UNIDADES NO LINEALES

Consideremos el caso en el cual la neurona de salida tiene una función de activación NO LINEAL

$g(z)$.

Podemos aplicar la misma regla de descenso por el gradiente, haciendo pequeños cambios en los pesos sinápticos, en la dirección contraria al gradiente de la función error.

$$E(\tilde{w}) = \frac{1}{2} \sum_{i}^{N} \sum_{\mu}^{P} \left(\xi_i^\mu - O_i^\mu\right)^2$$

$$= \frac{1}{2} \sum_{i} \sum_{\mu} \left(\xi_i^\mu - g(h_i^\mu)\right)^2$$

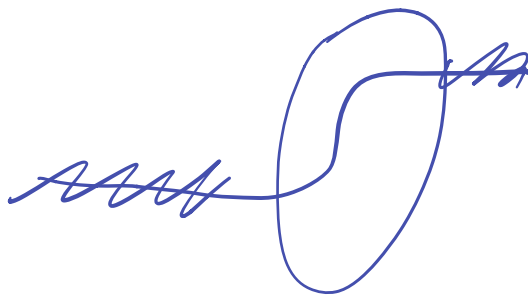
$$= \frac{1}{2} \sum_{i} \sum_{\mu} \left(\xi_i^\mu - g\left(\sum_k w_{ik} \xi_k^\mu\right)\right)^2$$

$$\frac{\partial E}{\partial w_{ik}} = \frac{1}{2} \sum_{i} \sum_{\mu} 2 \left(\xi_i^\mu - g(h_i^\mu)\right) \left(-g'(h_i^\mu)\right) \frac{dh_i^\mu}{dw_{ik}}$$

$$\frac{\partial E}{\partial w_{ik}} = \frac{1}{2}\sum_i \sum_\mu 2\left(\xi_i^\mu - g(h_i^\mu)\right)\left(-g'(h_i^\mu)\right)\xi_k^\mu$$

$$\boxed{\frac{\partial E}{\partial w_{ik}} = -\sum_i \sum_\mu \left(\xi_i^\mu - O_i^\mu\right) g'(h_i^\mu)\,\xi_k^\mu}$$

$$\frac{\partial E}{\partial w_{ik}} = -\sum_{\mu i} \delta_i^\mu\, \xi_k^\mu$$

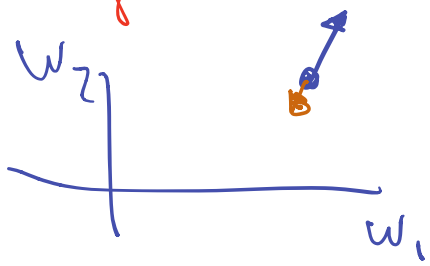


$$\delta_i^\mu \equiv \left[\xi_i^\mu - O_i^\mu\right] g'(h_i^\mu)$$

$$\Delta w_{ik} = \eta\, \delta_i^\mu\, \xi_k^\mu$$

Ahora podemos hacer lo mismo que con las neuronas de salida lineales: ENTRENAR LA RED COMENZANDO CON SINAPSIS $\{\vec{w}_i\}$ ALEATORIAS

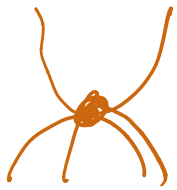
Los problemas del descenso por el gradiente



- La función error (o costo) tiene muchísimos mínimos locales, debido a la forma cuadrática de funciones fuertemente no lineales, a la alternancia de signos de los w_{ik} y su elasticidad
- El método de descenso por el gradiente (MDG) se puede atrapados en mínimos locales con valores altos de E.

- Las funciones no lineales son lentas de evaluar, y más lento es evaluar sus derivadas $g'(h)$.

