

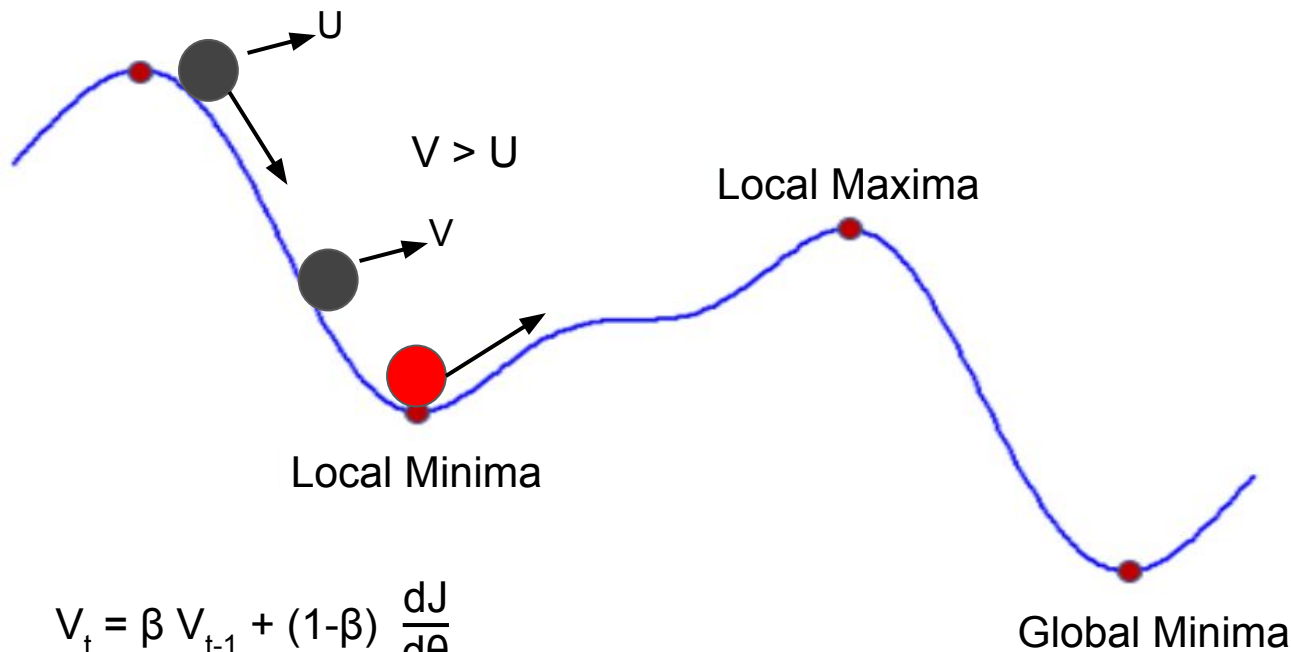
# Problems with Gradient Descent

# Problems with Gradient Descent

## 1. Getting stuck at local minima



# SGD with Momentum



$$V_t = \beta V_{t-1} + (1-\beta) \frac{dJ}{d\theta}$$

$$\theta_i = \theta_{i-1} - \alpha * V_t$$

# Problems with Gradient Descent

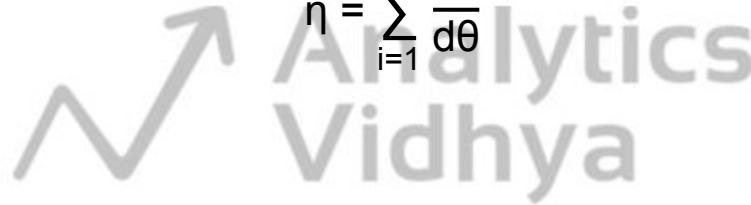
1. Getting stuck at local minima
2. Same Learning rate throughout the training process

# Problem: Same learning rate for all parameters



# Problem: Same learning rate for all parameters

$$\eta = \sum_{i=1}^{t-1} \frac{dJ}{d\theta}$$



# Problem: Same learning rate for all parameters

$$\eta = \sum_{i=1}^{t-1} \left[ \frac{dJ}{d\theta} \right]^2$$

# Problem: Same learning rate for all parameters

$$\eta = \sum_{i=1}^{t-1} \left[ \frac{dJ}{d\theta} \right]^2$$
$$\theta_i = \theta_{i-1} - \frac{\alpha}{\sqrt{\eta}} \frac{dJ}{d\theta}$$



