HW-2

EE16BTECH11023

HAVISH

SGD :-

The gradient updates proceed as follows in each case :-

For all cases, batch size = m,

$\odot \rightarrow$ Hadamard product

SGD :-

Gradient :- $\nabla_\theta (R(\theta)) = \nabla_\theta \left(\sum_{i=1}^{m} R_i(\theta) \right)$

update :- $\theta^{(r+1)} = \theta^{(r)} - (lr) \nabla_\theta (R(\theta))$

Momentum :-

$$v^{(0)} = 0$$

$$v^{(r)} = \alpha v^{(r-1)} - \epsilon \nabla_\theta^{(r)} R(\theta)$$

$$\theta^{(r)} = \theta^{(r-1)} + v^{(r)}$$

$v \rightarrow$ similar to velocity.

Clearly, effective learning rate depends on the parameters in previous step.

Nesterov :-

> Here, we apply an interim update on the parameter set an find gradient wrt that update.

$$\tilde{\theta}_{(r)} = \theta_{(r)} + \alpha v$$

$$g = \left(\frac{1}{m}\right) \nabla_{\theta} \sum R_i(\theta)$$

$$v = \alpha v - \varepsilon g$$

$$\theta_{(r+1)} = \theta_{(r)} + v_{(r)}$$

Ada-grad :-

Find the gradient

$$g = \frac{1}{m} \nabla_{\theta} \sum R_i(\theta)$$

$$r = r + (g \odot g) \rightarrow \text{square of gradient.}$$

update :-

$$\theta = \theta - \frac{\varepsilon}{\delta + \sqrt{r}} \cdot g$$

$\delta = 0 \quad (1e(-10))$

RMS - Prop :-

Small variation of the grad.

$$r^{(k+1)} = r^{(k)} \cdot (1-\rho)$$

$$g = \frac{1}{m} \nabla_\theta \sum R_i(\theta)$$

$$r^{(k+1)} = r^{(k)} (1-\rho) + \rho \cdot g \odot g$$

update :-

$$\theta^{(k+1)} = \theta^{(k)} - \frac{\varepsilon}{\sqrt{\delta + r^{(k+1)}}} \odot g$$

Adam :-

We decay the learning rate over batches using the 1st & 2nd moments.

So :-

$$g = \frac{1}{m} \nabla_\theta \left(\sum R_i(\theta) \right)$$

$t = t+1 \longrightarrow$ time step.

$s = \rho_1 s + (1-\rho_1) g$ (1st moment)

$r = \rho_2 r + (1-\rho_2) g \odot g$ (2nd ")

update :-

$$\theta = \theta - \varepsilon \frac{\hat{s}}{\sqrt{\hat{r}} + \delta}$$

where

$$\hat{s} = \frac{s}{1 - \rho_1^t}$$

$$\hat{r} = \frac{r}{1 - \rho_2^t}$$