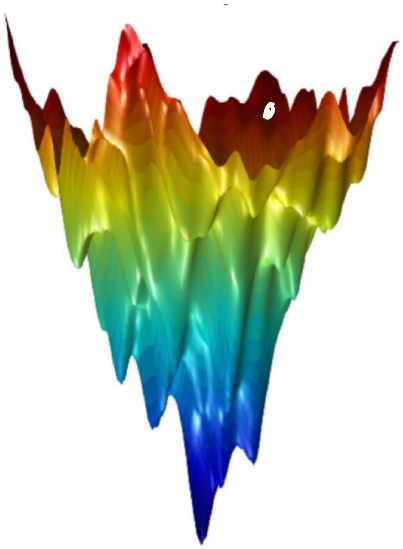
- El MDG es muy sensitivo al valor de la razón de aprendizaje η, y no es posible conocer exactamente su valor óptimo.

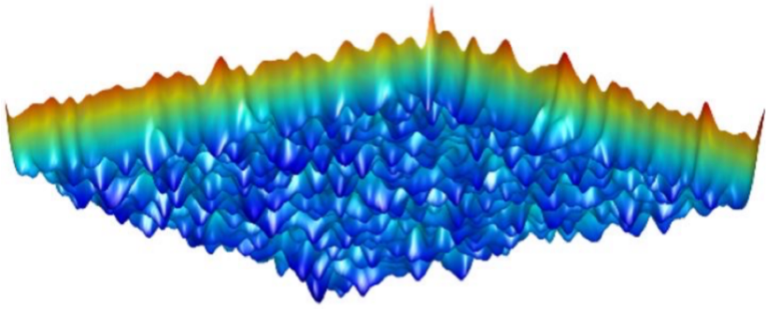


- El MDG depende FUERTEMENTE del valor inicial de los parámetros $\{\vec{w}_i\}$, por culpa de la rugosidad de E

- El MDG trata los $M \times N$ direcciones en $\mathbb{R}^N$ de la misma forma (con el mismo M), siendo que no todas las direcciones son igualmente importantes. Nos gustaría dar "pasos" largos en las direcciones en las cuales la función E varía poco, y "pasos" cortos en las que varía mucho.

- El método de descenso por el gradiente requiere de tiempos exponencialmente largos para poder escapar de puntos de ensilladura.





Podemos encontrar un $\bar{w}_{opt}$. Nos paramos en un punto cualquiera de $\mathbb{R}^{N+M}$ y denotamos fs $\bar{w} \in \mathbb{R}^{N+M}$ a ese punto

$$E(\underbrace{\bar{w} + \bar{v}}_{\bar{w}'}) = E(\bar{w}) + \partial_{\bar{w}} E(\bar{w})\bar{v} + \frac{1}{2}\partial^2_{\bar{w}} E(w)\bar{v}$$

Sea $\bar{w}_{min}$ el valor de $\bar{w}$ que minimiza $E(\bar{w})$.

$$\bar{w} = \bar{w}_{min} + \bar{v}$$

Por ser mínimo

$$\partial_{\bar{w}} E(\bar{w})_{min} = 0$$

Derivando con respecto a $\bar{v}$

$$\frac{\partial E}{\partial \bar{w}} \cdot \frac{\partial \bar{w}+\bar{v}}{\partial v} = \frac{\partial E}{\partial w} + \frac{\partial E}{\partial w} + 2\left(\frac{\partial^2 E}{\partial w^2}\right)\bar{v}$$

$$\bar{v} = -\left(\frac{\partial E}{\partial \bar{w}_m}\right) \cdot \frac{1}{\left(\frac{\partial^2 E}{\partial w^2}\right)}$$

$$\bar{w} = \bar{w}_{min} + \bar{v}$$

$$= \bar{w}_{min} + \left[\frac{1}{\left(\frac{\partial^2 E}{\partial \bar{w}^2}\right)}\right] \cdot \frac{\partial E}{\partial \bar{w}}$$

$$\eta_{OPT} = \frac{1}{\left(\frac{\partial^2 E(w)}{\partial \bar{w}^2}\right)}$$

oo Es es CARÍSIMO !! o

!!

