ADAM algorithm

A brief review of the ADAM algorithm used.

Initialization:

- Learning Rate (α)
- Exponential Decay Rates for Moment Estimates (β_1 and β_2): Parameters used to control the decay rates of moment estimates.
- Small Constant to Avoid Division by Zero (ϵ): A small value added to the denominator to prevent division by zero.
- Moment Vectors:
  - First Moment Estimate (m): A vector of zeros with the same dimension as the parameters.
  - Second Moment Estimate (v): A vector of zeros with the same dimension as the parameters.

For each t:

1. Compute Gradients: $g_t = \nabla f(x_t)$, where $f(x_t)$ is the objective function with respect to the parameters x_t at iteration t.
2. Update Biased First Moment Estimate:
$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$$
3. Update Biased Second Moment Estimate:
$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$$
4. Correct Bias in First Moment Estimate:
$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$
5. Correct Bias in Second Moment Estimate:
$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$
6. Update Parameters:
$$x_{t+1} = x_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t$$

Termination:

The stopping criterion used is the same used for the other methods.