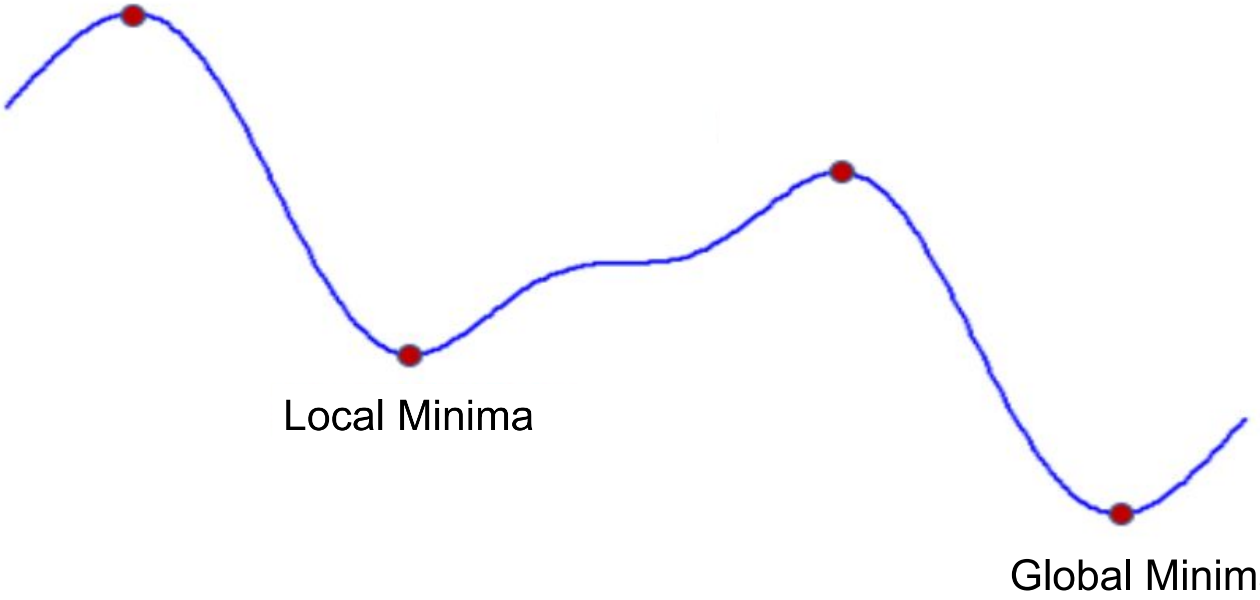
Problems with Gradient Descent

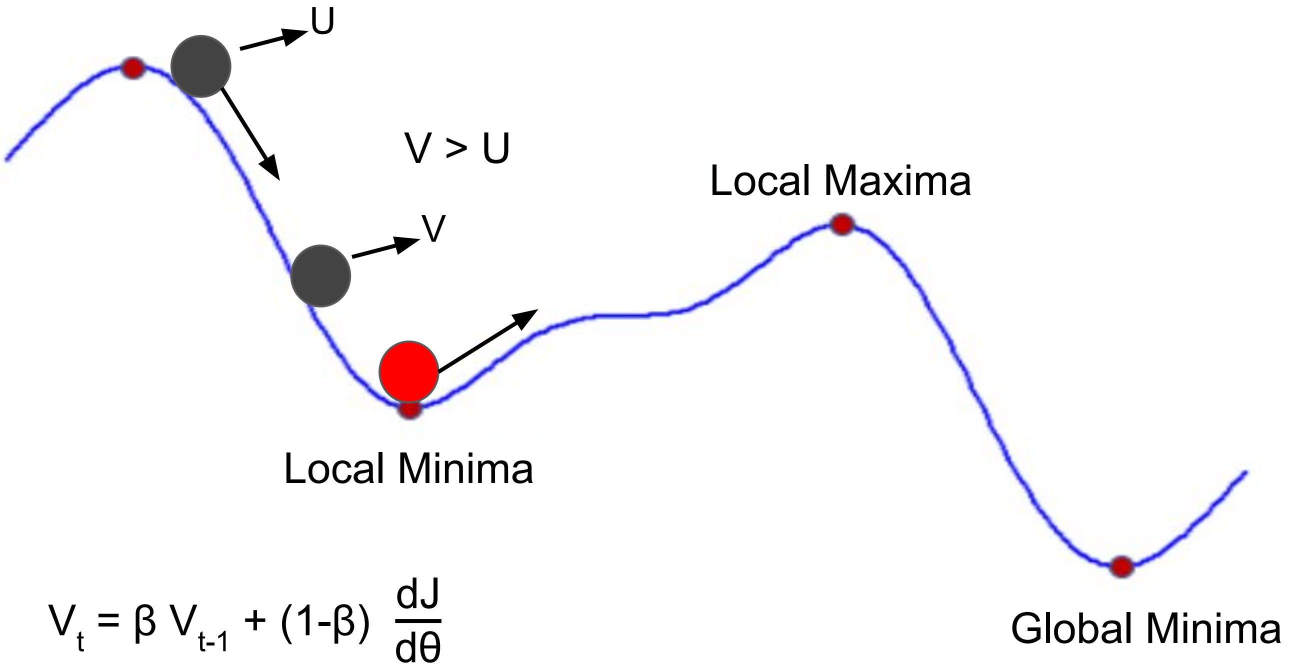
Problems with Gradient Descent

1. Getting stuck at local minima



Analytics Vidhya
Learn everything about analytics

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

$$\frac{dJ}{d\theta}$$

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

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

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

Problem: Same learning rate for all parameters

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

Always Increase

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

Problem: Same learning rate for all parameters

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

Always Increase

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

Always Decrease

Problem: Same learning rate for all parameters

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

$$\theta_i = \theta_{i-1} - \boxed{\frac{\alpha}{\sqrt{\eta}} \frac{dJ}{d\theta}} \rightarrow \text{Tends to Zero}$$

Problem: Same learning rate for all parameters

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

$$\theta_i = \theta_{i-1} - \boxed{\frac{\alpha}{\sqrt{\eta}} \frac{dJ}{d\theta}} \rightarrow \text{Tends to Zero}$$

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

Problem: Same learning rate for all parameters

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

RMSProp

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

Update equation: RMSProp

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

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

Update equation: RMSProp

High

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

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

Update equation: RMSProp

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

High

$$\theta_i = \theta_i - \boxed{\frac{\alpha}{\sqrt{\mu_t} + \varepsilon}}\, \frac{dJ}{d\theta}$$