Si fuésemos muy formales

$$E(\bar{w}+\bar{v}) \approx E(\bar{w}) + \vec{\nabla}_{\bar{w}} E(\bar{w}) \cdot \bar{v} + \frac{1}{2} \bar{v}^+ H(\bar{w}) \bar{v}$$

H : Hessiano de E

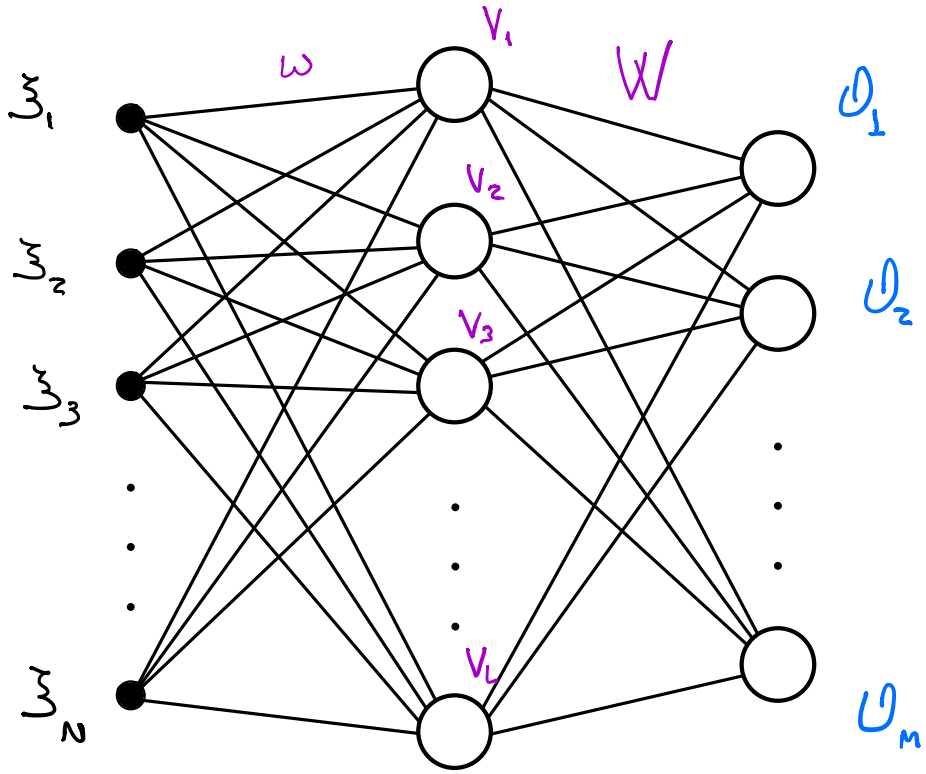
Si estamos en un mínimo $\nabla_{\bar{w}} E(\bar{w}+\bar{v}) = 0$

Derivando con respecto a $\bar{v}$

$$\nabla_{\bar{w}} E(\bar{w}) = - H(\bar{w})\, \bar{v}_{opt}$$

$$\bar{v}_{opt} \approx \bar{v}_t = H^{-1}(\bar{w}_t)\, \nabla_w (E(\bar{w}_t))$$

BACK PROPAGATION PARA UNA CAPA OCULTA



$$h_j^\mu = \sum_{k=1}^{N} w_{jk}\,\xi_k^\mu$$

$$V_j^\mu = g_1\left(h_j^\mu\right) = g_1\left(\sum_{k=1} w_{jk}\,\xi_k^\mu\right)$$

$$h_i^\mu = \sum_{j=1}^{L} W_{ij} V_j$$

$$O_i^\mu = g_2(h_i^\mu) = g_2\left(\sum_j W_{ij} V_j^\mu\right)$$

$$= g_2\left(\sum_j W_{ij} \cdot g_1\left(\sum_{k=1} w_{jk} \xi_k^\mu\right)\right)$$

$$E[\{\bar{w}\}] = \frac{1}{2} \sum_{\mu=1}^{P} \sum_{i=1}^{M} [\xi_i^\mu - O_i^\mu]^2$$

