Momentum Descent

{ Heavy ball method }

Plain gradient descent : $\theta_{t+1} = \theta_t - \eta \nabla J(\theta_t)$

But has ① Ill conditioned hessians.

② Oscillat across steep directions.

③ Move slowly along flat directions.

Momentum fixes this by accumulating velocity, not just reaching to the current gradient.

Analogy: (i) Gradient → Force

(ii) Momentum → Velocity

(iii) Parameters → Position

$v^* \in \mathbb{R}^d$

μ → Momentum

(β) Coefficient (0.9)

$$F = m\frac{dv}{dt} \Rightarrow F\,dt = m\,dv$$

Momentum is exponentially weighted moving average of past gradients.

* Algebraic form

θ_0 given , $v_0 = 0$

* Update Rule

$$v_t = \mu v_{t-1} + g_t = \mu(\mu v_{t-2} + g_{t-1}) + g_t = \ldots \sum_{k=0}^{t} \mu^k g_{t-k}$$

$$\theta_{t+1} = \theta_t - \eta v_t$$

$$v_t = \beta v_{t-1} + (1-\beta)\nabla_\theta J(\theta_t)$$

$$\theta_{t+1} = \theta_t - \eta v_t$$

* Example : $\{(1,2),(2,3),(3,5)\}$ → Fitting linear regression with momentum GD.

→ model: $y = wx + b$ $\quad \theta = \begin{pmatrix} w \\ b \end{pmatrix}$

$$\mathcal{L} = J(\theta) = \frac{1}{2m} \overset{m}{\underset{0}{\sum_i}} (\hat{y}_i - y)^2$$

$$= \frac{1}{2m} (\hat{y}_i - y)^T (\hat{y}_i - y)$$

$$= \frac{1}{2m} (x\theta - y)^T (x\theta - y)$$

$$= \frac{1}{2m} (\theta^T x^T x\theta - x^T\theta^T y - y^T x\theta$$
$$- y^T y)$$

$$= \frac{1}{2m} (\theta^T x^T x\theta - 2x^T\theta^T y - y^T y)$$

initialisations

$\theta = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad v = \begin{pmatrix} v_w \\ v_b \end{pmatrix} \in \mathbb{R}^2$

$\eta = 0.1$

$\beta = 0.9$

$x = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 2}$

$y = \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix}$

$\hat{y} = x\theta$

$$\nabla_\theta J(\theta) = \frac{1}{m} x^T (x\theta - y)$$

* Update rules :

$$\boxed{\begin{aligned} v_t &= \beta v_{t-1} + (1-\beta)\nabla J(\theta_t) \\ \theta_{t+1} &= \theta_t - \eta v_t \end{aligned}}$$

* Predictions : $x\theta = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \hat{y}_0 \qquad e_0 = \begin{pmatrix} -2 \\ -3 \\ -5 \end{pmatrix}$

1st iteration

$$\nabla J(\theta) = \frac{1}{3} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -2 \\ -3 \\ -5 \end{pmatrix} = \begin{pmatrix} -7.667 \\ -0.333 \end{pmatrix}$$

$$v_1 = (0.9)\begin{pmatrix} 0 \\ 0 \end{pmatrix} - (1-0.9)\begin{pmatrix} -7.667 \\ -0.333 \end{pmatrix} = \begin{pmatrix} -0.767 \\ -0.333 \end{pmatrix}$$

$$\theta_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 0.1 \begin{pmatrix} 0.767 \\ 0.333 \end{pmatrix} = \begin{pmatrix} 0.0767 \\ 0.0333 \end{pmatrix}$$