Optimizer

Linh Le

November 2024

1 Gradient Descent:

$$f(w_1, w_2) = 0.1w_1^2 + 2w_2^2$$

Gradient Descent:

$$W = W - \alpha * \nabla W$$

Initial Conditions:

- Weights: $w_1 = -5, w_2 = -2$
- Learning rate: $\alpha = 0.4$
- Epochs: 2

Epoch 1:

Step 1: Compute Gradient

$$dw_1 = 0.2 * w_1 = 0.2 * (-5) = -1$$
$$dw_2 = 4 * w_2 = 4 * (-2) = -8$$

Step 2: Update w_1 and w_2

$$w_1 = w_1 - \alpha * dw_1 = -5 - 0.4 * (-1) = -4.6$$
$$w_2 = w_2 - \alpha * dw_2 = -2 - 0.4 * (-8) = 1.2$$

Epoch 2:

Step 1: Compute Gradient

$$dw_1 = 0.2 * w_1 = 0.2 * (-4.6) = -0.92$$
$$dw_2 = 4 * w_2 = 4 * 1.2 = 4.8$$

Step 2: Update w_1 and w_2

$$w_1 = w_1 - \alpha * dw_1 = -4.6 - 0.4 * (-0.92) = -4.232$$
$$w_2 = w_2 - \alpha * dw_2 = 1.2 - 0.4 * 4.8 = -0.7$$

2 Gradient Descent + Momentum:

$$f(w_1, w_2) = 0.1w_1^2 + 2w_2^2$$

Gradient Descent + Momentum:

$$V_t = \beta V_{t-1} + (1 - \beta)dW_t$$
$$W_t = W_t - \alpha * V_t$$

Initial Conditions:

- Weights: $w_1 = -5, w_2 = -2$
- Momentum vector: $v_1 = 0, v_2 = 0$
- Learning rate: $\alpha = 0.6$
- Momentum hyperparameter: $\beta = 0.5$
- Epochs: 2

Epoch 1

Step 1: Compute Gradient

$$dw_1 = 0.2w_1 = 0.2 * (-5) = -1$$
$$dw_2 = 4w_2 = 4 * (-2) = -8$$

Step 2: Update v_1 and v_2

$$v_1 = \beta v_1 + (1 - \beta) * dw_1 = 0.5 * 0 + (1 - 0.5) * (-1.0) = -0.5$$
$$v_2 = \beta v_2 + (1 - \beta) * dw_2 = 0.5 * 0 + (1 - 0.5) * (-8.0) = -4.0$$

Step 3: Update w_1 and w_2

$$w_1 = w_1 - \alpha * v_1 = -5 - 0.6 * (-0.5) = -4.7$$
$$w_2 = w_2 - \alpha * v_2 = -2 - 0.6 * (-4.0) = 0.4$$

Epoch 2

Step 1: Compute Gradient

$$dw_1 = 0.2 * w_1 = 0.2 * (-4.7) = -0.94$$

$$dw_2 = 4 * w_2 = 4 * 0.4 = 1.6$$

Step 2: Update v_1 and v_2

$$v_1 = \beta v_1 + (1 - \beta) * dw_1 = 0.5 * (-0.5) + (1 - 0.5) * (-0.94) = -0.72$$

$$v_2 = \beta v_2 + (1 - \beta) * dw_2 = 0.5 * (-4.0) + (1 - 0.5) * 1.6 = -1.2$$

Step 3: Update w_1 and w_2

$$w_1 = w_1 - \alpha * v_1 = -4.7 - 0.6 * (-0.72) = -4.268$$

$$w_2 = w_2 - \alpha * v_2 = 0.4 - 0.6 * (-1.2) = 1.12$$

3 RMSProp:

$$f(w_1, w_2) = 0.1w_1^2 + 2w_2^2$$

RMSProp:

$$S_t = \gamma S_{t-1} + (1 - \gamma) dW_t^2$$

$$W_t = W_t - \frac{\alpha * dW_t}{\sqrt{S_t + \epsilon}}$$

Initial Conditions:

