REDES NEURONALES 2024

Clase 22 parte 2

Jueves 7 de noviembre 2024

FAMAF, Universidad Nacional de Córdoba

Instituto de Física Enrique Gaviola (UNC-CONICET)

M14/11/23 c23p2

4. AGREGAMOS MEMORIA Y ADAPTACIÓN AL MDG Estocástico

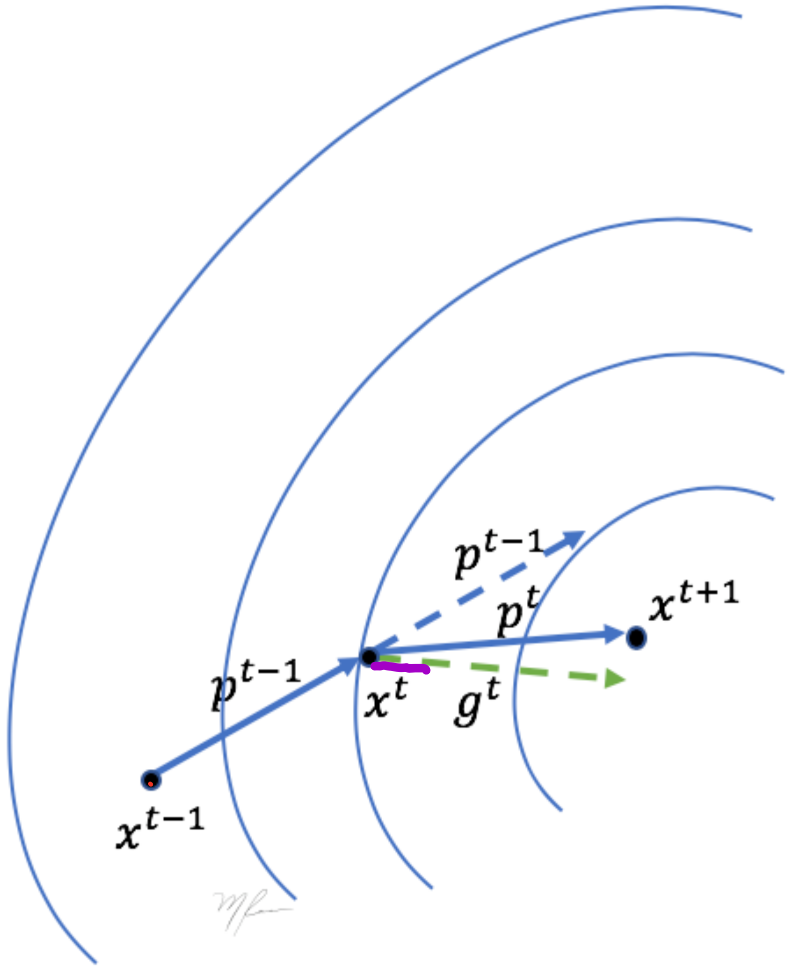
Agregamos memoria

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} + \Delta w_{ij}^{(t)}$$

$$\Delta w_{ij}^{(t)} = \gamma \, \Delta w_{ij}^{(t-1)} - \eta \, \frac{\partial E}{\partial w_{ij}^{(t)}}$$

Términos de momento

$$0 \leq \gamma \leq 1$$ nuevo híper-parámetro


p^{t-1}
x^{t+1}
p^t
p^{t-1}
x^t
g^t
x^{t-1}

Esto le permite al método ganar fuerza cuando el incremento es persistente y pequeño. Si andamos por una zona "plana" de la función loss y el parametro del momento es próximo a uno:

$$\Delta w_{ij}^{(t)} \approx \Delta w_{ij}^{(t-1)}$$

$$(1-\gamma)\,\Delta w_{ij}^{(t)} = -\eta \frac{\partial E}{\partial w_{ij}}$$

$$\Delta w_{ij} = -\underbrace{\frac{\eta}{(1-\gamma)}}_{\text{avanzamos con un paso más largo}} \frac{\partial E}{\partial w_{ij}}$$

Podemos comenzar con $\gamma = 0{,}5$ y aproximarlo a medida que descendemos hasta llegar a $\gamma = 0{,}99$.

