

Optimizers

Gradient Descent

Parameter Learning Rate Gradient/ Slope

The diagram shows the equation $\theta_i = \theta_i - \eta \frac{\partial J}{\partial \theta_i}$. Three blue arrows point from labels to parts of the equation: one from 'Parameter' to the first θ_i , one from 'Learning Rate' to η , and one from 'Gradient/ Slope' to $\frac{\partial J}{\partial \theta_i}$.

$$\theta_i = \theta_i - \eta \frac{\partial J}{\partial \theta_i}$$

Entire Training set
(m)

Batch Gradient
Descent

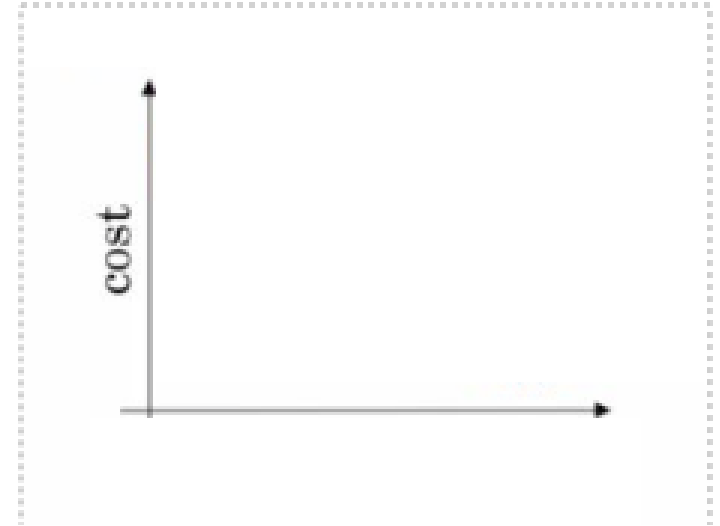
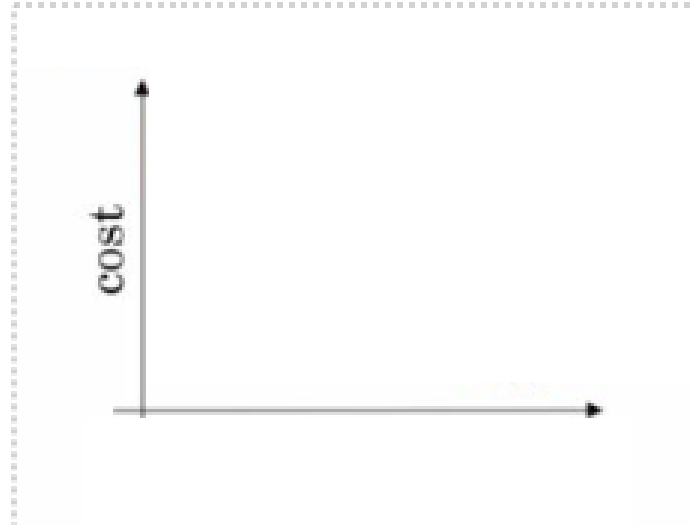
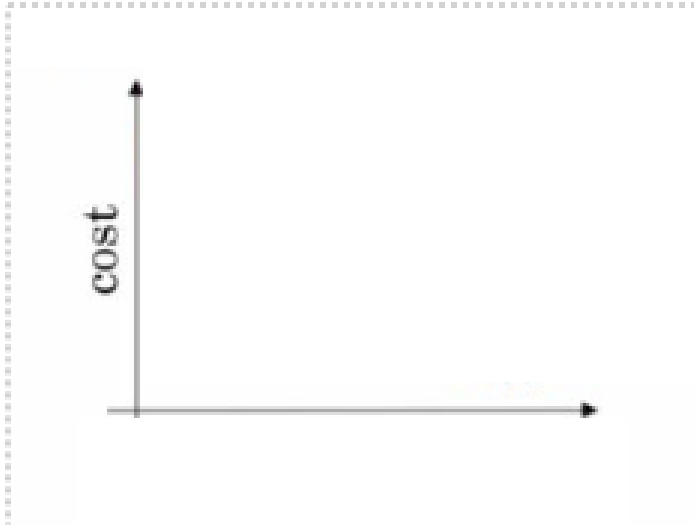
Single
Observation (1)

