Bias Correction in Adam Optimization

Detailed Bias Correction Formulation

Let's define the key terms:

- g_t: Gradient at time step t (varies over time, no constant assumption).
- m_t: First moment estimate (exponential moving average of gradients).
- v_t: Second moment estimate (exponential moving average of squared gradients).
- β_1: Decay rate for the first moment (e.g., 0.9).
- β_2: Decay rate for the second moment (e.g., 0.999).
- η: Learning rate.
- ϵ: Small constant to prevent division by zero.

The update rules in Adam are:

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

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

With initial conditions:

$$m_0 = 0, \quad v_0 = 0 \tag{3}$$

The parameter update in Adam uses:

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

where $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected estimates.

Formulating the Moving Averages

First, let's express m_t and v_t explicitly as weighted sums of past gradients.

For m_t, starting from $m_0 = 0$:

$$m_1 = \beta_1 m_0 + (1 - \beta_1) g_1 = (1 - \beta_1) g_1 \tag{5}$$

$$m_2 = \beta_1 m_1 + (1 - \beta_1) g_2 = \beta_1 (1 - \beta_1) g_1 + (1 - \beta_1) g_2 \tag{6}$$

$$m_3 = \beta_1 m_2 + (1 - \beta_1) g_3 = \beta_1^2 (1 - \beta_1) g_1 + \beta_1 (1 - \beta_1) g_2 + (1 - \beta_1) g_3 \tag{7}$$

Generalizing:

$$m_t = (1 - \beta_1) \sum_{i=1}^{t} \beta_1^{t-i} g_i \tag{8}$$

Here, β_1^{t-i} is the weight of gradient g_i at time i, decaying exponentially as the time difference $t - i$ increases.

Similarly, for v_t:

$$v_t = (1 - \beta_2) \sum_{i=1}^{t} \beta_2^{t-i} g_i^2 \tag{9}$$

Identifying the Bias

The moving average m_t is a weighted sum of gradients up to time t. The weights are:

$$w_i = (1 - \beta_1)\beta_1^{t-i}, \quad i = 1, 2, ..., t \tag{10}$$

Sum of weights:

$$\sum_{i=1}^{t} w_i = (1 - \beta_1) \sum_{i=1}^{t} \beta_1^{t-i} = (1 - \beta_1) \sum_{k=0}^{t-1} \beta_1^k \tag{11}$$

$$= (1 - \beta_1) \cdot \frac{1 - \beta_1^t}{1 - \beta_1} = 1 - \beta_1^t \tag{12}$$

Since $\beta_1 < 1$, $0 < \beta_1^t < 1$ for finite t, so the total weight $1 - \beta_1^t < 1$. This means m_t doesn't fully represent the average of past gradients—its weights don't sum to 1, underrepresenting the true mean, especially when t is small (e.g., $\beta_1 = 0.9$, $t = 1$, $1 - 0.9 = 0.1$).

Compare this to an unbiased moving average over t steps:

$$\text{Unbiased mean} = \frac{1}{t} \sum_{i=1}^{t} g_i \tag{13}$$

The weights in m_t decay exponentially and sum to less than 1, biasing m_t toward zero due to $m_0 = 0$.

Bias Correction

To make m_t an unbiased estimator of a weighted mean, we normalize by the sum of weights:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} = \frac{(1 - \beta_1) \sum_{i=1}^{t} \beta_1^{t-i} g_i}{1 - \beta_1^t} \tag{14}$$

Now, the effective weights $\frac{(1-\beta_1)\beta_1^{t-i}}{1-\beta_1^t}$ sum to 1:

$$\sum_{i=1}^{t} \frac{(1 - \beta_1)\beta_1^{t-i}}{1 - \beta_1^t} = \frac{(1 - \beta_1) \sum_{i=1}^{t} \beta_1^{t-i}}{1 - \beta_1^t} = \frac{1 - \beta_1^t}{1 - \beta_1^t} = 1 \tag{15}$$

Similarly:

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} = \frac{(1 - \beta_2) \sum_{i=1}^{t} \beta_2^{t-i} g_i^2}{1 - \beta_2^t} \tag{16}$$