Low

Update equation: RMSProp

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

High

Low

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

Low

Update equation: RMSProp

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

High

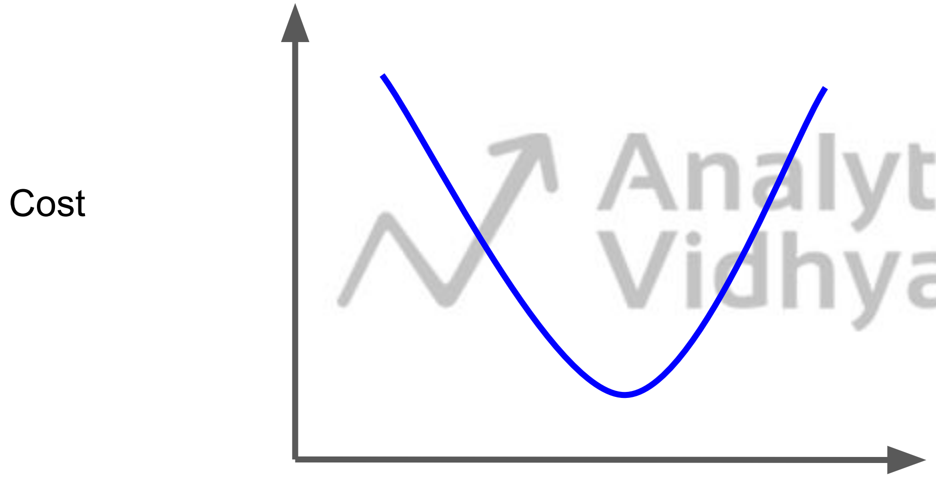
Low

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

Low

High

Understanding RMSProp

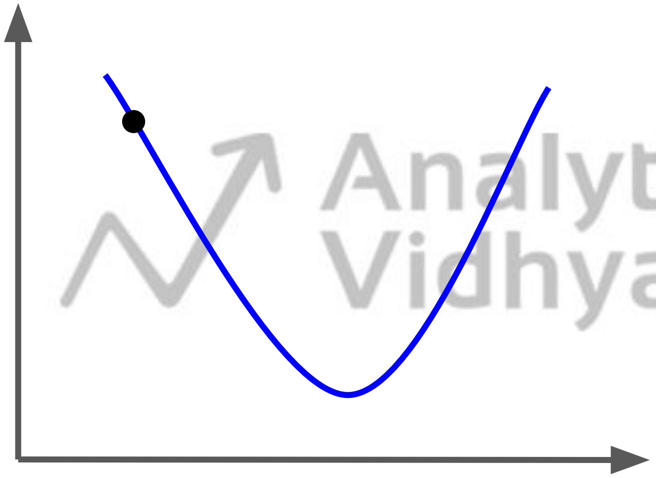


$$\mu_t = \beta\,\mu_{t-1} + (1-\beta)\left[\frac{dJ}{d\theta}\right]^2$$

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

Understanding RMSProp

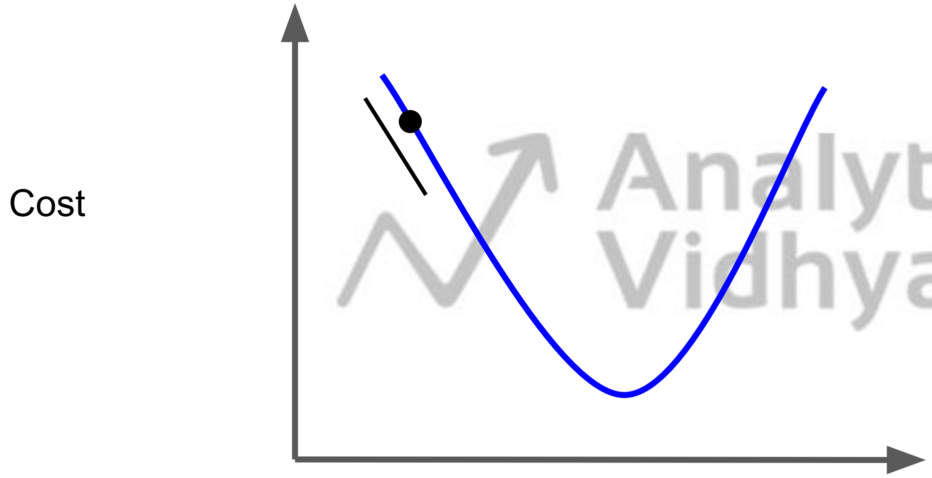
Cost



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

