Análisis Matemático para Inteligencia Artificial

Martín Errázquin (merrazquin@fi.uba.ar)

Especialización en Inteligencia Artificial

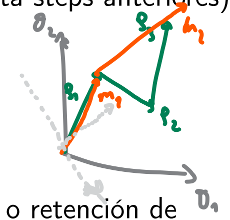
Gradient Descent: extensiones

Momentum

$g_t = \widehat{\nabla J(\theta_t)}$

$g_1, g_2, g_3, g_4, \dots$ EMA $\rightarrow v_1, v_2, v_3, v_4$

Idea: adaptar el γ según consistencia (tener en cuenta steps anteriores) $\rightarrow$ agregar memoria.



$$\begin{cases} v_t = \alpha v_{t-1} - \gamma \cdot g \\ \theta_{t+1} = \theta_t + v_t \end{cases}$$

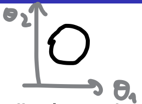
- $\alpha \in (0,1)$ es la *viscosidad* (en términos físicos) o retención de memoria de valores anteriores.

Observar que

$$\theta_{t+1} = \theta_t - \gamma(g_t + \alpha g_{t-1} + \alpha^2 g_{t-2} + \dots) = \theta_t - \gamma \sum_{i=0}^{t} \alpha^i g_{t-i}$$

GD: $\Delta\theta = -\gamma \cdot g_t$

MOM: $\Delta\theta = -\gamma \cdot \mathrm{EMA}(g_t) = -\gamma \cdot \sum_{i=0}^{t} \alpha^i g_{t-i}$



Idea: "reescalar" el gradiente para tener más estabilidad. El reescalamiento se hace a nivel de *feature* para que variaciones grandes sobre un feature no anulen a otros que aún no variaron.

g/np.sqrt(s + eps)

$$\begin{cases} s_t = \lambda s_{t-1} + (1-\lambda) g^2 \\ \theta_{t+1} = \theta_t - \frac{\gamma}{\sqrt{s_t + \epsilon}} \odot g \end{cases}$$

eps= 1e-6

con ^{2} y $\sqrt{}$ aplicados *element-wise*, e.g. $g^2 = g \odot g = (g_1^2, g_2^2, \ldots, g_n^2)$.

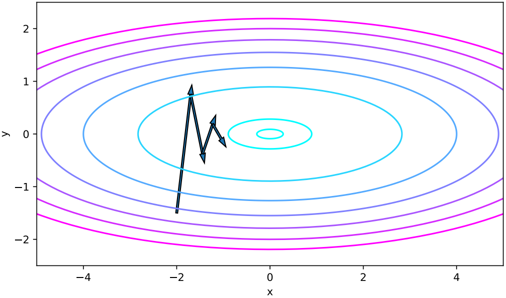
- $\lambda \in (0,1)$ es la retención de memoria de valores anteriores.
- $0 < \epsilon \ll 1$ es una constante para estabilidad numérica. Valores típicos rondan 10^{-6}.

GD: $\Delta\theta = -\gamma \cdot g_t = -\gamma \cdot \left(g_t^{(1)}, g_t^{(2)}, \ldots, g_t^{(n)}\right)$

RMSProp: $\Delta\theta = -\gamma \cdot \frac{g_t}{\sqrt{s_t}} = -\gamma \cdot \left(\frac{g_t^{(1)}}{\sqrt{s_t^{(1)}}}, \frac{g_t^{(2)}}{\sqrt{s_t^{(2)}}}, \ldots, \frac{g_t^{(n)}}{\sqrt{s_t^{(n)}}}\right)$

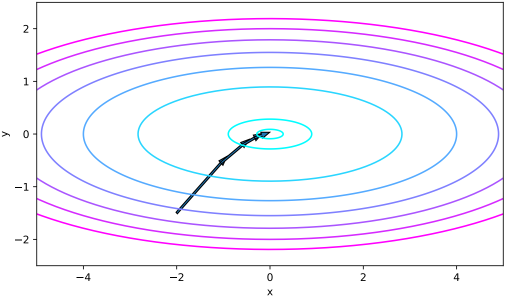
Visualización GD vs RMSProp

GD



$\left(\frac{x}{3}\right)^2 + y^2$

RMSProp



Adam "el que se usa"

Idea: Momentum y RMSProp hacen cosas distintas y ambas están buenas ¡Mezclemos!

$$
\begin{cases}
v_t = \beta_1 v_{t-1} + (1-\beta_1) g & \text{Mom.} \\
s_t = \beta_2 s_{t-1} + (1-\beta_2) g^2 & \text{RMSP.} \\
v'_t = \frac{v_t}{1-\beta_1^t} & \text{rescaling} \\
s'_t = \frac{s_t}{1-\beta_2^t} & \\
\theta_{t+1} = \theta_t - \frac{\gamma}{\underbrace{\sqrt{s'_t+\epsilon}}_{\text{RMSP.}}} \odot v'_t & \text{Mom.}
\end{cases}
$$

- $\beta_1, \beta_2 \in (0,1)$ son la retención de memoria de valores anteriores de media y variabilidad del gradiente. Valores default son $\beta_1 = 0.99, \beta_2 = 0.999$.
- $0 < \epsilon \ll 1$ es una constante para estabilidad numérica. Valor default es 10^{-8}.