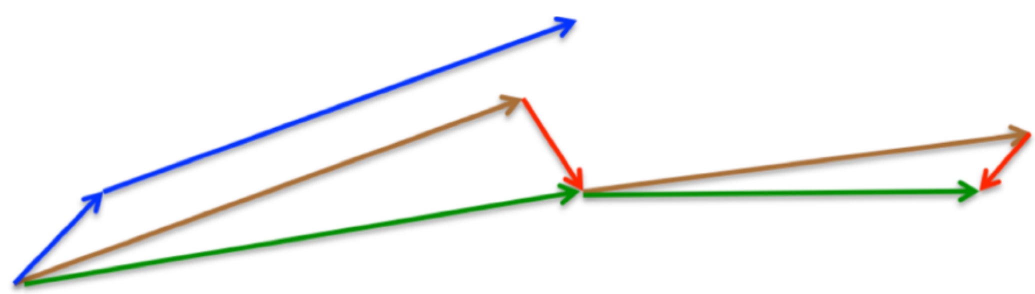
El método de momentos amortigua las oscilaciones en las direcciones de alta curvatura y aumenta la velocidad en direcciones de gradiente pequeño.

El método de descenso por el gradiente acelerado de Nesterov

Primero damos un salto grande en la dirección del gradiente acumulado previo. Ahora nos paramos en este punto y es en este punto en el cual calculamos el gradiente. Ahora sumamos ambos.

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} + \Delta w_{ij}^{(t)}$$

$$\Delta w_{ij}^{(t)} = \gamma_t \Delta w_{ij}^{(t-1)} + \eta \frac{\partial E}{\partial w_{ij}}\left(w_{ij}^{(t-1)} + \gamma_t \Delta w_{ij}^{(t)}\right)$$



Métodos adaptativos

Ahora queremos crear métodos adaptativos, en los cuales cada dimensión del problema tenga su propio criterio de descenso por el gradiente, conforme a la historia del descenso.

$$\Delta w_{ij}(t) = -\eta \overbrace{g_{ij}(t)}^{\eta^{ef}_{ij}} \frac{\partial E}{\partial w_{ij}}$$

universal

específico

Inicialmente $g_{ij}(0) = 1 \quad \forall_i \ y \ \forall_j$

If $\left[\frac{\partial E(t-1)}{\partial w_{ij}} * \frac{\partial E(t)}{\partial w_{ij}}\right] > 0$ then

$$g_{ij}(t) = g_{ij}(t-1) + 0.05$$

else

$$g_{ij}(t) = 0.95 * g_{ij}(t-1)$$

end

Hinton et al

Esto lo hacemos para cada mini batch

- Limitamos los valores de $g_{ij}(t)$

$$0.1 \leq g_{ij}(t) \leq 10 \quad \forall t,i,j$$

$$0.01 \leq g_{ij}(t) \leq 100 \quad \forall t,i,j$$

- usamos mini batches NO MUY PEQUEÑOS. Esto evita que no haya cambios de signo en el gradiente solo porque tenemos "muestras" pequeños. Recuerden que el error es una "función aleatoria" (de muchas variables aleatorias).

- Mezclamos momento con pesos adaptativos.

- lo adaptativo se hace "por eje"

ADAGRAD (Adaptative gradiente algoritmo)

$$W_{ij}^{t+1} = W_{ij}^{t} - \eta_{ij}^{t} \frac{\partial E}{\partial w_{ij}^{t}}$$

$$\eta_{ij}^{t} = \frac{\eta}{\sqrt{\alpha_t + \varepsilon}}$$ $\varepsilon > 0$ pequeño

$$\alpha_{ij}^{t} = \sum_{k=1}^{t} \left(\frac{\partial E}{\partial w_{ij}^{k}} \right)^2$$

ADADELTA

$$W_{ij}^{t+1} = W_{ij}^{t} - \eta_{ij}^{t} \frac{\partial E}{\partial w_{ij}^{t}}$$

$$\eta_{ij}^{t} = \frac{\eta}{\sqrt{S_{ij}^{t} + \varepsilon}}$$

$$S_{ij}^{t} = \beta S_{ij}^{t-1} + (1-\beta) \cdot \left(\frac{\partial E}{\partial w_{ij}^{t-1}} \right)^2$$

RPROP (Back Propagation Resciliente)