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

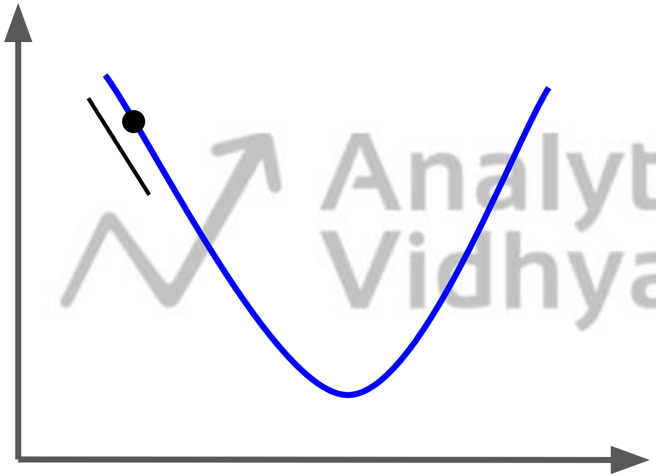
Understanding RMSProp



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

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

Understanding RMSProp

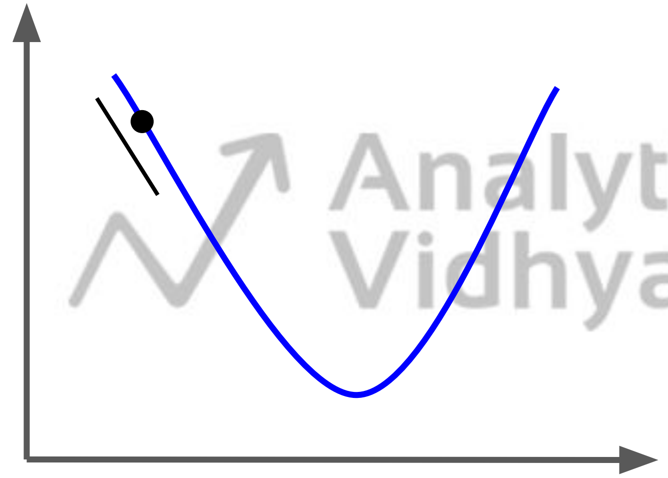


$$\mu_t = \beta\, \mu_{t-1} + (1-\beta) \overbrace{\boxed{\left[\frac{dJ}{d\theta}\right]^2}}^{\text{High}}$$

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

Understanding RMSProp

Cost



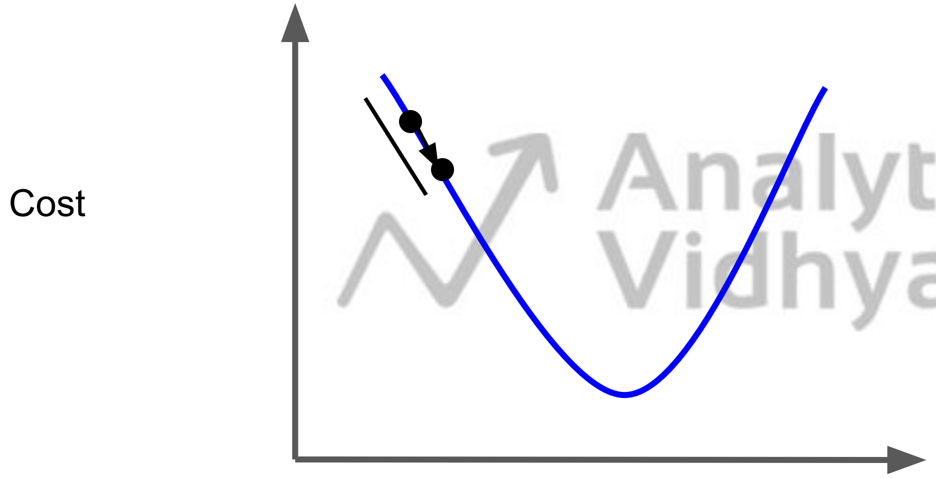
High

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

$$\theta_i = \theta_i - \boxed{\frac{\alpha}{\sqrt{\mu_t + \varepsilon}}}\,\frac{dJ}{d\theta}$$

