REDES NEURONALES 2024

Clase 18 parte 1

Jueves 23 de octubre 2024

FAMAF, Universidad Nacional de Córdoba

Instituto de Física Enrique Gaviola (UNC-CONICET)

J26/10/23 c19

Aunque quizá no sea obvio, debemos saber que el Método de Descenso por el Gradiente es muy limitado como método de optimización global de la función loss *E({W})*. Veamos porqué.

1.- La función loss o error tiene ahora muchísimos mínimos locales debido a la no-linealidad de las funciones de activación sumado al hecho de que se elevan al cuadrado y a la alternancia y aleatoriedad de los signos de las sinapsis.

2.- El método del descenso por el gradiente (MDG) se queda atrapado en el primer mínimo local que encuentra, y es muy poco probable que este sea el que tiene el menor valor de la función loss E.

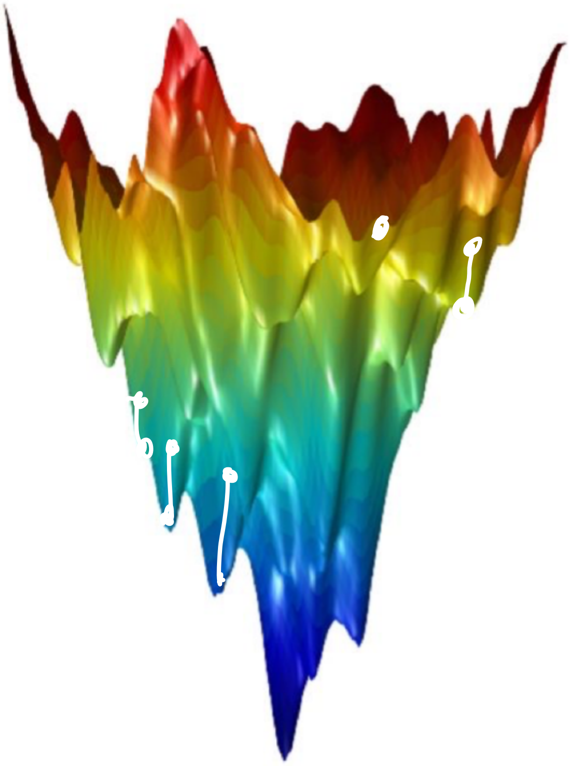
3.- Las funciones de activación exponenciales y sus derivadas son lentas de calcular en punto flotante.

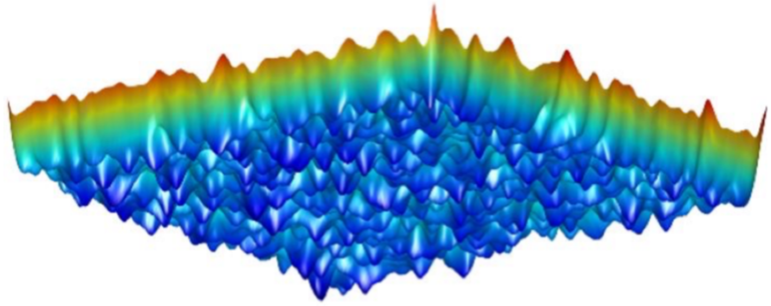
4.- El MDG depende fuertemente del valor inicial aleatorio de los parámetros del problema (las sinapsis y el umbral) debido a la rugosidad de *E({W})*.

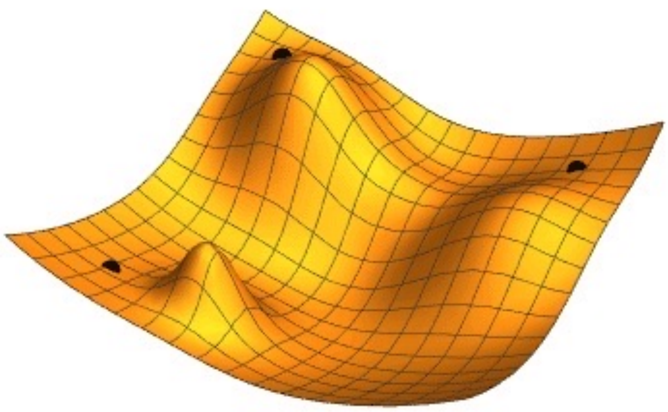
5.- El MDG es muy sensitivo al valor de la razón de aprendizaje y este no puede determinarse en forma óptima a priori.