Tenemos presente que el gradiente puede ser muy diferente para pesos diferentes y esto puede cambiar a lo largo del descenso por el gradiente

Combinamos la idea de adaptar solo por el signo con la idea de adaptar pesos por separado

IF $\left[\frac{\partial E(t-1)}{\partial w_{ij}} \cdot \frac{\partial E(t)}{\partial w_{ij}}\right] > 0$ then

$$g_{ij}(t+1) = g_{ij}(t) * 1.2$$

else

$$g_{ij}(t+1) = g_{ij}(t) * 0.5$$

end

Este método anda mal para minibatches

RMSprop (Hinton)

Buscamos un método tipo RPROP pero que nos permite usar mini batches.

$$\text{Mean Square}(w_{ij}, t) = 0{,}9\ \text{Mean Square}(w_{ij}, t-1) + 0{,}1\left(\frac{\partial E}{\partial w_{ij}}\right)^2$$

$$S_{ij}^{t}$$

$$g_t = \frac{\partial E}{\partial w_{ij}^{(t)}}$$

$$S_t^{ij} = \beta\, S_{t-1}^{ij} + (1-\beta)\left(g_t^{ij}\right)^2$$

$$\Delta w_{ij}^{(t)} = -\eta\,\frac{g_t}{\sqrt{S_t^{ij} + \epsilon}}$$

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} + \Delta w_{ij}^{(t)}$$

$$\boxed{\begin{array}{l}\beta \approx 0.9\\ \epsilon \approx 10^{-8}\\ \eta \approx 10^{-3}\end{array}}$$

ADAM (Adaptative Moment Estimation)

