

# AdamW 完整的更新公式

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

其中:

- $\theta_{t+1}$ : 下一次迭代的参数值。这是优化器在当前步骤更新后, 模型将使用的新的权重或偏置。
- $\theta_t$ : 当前迭代的参数值
- $\eta$ : 学习率(Leanning Rate)。它控制了模型在梯度方向上更新的步长。学习率越大, 参数更新越快, 但也可能导致训练不稳定; 学习率越小, 训练越稳定, 但收敛速度可能变慢。
- $\lambda$ : 权重衰减系数(Weight Decay Coefficient)。这个一个超参数, 用于控制权重衰减的强度。 $\lambda$  越大, 权重被推向零的力度越大, 模型正则化效果越好。
- $\epsilon$ : 一个很小的常数, 通常为  $\frac{1}{10^8}$ 。它的作用是防止分母为零, 从而提高数值稳定性。
- $\hat{m}_t$ : 偏差修正后的一阶矩估计。它是一阶矩  $m_t$  经过偏差修正后的版本。
  - $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$ : 一阶矩(First Moment)。它是梯度的指数移动平均, 可以看作梯度的"动量"。
  - $\beta_1$ : 用于计算一阶矩的超参数, 通常设置为 0.9。
  - $g_t$ : 原始损失函数对参数的梯度, 在反向传播时已经计算出来。
- $\hat{v}_t$ : 偏差修正后的二阶矩估计。它由二阶矩  $v_t$  经过偏差修正的得到。
  - $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$ : 二阶矩 (Second Moment)。它是梯度平方的指数移动平均, 用于估计梯度的方差。
  - $\beta_2$ : 用于计算二阶矩的超参数, 通常设置为 0.999

## 偏差修正的含义

在 Adam 和 AdamW 优化器中, 偏差修正(Bias Correction) 是一个关键步骤。

- **一阶矩偏差修正**: 在训练初期,  $m_t$  的值会倾向于零, 因为它从零开始初始化。 $\frac{1}{1-\beta_1^t}$  这一修正项可以抵消这个偏差, 确保  $m_t$  的估计值在训练初期就能准确反映梯度的真实平均值。
- **二阶矩偏差修正**: 同理,  $v_t$  也会有类似的偏差, 修正项  $\frac{1}{1-\beta_2^t}$  同样能够纠正这一问题。

我们可以将 AdamW 的更新公式分解为两个主要部分:

- Adam 更新项:

$$- \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t$$

- 权重衰减项:

$$-\eta\lambda\theta_t$$

这部分是 AdamW 优化器独有的，它独立于 Adam 更新项，直接将当前的参数值按比例减小，以实现正则化。

# 具体实现过程

## 初始化

设置如下超参数：

- $\eta$ : 学习率,比如 0.0001。
- $\lambda$ : 权重衰减系统，比如 0.01。
- $\beta_1$ : 计算一阶矩超参数，通常设置为 0.9。
- $\beta_2$ : 计算二阶矩超参数，通常设置为 0.999。
- $\epsilon$ : 小常数，防止分母为零。

还包含一些内部参数：

- $t$ : 时间步，用于偏差修正，每次更新自动+1,初始为零。
- $m\_states$ : 存储各个可更新参数的一阶矩估计矩阵，形状与可更新参数矩阵一致。
- $v\_states$ : 存储各个可更新参数的二阶矩估计矩阵，形状与可更新参数矩阵一致。

## 更新过程

在一次训练的反向传播之后，将需要将所有具有可学习参数层的权重参数、偏置参数(可选)、权重梯度、偏置梯度(可选) 传给 AdamW 进行梯度更新。

并且，更新权重参数和偏置参数都是逐元素更新，也就是对参数中的每个元素应用相同的计算方法。以下举例，我们只考虑其中一个元素的更新过程。

1. 计算偏差修正因子,这个计算与参数无关，所以可以预先计算：

$$t = t + 1$$

$$bias_1 = 1.0 - \beta_1^t$$

$$bias_2 = 1.0 - \beta_2^t$$

2. 权重衰减

首先要确保当前更新参数的  $[m\_states, v\_states]$  存在，并且与参数矩阵形状一致。这个过程可以在第一次更新的时候完成，不需要预先初始化。

注意，权重衰减通常只应用于权重，不应用于偏置。

计算公式：

$$\theta_{temp} = \theta_t - \eta \lambda \theta_t$$

或者等价形式：

$$\theta_{temp} = \theta_t (1 - \eta \lambda)$$

实际计算时， $\theta_t$  原地操作，所以  $\theta_t = \theta_{temp}$

### 3. 参数更新

首先取出缓存的上次更新的一阶矩动量  $m_{t-1}$  和二阶矩动量  $v_{t-1}$ （对应于当前参数矩阵的当前元素），然后更新  $m_t$  和  $v_t$ ：

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

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

将以上  $m_t$  和  $v_t$  更新到缓存中，以便下次使用。

偏差修正：

$$\hat{m}_t = \frac{m_t}{1.0 - \beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1.0 - \beta_2^t}$$

最后的参数更新：

$$\theta_{t+1} = \theta_{temp} - \eta \left( \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \right)$$

在更新偏置参数时，因为没有权重衰减，所以直接用  $\theta_t$  替代  $\theta_{temp}$ 。最后将  $\theta_{t+1}$  写入原参数即可。