## Problem: Same learning rate for all parameters

$$\eta = \sum_{i=1}^{t-1} \left[ \frac{dJ}{d\theta} \right]^2$$

$$\theta_i = \theta_{i-1} - \frac{\alpha}{\sqrt{\eta}} \frac{dJ}{d\theta}$$

# Problem: Same learning rate for all parameters

$$\eta = \sum_{i=1}^{t-1} \left[ \frac{dJ}{d\theta} \right]^2$$

Always  
Increase

$$\theta_i = \theta_{i-1} - \frac{\alpha}{\sqrt{\eta}} \frac{dJ}{d\theta}$$

# Problem: Same learning rate for all parameters

Diagram illustrating the problem of using a constant learning rate for all parameters. The equation for parameter update is shown:

$$\theta_i = \theta_{i-1} - \frac{\alpha}{\sqrt{\eta}} \frac{dJ}{d\theta}$$

The term  $\eta = \sum_{i=1}^{t-1} \left[ \frac{dJ}{d\theta} \right]^2$  is highlighted in a red box, with an arrow pointing to the text "Always Increase".

The term  $\frac{\alpha}{\sqrt{\eta}}$  is highlighted in a red box, with an arrow pointing to the text "Always Decrease".

# Problem: Same learning rate for all parameters

$$\eta = \sum_{i=1}^{t-1} \left[ \frac{dJ}{d\theta} \right]^2$$
$$\theta_i = \theta_{i-1} - \frac{\alpha}{\sqrt{\eta}} \frac{dJ}{d\theta}$$

Tends to Zero

# Problem: Same learning rate for all parameters

$$\eta = \sum_{i=1}^{t-1} \left[ \frac{dJ}{d\theta} \right]^2$$
$$\theta_i = \theta_{i-1} - \frac{\alpha}{\sqrt{\eta}} \frac{dJ}{d\theta}$$

Tends to Zero

$$\theta_i \approx \theta_{i-1}$$

# Problem: Same learning rate for all parameters

$$V_t = \beta V_{t-1} + (1-\beta) \frac{dJ}{d\theta}$$

Analytics Vidhya

# RMSProp

$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

 Analytics  
Vidhya

# Update equation: RMSProp

$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$



# Update equation: RMSProp

$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

An arrow points from the boxed term  $\left[ \frac{dJ}{d\theta} \right]^2$  to the word **High** in a blue box.

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$

# Update equation: RMSProp

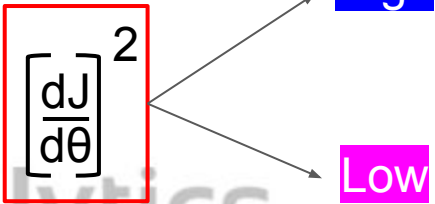
$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

