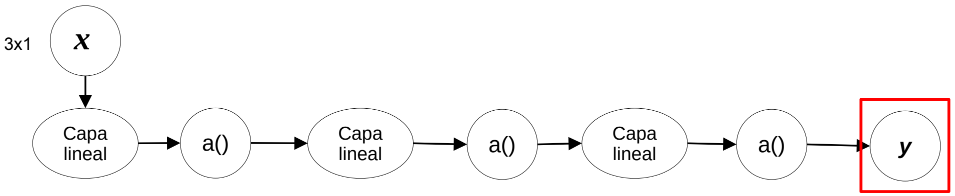
5.2 Optimización con redes profundas

- 5.2.1 Backpropagation
- 5.2.2 Inicialización de parámetros
- 5.2.3 Regularización

5.2 Optimización con redes profundas

- **5.2.1 Backpropagation**
- 5.2.2 Inicialización de parámetros
- 5.2.3 Regularización

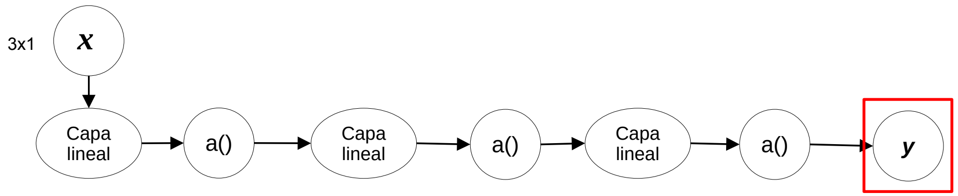
Ejemplo



$\mathbf{z}_0=\boldsymbol{\Omega}_0\mathbf{x}+\beta_0$ (4x1, 4x3, 4x1) $\quad \mathbf{h}_1=a(\mathbf{z}_0)$ (4x1) $\quad \mathbf{z}_1=\boldsymbol{\Omega}_1\mathbf{h}_1+\beta_1$ (2x1, 2x4, 2x1) $\quad \mathbf{h}_2=a(\mathbf{z}_1)$ (2x1) $\quad \mathbf{z}_2=\boldsymbol{\Omega}_2\mathbf{h}_2+\beta_2$ (3x1, 3x2, 3x1) $\quad \mathbf{h}_3=a(\mathbf{z}_2)$ (3x1) $\quad \mathbf{y}=\boldsymbol{\Omega}_3\mathbf{h}_3+\beta_3$ (2x1, 2x3, 2x1)

- Comenzamos con la *derivada de la función de pérdida* con respecto a ***y*:** $\frac{\partial J}{\partial \mathbf{y}}$

Ejemplo

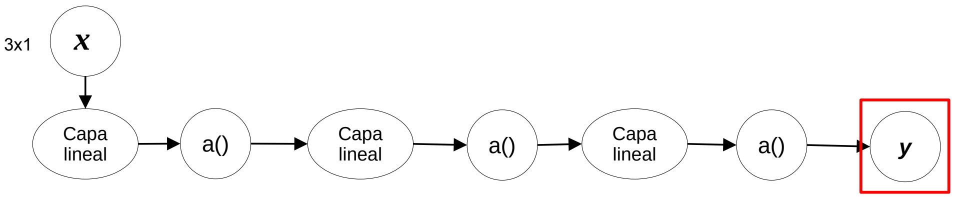


$\mathbf{z}_0=\boldsymbol{\Omega}_0\mathbf{x}+\beta_0$ (4x1, 4x3, 4x1) $\quad \mathbf{h}_1=a(\mathbf{z}_0)$ (4x1) $\quad \mathbf{z}_1=\boldsymbol{\Omega}_1\mathbf{h}_1+\beta_1$ (2x1, 2x4, 2x1) $\quad \mathbf{h}_2=a(\mathbf{z}_1)$ (2x1) $\quad \mathbf{z}_2=\boldsymbol{\Omega}_2\mathbf{h}_2+\beta_2$ (3x1, 3x2, 3x1) $\quad \mathbf{h}_3=a(\mathbf{z}_2)$ (3x1) $\quad \mathbf{y}=\boldsymbol{\Omega}_3\mathbf{h}_3+\beta_3$ (2x1, 2x3, 2x1)

- Comenzamos con la *derivada de la función de pérdida* con respecto a ***y*:** $\frac{\partial J}{\partial \mathbf{y}}$

$$\boxed{\frac{\partial \mathbf{z}_k}{\partial \mathbf{h}_k}=\boldsymbol{\Omega}_k^T \qquad \frac{\partial a(\mathbf{z}_k)}{\partial \mathbf{z}_k}=\begin{bmatrix} I(z_{k,1}>0) & 0 & 0 \\ 0 & I(z_{k,2}>0) & 0 \\ 0 & 0 & I(z_{k,3}>0) \end{bmatrix} \qquad \frac{\partial \mathbf{z}_k}{\partial \boldsymbol{\Omega}_k}=\mathbf{h}_k^T \qquad \frac{\partial \mathbf{z}_k}{\partial \boldsymbol{\beta}_k}=\mathbf{I}}$$

Ejemplo



$\boldsymbol{z}_0=\boldsymbol{\Omega}_0\boldsymbol{x}+\beta_0$ (4x1, 4x3, 4x1) $\quad \boldsymbol{h}_1=a(\boldsymbol{z}_0)$ (4x1) $\quad \boldsymbol{z}_1=\boldsymbol{\Omega}_1\boldsymbol{h}_1+\beta_1$ (2x1, 2x4, 2x1) $\quad \boldsymbol{h}_2=a(\boldsymbol{z}_1)$ (2x1) $\quad \boldsymbol{z}_2=\boldsymbol{\Omega}_2\boldsymbol{h}_2+\beta_2$ (3x1, 3x2, 3x1) $\quad \boldsymbol{h}_3=a(\boldsymbol{z}_2)$ (3x1) $\quad \boldsymbol{y}=\boldsymbol{\Omega}_3\boldsymbol{h}_3+\beta_3$ (2x1, 2x3, 2x1)

- Comenzamos con la *derivada de la función de pérdida* con respecto a ***y*:** $\frac{\partial J}{\partial \boldsymbol{y}}$
- Y vamos hacia atrás en por la red:

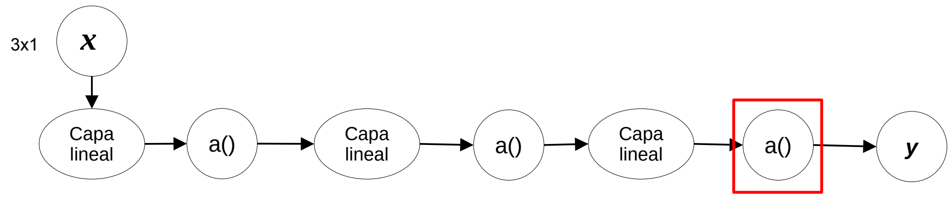
$$\frac{\partial L}{\partial \boldsymbol{h}_3}=\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{h}_3}\cdot\frac{\partial L}{\partial \boldsymbol{y}}=\boldsymbol{\Omega}_3^T\cdot\frac{\partial L}{\partial \boldsymbol{y}}$$

$$\left.\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\beta}_3}&=\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\beta}_3}\cdot\frac{\partial L}{\partial \boldsymbol{y}}=\frac{\partial L}{\partial \boldsymbol{y}}\\ \frac{\partial L}{\partial \boldsymbol{\Omega}_3}&=\frac{\partial L}{\partial \boldsymbol{y}}\cdot\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\Omega}_3}\cdot=\frac{\partial L}{\partial \boldsymbol{y}}\cdot\boldsymbol{h}_3^T\end{aligned}\right\}$$

Derivadas de la función de pérdida con respecto a a los parámetros

$$\boxed{\frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{h}_k}=\boldsymbol{\Omega}_k^T \qquad \frac{\partial a(\boldsymbol{z}_k)}{\partial \boldsymbol{z}_k}=\begin{bmatrix} I(z_{k,1}>0) & 0 & 0\\ 0 & I(z_{k,2}>0) & 0\\ 0 & 0 & I(z_{k,3}>0)\end{bmatrix} \qquad \frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{\Omega}_k}=\boldsymbol{h}_k^T \qquad \frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{\beta}_k}=\boldsymbol{I}}$$

