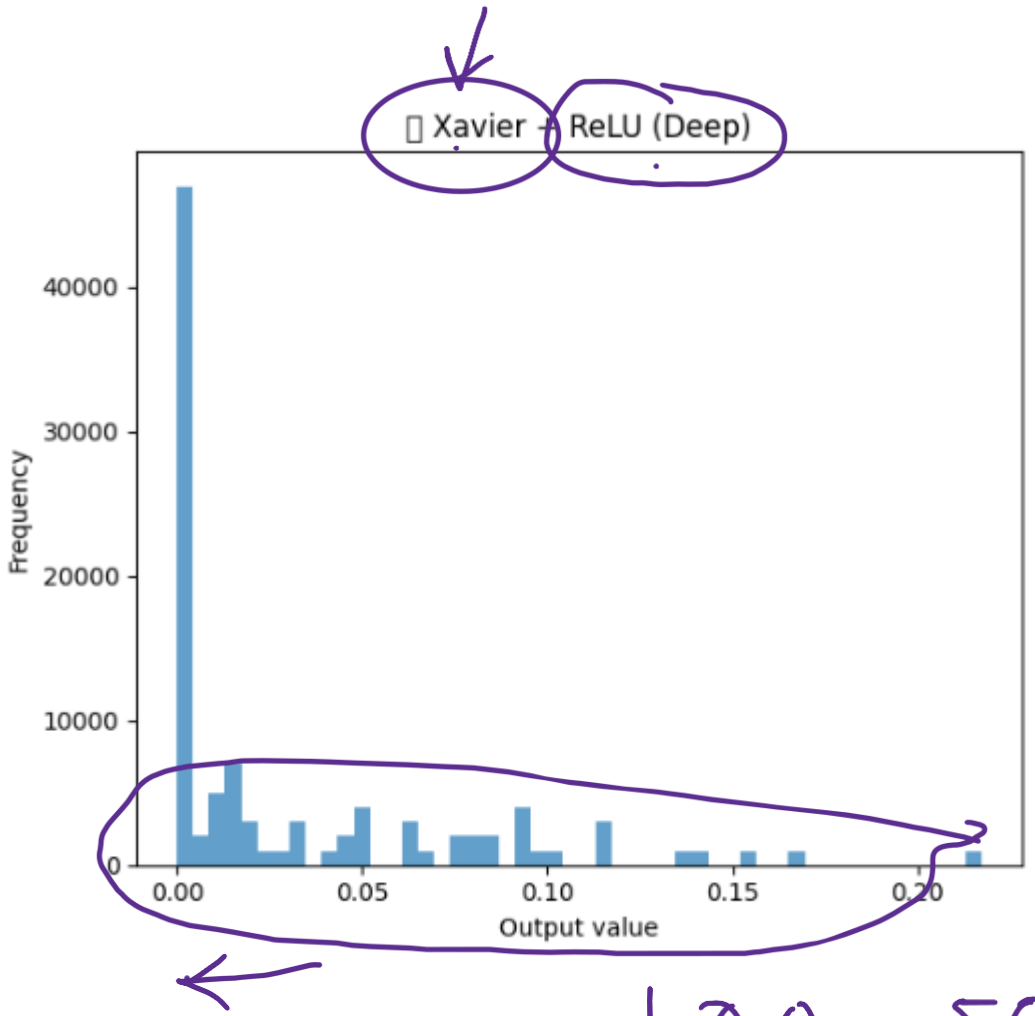
he ReLu



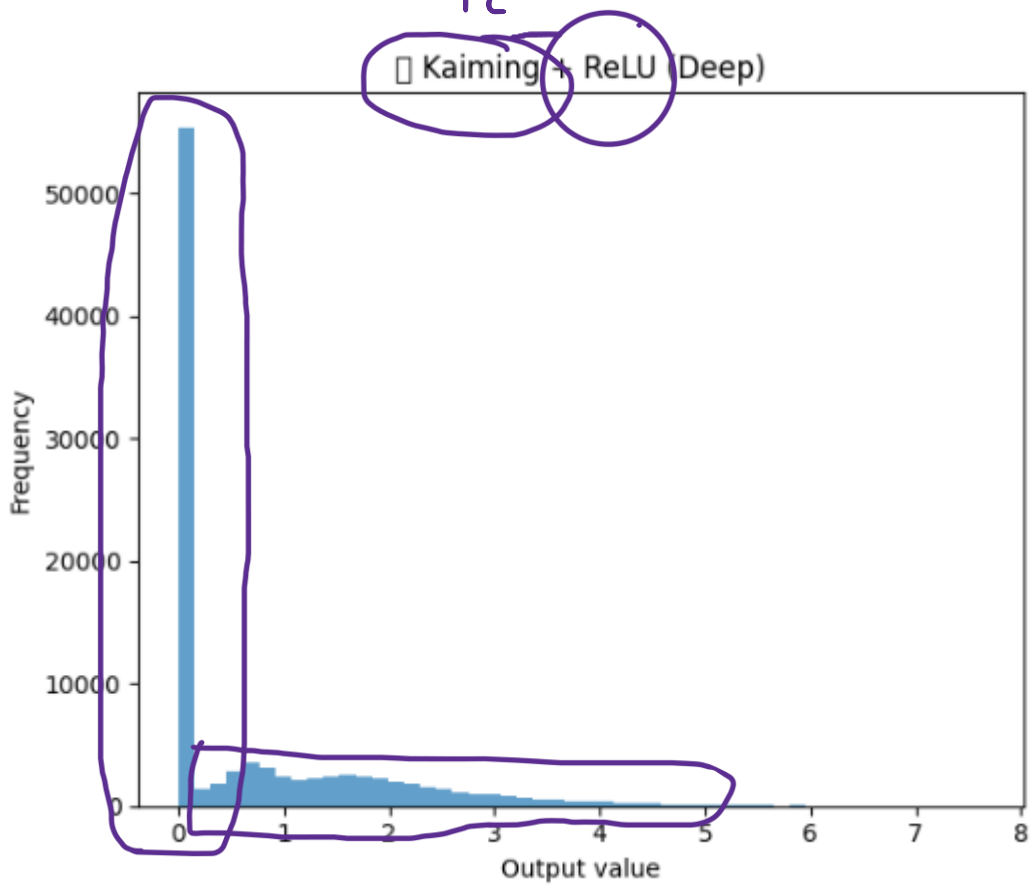


Experimentos

ReLu = max(0, x)



he



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_