Mathematical Formulation of ADAM Optimizer with Example

1 Introduction

ADAM (Adaptive Moment Estimation) is an optimization algorithm that combines the advantages of both momentum and Root Mean Square Propagation (RMSprop). It is widely used in deep learning.

RMSprop (Root Mean Square Propagation) is an adaptive learning rate optimization algorithm commonly used for training neural networks. It is designed to address the limitations of standard gradient descent by adjusting the learning rate for each parameter individually, preventing drastic oscillations and ensuring more stable convergence.

2 Mathematical Formulation

Given a stochastic objective function $f(\theta)$ where θ represents parameters, ADAM updates the parameters iteratively using the following equations:

1. Initialize parameters
   - Initialize first moment estimate $m_0 = 0$
   - Initialize second moment estimate $v_0 = 0$
   - Set hyperparameters:
     - Learning rate α
     - Decay rates β_1, β_2 (default: 0.9, 0.999)
     - Small constant $\epsilon =$ (default: 10^-8)
   - At each time step t:
     - Compute gradient: $g_t = \nabla_\theta f(\theta_t)$
     - Update biased first moment estimate (momentum term):

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{1}$$

     - Update biased second moment estimate (RMS term):

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

– Compute bias-corrected moments:

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{3}$$

– Update parameters:

$$\theta_{t+1} = \theta_t - \frac{\alpha \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{4}$$

- Repeat until convergence.

3 Simple Example

Consider a simple function:

$$f(\theta) = \theta^2 \tag{5}$$

with initial parameter $\theta_0 = 2$, learning rate $\alpha = 0.1$, and ADAM parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$.

For the first iteration:

(a) Compute gradient: $g_1 = 2(2) = 4$

(b) Compute first moment: $m_1 = 0.9(0) + 0.1(4) = 0.4$

(c) Compute second moment: $v_1 = 0.999(0) + 0.001(16) = 0.016$

(d) Bias correction: $\hat{m}_1 = \frac{0.4}{1-0.9} = 4$, $\hat{v}_1 = \frac{0.016}{1-0.999} = 16$

(e) Parameter update:

$$\theta_1 = 2 - \frac{0.1 \times 4}{\sqrt{16} + 10^{-8}} = 2 - 0.1 = 1.9 \tag{6}$$

Thus, $\theta_1 = 1.9$.

4 Conclusion

The ADAM optimizer adaptively updates learning rates using first and second moment estimates. It is effective in minimizing loss functions in machine learning applications.