Ejemplo



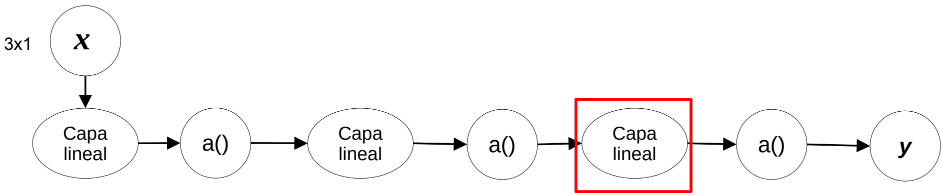
$\boldsymbol{z}_0=\boldsymbol{\Omega}_0\boldsymbol{x}+\beta_0$ (4x1, 4x3, 4x1) $\quad \boldsymbol{h}_1=a(\boldsymbol{z}_0)$ (4x1) $\quad \boldsymbol{z}_1=\boldsymbol{\Omega}_1\boldsymbol{h}_1+\beta_1$ (2x1, 2x4, 2x1) $\quad \boldsymbol{h}_2=a(\boldsymbol{z}_1)$ (2x1) $\quad \boldsymbol{z}_2=\boldsymbol{\Omega}_2\boldsymbol{h}_2+\beta_2$ (3x1, 3x2, 3x1) $\quad \boldsymbol{h}_3=a(\boldsymbol{z}_2)$ (3x1) $\quad \boldsymbol{y}=\boldsymbol{\Omega}_3\boldsymbol{h}_3+\beta_3$ (2x1, 2x3, 2x1)

- Comenzamos con la *derivada de la función de pérdida* con respecto a ***y*:** $\frac{\partial J}{\partial \boldsymbol{y}}$
- Y vamos hacia atrás en por la red:

$$\frac{\partial L}{\partial \boldsymbol{z}_2}=\frac{\partial \boldsymbol{h}_3}{\partial \boldsymbol{z}_2}\cdot\frac{\partial L}{\partial \boldsymbol{h}_3}=\begin{bmatrix} I(z_{2,1}>0) & 0 & 0\\ 0 & I(z_{2,2}>0) & 0\\ 0 & 0 & I(z_{2,3}>0)\end{bmatrix}\cdot\frac{\partial L}{\partial \boldsymbol{h}_3}$$

$$\boxed{\frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{h}_k}=\boldsymbol{\Omega}_k^T \qquad \frac{\partial a(\boldsymbol{z}_k)}{\partial \boldsymbol{z}_k}=\begin{bmatrix} I(z_{k,1}>0) & 0 & 0\\ 0 & I(z_{k,2}>0) & 0\\ 0 & 0 & I(z_{k,3}>0)\end{bmatrix} \qquad \frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{\Omega}_k}=\boldsymbol{h}_k^T \qquad \frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{\beta}_k}=\boldsymbol{I}}$$

Ejemplo



$\boldsymbol{z}_0=\boldsymbol{\Omega}_0\boldsymbol{x}+\beta_0$ (4x1, 4x3, 4x1) $\quad \boldsymbol{h}_1=a(\boldsymbol{z}_0)$ (4x1) $\quad \boldsymbol{z}_1=\boldsymbol{\Omega}_1\boldsymbol{h}_1+\beta_1$ (2x1, 2x4, 2x1) $\quad \boldsymbol{h}_2=a(\boldsymbol{z}_1)$ (2x1) $\quad \boldsymbol{z}_2=\boldsymbol{\Omega}_2\boldsymbol{h}_2+\beta_2$ (3x1, 3x2, 3x1) $\quad \boldsymbol{h}_3=a(\boldsymbol{z}_2)$ (3x1) $\quad \boldsymbol{y}=\boldsymbol{\Omega}_3\boldsymbol{h}_3+\beta_3$ (2x1, 2x3, 2x1)

- Comenzamos con la *derivada de la función de pérdida* con respecto a ***y*:** $\frac{\partial J}{\partial \boldsymbol{y}}$
- Y vamos hacia atrás en por la red:

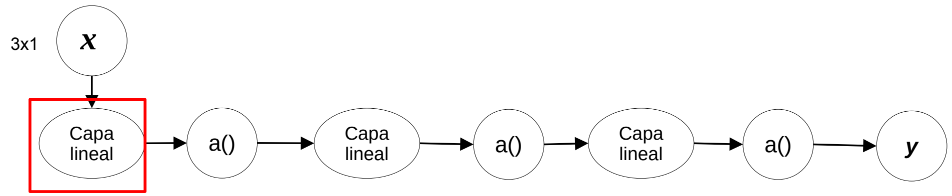
$$\frac{\partial L}{\partial \boldsymbol{h}_2}=\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{h}_2}\cdot\frac{\partial L}{\partial \boldsymbol{z}_2}=\boldsymbol{\Omega}_2^T\cdot\frac{\partial L}{\partial \boldsymbol{z}_2}$$

$$\left.\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\beta}_2}&=\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{\beta}_2}\cdot\frac{\partial L}{\partial \boldsymbol{z}_2}=\frac{\partial L}{\partial \boldsymbol{\beta}_2}\\ \frac{\partial L}{\partial \boldsymbol{\Omega}_2}&=\frac{\partial L}{\partial \boldsymbol{z}_2}\cdot\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{\Omega}_2}\cdot=\frac{\partial L}{\partial \boldsymbol{z}_2}\cdot\boldsymbol{h}_2^T\end{aligned}\right\}$$

Derivadas de la función de pérdida con respecto a a los parámetros

$$\boxed{\frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{h}_k}=\boldsymbol{\Omega}_k^T \qquad \frac{\partial a(\boldsymbol{z}_k)}{\partial \boldsymbol{z}_k}=\begin{bmatrix} I(z_{k,1}>0) & 0 & 0\\ 0 & I(z_{k,2}>0) & 0\\ 0 & 0 & I(z_{k,3}>0)\end{bmatrix} \qquad \frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{\Omega}_k}=\boldsymbol{h}_k^T \qquad \frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{\beta}_k}=\boldsymbol{I}}$$

Ejemplo



$\boldsymbol{z}_0=\boldsymbol{\Omega}_0\boldsymbol{x}+\beta_0$ (4x1, 4x3, 4x1) $\quad \boldsymbol{h}_1=a(\boldsymbol{z}_0)$ (4x1) $\quad \boldsymbol{z}_1=\boldsymbol{\Omega}_1\boldsymbol{h}_1+\beta_1$ (2x1, 2x4, 2x1) $\quad \boldsymbol{h}_2=a(\boldsymbol{z}_1)$ (2x1) $\quad \boldsymbol{z}_2=\boldsymbol{\Omega}_2\boldsymbol{h}_2+\beta_2$ (3x1, 3x2, 3x1) $\quad \boldsymbol{h}_3=a(\boldsymbol{z}_2)$ (3x1) $\quad \boldsymbol{y}=\boldsymbol{\Omega}_3\boldsymbol{h}_3+\beta_3$ (2x1, 2x3, 2x1)

- Comenzamos con la *derivada de la función de pérdida* con respecto a ***y*:** $\frac{\partial J}{\partial \boldsymbol{y}}$
- Y vamos hacia atrás en por la red:

$$\frac{\partial L}{\partial \boldsymbol{x}}=\frac{\partial \boldsymbol{z}_0}{\partial \boldsymbol{x}}\cdot\frac{\partial L}{\partial \boldsymbol{z}_0}=\boldsymbol{\Omega}_0^T\cdot\frac{\partial L}{\partial \boldsymbol{z}_0}$$

$$\left.\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\beta}_0}&=\frac{\partial \boldsymbol{z}_0}{\partial \boldsymbol{\beta}_0}\cdot\frac{\partial L}{\partial \boldsymbol{z}_0}=\frac{\partial L}{\partial \boldsymbol{\beta}_0}\\ \frac{\partial L}{\partial \boldsymbol{\Omega}_0}&=\frac{\partial L}{\partial \boldsymbol{z}_0}\cdot\frac{\partial \boldsymbol{z}_0}{\partial \boldsymbol{\Omega}_0}=\frac{\partial L}{\partial \boldsymbol{z}_0}\cdot\boldsymbol{x}^T\end{aligned}\right\}$$

Derivadas de la función de pérdida con respecto a a los parámetros

$$\boxed{\frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{h}_k}=\boldsymbol{\Omega}_k^T \qquad \frac{\partial a(\boldsymbol{z}_k)}{\partial \boldsymbol{z}_k}=\begin{bmatrix} I(z_{k,1}>0) & 0 & 0\\ 0 & I(z_{k,2}>0) & 0\\ 0 & 0 & I(z_{k,3}>0)\end{bmatrix} \qquad \frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{\Omega}_k}=\boldsymbol{h}_k^T \qquad \frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{\beta}_k}=\boldsymbol{I}}$$

5.2 Optimización con redes profundas

- 5.2.1 Backpropagation
- 5.2.2 Inicialización de parámetros
- 5.2.3 Normalización
- 5.2.4 Regularización

5.2.1 Inicialización de parámetros

- Necesidad de la inicialización
- Inicialización de Xavier y de He

Inicialización

- Las preactivaciones de una capa lineal:

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

- ¿Cómo inicializamos los sesgos y los pesos?
- Equivalente a elegir un punto de arranque de la optimización en el regresor lineal/Gabor que hemos visto en el Tema 2.