Stochastic Gradient Descent

$1 < x < m$

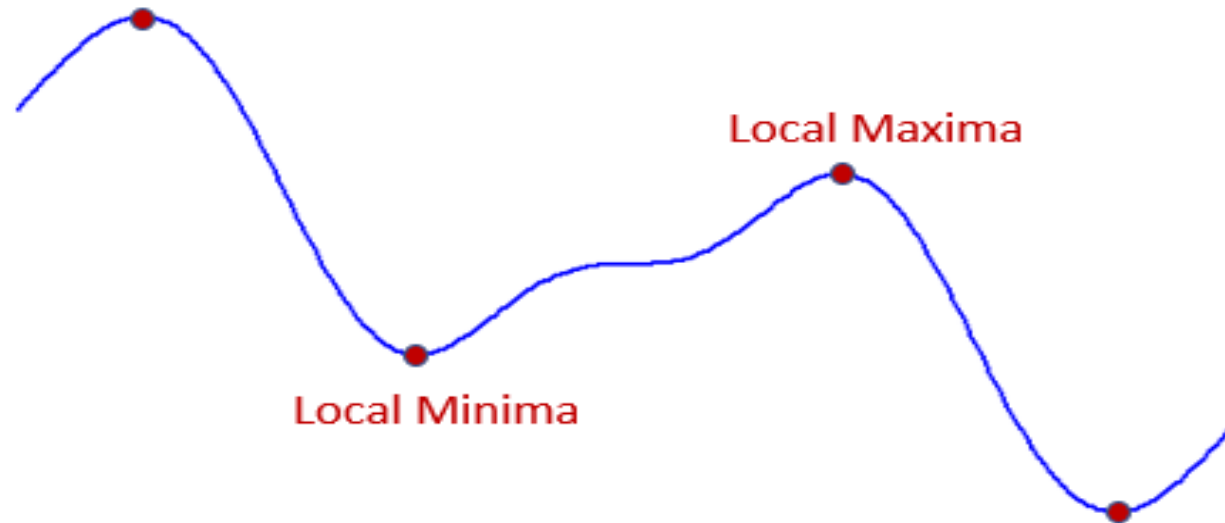
Mini - Batch Gradient
Descent

Gradient Descent



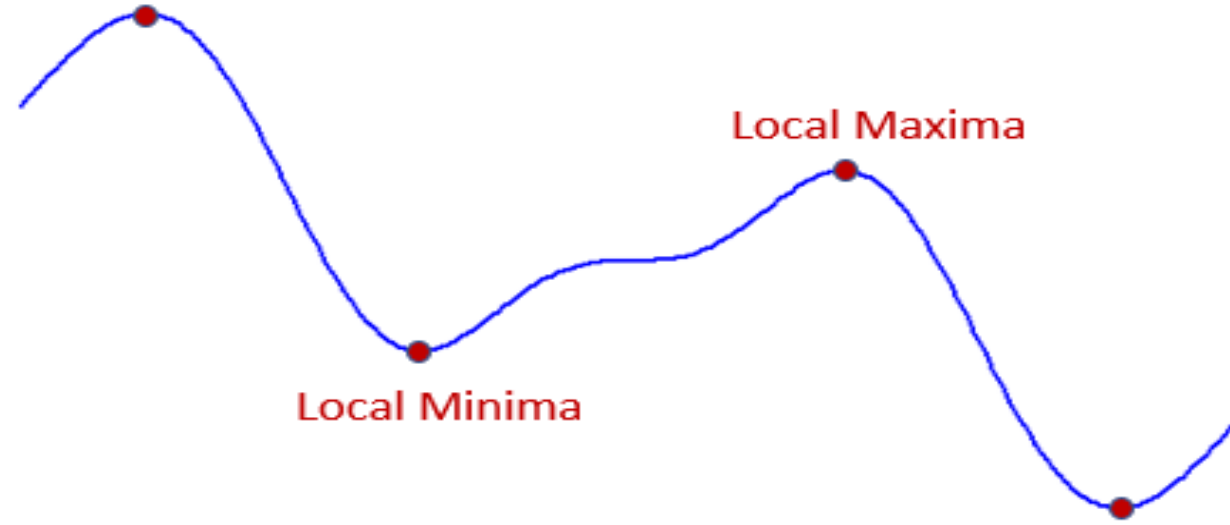
Problems of Gradient Descent

1. Getting stuck at Local Minima



$$\theta_i = \theta_i - \eta \frac{\partial J}{\partial \theta_i}$$

Problems of Gradient Descent



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

$$\theta_i = \theta_i - \eta v_t$$

Problems of Gradient Descent

2. Same Learning Rate for all parameters

n	n	n	n
gender	age	hypertension	does_smoke
1	3	0	0
1	58	1	1
0	8	0	0
0	70	0	1
1	14	0	0
0	47	0	0
0	52	0	1
0	75	0	0
0	32	0	1

stroke
0
1
0
1
0
0
1
0
0

Problems of Gradient Descent

2. Same Learning Rate for all parameters

n1	n2	n3	n4
gender	age	hypertension	does_smoke
1	3	0	0
1	58	1	1
0	8	0	0
0	70	0	1
1	14	0	0
0	47	0	0
0	52	0	1
0	75	0	0
0	32	0	1

stroke
0
1
0
1
0
0
1
0
0

Problems of Gradient Descent

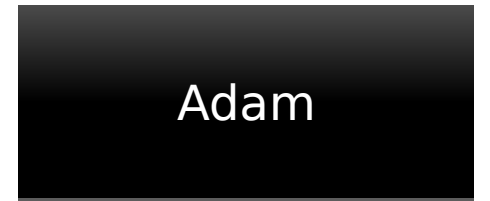
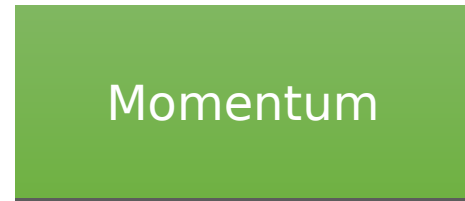
2. Same Learning Rate for all parameters

RMSProp

$$\mu_t = \beta \mu_{t-1} + (1 - \beta) \left(\frac{\partial J}{\partial \theta_{t,i}} \right)^2$$

$$\theta_{t,i} = \theta_{t,i} - \frac{\eta}{\sqrt{\mu_t + \epsilon}} \frac{\partial J}{\partial \theta_{t,i}}$$

Adam



Exponential Weighted
Sum of Past gradients

Exponential Weighted
Sum of Squares of
Past Gradients

Thank You