Low

Understanding RMSProp



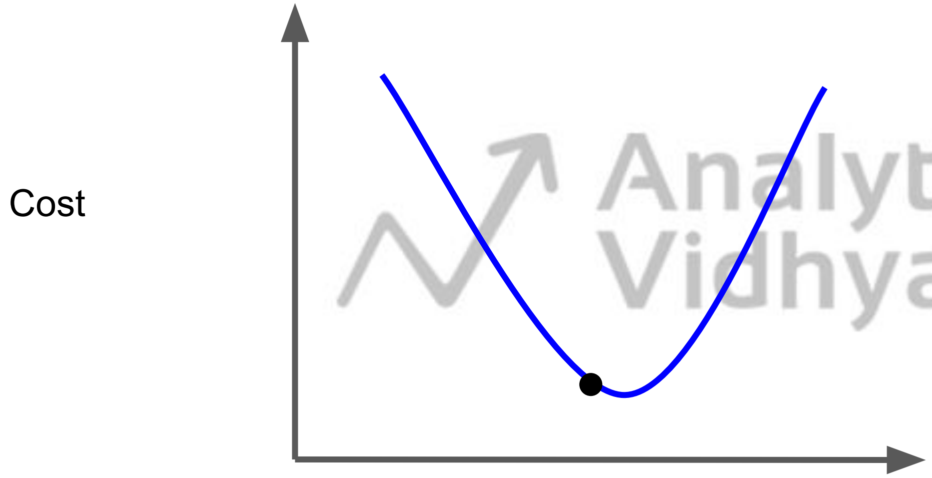
High

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

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

Low

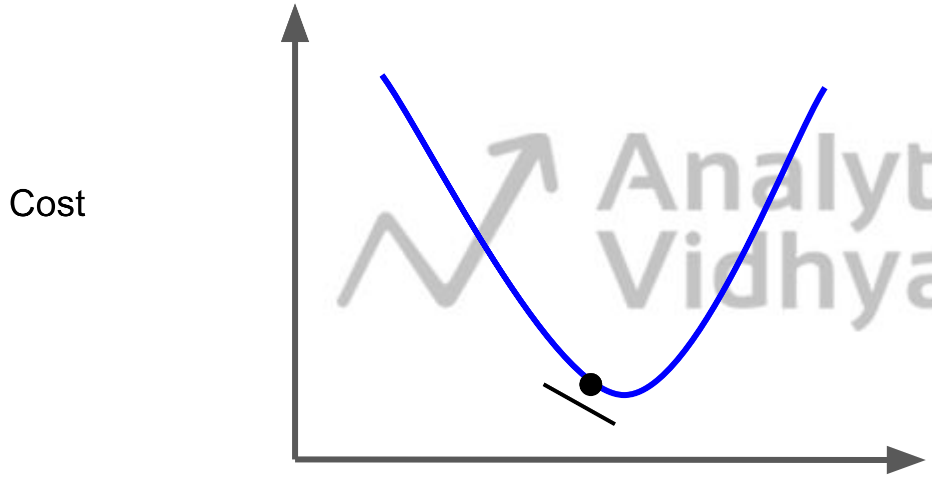
Understanding RMSProp



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

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

Understanding RMSProp

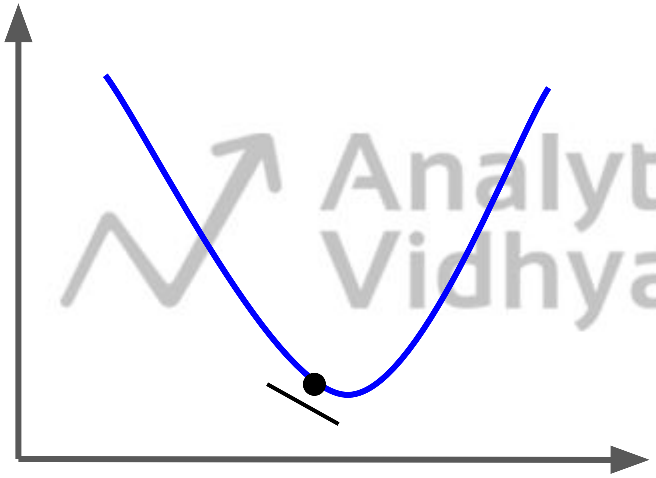


$$\mu_t = \beta\,\mu_{t-1} + (1-\beta)\left[\frac{dJ}{d\theta}\right]^2$$

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