Inicialización: Pregunta 1

- Las preactivaciones de una capa lineal:

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

- ¿Podemos inicializarlos todos a 0?

Inicialización: Pregunta 2

- Las preactivaciones de una capa lineal:

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

- Pensemos en una red con 2 unidades ocultas. Si entrenamos la red hasta el final e intercambiamos los pesos de entrada y salida entre las dos unidades ¡Tendremos el mismo resultado! (simetría de la red).
  - ¿Qué pasaría si inicializamos los pesos de las dos unidades ocultas con los mismos valores?

Inicialización

- Las preactivaciones de una capa lineal:

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

preactivación i-ésima en esa capa:

$$\boldsymbol{z}_k(i) = \sum_{j=1}^{D_h} w_{ij} \boldsymbol{h}_k(j) + \boldsymbol{\beta}_k(i)$$

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \boldsymbol{0}$

preactivación i-ésima en esa capa:

$$\boldsymbol{z}_k(i) = \sum_{j=1}^{D_h} w_{ij} \boldsymbol{h}_k(j) + \cancel{\boldsymbol{\beta}_k(i)}$$

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \boldsymbol{0}$

preactivación i-ésima en esa capa:

$$\boldsymbol{z}_k(i) = \sum_{j=1}^{D_h} w_{ij} \boldsymbol{h}_k(j)$$

Repaso esperanza matemática: E[ ]

$$E[g[x]]=\int_{-\infty}^{+\infty} g[x]Pr(x)dx$$

- Interpretación: es el valor promedio de g[x] cuando tenemos en cuenta la probabilidad de x
- En el caso discreto tenemos g[x] para cada valor discreto de x:

$$E[g[x]]=\sum_{i=1}^{N} g[x_i]Pr(x_i)$$

Repaso esperanza matemática: E[ ]

Function g[•]	Expectation
x	mean, μ
x^k	kth moment about zero
$(x-\mu)^k$	kth moment about the mean
$(x-\mu)^2$	variance
$(x-\mu)^3$	skew
$(x-\mu)^4$	kurtosis

Table B.1 Special cases of expectation. For some functions g[x], the expectation $\mathbb{E}$[g[x]] is given a special name. Here we use the notation μ_x to represent the mean with respect to random variable x.

Repaso esperanza matemática: E[ ]

- Propiedades de la Esperanza:

$$\begin{aligned}
\mathbb{E}\Big[k\Big] &= k \\
\mathbb{E}\Big[k \cdot \mathrm{g}[x]\Big] &= k \cdot \mathbb{E}\Big[\mathrm{g}[x]\Big] \\
\mathbb{E}\Big[\mathrm{f}[x] + \mathrm{g}[x]\Big] &= \mathbb{E}\Big[\mathrm{f}[x]\Big] + \mathbb{E}\Big[\mathrm{g}[x]\Big] \\
\mathbb{E}\Big[\mathrm{f}[x]g[y]\Big] &= \mathbb{E}\Big[\mathrm{f}[x]\Big]\mathbb{E}\Big[\mathrm{g}[y]\Big] \qquad \text{if} \quad x, y \quad \text{independent}
\end{aligned}$$

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \mathbf{0}$

preactivación i-ésima en esa capa:

$$\boldsymbol{z}_k(i)=\sum_{j=1}^{D_h} w_{ij}\boldsymbol{h}_k(j)$$

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \mathbf{0}$

Elevando al cuadrado en ambos lados:

$$\boldsymbol{z}_k(i)^2 = \left(\sum_{j=1}^{D_h} w_{ij} \boldsymbol{h}_k(j) \right)^2$$

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \mathbf{0}$

Aplicando la esperanza en ambos lados:

$$E[\boldsymbol{z}_k(i)^2] = E\left[\left(\sum_{j=1}^{D_h} w_{ij} \boldsymbol{h}_k(j) \right)^2 \right]$$

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \mathbf{0}$
- Pesos, **Ω**$_k$, con distribución Normal: media 0 y varianza σ^2_Ω

$$E[\boldsymbol{z}_k(i)^2] = E\left[\left(\sum_{j=1}^{D_h} w_{ij} \boldsymbol{h}_k(j) \right)^2 \right]$$

La varianza es la esperanza de (x- media)^2 – pero aquí la media es 0

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \mathbf{0}$
- Pesos, **Ω**$_k$, con distribución Normal: media 0 y varianza σ^2_Ω

varianza de la preactivación i-ésima en esa capa:

$$\sigma^2_{z_k(i)} = E\left[\left(\sum_{j=1}^{D_h} w_{ij} \boldsymbol{h}_k(j) \right)^2 \right]$$

La varianza es la esperanza de (x- media)^2 – pero aquí la media es 0

Varianza de la preactivación

$$\sigma^2_{z_k(i)} = E\left[\left|\sum_{j=1}^{D_h} w_{ij} \boldsymbol{h_k}(j)\right|^2\right]$$

- Recordemos que:

$$\begin{aligned}(a+b)^2 &= a^2+2ab+b^2\\ (a+b+c)^2 &= a^2+b^2+c^2+2bc+2ca+2ab\\ &\vdots\\ \left(\sum_i a_i\right)^2 &= \sum_i a_i^2 + 2\sum_{i<j} a_i a_j\end{aligned}$$

Varianza de la preactivación

$$\sigma^2_{z_k(i)} = E\left[\sum_{j=1}^{D_h} (w_{ij} \boldsymbol{h_k}(j))^2 + 2\cdot\sum_{r<t} (w_{ir} \boldsymbol{h_k}(r) w_{it} \boldsymbol{h_k}(t))\right]$$

- Recordemos que:

$$\begin{aligned}(a+b)^2 &= a^2+2ab+b^2\\ (a+b+c)^2 &= a^2+b^2+c^2+2bc+2ca+2ab\\ &\vdots\\ \left(\sum_i a_i\right)^2 &= \sum_i a_i^2 + 2\sum_{i<j} a_i a_j\end{aligned}$$

Varianza de la preactivación

$$\sigma^2_{z_k(i)} = E\left[\sum_{j=1}^{D_h} (w_{ij} \boldsymbol{h_k}(j))^2 + 2\cdot\sum_{r<t} (w_{ir} \boldsymbol{h_k}(r) w_{it} \boldsymbol{h_k}(t))\right]$$

- Y con las propiedades de la esperanza:

$$\begin{aligned}\mathbb{E}\Big[k\Big] &= k\\ \mathbb{E}\Big[k\cdot \mathrm{g}[x]\Big] &= k\cdot\mathbb{E}\Big[\mathrm{g}[x]\Big]\\ \Rightarrow\quad \mathbb{E}\Big[\mathrm{f}[x]+\mathrm{g}[x]\Big] &= \mathbb{E}\Big[\mathrm{f}[x]\Big]+\mathbb{E}\Big[\mathrm{g}[x]\Big]\\ \mathbb{E}\Big[\mathrm{f}[x]g[y]\Big] &= \mathbb{E}\Big[\mathrm{f}[x]\Big]\mathbb{E}\Big[\mathrm{g}[y]\Big] \quad \text{if} \quad x,y \quad \text{independent}\end{aligned}$$

Varianza de la preactivación

$$\sigma^2_{z_k(i)} = \sum_{j=1}^{D_h} E[(w_{ij} \boldsymbol{h_k}(j))^2] + 2\cdot\sum_{r<t} E[w_{ir} \boldsymbol{h_k}(r) w_{it} \boldsymbol{h_k}(t)]$$

- Y con las propiedades de la esperanza:

$$\begin{aligned}\mathbb{E}\Big[k\Big] &= k\\ \mathbb{E}\Big[k\cdot \mathrm{g}[x]\Big] &= k\cdot\mathbb{E}\Big[\mathrm{g}[x]\Big]\\ \Rightarrow\quad \mathbb{E}\Big[\mathrm{f}[x]+\mathrm{g}[x]\Big] &= \mathbb{E}\Big[\mathrm{f}[x]\Big]+\mathbb{E}\Big[\mathrm{g}[x]\Big]\\ \mathbb{E}\Big[\mathrm{f}[x]g[y]\Big] &= \mathbb{E}\Big[\mathrm{f}[x]\Big]\mathbb{E}\Big[\mathrm{g}[y]\Big] \quad \text{if} \quad x,y \quad \text{independent}\end{aligned}$$

Varianza de la preactivación

$$\sigma^2_{z_k(i)} = \sum_{j=1}^{D_h} E[(w_{ij}\boldsymbol{h_k}(j))^2] + 2\cdot\sum_{r<t} E[w_{ir}\boldsymbol{h_k}(r)\, w_{it}\boldsymbol{h_k}(t)]$$