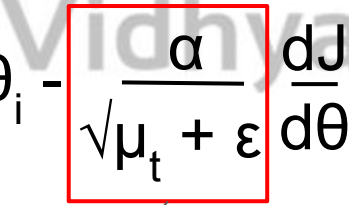
High

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$

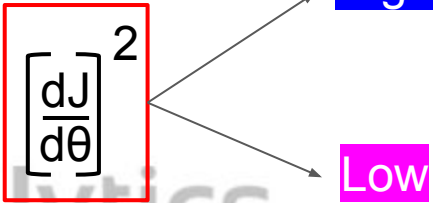
Low

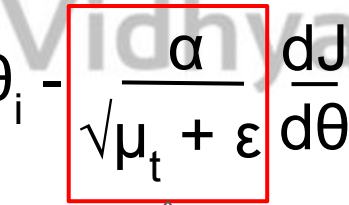
# Update equation: RMSProp

$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$


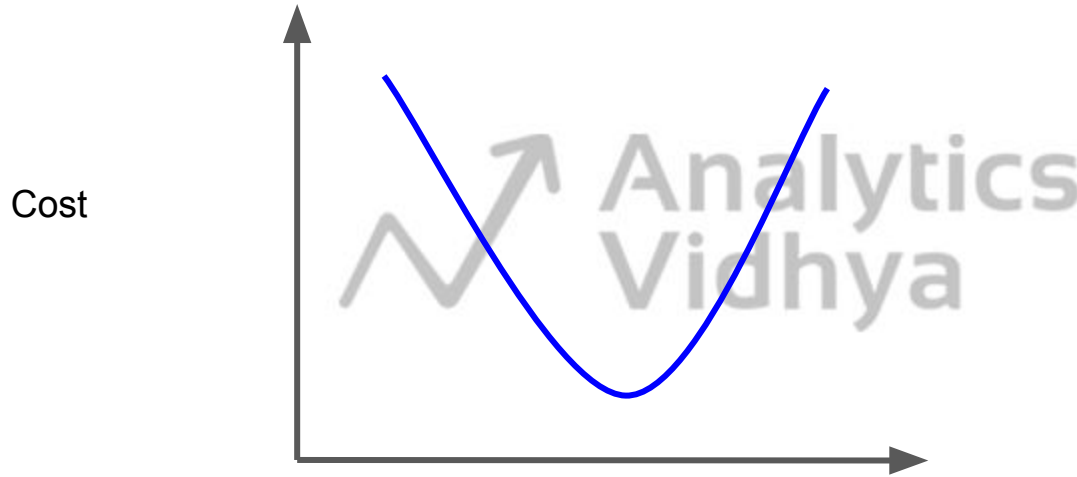
$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$


