Gradient Descent

$$\theta := \theta - \alpha \, d\theta$$ hyperparameter: α

Gradient Descent with Momentum

$$v_\theta := \beta v_\theta + (1-\beta) d\theta$$ hyperparameter: α, β

$$\theta := \theta - \alpha v_\theta$$ common value: 0.9

(some literature

$$v_\theta := \beta v_\theta + d\theta$$

)

RMSprop (Root Mean Square)

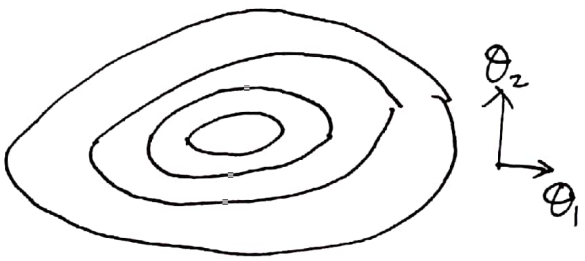
element-wise ↓

$$S_\theta := \beta S_\theta + (1-\beta) d\theta^2$$

$$\theta := \theta - \alpha \frac{d\theta}{\sqrt{S_\theta + \varepsilon}}$$

ε ↓ ensure numerical stability

intuition:



in θ_1 dimention,

$d\theta_1$ smaller → S_{θ_1} smaller → $\frac{1}{\sqrt{S_{\theta_1}}}$ larger

in θ_2 dimention

$d\theta_2$ larger → S_{θ_2} larger → $\frac{1}{\sqrt{S_{\theta_2}}}$ smaller

Adam Optimization (Adaptive Moment Estimation)

(gradient descent with momentum + RMS prop gradient descent)

$$V_{\theta} := \beta_1 V_{\theta} + (1-\beta_1)\, d\theta$$

$$S_{\theta} := \beta_2 S_{\theta} + (1-\beta_2)\, d\theta^2$$

$$V_{\theta_{corrected}} = \frac{V_{\theta}}{1-\beta_1^t}$$

$$S_{\theta_{corrected}} = \frac{S_{\theta}}{1-\beta_2^t}$$

$$\theta := \theta - \alpha \frac{V_{\theta_{corrected}}}{\sqrt{S_{\theta_{corrected}}} + \varepsilon}$$

Hyperparameters:

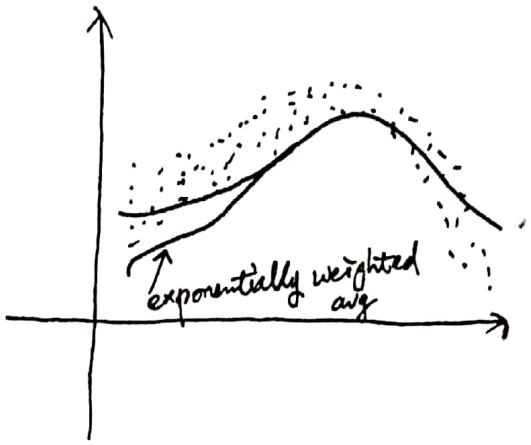
- α
- β_1 0.9
- β_2 0.999
- ε : 10^{-8}

Learning Rate Decay

- $\alpha = \frac{\alpha_0}{1 + (\text{decay\_rate})^{\text{epoch\_num}}}$
- $\alpha = 0.95^{\text{epoch\_num}} * \alpha_0$
- $\alpha = \frac{k}{\sqrt{\text{epoch\_num}}}\, \alpha_0$
- manually adjust α

Exponentially Weighted Average

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



Bias Correction

$$\frac{v_t}{1-\beta^t}$$

correct the initial phase