– Y con las propiedades de la esperanza:

$$\begin{aligned} \mathbb{E}\Big[k\Big] &= k \\ \mathbb{E}\Big[k\cdot \mathrm{g}[x]\Big] &= k\cdot\mathbb{E}\Big[\mathrm{g}[x]\Big] \\ \mathbb{E}\Big[\mathrm{f}[x]+\mathrm{g}[x]\Big] &= \mathbb{E}\Big[\mathrm{f}[x]\Big]+\mathbb{E}\Big[\mathrm{g}[x]\Big] \\ \Rightarrow\quad \mathbb{E}\Big[\mathrm{f}[x]g[y]\Big] &= \mathbb{E}\Big[\mathrm{f}[x]\Big]\mathbb{E}\Big[\mathrm{g}[y]\Big] \qquad \text{if} \quad x, y \quad \text{independent} \end{aligned}$$

Varianza de la preactivación

$$\sigma^2_{z_k(i)} = \sum_{j=1}^{D_h} E[w_{ij}^2]E[\boldsymbol{h_k}(j)^2] + 2\cdot\sum_{r<t} E[w_{ir}]E[\boldsymbol{h_k}(r)]E[w_{it}]E[\boldsymbol{h_k}(t)]$$

– Y con las propiedades de la esperanza:

$$\begin{aligned} \mathbb{E}\Big[k\Big] &= k \\ \mathbb{E}\Big[k\cdot \mathrm{g}[x]\Big] &= k\cdot\mathbb{E}\Big[\mathrm{g}[x]\Big] \\ \mathbb{E}\Big[\mathrm{f}[x]+\mathrm{g}[x]\Big] &= \mathbb{E}\Big[\mathrm{f}[x]\Big]+\mathbb{E}\Big[\mathrm{g}[x]\Big] \\ \Rightarrow\quad \mathbb{E}\Big[\mathrm{f}[x]g[y]\Big] &= \mathbb{E}\Big[\mathrm{f}[x]\Big]\mathbb{E}\Big[\mathrm{g}[y]\Big] \qquad \text{if} \quad x, y \quad \text{independent} \end{aligned}$$

Inicialización

- Las preactivaciones de una capa lineal:

$$\begin{aligned} \boldsymbol{z_k} &= \boldsymbol{\Omega_k h_k} + \boldsymbol{\beta_k} \\ &= \boldsymbol{\Omega_k a}[\boldsymbol{z_{k-1}}] + \boldsymbol{\beta_k} \end{aligned}$$

- Establecer todos los sesgos a 0: $\boldsymbol{\beta_k} = \boldsymbol{0}$
- Pesos, $\boldsymbol{\Omega}_k$, con distribución Normal: media 0 y varianza σ^2_Ω

$$\sigma^2_{z_k(i)} = \sum_{j=1}^{D_h} E[w_{ij}^2]E[\boldsymbol{h_k}(j)^2] + 2\cdot\cancel{\sum_{r<t} E[w_{ir}]E[\boldsymbol{h_k}(r)]E[w_{it}]E[\boldsymbol{h_k}(t)]}$$

Inicialización

- Las preactivaciones de una capa lineal:

$$\begin{aligned} \boldsymbol{z_k} &= \boldsymbol{\Omega_k h_k} + \boldsymbol{\beta_k} \\ &= \boldsymbol{\Omega_k a}[\boldsymbol{z_{k-1}}] + \boldsymbol{\beta_k} \end{aligned}$$

- Establecer todos los sesgos a 0: $\boldsymbol{\beta_k} = \boldsymbol{0}$
- Pesos, $\boldsymbol{\Omega}_k$, con distribución Normal: media 0 y varianza σ^2_Ω

$$\sigma^2_{z_k(i)} = \sum_{j=1}^{D_h} E[w_{ij}^2]\cdot E[\boldsymbol{h_k}(j)^2]$$

La varianza es la esperanza de (x- media)^2 – pero aquí la media es 0

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \boldsymbol{0}$
- Pesos, **Ω**k, con distribución Normal: media 0 y varianza σ_Ω^2

$$\sigma^2_{z_k(i)} = \sum_{j=1}^{D_h} \sigma_\Omega^2 \cdot E[\boldsymbol{h}_k(j)^2]$$

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \boldsymbol{0}$
- Pesos, **Ω**k, con distribución Normal: media 0 y varianza σ_Ω^2

$$\sigma^2_{z_k(i)} = \sigma_\Omega^2 \sum_{j=1}^{D_h} E[\boldsymbol{h}_k(j)^2]$$

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \boldsymbol{0}$
- Pesos, **Ω**k, con distribución Normal: media 0 y varianza σ_Ω^2
- Varianza de las activaciones en la capa k es σ_k^2

$$\sigma^2_{z_k(i)} = \sigma_\Omega^2 \sum_{j=1}^{D_h} E[\boldsymbol{h}_k(j)^2]$$

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \boldsymbol{0}$
- Pesos, **Ω**k, con distribución Normal: media 0 y varianza σ_Ω^2
- Varianza de las activaciones en la capa k es σ_k^2

$$\sigma^2_{z_k(i)} = \sigma_\Omega^2 \sum_{j=1}^{D_h} \sigma_k^2$$

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \mathbf{0}$
- Pesos, **Ω**k, con distribución Normal: media 0 y varianza σ_Ω^2
- Varianza de las activaciones en la capa k es σ_k^2

$$\sigma^2_{z_k(i)} = D_h \cdot \sigma_\Omega^2 \cdot \sigma_k^2$$

Inicialización

- Las preactivaciones de una capa lineal:

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

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \mathbf{0}$
- Pesos, **Ω**k, con distribución Normal: media 0 y varianza σ_Ω^2
- Varianza de las activaciones en la capa k es σ_k^2
- Todas las preactivaciones de la capa k comparten varianza:

$$\sigma^2_{z_k} = D_h \cdot \sigma_\Omega^2 \cdot \sigma_k^2$$

5.2.1 Inicialización de parámetros

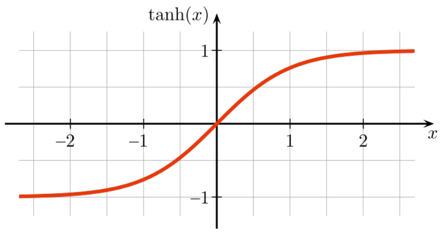
- Necesidad de la inicialización
- Inicialización de Xavier y de He

Understanding the difficulty of training deep feedforward neural networks. Xavier Glorot, Yoshua Bengio. AISTATS 2010: 249-256

Delving Deep into Rectifiers:Surpassing Human-Level Performance on ImageNet Classification
Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun. ICCV 2015

Activación tanh

- La función de activación tanh tiene la forma:



$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

- Si suponemos que al comienzo del entrenamiento las preactivaciones siempre se encuentran en la parte lineal:

$$\mathbf{tanh}[\boldsymbol{z}_k] = \boldsymbol{z}_k \Rightarrow \sigma^2_{z_k} = \sigma^2_{k+1} \Rightarrow \sigma^2_{k+1} = D_h \cdot \sigma_\Omega^2 \cdot \sigma_k^2$$

$$\Rightarrow \sigma^2_{z_{k-1}} = \sigma_k^2 \Rightarrow \sigma^2_{z_k} = D_h \cdot \sigma_\Omega^2 \cdot \sigma^2_{z_{k-1}}$$

Inicialización: Pregunta 3

- Las preactivaciones de una capa lineal:

$$\boldsymbol{z}_k = \boldsymbol{\Omega}_k \boldsymbol{h}_k + \boldsymbol{\beta}_k$$
$$= \boldsymbol{\Omega}_k \boldsymbol{a}[\boldsymbol{z}_{k-1}] + \boldsymbol{\beta}_k$$

- Establecer todos los sesgos a 0: $\boldsymbol{\beta}_k = \boldsymbol{0}$
- Pesos, $\boldsymbol{\Omega}_k$, con distribución Normal: media 0 y varianza σ_Ω^2
  - ¿Qué pasará al movernos por la red si la varianza de los pesos de cada capa es tal que $D_h \cdot \sigma_\Omega^2 < 1$?
  - ¿Qué pasará al movernos por la red si la varianza es tal que $D_h \cdot \sigma_\Omega^2 > 1$?