- Weights: $w_1 = -5, w_2 = -2$
- Running average of squared gradients: $s_1 = 0, s_2 = 0$
- Learning rate: $\alpha = 0.3$
- Decay rate: $\gamma = 0.9$
- Small constant: $\epsilon = 10^{-6}$
- Epochs: 2

Epoch 1

Step 1: Compute Gradient

$$dw_1 = 0.2 * w_1 = 0.2 * (-5) = -1$$

$$dw_2 = 4 * w_2 = 4 * (-2) = -8$$

Step 2: Update s_1 and s_2

$$s_1 = \gamma s_1 + (1-\gamma) dw_1^2 = 0.9 * 0 + (1-0.9) * (-1)^2 = 0.1$$

$$s_2 = \gamma s_2 + (1-\gamma) dw_2^2 = 0.9 * 0 + (1-0.9) * (-8)^2 = 6.4$$

Step 3: Update w_1 and w_2

$$w_1 = w_1 - \frac{\alpha * dw_1}{\sqrt{s_1 + \epsilon}} = -5 - \frac{0.3 * (-1)}{\sqrt{0.1 + 10^{-6}}} = -4.05$$

$$w_2 = w_2 - \frac{\alpha * dw_2}{\sqrt{s_2 + \epsilon}} = -2 - \frac{0.3 * (-8)}{\sqrt{6.4 + 10^{-6}}} = -1.05$$

Epoch 2

Step 1: Compute Gradient

$$dw_1 = 0.2 * w_1 = 0.2 * (-4.05) = -0.81$$

$$dw_2 = 4 * w_2 = 4 * (-1.05) = -4.2$$

Step 2: Update s_1 and s_2

$$s_1 = \gamma s_1 + (1-\gamma) dw_1^2 = 0.9 * 0.1 + (1-0.9) * (-0.81)^2 = 0.16$$

$$s_2 = \gamma s_2 + (1-\gamma) dw_2^2 = 0.9 * 6.4 + (1-0.9) * (-4.2)^2 = 7.5$$

Step 3: Update w_1 and w_2

$$w_1 = w_1 - \frac{\alpha * dw_1}{\sqrt{s_1 + \epsilon}} = -4.05 - \frac{0.3 * (-0.81)}{\sqrt{0.16 + 10^{-6}}} = -3.43$$

$$w_2 = w_2 - \frac{\alpha * dw_2}{\sqrt{s_2 + \epsilon}} = -1.05 - \frac{0.3 * (-4.2)}{\sqrt{7.5 + 10^{-6}}} = -0.59$$

4 Adam:

$$f(w_1, w_2) = 0.1w_1^2 + 2w_2^2$$

Adam:

$$V_t = \beta_1 V_{t-1} + (1 - \beta_1) dW_t$$

$$S_t = \beta_2 S_{t-1} + (1 - \beta_2) dW_t^2$$

$$V_{corr} = \frac{V_t}{1 - \beta_1^t}$$

$$S_{corr} = \frac{S_t}{1 - \beta_2^t}$$

$$W_t = W_t - \alpha * \frac{V_{corr}}{\sqrt{S_{corr}} + \epsilon}$$

Initial Conditions:

- Weights: $w_1 = -5, w_2 = -2$
- Momentum: $v_1 = 0, v_2 = 0$
- Running average of squared gradients: $s_1 = 0, s_2 = 0$
- Decay rates: $\beta_1 = 0.9, \beta_2 = 0.999$
- Learning rate: $\alpha = 0.2$
- Small constant: $\epsilon = 10^{-6}$
- Epochs: 2