$$= \frac{1}{2} \sum_{\mu=1}^{P} \sum_{i=1}^{M} \left[\xi_i^\mu - g_2\left(\sum_j W_{ij} \cdot g_1\left(\sum_{k=1} w_{jk} \xi_k^\mu\right)\right)\right]^2$$

$$W_{ij}^{\text{nuevo}} = W_{ij}^{\text{anterior}} + \Delta W_{ij}$$

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

$$= \eta \sum_{\mu=1}^{P} \frac{2}{2} [\zeta_i^\mu - O_i^\mu] g'(h_i^\mu) V_j^\mu$$

$$= \eta \sum_{\mu=1}^{P} \delta_i^\mu V_j^\mu$$

$$\delta_i^\mu \equiv g'(h_i^\mu)[\zeta_i^\mu - O_i^\mu]$$

Entonces, presentamos el elemento μ del conjunto de entrenamiento, $\vec{\xi}^\mu$, al perceptron con una capa oculta, y obtenemos el resultado $\vec{O}^\mu$.

Con $\vec{O}^\mu$ calculamos el error cuadrático $E(\{w\})$,

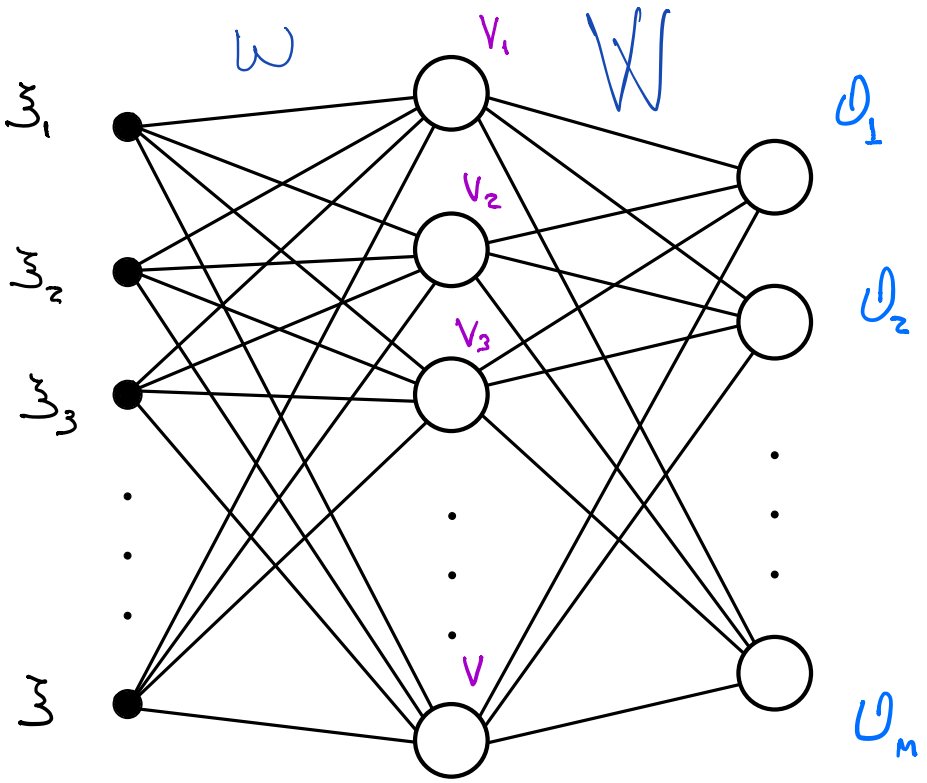
$$\Delta \omega_{jk} = -\eta \frac{\partial E}{\partial \omega_{jk}}$$

$$= \eta \sum_{\mu}^{P} \left[\zeta_i^\mu - O_i^\mu \right] g_1'(h_i^\mu) W_{ij} g_2'(h_j^\mu) \xi_k^\mu$$