# Update equation: RMSProp

$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$


$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$


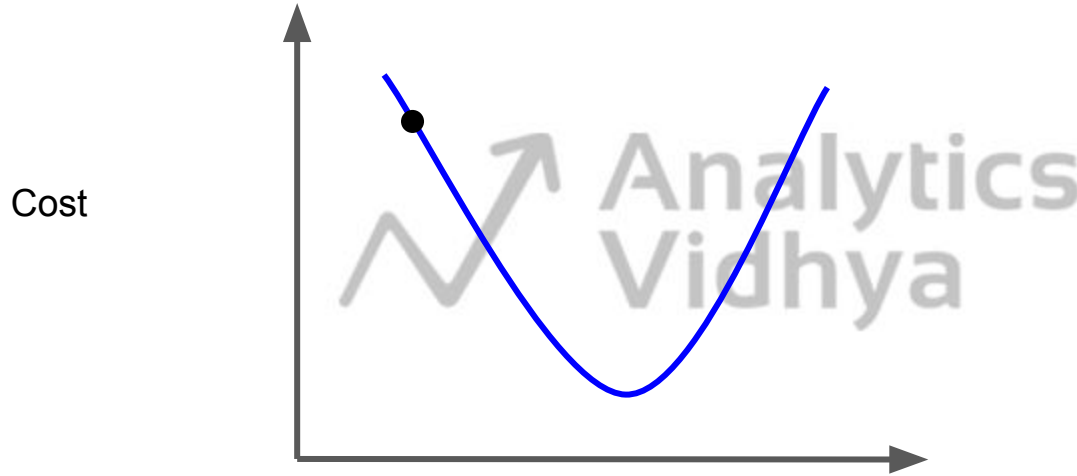
# Understanding RMSProp



$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$

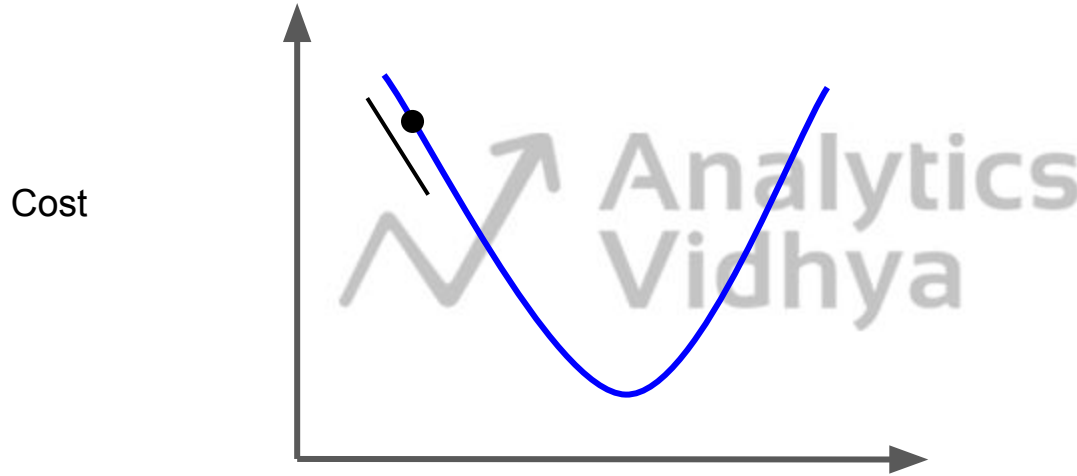
# Understanding RMSProp



$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$

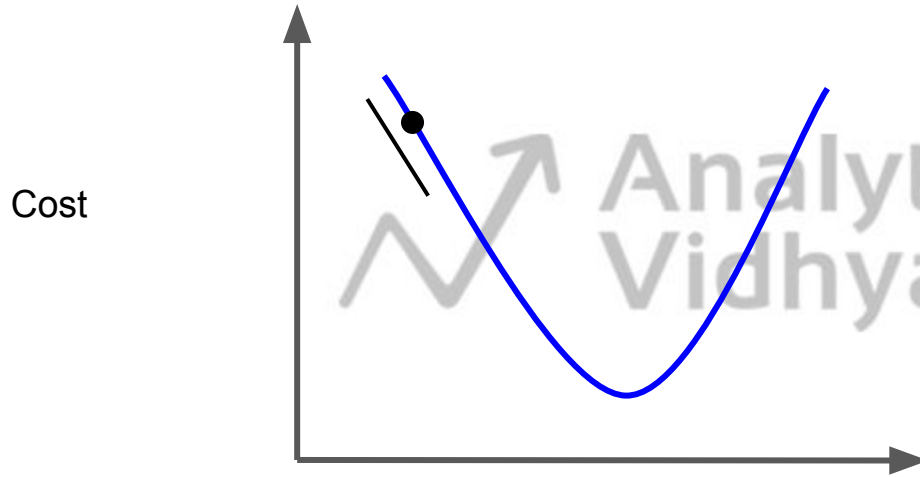
# Understanding RMSProp



$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$

# Understanding RMSProp



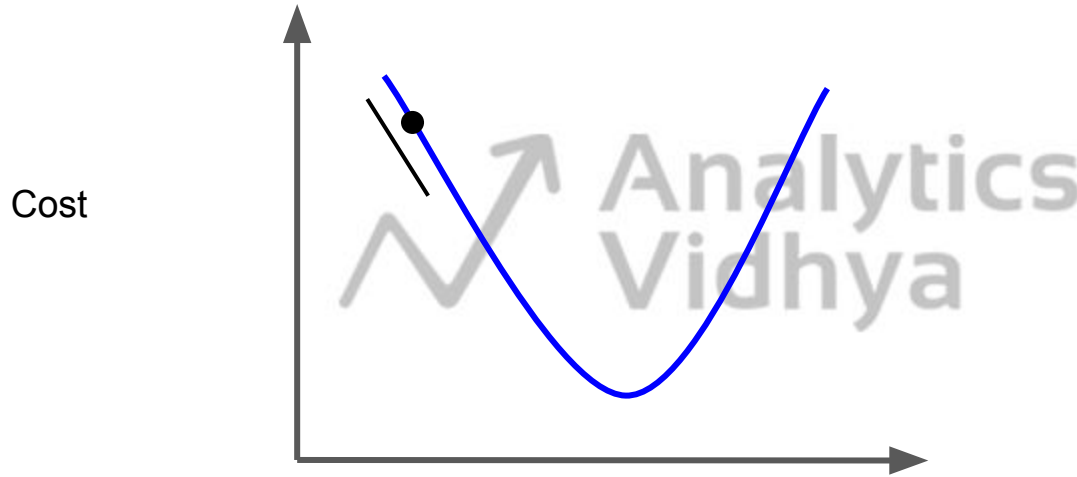
$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