$$\sigma_{k+1}^2 = D_h \cdot \sigma_\Omega^2 \cdot \sigma_k^2$$

Problemas numéricos con el gradiente

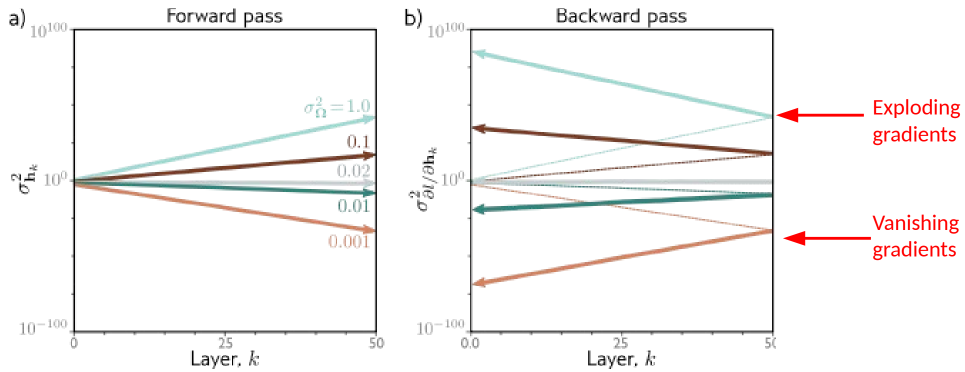


Figure 7.4 Weight initialization. Consider a deep network with 50 hidden layers and $D_h = 100$ hidden units per layer. The network has a 100 dimensional input $\mathbf{x}$ initialized with values from a standard normal distribution, a single output fixed at $y = 0$, and a least squares loss function. The bias vectors $\boldsymbol{\beta}_k$ are initialized to zero and the weight matrices $\boldsymbol{\Omega}_k$ are initialized with a normal distribution with mean zero and five different variances $\sigma^2_\Omega \in \{0.001, 0.01, 0.02, 0.1, 1.0\}$. a)

Relación entre varianzas (asume tanh)

- La relación entre la varianza de las activaciones en la capa k, y la varianza de las activaciones en la capa k+1 es:

$$\sigma_{k+1}^2 = D_h \cdot \sigma_\Omega^2 \cdot \sigma_k^2$$

- La relación entre la varianza de los gradientes en la capa oculta k+1 y la varianza de los gradientes en capa k es:

$$\sigma_k^2 = D_{h'} \cdot \sigma_\Omega^2 \cdot \sigma_{k+1}^2$$

Inicialización de Xavier

- Forward pass: se busca que la varianza de las activaciones en la capa k+1 sea igual que la varianza de las activaciones en la capa k:

$$\sigma_{k+1}^2 = D_h \cdot \sigma_\Omega^2 \cdot \sigma_k^2 \Rightarrow \boxed{\sigma_\Omega^2 = \frac{1}{D_h}}$$

Nº de unidades en la capa k

- Backward pass: se busca que la varianza de los gradientes en la capa k sea igual que la varianza de los gradientes en la capa k+1:

$$\sigma_k^2 = D_{h'} \cdot \sigma_\Omega^2 \cdot \sigma_{k+1}^2 \Rightarrow \boxed{\sigma_\Omega^2 = \frac{1}{D_{h'}}}$$

Nº de unidades en la capa k+1

Inicialización de Xavier

- Si $D_h \neq D_{h'}$ entonces podemos usar $(D_h+D_{h'})/2$ como número de unidades:

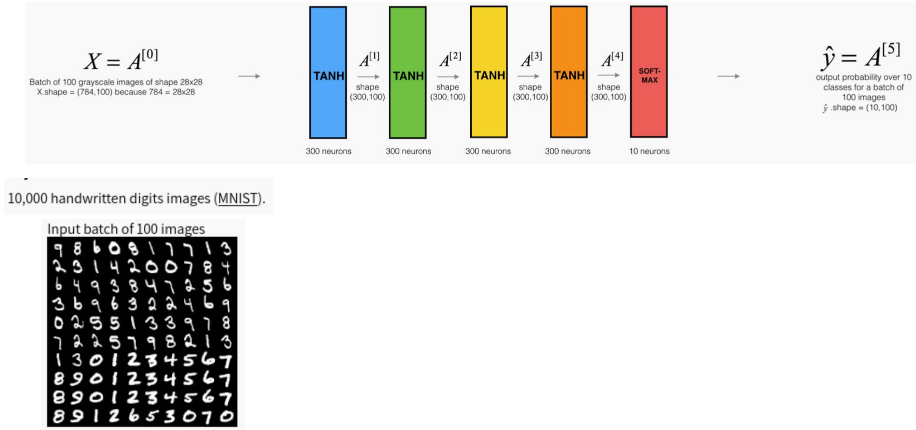
$$\sigma_\Omega^2 = \frac{1}{\frac{D_h + D_{h'}}{2}} = \frac{2}{\boxed{D_h + D_{h'}}}$$

N° de entradas + N° de salidas de la capa K

- Si esta varianza se utiliza en una distribución uniforme, U(a, -a) que tiene varianza $a^2/3$ → tendremos que usar una uniforme:

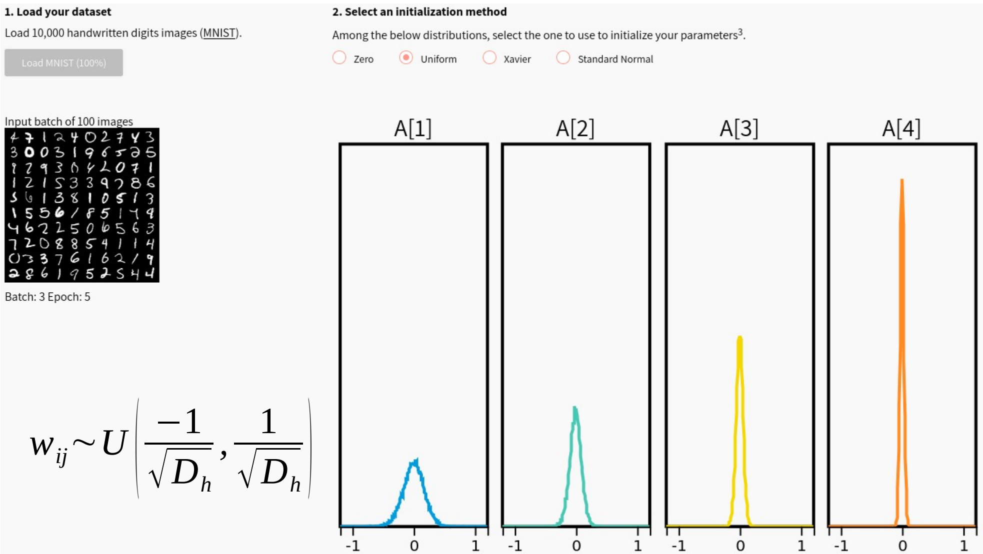
$$U\left(-\sqrt{\frac{6}{D_h + D_{h'}}}, \sqrt{\frac{6}{D_h + D_{h'}}}\right)$$

Otras Inicializaciones vs Xavier



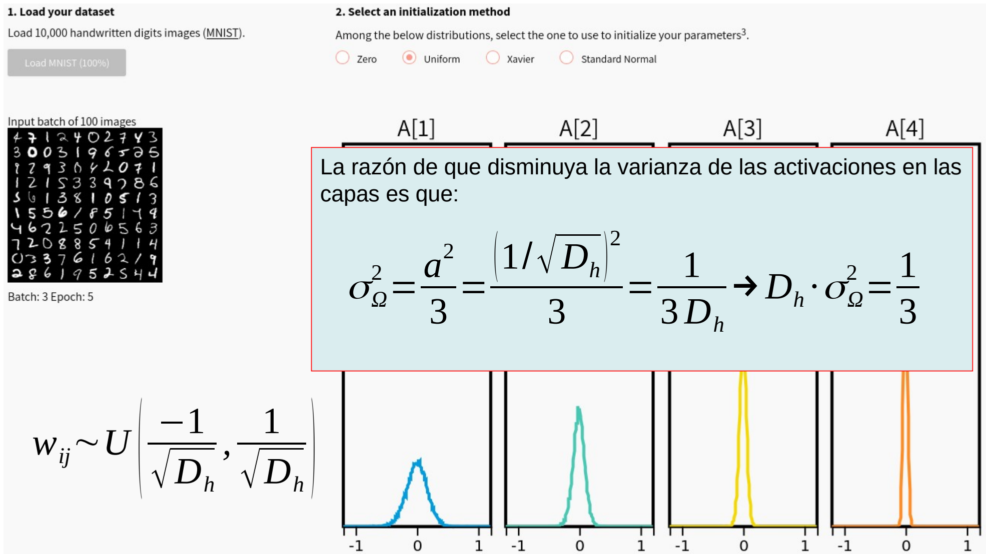
https://www.deeplearning.ai/ai-notes/initialization/index.html