$$g_t = \frac{\partial E^t}{\partial w_{ij}}$$

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1}$$

$$s_t = \beta_2 s_{t-1} + (1-\beta_2) g_t^2$$

$$\hat{s}_t = \frac{s_t}{(1-\beta_2)}$$

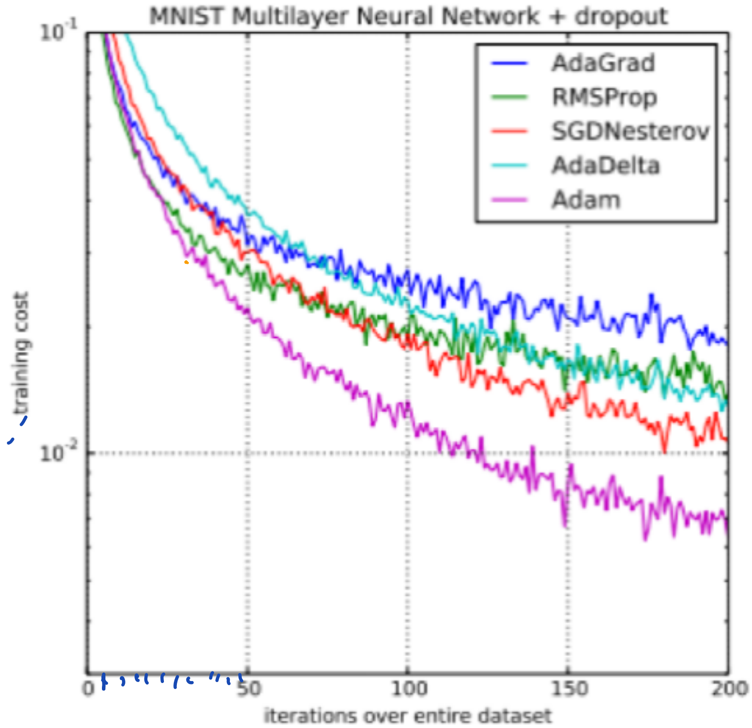
$$E[m_t] = E[g_t]$$

$$E[s_t] = E[g_t^2]$$

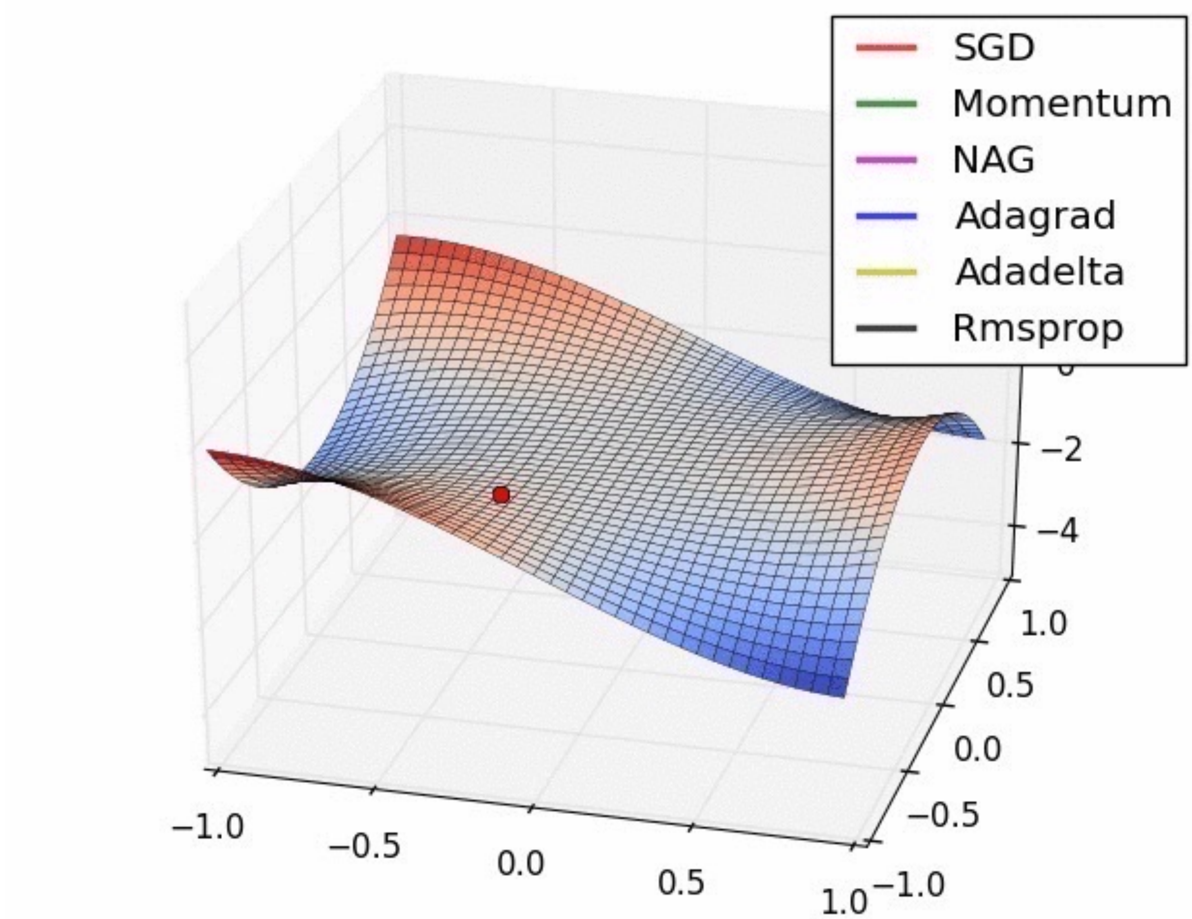
$$w_{ij}^{t+1} = w_{ij}^t + \Delta w_{ij}^t$$

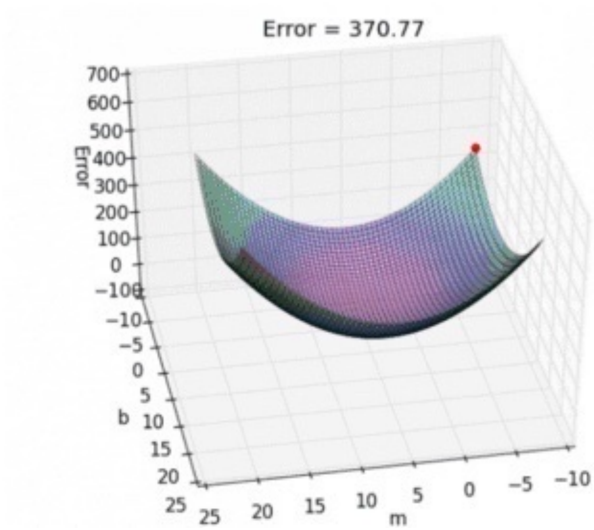
$$\Delta w_{ij}^t = -\eta \frac{\hat{m}_t}{\sqrt{\hat{s}_t + \epsilon}}$$