6.- El MDG trata de la misma forma a todas las direcciones en el espacio N dimensional en el cual viven los vectores $\overline{W}$. Nos gustaría tener un método que dé pasos cortos en direcciones íngrimes del paisaje rugoso de la función $E(\{\overline{W}\})$.

7.- El MDG requiere de tiempos exponenciales en N para escapar de los puntos de ensilladura, los cuales proliferan a medida que N crece.







En realidad podemos buscar un η óptimo η_{opt}, pero eso es muy caro en cálculo computacional. Para ello necesitamos realizar una expansión de Taylor de la función $E(\overline{W})$ alrededor del mínimo. Nos paramos en un punto cualquiera de $\mathbb{R}^{N\times M}$ que denotamos $\overline{W}$.

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

Sea W_{min} el punto que minimiza $E(W)$.

$$\overline{W} = \overline{W}_{min} + \overline{V}$$

Si es un mínimo, todas las componentes del,gradiente son nulas:

$$\partial_{\overline{w}} E(\overline{w}) = 0$$

Derivando con respecto a $\overline{v}$

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

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

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

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

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

ESTO ES CARÍSIMO EN TÉRMINOS COMPUTACIONALES

Si fuésemos muy formales

$$E(\bar{w}+\bar{v}) \approx E(\bar{w}) + \nabla_{\bar{w}} E(\bar{w}) \cdot \bar{v} + \frac{1}{2} \bar{v}^{t} H(\bar{w}) \bar{v}$$

donde H es el Hessiano de *E*

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

Derivando con respecto a $\bar{v}$

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

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

Si estamos en en un mínimo de E : $\bar{\nabla}_w E(\bar{w} + \tilde{v}) = 0$

Derivando con respecto a $\bar{v}$

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

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

$$\eta < \frac{2}{\lambda_{max}}$$

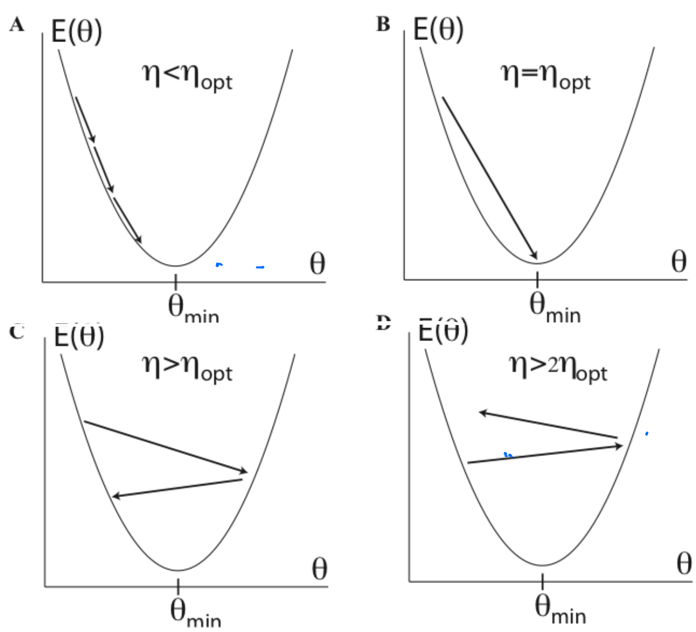


FIG. 8 **Effect of learning rate on convergence**. For a one dimensional quadratic potential, one can show that there exists four different qualitative behaviors for gradient descent (GD) as a function of the learning rate η depending on the relationship between η and $\eta_{\text{opt}} = [\partial_\theta^2 E(\theta)]^{-1}$. (a) For $\eta < \eta_{\text{opt}}$, GD converges to the minimum. (b) For $\eta = \eta_{\text{opt}}$, GD converges in a single step. (c) For $\eta_{\text{opt}} < \eta < 2\eta_{\text{opt}}$, GD oscillates around the minima and eventually converges. (d) For $\eta > 2\eta_{\text{opt}}$, GD moves away from the minima. This figure is adapted from (LeCun *et al.*, 1998b).