Otras Inicializaciones vs Xavier



$$w_{ij} \sim U\left(\frac{-1}{\sqrt{D_h}}, \frac{1}{\sqrt{D_h}}\right)$$

Otras Inicializaciones vs Xavier

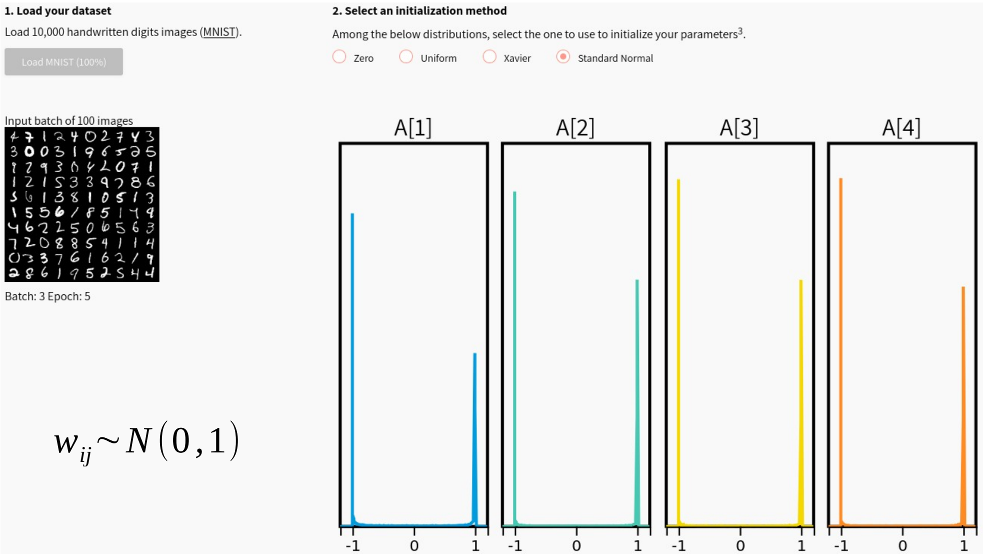


$$w_{ij} \sim U\left(\frac{-1}{\sqrt{D_h}}, \frac{1}{\sqrt{D_h}}\right)$$

La razón de que disminuya la varianza de las activaciones en las capas es que:

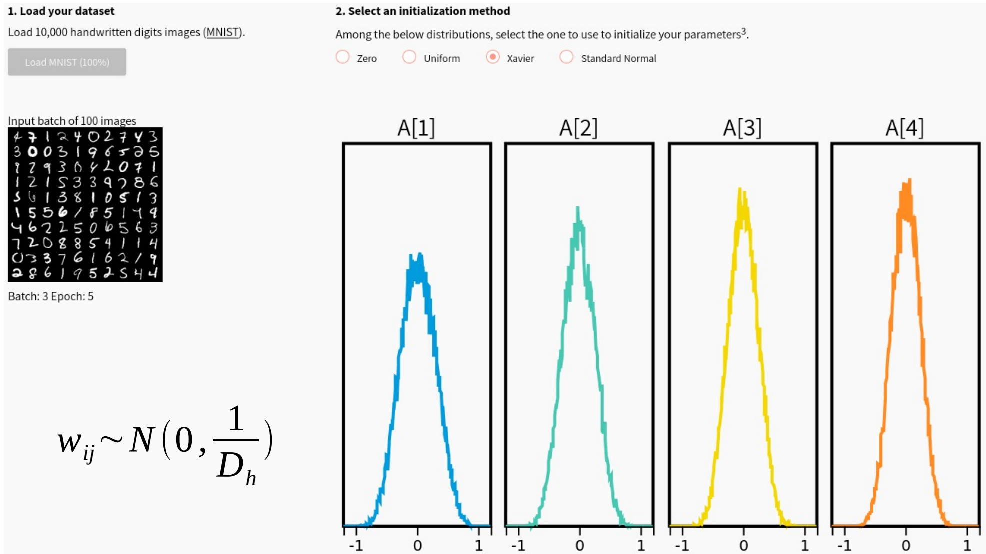
$$\sigma_\Omega^2 = \frac{a^2}{3} = \frac{\left(1/\sqrt{D_h}\right)^2}{3} = \frac{1}{3D_h} \rightarrow D_h \cdot \sigma_\Omega^2 = \frac{1}{3}$$

Otras Inicializaciones vs Xavier



$$w_{ij} \sim N(0,1)$$

Otras Inicializaciones vs Xavier



$$w_{ij} \sim N\left(0, \frac{1}{D_h}\right)$$

Relación entre varianzas (asume ReLU)

- ¿Y qué pasa si tenemos ReLU como función de activación?

$$\begin{aligned}
\sigma_{z_k}^2 &= \sigma_\Omega^2 \sum_{j=1}^{D_h} E[\boldsymbol{h_k}(j)^2] \\
&= \sigma_\Omega^2 \sum_{j=1}^{D_h} E[ReLU(\boldsymbol{z_{k-1}}(j))^2] \\
&= \sigma_\Omega^2 \sum_{j=1}^{D_h} \int_{-\infty}^{\infty} ReLU(\boldsymbol{z_{k-1}}(j))^2 Pr(\boldsymbol{z_{k-1}}(j)) dz_k(j) \\
&= \sigma_\Omega^2 \sum_{j=1}^{D_h} \int_{0}^{\infty} \boldsymbol{z_{k-1}}(j)^2 Pr(\boldsymbol{z_{k-1}}(j)) dz_k(j) \\
&= \sigma_\Omega^2 \sum_{j=1}^{D_h} \frac{\sigma_{z_{k-1}}^2}{2} = \frac{D_h \sigma_\Omega^2 \sigma_{z_{k-1}}^2}{2}
\end{aligned}$$

Relación entre varianzas (asume ReLU)

- La relación entre la varianza de las preactivaciones en la capa k, y la varianza de las preactivaciones en la capa k-1 es:

$$\sigma_{z_k}^2 = \frac{1}{2} \cdot D_h \cdot \sigma_\Omega^2 \cdot \sigma_{z_{k-1}}^2$$

- La relación entre la varianza de los gradientes de las preactivaciones en la capa k-1 y la varianza de los gradient

$$\sigma_{z_{k-1}}^2 = \frac{1}{2} \cdot D_{h'} \cdot \sigma_\Omega^2 \cdot \sigma_{z_k}^2$$

Inicialización de He (asume ReLU)

- Forward pass: se busca que la varianza de las preactivaciones en la capa k sea igual que la varianza de las preactivaciones en la capa k-1:

$$\sigma^2_{z_k} = \frac{1}{2} \cdot D_h \cdot \sigma^2_{\Omega} \cdot \sigma^2_{z_{k-1}} \Rightarrow \boxed{\sigma^2_{\Omega} = \frac{2}{D_h}}$$

N° de unidades en la capa k

- Backward pass: se busca que la varianza de los gradientes en la capa k-1 sea igual que la varianza de los gradientes en la capa k:

$$\sigma^2_{z_{k-1}} = \frac{1}{2} \cdot D_{h'} \cdot \sigma^2_{\Omega} \cdot \sigma^2_{z_k} \Rightarrow \boxed{\sigma^2_{\Omega} = \frac{2}{D_{h'}}}$$

N° de unidades en la capa k

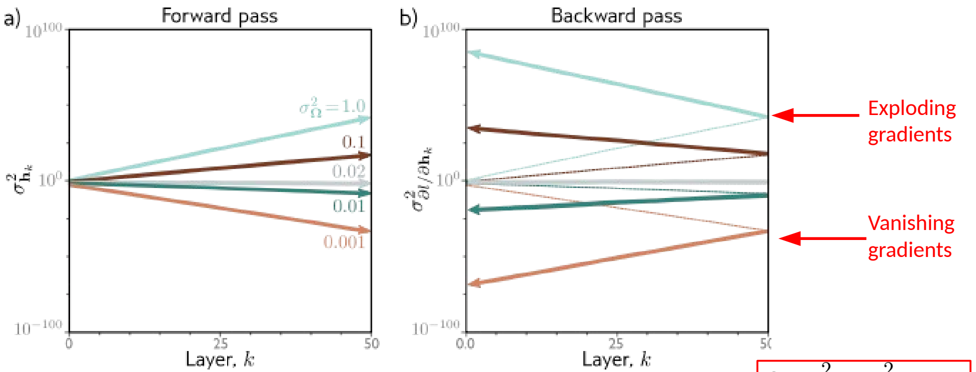
Inicialización de He (asume ReLU)

- Si $D_h \neq D_{h'}$ entonces podemos usar $(D_h+D_{h'})/2$ como número de unidades:

$$\sigma^2_{\Omega} = \frac{2}{\frac{D_h + D_{h'}}{2}} = \frac{4}{\boxed{D_h + D_{h'}}}$$

N° de entradas + N° de salidas de la capa K

Problemas numéricos con el gradiente



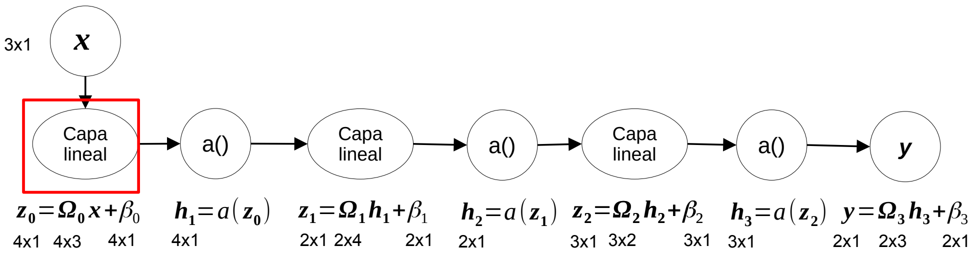
$$\boxed{\sigma^2_\Omega = \frac{2}{D_h} = \frac{2}{100} = 0.02}$$

Figure 7.4 Weight initialization. Consider a deep network with 50 hidden layers and $D_h = 100$ hidden units per layer. The network has a 100 dimensional input $\mathbf{x}$ initialized with values from a standard normal distribution, a single output fixed at $y = 0$, and a least squares loss function. The bias vectors $\boldsymbol{\beta}_k$ are initialized to zero and the weight matrices $\boldsymbol{\Omega}_k$ are initialized with a normal distribution with mean zero and five different variances $\sigma^2_\Omega \in \{0.001, 0.01, 0.02, 0.1, 1.0\}$. a)

5.2 Optimización con redes profundas

- 5.2.1 Backpropagation
- 5.2.2 Inicialización de parámetros
- 5.2.3 Normalización
- 5.2.4 Regularización

El problema de los valores grandes ...