High

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$



# Understanding RMSProp



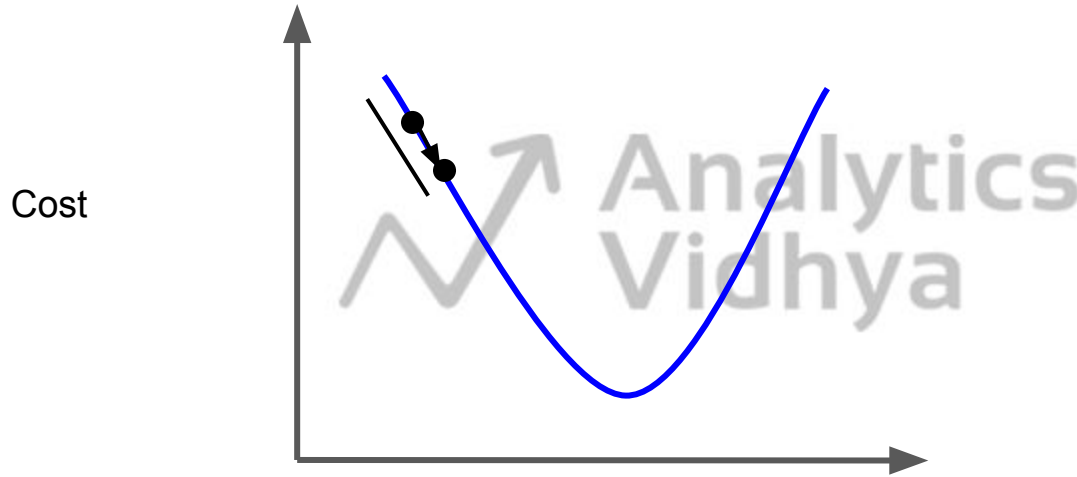
$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

High

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$

Low

# Understanding RMSProp



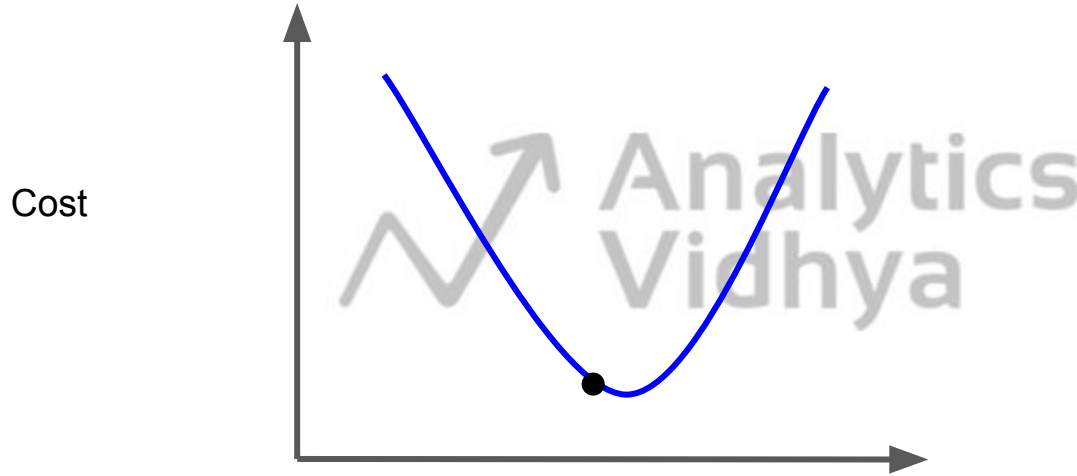
$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

