Gradient descent methods in NN

Cost function: $J(\theta)$, $\theta \in \mathbb{R}^d$ NN params

Gradient: $\nabla_\theta J(\theta) = \begin{bmatrix} \frac{\partial J}{\partial \theta_1} \\ \vdots \\ \frac{\partial J}{\partial \theta_d} \end{bmatrix}$

Generally speaking, $\theta^{(t+1)} = \theta^{(t)} - \Delta^{(t)}$

① Gradient descent

$$\Delta^{(t)} = \eta \nabla_\theta J(\theta^{(t)}) \leftarrow \eta : \text{learning rate}$$

② Momentum

$$\Delta^{(t)} = \gamma \Delta^{(t-1)} + \eta \nabla_\theta J(\theta^{(t)})$$

γ : decay factor, 0.9

③ Adagrad

$$\Delta^{(t)} = \frac{\eta}{\sqrt{G^{(t)} + \varepsilon I}} \nabla_\theta J(\theta^{(t)})$$

$$G^{(t)} = \sum^{t} \begin{bmatrix} \frac{\partial J}{\partial \theta_1} & & \\ & \ddots & \end{bmatrix}^2$$

$$i=1 \quad \left[\quad \cdots \quad \frac{\partial J}{\partial \theta_d} \right]_{\theta=\theta^{(i)}}$$

④ RMSprop

$$\Delta^{(t)} = \frac{\eta}{\sqrt{G^{(t)} + \varepsilon I}} \nabla_\theta J(\theta^{(t)})$$

$$G^{(t)} = \gamma G^{(t-1)} + (1-\gamma) \begin{bmatrix} \frac{\partial J}{\partial \theta_1} & & \\ & \ddots & \\ & & \frac{\partial J}{\partial \theta_d} \end{bmatrix}^2_{\theta=\theta^{(t)}}$$

0.9

⑤ Adam

$$m^{(t)} = \beta_1 m^{(t-1)} + (1-\beta_1) \nabla_\theta J(\theta^{(t)})$$

$$v^{(t)} = \beta_2 v^{(t-1)} + (1-\beta_2) \left(\nabla_\theta J(\theta^{(t)})\right)^2$$

element wise

$$\hat{m}^{(t)} = \frac{m^{(t)}}{1-\beta_1^t}, \quad \hat{v}^{(t)} = \frac{v^{(t)}}{1-\beta_2^t}$$

$$\Delta^{(t)} = \frac{\eta}{\sqrt{\hat{v}^{(t)}} + \varepsilon} \hat{m}^{(t)}$$

Typical values: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$