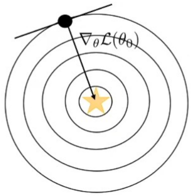


- Recordemos: $\frac{\partial L}{\partial \Omega_0}=\frac{\partial L}{\partial z_0}\cdot\frac{\partial z_0}{\partial \Omega_0}=\frac{\partial L}{\partial z_0}\cdot x^T$
- ¿Qué pasa si en **x**=(x_1, x_2, x_3)^T uno de los x_i tiene una magnitud mucho mayor que los otros?

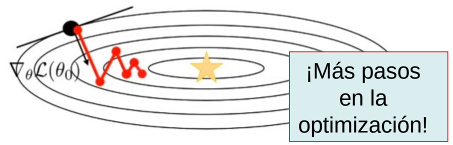
Las componentes en el gradiente con respecto a $\mathbf{\Omega_0}$ tendrán magnitudes muy diferentes (son el resultado de un producto con $\mathbf{x}^T$)

El problema de los valores grandes ...

- Con magnitudes iguales en los gradientes tendremos una función de coste a optimizar:

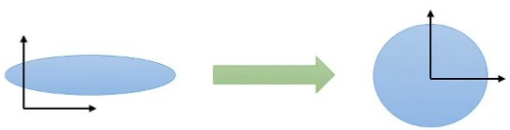


- Con magnitudes muy diferentes en los gradientes tendremos:



El problema de los valores grandes ...

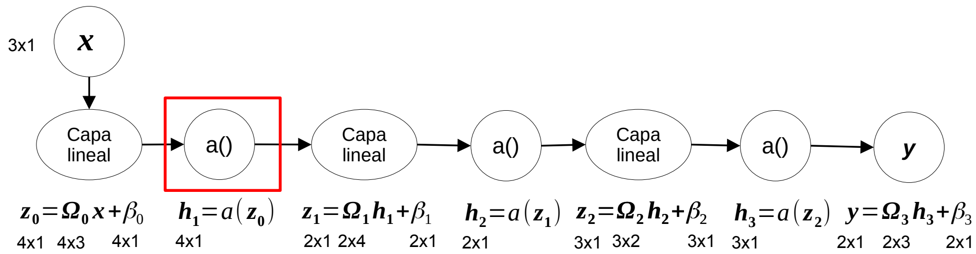
- Estandarizar entradas y etiquetas (regresión) con *μ=0* y *σ=1*:



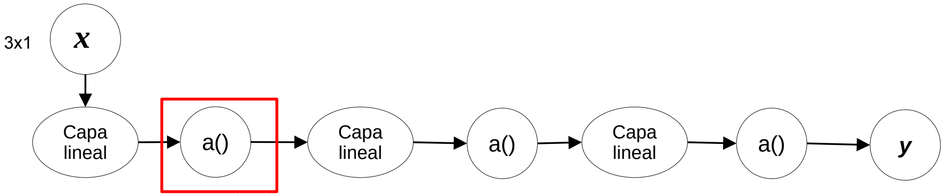
- Para hacer *μ=0:* $\bar{x}_i=x_i-E[x]$ $E[x]\approx\frac{1}{N}\cdot\sum_{i=1}^{N}x_i$
- *Para σ=1:* $\bar{x}_i=\frac{x_i-E[x]}{\sqrt{E[(x-E[x])^2]}}$

El problema de los valores grandes ...

- Estandarizar entradas y etiquetas (regresión) con *μ=0* y *σ=1*
- ¿Qué pasa si empezamos a tener activaciones con magnitudes muy diferentes en cualquier capa?



¿Estandarizar las activaciones?



$\boldsymbol{z}_0=\boldsymbol{\Omega}_0\boldsymbol{x}+\beta_0$ (4x1 4x3 4x1) $\boldsymbol{h}_1=a(\boldsymbol{z}_0)$ (4x1) $\boldsymbol{z}_1=\boldsymbol{\Omega}_1\boldsymbol{h}_1+\beta_1$ (2x1 2x4 2x1) $\boldsymbol{h}_2=a(\boldsymbol{z}_1)$ (2x1) $\boldsymbol{z}_2=\boldsymbol{\Omega}_2\boldsymbol{h}_2+\beta_2$ (3x1 3x2 3x1) $\boldsymbol{h}_3=a(\boldsymbol{z}_2)$ (3x1) $\boldsymbol{y}=\boldsymbol{\Omega}_3\boldsymbol{h}_3+\beta_3$ (2x1 2x3 2x1)

- Media y desviación típica del componente *i* del vector de activaciones $\boldsymbol{h}_1$ (N datos de entrenamiento):

$$\boldsymbol{\mu}_1(i)=\frac{1}{N}\cdot\sum_{j=1}^{N}\boldsymbol{h}_1(i,j) \qquad \boldsymbol{\sigma}_1(i)=\sqrt{\frac{1}{N}\cdot\sum_{j=1}^{N}(\boldsymbol{h}_1(i,j)-\boldsymbol{\mu}_1(i))^2}$$

Dependen de $\boldsymbol{\Omega}_0$ y $\boldsymbol{\beta}_0$ y ¡hay que recalcular media y varianza en cada paso de cálculo de gradiente!

- La activación normalizada: $\bar{\boldsymbol{h}}_1(i)=\frac{\boldsymbol{h}_1(i)-\boldsymbol{\mu}_1(i)}{\boldsymbol{\sigma}_1(i)}$
- Y la preactivación siguiente: $\boldsymbol{z}_1=\boldsymbol{\Omega}_1\bar{\boldsymbol{h}}_1+\beta_1$

Batch normalization

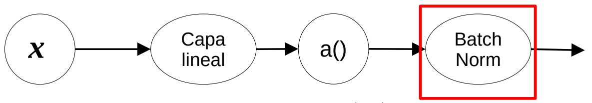
- Media y varianza del componente *i* del vector de activaciones $\boldsymbol{h}_1$ calculado con los B datos del mini-batch:

$$\boldsymbol{\mu}_1(i)\approx\frac{1}{B}\cdot\sum_{j=1}^{B}\boldsymbol{h}_1(i,j) \qquad \boldsymbol{\sigma}_1(i)\approx\sqrt{\frac{1}{B}\cdot\sum_{j=1}^{B}(\boldsymbol{h}_1(i,j)-\boldsymbol{\mu}_1(i))^2}$$

- La activación normalizada se modifica con dos parámetros que se aprenden:

$$\bar{\boldsymbol{h}}_1(i) = \frac{\boldsymbol{h}_1(i)-\boldsymbol{\mu}_1(i)}{\boldsymbol{\sigma}_1(i)}\cdot\gamma + \beta$$

Batch normalization



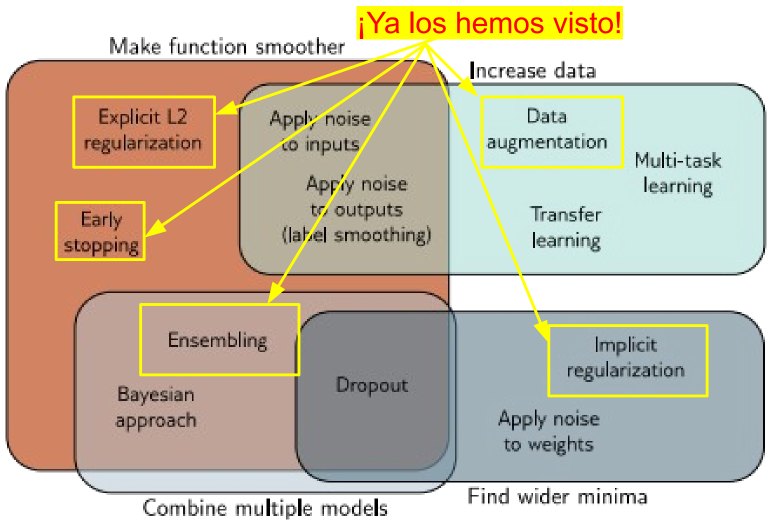
3x1 $\boldsymbol{z}_0=\boldsymbol{\Omega}_0\boldsymbol{x}+\beta_0$ (4x1 4x3 4x1) $\boldsymbol{h}_1=a(\boldsymbol{z}_0)$ (4x1)

- Hay una capa adicional de normalización del mini-batch (batch norm)
- Sus parámetros se entrenan con backpropagation.
- ¿Qué se hace en inferencia con la media y la varianza de cada capa?