$$= \eta \sum_{\mu}^{P} \delta_i^\mu W_{ij} g'(h_j^\mu) \xi_k^\mu$$

$$= \eta \sum_{\mu}^{P} \delta_j^\mu \xi_k^\mu$$

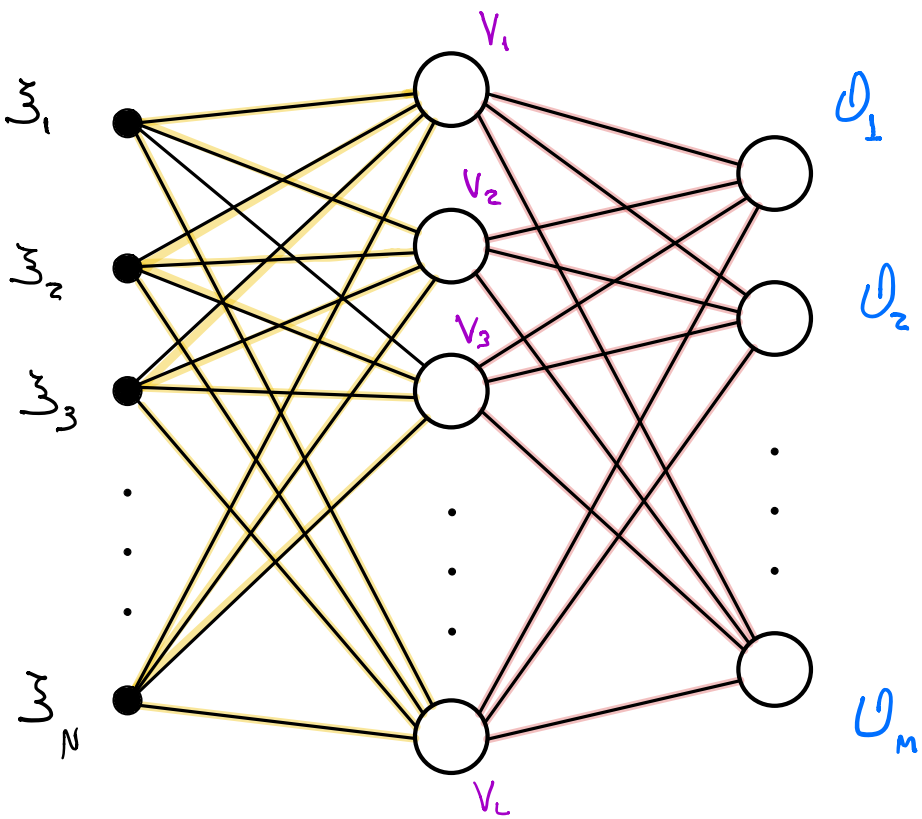
$$\delta_j^\mu = g_2'(h_j^\mu) \sum_i W_{ij} \delta_i^\mu$$



$\vec{\xi}^{\mu} \longrightarrow \vec{V}^{\mu} \longrightarrow \vec{O}^{\mu} \longrightarrow E$

$\{\vec{w}_k\}$ ← $\vec{\delta}^{\mu}$ ← $\{\vec{W}_i\}$ ← $\vec{\delta}^{\mu}$ ← $2E$

$k = 1, 2, \ldots, L$ $\qquad$ $i = 1, 2, \ldots, M$



¿cuantos acoplamientos tenemos?

$$(N \times L) + (L \times M) = L \times (N+M)$$

EL ALGORITMO

- presento $\vec{\xi}^{\mu}$, calculo $\vec{V}^{\mu}$, luego $\vec{O}^{\mu}$.
- con $\vec{O}^{\mu}$ y $\vec{\xi}^{\mu}$ calculamos $E(\{\vec{w}\})$
- con E calculo los δ_i^{μ} $(i=1,\dots,M)$ y con ellos actualizo los pesos W_{ij}

$$W_{ij}^{\text{nuevo}} = W_{ij}^{\text{anterior}} + \Delta W_{ij}$$

con los δ_i^{μ} calculo los δ_j^{μ} $(j=1,2,\dots,L)$ y con ellos actualizamos los pesos w_{jk}

$$w_{jk}^{\text{nuevo}} = w_{jk}^{\text{anterior}} + \Delta w_{jk}$$

EL ALGORITMO

- $\mu = 1$

WHILE $(\mu \leq P)$ repeat

1.- presento $\vec{\xi}^{\mu}$, calculo $\vec{V}^{\mu}$, luego $\vec{O}^{\mu}$.

