

Análisis Matemático para Inteligencia Artificial

Martín Errázquin (merrazquin@fi.uba.ar)

Especialización en Inteligencia Artificial

Gradient Descent: extensiones

Momentum

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

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

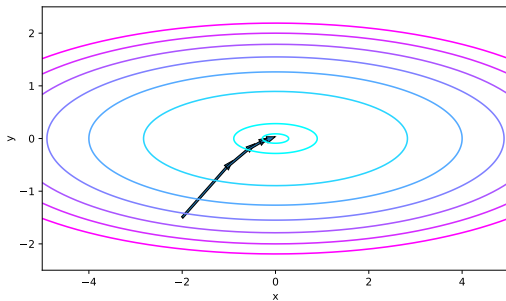
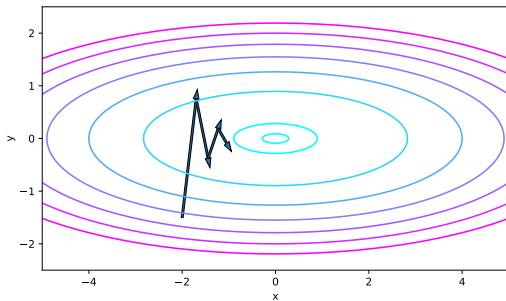
Idea: "reescalar" el gradiente para tener más estabilidad. El reescalamiento se hace a nivel de *feature* para que variaciones grandes sobre un feature no anulen a otros que aún no variaron.

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

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

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

Visualización GD vs RMSProp



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

$$\left\{ \begin{array}{l} v_t = \beta_1 v_{t-1} + (1 - \beta_1)g \\ s_t = \beta_2 s_{t-1} + (1 - \beta_2)g^2 \\ v'_t = \frac{v_t}{1 - \beta_1^t} \\ s'_t = \frac{s_t}{1 - \beta_2^t} \\ \theta_{t+1} = \theta_t - \frac{\gamma}{\sqrt{s'_t + \epsilon}} \odot v'_t \end{array} \right.$$

- $\beta_1, \beta_2 \in (0, 1)$ 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} .