Epoch 1

Step 1: Compute Gradient

$$dw_1 = 0.2 * w_1 = 0.2 * (-5) = -1$$

$$dw_2 = 4 * w_2 = 4 * (-2) = -8$$

Step 2: Update v_1 and v_2

$$v_1 = \beta_1 v_1 + (1 - \beta_1) dw_1 = 0.9 * 0 + (1 - 0.9) * (-1) = -0.1$$

$$v_2 = \beta_1 v_2 + (1 - \beta_1) dw_2 = 0.9 * 0 + (1 - 0.9) * (-8) = -0.8$$

Step 3: Update s_1 and s_2

$$s_1 = \beta_2 s_1 + (1-\beta_2) dw_1^2 = 0.999 * 0 + 0.001 * (-1)^2 = 0.001$$
$$s_2 = \beta_2 s_2 + (1-\beta_2) dw_2^2 = 0.999 * 0 + 0.001 * (-8)^2 = 0.064$$

Step 4: Bias correction for v_1, v_2 and s_1, s_2

t = 1

$$v_{corr_1} = \frac{v_1}{1-\beta_1^1} = \frac{-0.1}{1-0.9^1} = -1$$
$$v_{corr_2} = \frac{v_2}{1-\beta_1^1} = \frac{-0.8}{1-0.9^1} = -8$$
$$s_{corr_1} = \frac{s_1}{1-\beta_2^1} = \frac{0.001}{1-0.999^1} = 1$$
$$s_{corr_2} = \frac{s_2}{1-\beta_2^1} = \frac{0.064}{1-0.999^1} = 64$$

Step 5: Update w_1 and w_2

$$w_1 = w_1 - \alpha * \frac{v_{corr_1}}{\sqrt{s_{corr_1}} + \epsilon} = -5 - 0.2 * \frac{-1}{\sqrt{1} + 10^{-6}} = -4.8$$
$$w_2 = w_2 - \alpha * \frac{v_{corr_2}}{\sqrt{s_{corr_1}} + \epsilon} = -2 - 0.2 * \frac{-8}{\sqrt{64} + 10^{-6}} = -1.8$$

Epoch 2

Step 1: Compute Gradient

$$dw_1 = 0.2 * w_1 = 0.2 * (-4.8) = -0.96$$
$$dw_2 = 4 * w_2 = 4 * (-1.8) = -7.2$$

Step 2: Update v_1 and v_2

$$v_1 = \beta_1 v_1 + (1-\beta_1) dw_1 = 0.9 * (-0.1) + (1-0.9) * (-0.96) = -0.186$$
$$v_2 = \beta_1 v_2 + (1-\beta_1) dw_2 = 0.9 * (-0.8) + (1-0.9) * (-7.2) = -1.44$$

Step 3: Update s_1 and s_2

$$s_1 = \beta_2 s_1 + (1-\beta_2) dw_1^2 = 0.999 * 0.001 + 0.001 * (-0.96)^2 = 0.00192$$
$$s_2 = \beta_2 s_2 + (1-\beta_2) dw_2^2 = 0.999 * 0.064 + 0.001 * (-7.2)^2 = 0.1157$$

Step 4: Bias correction for v_1, v_2 and s_1, s_2

t = 2

$$v_{corr_1} = \frac{v_1}{1-\beta_1^1} = \frac{-0.186}{1-0.9^2} = -0.98$$

$$v_{corr_2} = \frac{v_2}{1-\beta_1^1} = \frac{-1.44}{1-0.9^2} = -7.58$$

$$s_{corr_1} = \frac{s_1}{1-\beta_2^1} = \frac{0.00192}{1-0.999^2} = 0.96$$

$$s_{corr_2} = \frac{s_2}{1-\beta_2^1} = \frac{0.1157}{1-0.999^2} = 57.92$$

Step 5: Update w_1 and w_2

$$w_1 = w_1 - \alpha * \frac{v_{corr_1}}{\sqrt{s_{corr_1}} + \epsilon} = -4.8 - 0.2 * \frac{-0.98}{\sqrt{0.96} + 10^{-6}} = -4.6$$

$$w_2 = w_2 - \alpha * \frac{v_{corr_2}}{\sqrt{s_{corr_1}} + \epsilon} = -1.8 - 0.2 * \frac{-7.58}{\sqrt{57.92} + 10^{-6}} = -1.6$$