2.- Con $\vec{O}^{\mu}$ y $\vec{S}^{\mu}$ calculamos $E(\{\vec{w}\})$

3.- Con E calculo los δ_i^{μ} $(i=1,\dots,M)$ y con ellos actualizo los pesos W_{ij}

$$W_{ij}^{\text{nuevo}} = W_{ij}^{\text{anterior}} + \Delta W_{ij} \qquad \Delta W_{ij} = \sum_{\mu}^{M} \omega_{ij}$$

4.- Con los δ_i^{μ} calculo los δ_j^{μ} $(j=1,2,\dots,L)$ y con ellos actualizamos los pesos ω_{jk}

$$\omega_{jk}^{\text{nuevo}} = \omega_{jk}^{\text{anterior}} + \Delta\omega_{jk}$$

$\mu = \mu + 1$

Hay diferentes formas de hacer el proceso

- **en línea**: luego de presentar el ejemplo redefino todos los acoplamientos sinápticos.

- **en batch**: luego de presentar el ejemplo almacena la variación de los acoplamientos sinápticos pero actualizo cuando terminé de mostrarle los P elementos del conjunto de entrenamiento.

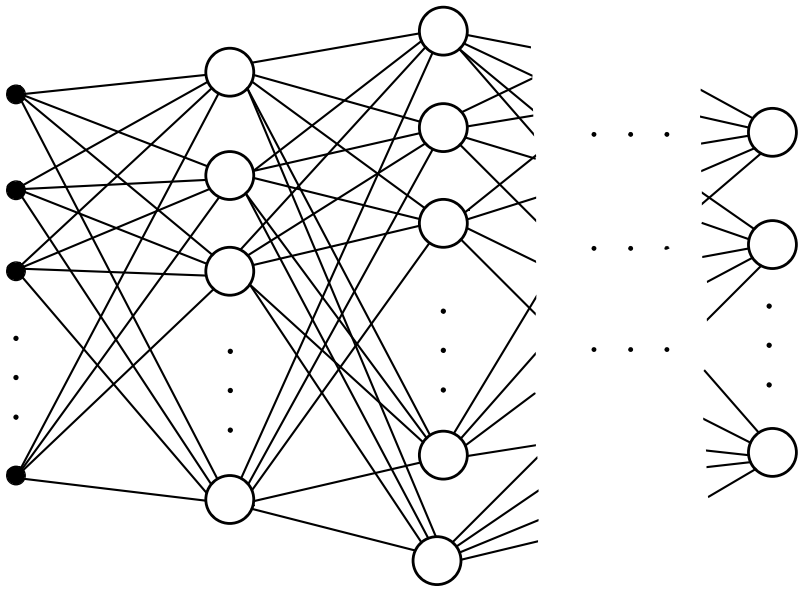
$$\Delta w_{ij} = \Delta w_{ij}^{1} + \Delta w_{ij}^{2} + \cdots + \Delta w_{ij}^{P}$$

$$\Delta w_{jk} = \Delta w_{jk}^{1} + \Delta w_{jk}^{2} + \cdots + \Delta w_{jk}^{P}$$

Veremos que hay se usa la técnica MINIBATCHES

En la próxima clase veremos MUCHOS trucos para superar, heurísticamente, la dificultad intrínseca de tener que minimizar una función intrincada y compleja como es E.

Ahora podemos generalizar el método de back propagation a cualquier número de capas ocultas



Si queremos estudiar los acoplamientos w entre la capa q (anterior) y p (posterior) usamos la regla

$$\Delta w_{pq} = \eta \sum_{\mu} \Delta \delta \times V$$

p → posterior

V → anterior

μ → patrones del conjunto de entrenamiento