Understanding RMSProp

Cost

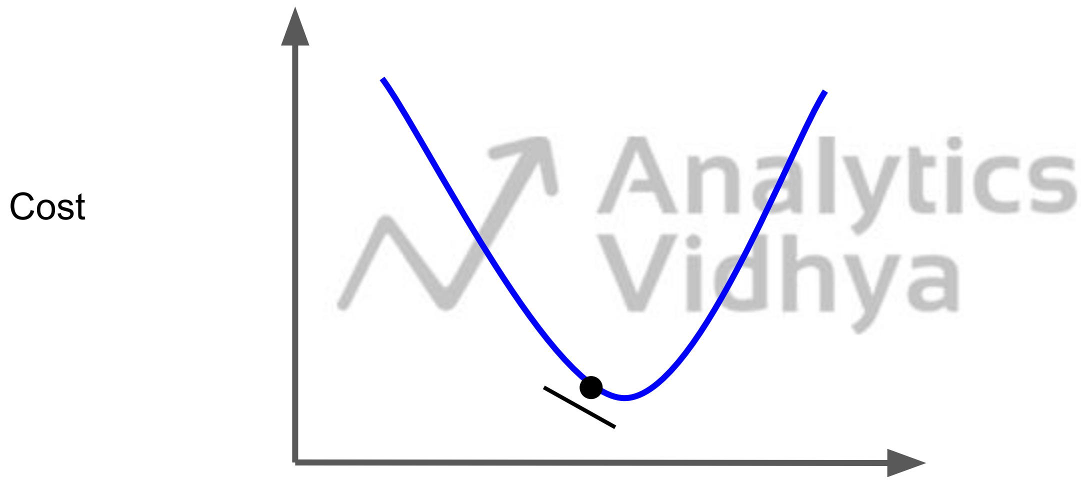


Low

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

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

Understanding RMSProp



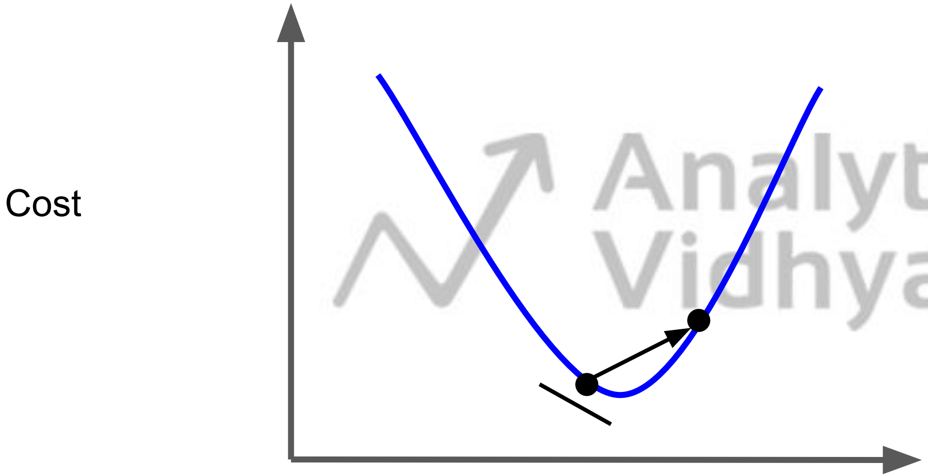
Low

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

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

High

Understanding RMSProp



Low

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

$$\theta_i = \theta_i - \boxed{\frac{\alpha}{\sqrt{\mu_t + \varepsilon}}}\,\frac{dJ}{d\theta}$$

High

Thank You