ADAM Optimization

ADAM

- $m_1 = \beta_1 \cdot m_0 + (1 - \beta_1) \cdot f'(x)$
- $v_1 = \beta_2 \cdot v_0 + (1 - \beta_2) \cdot (f'(x))^2$

4. Correct the bias:
   - $\hat{m}_1 = \frac{m_1}{1-\beta_1^t}$
   - $\hat{v}_1 = \frac{v_1}{1-\beta_2^t}$

$$w_{t+1} = w_t - \alpha \cdot \frac{\hat{m}_1}{\sqrt{\hat{v}_1}+\epsilon}$$

ADAM

$$m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * \nabla w_t$$

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

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \qquad \hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} * \hat{m}_t$$

$\eta : Initial\ Learning\ rate$
$g_t : Gradient\ at\ time\ t\ along\ \omega^j$
$\nu_t : Exponential\ Average\ of\ gradients\ along\ \omega_j$
$s_t : Exponential\ Average\ of\ squares\ of\ gradients\ along\ \omega_j$
$\beta_1, \beta_2 : Hyperparameters$

ADAM

- β_1: 0.9
- β_2: 0.999

- Given:Weight (w) = 0.5, Input (x,y)=(3,4), learning rate =0.1 ϵ = 10^{-8}
- Our goal is to minimize the mean squared error (MSE) loss function:
- L = 1/n* (y_pred - y)^2
- y_pred = 0.5×3=1.5
- gt=dL_dw = 2/1 * (y_pred - y) * x= -15
- $gt^2=(-15)^2$ =225
- $m1$=0.9*0+(1−0.9)*-15= −1.5.
- $v1$=0.999*0+(1−0.999)* 225= 0.225

$$\hat{m}_2 = \frac{-1.5}{1-0.9^2}$$

$$\hat{m}_2 = \frac{-1.5}{1-0.81}$$

$$\hat{m}_2 = \frac{-1.5}{0.19}$$

$$\hat{m}_2 \approx -7.8947$$

$$\hat{v}_1 = \frac{0.225}{1-0.999^2}$$

$$\hat{v}_1 = \frac{0.225}{1-0.998001}$$

$$\hat{v}_1 = \frac{0.225}{0.001999}$$

$$\hat{v}_1 \approx 112.811$$

$$w_{t+1} = w_t - \alpha \cdot \frac{\hat{m}_1}{\sqrt{\hat{v}_1}+\epsilon}$$

$$0.5 - \frac{0.1\times(-7.8947)}{\sqrt{112.811}+10^{-8}}$$

≈0.4257

Early Stopping

In Regularization by Early Stopping, we stop training the model when the performance on the validation set is getting worse- increasing loss decreasing accuracy, or poorer scores of the scoring metric.

- Early stopping essentially returns the set of parameters that were used at this point and so is equivalent to stopping training at that point.
- The final parameters returned will enable the model to have low variance and better generalization.
- The model at the time the training is stopped will have a better generalization performance than the model with the least training error.