Q(5) Adam is suitable for large models with complex datasets, where there are noisy gradients or vanishing gradients/exploding gradients.

Mathematical Formulation

$f_t \rightarrow$ hypothesis at t^{th} step

$\theta \rightarrow$ parameters

Repeat till convergence:

$$g_t = \nabla_\theta f_t(\theta_{t-1})$$

[redacted]

$$m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) \cdot g_t$$

[update momentum (weighted avg of gradients)]

$$v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) \cdot g_t^2$$

[update second moment (variance of gradients)]

$$\left.\begin{aligned} \hat{m}^t &\leftarrow \frac{m_t}{1-\beta_1^t} \\ \hat{v}^t &\leftarrow \frac{v_t}{1-\beta_2^t} \end{aligned}\right\} \text{(Bias-correction)}$$

$$\theta_t \leftarrow \theta_{t-1} - \frac{\alpha \cdot \hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}.$$