Introduction to Data Science

Gradient Descent

Arijit Mondal

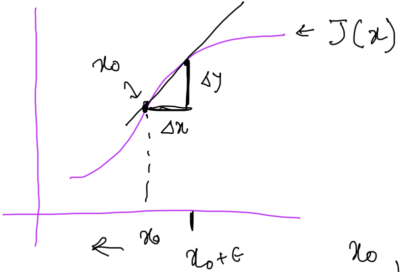
Dept. of Computer Science & Engineering

Indian Institute of Technology Patna

arijit@iitp.ac.in

Derivative

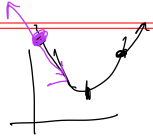
derivative = $\frac{\Delta y}{\Delta x}$



$J(x_0) < J(x_0+\epsilon)$

$J(x_0) > J(x_0-\epsilon)$

$x_0, \quad \left(x_0 - \frac{\partial y}{\partial x}\right)$

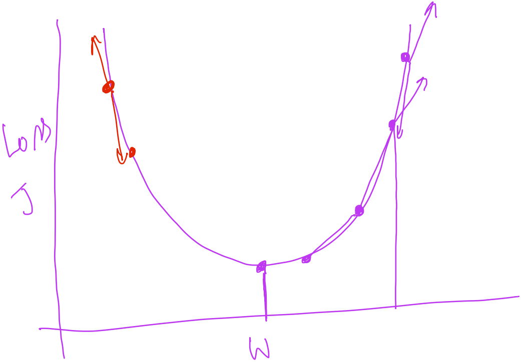


$x_0^0, \quad \frac{\partial x}{\partial y}$

new $x_0^1 \leftarrow x_0^0 - \frac{\partial x}{\partial y}$

x_0^2

Gradient descent



w_1, w_2

w_1, w_2

$$w^{t+1} = w^t - \alpha \frac{\partial J(w)}{\partial w}$$

Till you reach the convergence

$$|w^{t+1} - w^t| \le \delta$$

$$w_1 = w_1 - \alpha \frac{\partial J(w_1, w_2)}{\partial w_1}$$

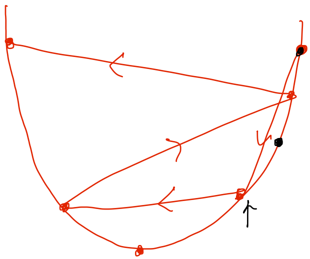
$$w_2 = w_2 - \alpha \frac{\partial J(w_1, w_2)}{\partial w_2}$$

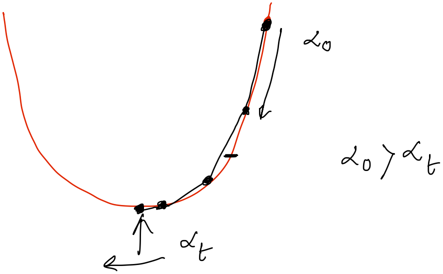
$$|w_1^{t+1} - w_1^t| \le \delta$$

$$|w_2^{t+1} - w_2^t| \le \delta$$

α = Learning rate $(0, 1)$

Learning rate





Example: Gradient descent

- Consider the following pair (x, y) of points - $(1, 2)$, $(2, 4)$, $(3, 6)$, $(4, 8)$
- Let us try to fit a curve as follows $\hat{y} = w \times x$ where w is initialized with 4, learning rate as 0.1, MSE is used as cost function

Example: Stochastic gradient descent

- Consider the following pair (x, y) of points - $(1, 2)$, $(2, 4)$, $(3, 6)$, $(4, 8)$ 10^6
- Let us try to fit a curve as follows $\hat{y} = w \times x$ where w is initialized with 4, learning rate as 0.1, MSE is used as cost function

$J(w)$ MSE $= \frac{1}{2}(wx_i - y_i)^2$

$J'(w) = (wx_i - y_i)x_i \leftarrow w = 4,\ x_i = 1,\ y_i = 2$

$w = 3.8 \quad x_i = 2 \quad y_i = 4$

Point	Derivative	New w
(1,2)	2	3.8
(2,4)	7.2	3.08
3,6		

$new\_w = w - \alpha \frac{\partial J}{\partial w}$

SGD ⟷ GD

SGD → 1 point; Minibatch; GD → All points

8–10 steps → 2.0