High

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$

Low

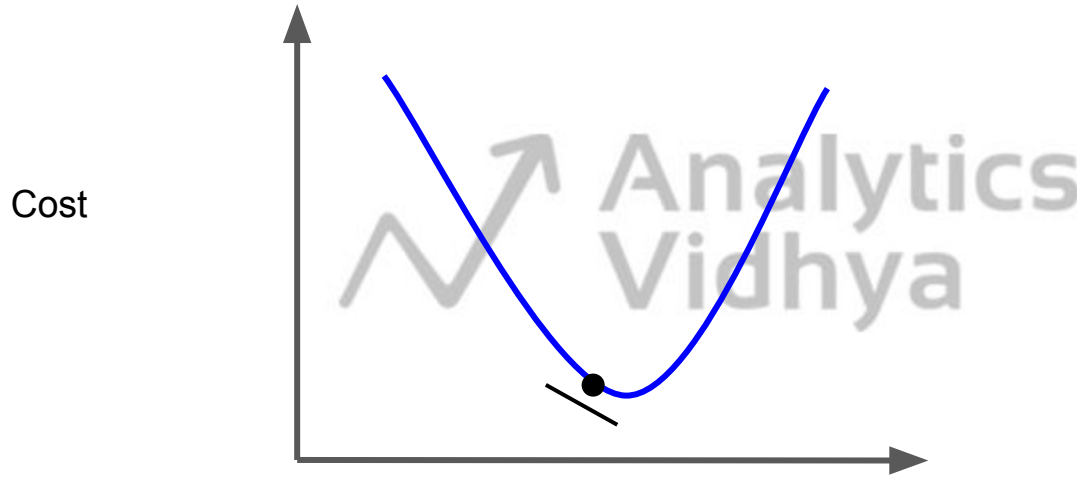
# Understanding RMSProp



$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$

# Understanding RMSProp

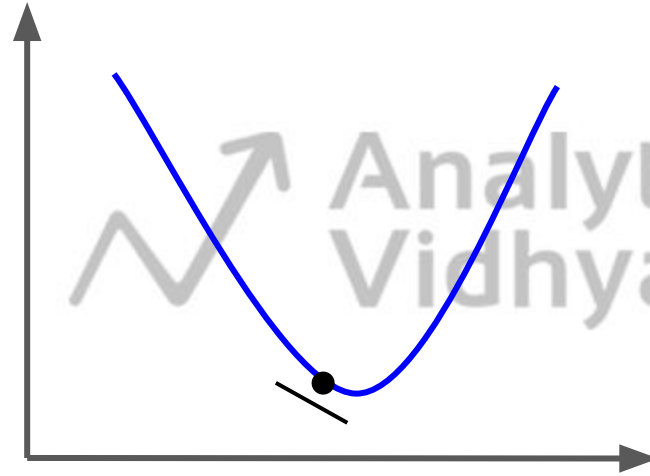


$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$

# Understanding RMSProp

Cost

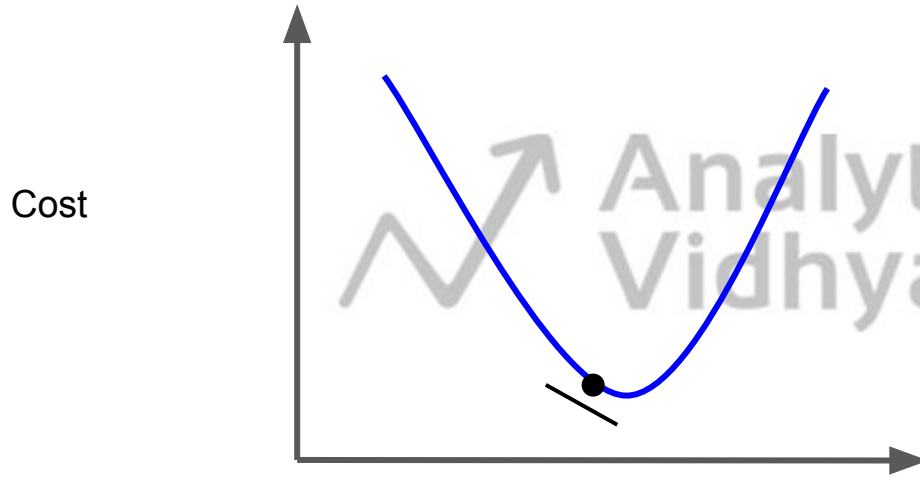


$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

Low

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t + \epsilon}} \frac{dJ}{d\theta}$$

# Understanding RMSProp



$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

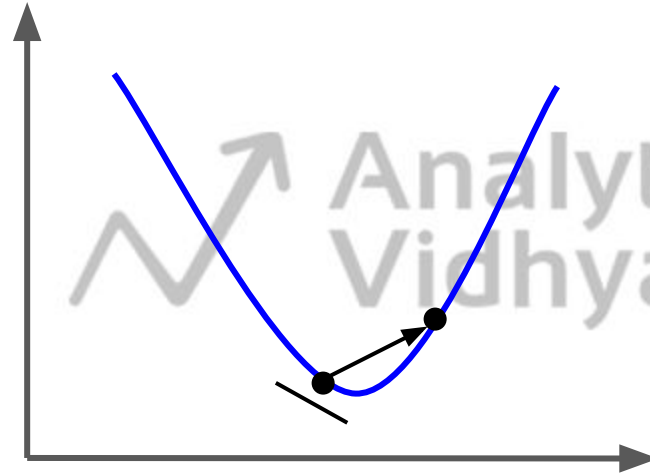
Low

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$

High

# Understanding RMSProp

Cost



$$\mu_t = \beta \mu_{t-1} + (1-\beta) \left[ \frac{dJ}{d\theta} \right]^2$$

Low

$$\theta_i = \theta_i - \frac{\alpha}{\sqrt{\mu_t} + \epsilon} \frac{dJ}{d\theta}$$

High



Thank You