5.2 Optimización con redes profundas

- 5.2.1 Backpropagation
- 5.2.2 Inicialización de parámetros
- 5.2.3 Normalización
- 5.2.4 Regularización

Métodos de regularización



Recordatorio Bagging (Classifier Ensemble)

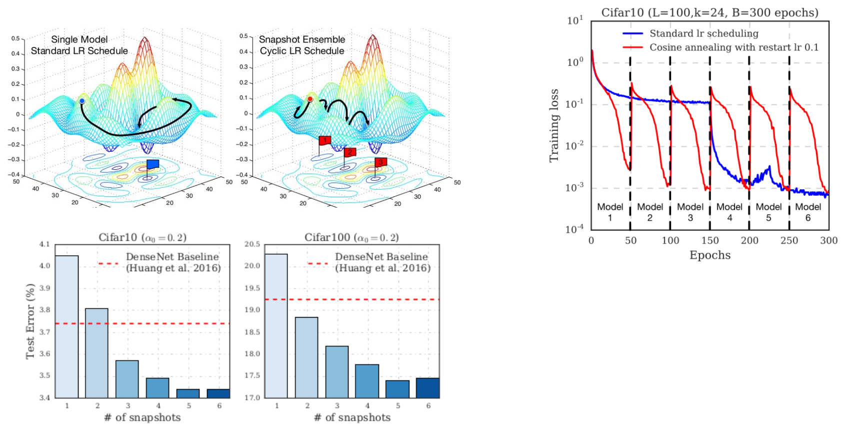
- Algoritmo **B**ootstrapping **Agg**regat**ing** (Leo Breiman 1996)
  - Para $i = 1 \dots M$
    - Obtener $n^* < n$ muestra de $\mathcal{D}$ con reemplazamiento.
    - Aprender un clasificador C_i sobre la nueva muestra.
  - El clasificador final es una votación de los $C_i, \dots, C_M$.
- El clasificador resultante del Bagging:
  - Incrementa la estabilidad del clasificador
  - Reduce la varianza:
    - Mejora las técnicas de varianza alta (Redes de neuronas, Árboles de decisión)
    - No mejora las técnicas de varianza baja (Un clasificador lineal)

Bagging de redes de neuronas

- En el entrenamiento de las redes existen muchas fuentes de aleatoriedad:
  - Inicialización aleatoria
  - Permutación aleatoria del mini-batch
  - Descenso de Gradiente Estocástico
- Se podría hacer Bagging de la misma red entrenada sobre los mismos datos cambiando la semilla del generador de números aleatorios:
  - Se necesitan M entrenamientos ¡muy costoso!

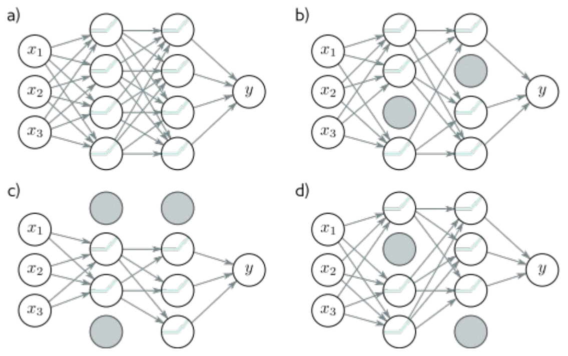
Bagging de redes de neuronas

- Tomar parámetros de diferentes instantes de un único entrenamiento (snapshots) como los modelos a usar en Bagging



Snapshot Ensembles: Train 1, Get M for Free. Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, Kilian Q. Weinberger: ICLR 2017

Dropout: apagar neuronas aleatoriamente



Aleatoriamente algunas activaciones se igualan a 0 en la paso hacia adelante (*forward pass*)

En el entrenamiento:

Para toda $h_{i,j}$, cambiarla por $h_{i,j} \cdot m_{i,j}$

$m_{i,j}$ ~Bernoulli(0.5) (1 con probabilidad 0.5, 0 en otro caso)

Dropout: apagar neuronas aleatoriamente

```
p = 0.5 # probability of keeping a unit active. higher = less dropout

def train_step(X):
  """ X contains the data """

  # forward pass for example 3-layer neural network
  H1 = np.maximum(0, np.dot(W1, X) + b1)
  U1 = np.random.rand(*H1.shape) < p # first dropout mask
  H1 *= U1 # drop!
  H2 = np.maximum(0, np.dot(W2, H1) + b2)
  U2 = np.random.rand(*H2.shape) < p # second dropout mask
  H2 *= U2 # drop!
  out = np.dot(W3, H2) + b3

  # backward pass: compute gradients... (not shown)
  # perform parameter update... (not shown)
```

Andrej Karpathy

Dropout en inferencia ...

En el entrenamiento:

Para toda $h_{i,j}$, cambiarla por $h_{i,j} \cdot m_{i,j}$

$m_{i,j}$ ~Bernoulli(0.5) (1 con probabilidad 0.5, 0 en otro caso)

En inferencia:

- Podríamos realizar la inferencia muchas veces con una máscaras de dropout generada cada vez y hacer Bagging explícito con los resultados.

  ¡Eso es costoso!

Dropout en inferencia ...

En el entrenamiento:

Para toda $h_{i,j}$, cambiarla por $h_{i,j} \cdot m_{i,j}$

$m_{i,j}$ ~Bernoulli(0.5) (1 con probabilidad 0.5, 0 en otro caso)

En inferencia (solución 1):

- ¿Y si usamos todas las neuronas ($m_{i,j}$=1) en inferencia?

  Entrenamiento: en promedio ½ de las activaciones son 0 en cada capa lineal.

  Inferencia: Ninguna de ellas es 0, con lo que $\mathbf{z}_k(i) = \Sigma_j w_{ij} \cdot \boldsymbol{h}_k(j)$ se multiplicarán por 2.

  – Solución: $\boldsymbol{\Omega}_k(j) = 1/2 \cdot \boldsymbol{\Omega}_k(j)$ (dividir los pesos por

Dropout en inferencia … otra manera

En el entrenamiento (solución 2):

Para toda $h_{i,j}$, cambiarla por $h_{i,j} \cdot m_{i,j}$

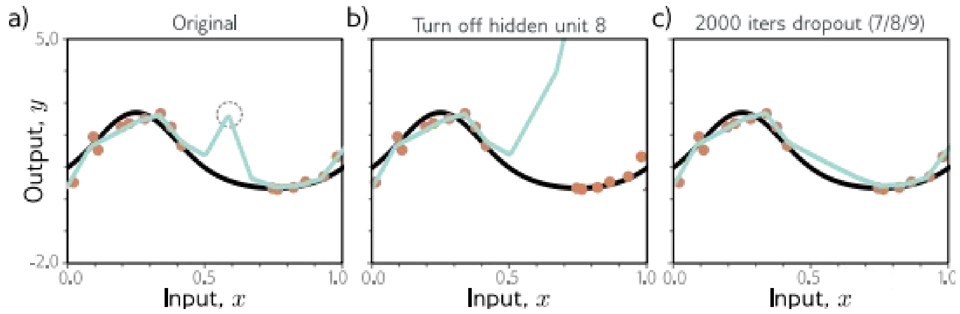
$m_{i,j}$ ~Bernoulli(p)/p (1/p con probabilidad p, 0 en otro caso)

```
p = 0.5 # probability of keeping a unit active. higher = less dropout

def train_step(X):
  # forward pass for example 3-layer neural network
  H1 = np.maximum(0, np.dot(W1, X) + b1)
  U1 = (np.random.rand(*H1.shape) < p) / p # first dropout mask. Notice /p!
  H1 *= U1 # drop!
  H2 = np.maximum(0, np.dot(W2, H1) + b2)
  U2 = (np.random.rand(*H2.shape) < p) / p # second dropout mask. Notice /p!
  H2 *= U2 # drop!
  out = np.dot(W3, H2) + b3

  # backward pass: compute gradients... (not shown)
  # perform parameter update... (not shown)

def predict(X):
  # ensembled forward pass
  H1 = np.maximum(0, np.dot(W1, X) + b1) # no scaling necessary
  H2 = np.maximum(0, np.dot(W2, H1) + b2)
  out = np.dot(W3, H2) + b3
```

test time is unchanged!

Andrej Karpathy

Dropout



Puede eliminar puntos estimados que están lejos de los datos y no contribuyen a la pérdida de entrenamiento.