Momentum – GD with Momentum

Saturday, October 9, 2021 12:11 PM

What – Momentum is a force, comes from Physics. We use to accelerate our GD

$m = \beta m + \eta \nabla_\theta J(\theta)$, $\theta = \theta - m$ $\Rightarrow$ $m = \beta m + \eta \frac{dL}{dW}\Big|_{W_0}$

$W_N = W_0 - m \Rightarrow$ (1)

$W_N = W_0 - \eta \frac{dL}{dW}$ → GD

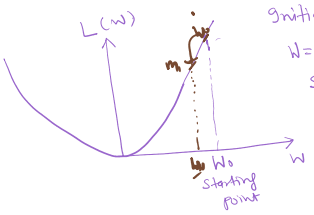
$W_n = W_0 - \beta m - \eta \frac{dL}{dW}$ (βm → New term)

β – coeff. of momentum

m – A fixed variable

Let's take β=0

$W_n = W_0 - \eta \frac{dL}{dW}$ → Gradient Descent normal



Initialization ⇒

$W = W_0$, $m_0 = 0$, $\beta = 0.9$ default

Step 1 ⇒

$m_1 = \beta_0 m_0 + \eta \frac{dL}{dW}\Big|_{W=W_0}$

$m_1 = \eta \frac{dL}{dW}\Big|_{W=W_0}$ if $m_0 = 0$ → (1)

$W_1 = W_0 - m_1$

$\Rightarrow W_1 = W_0 - \eta \frac{dL}{dW}\Big|_{W=W_0}$ → (1) GD only

Step 2 ⇒ $m_1 = \eta \frac{dL}{dW}\Big|_{W=W_0}$, $W_1 = W_0 - \eta \frac{dL}{dW}\Big|_{W=W_0}$, $\beta = 0.9$

$m_2 = \beta m_1 + \eta \frac{dL}{dW}\Big|_{W=W_1}$

$m_2 = \beta \cdot \eta \frac{dL}{dW}\Big|_{W=W_0} + \eta \frac{dL}{dW}\Big|_{W=W_1}$

$m_2 = \eta \left[\beta \cdot \frac{dL}{dW}\Big|_{W=W_0} + \frac{dL}{dW}\Big|_{W=W_1} \right]$ → (m2)

0.9 (fixed value) · past gradient; current gradient (100%)

90% past gradient

$W_2 = W_1 - m_2 = W_1 - \eta \left[\beta \frac{dL}{dW}\Big|_{W=W_0} + \frac{dL}{dW}\Big|_{W=W_1} \right]$

Let's take a special case $\beta = 0$

$W_2 = W_1 - \eta \frac{dL}{dW}\Big|_{W=W_1}$ — Simple GD weight updates → (A)

$\beta = 0.5$

$W_2 = W_1 - \eta \left[0.5 \frac{dL}{dW}\Big|_{W=W_0} + \frac{dL}{dW}\Big|_{W=W_1} \right]$ → (B)

Q – Which scenario will have more weight updates ?? A OR B

Step 3 ⇒

$m_3 = \beta m_2 + \eta \frac{dL}{dW}\Big|_{W=W_2}$ $W = W_2$, $\beta = 0.9$, $m_2 =$ (m2)

$m_2 = \eta \left[\beta \cdot \frac{dL}{dW}\Big|_{W=W_0} + \frac{dL}{dW}\Big|_{W=W_1} \right]$ → (m2)

$m_3 = \beta \left[\eta \left[\beta \cdot \frac{dL}{dW}\Big|_{W=W_0} + \frac{dL}{dW}\Big|_{W=W_1} \right] \right] + \eta \frac{dL}{dW}\Big|_{W=W_2}$

$m_3 = \beta \cdot \eta \left[\beta \cdot \frac{dL}{dW}\Big|_{W=W_0} + \frac{dL}{dW}\Big|_{W=W_1} \right] + \eta \frac{dL}{dW}\Big|_{W=W_2}$

$m_3 = \eta \left[\beta^2 \frac{dL}{dW}\Big|_{W=W_0} + \beta \cdot \frac{dL}{dW}\Big|_{W=W_1} + \frac{dL}{dW}\Big|_{W=W_2} \right]$ → (IV)

81% of W_0 · 90% of W_1 · 100% of W_2

$$m_3 = \eta \left[\beta^2 \left.\frac{dL}{dW}\right|_{W=W_0} + \beta \left.\frac{dL}{dW}\right|_{W=W_1} + \left.\frac{dL}{dW}\right|_{W=W_2} \right] \rightarrow (IV)$$

β^2 ↓ 81% of W_0 ; β 90% of W_1 ; 100% of W_2 → even if this becomes 0

Challenges above ??

1) β - Additional Parameter to tune - 0.9
2) Oscillations closer to local or global minima
   ↳ take $\frac{dL}{dW} = 0$