🚀 Adam Optimizer (Adaptive Moment Estimation)

Adam is an advanced optimization algorithm that **combines the advantages of both Momentum and RMSProp**. It adapts the learning rate for each parameter by using estimates of **first and second moments of the gradients**.

🔍 Core Components

1. **Momentum (First Moment Estimate)**

   Helps in **smoothing** the gradient by taking a moving average of past gradients:

$$V_{d\omega} = \beta_1 V_{d\omega} + (1 - \beta_1)\frac{\partial L}{\partial \omega}$$

$$V_{db} = \beta_1 V_{db} + (1 - \beta_1)\frac{\partial L}{\partial b}$$

   - β_1 is typically set around **0.9**
   - This captures the **direction** (momentum) of gradients

2. **RMSProp (Second Moment Estimate)**

   Maintains an exponentially weighted average of squared gradients for adaptive learning rate scaling:

$$S_{d\omega} = \beta_2 S_{d\omega} + (1 - \beta_2)\left(\frac{\partial L}{\partial \omega}\right)^2$$

$$S_{db} = \beta_2 S_{db} + (1 - \beta_2)\left(\frac{\partial L}{\partial b}\right)^2$$

   - β_2 is typically set around **0.999**
   - Prevents large updates by normalizing gradients

⚙️ Final Parameter Update Rules

Weights and biases are updated using both moment estimates:

$$\omega_t = \omega_{t+1} - \frac{\eta \cdot V_{d\omega}}{\sqrt{S_{d\omega}} + \epsilon}$$

$$b_t = b_{t+1} - \frac{\eta \cdot V_{db}}{\sqrt{S_{db}} + \epsilon}$$

Where:

- η: learning rate
- ϵ: small constant to prevent division by zero (e.g., 10^{-8})

✅ **Advantages of Adam**

- Combines **momentum's smoothing** and **RMSProp's adaptive learning rate**
- Works well in practice and is widely used
- Minimal hyperparameter tuning required

📌 Initialization & Iterative Steps

1. Initialize:

   $V_{d\omega}, V_{db}, S_{d\omega}, S_{db} = 0$
2. On each iteration t with a mini-batch:
   - Compute gradients:
     $\frac{\partial L}{\partial \omega}, \frac{\partial L}{\partial b}$
   - Update momentum and squared gradients
   - Apply the final update rule to weights and biases