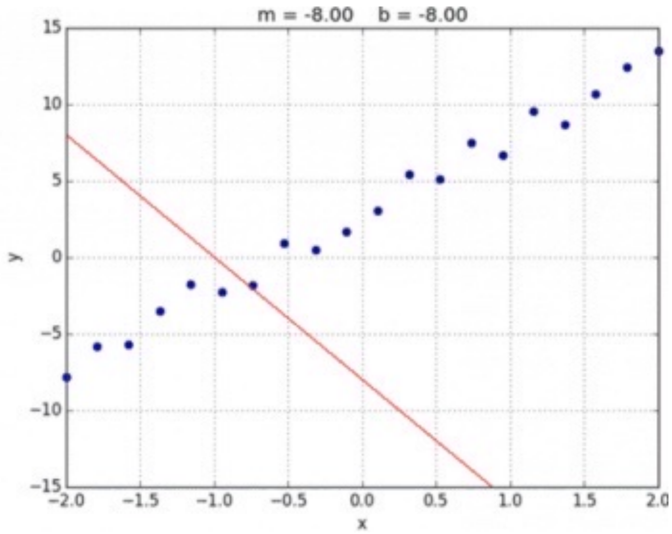
$$\boxed{\begin{aligned} \beta_1 &\approx 0.9 \\ \beta_2 &\approx 0.99 \\ \epsilon &\approx 10^{-8} \\ \eta &\approx 10^{-3} \end{aligned}}$$

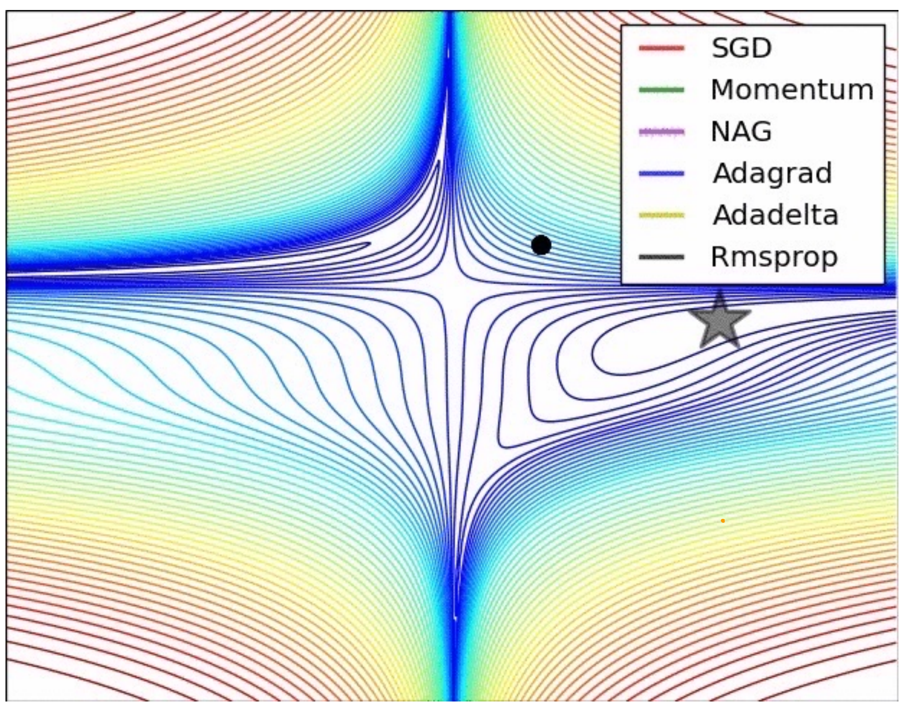


Diferentes métodos de optimizar el descenso por el gradiente (problema MNIST)




Error = 370.77
Error
700
600
500
400
300
200
100
0
−100
−10
−5
0
5
b
10
15
20
25
25
20
15
10
5
0
−5
−10
m


m = -8.00   b = -8.00
15
10
5
y
0
−5
−10
−15
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
x


SGD
Momentum
NAG
Adagrad
Adadelta
Rmsprop