Optimization Method

12/04/2021
NYU Shanghai

- Momentum

$$t=0, \quad \text{set } v_0 = 0$$

$$t>0, \quad v_t \leftarrow \gamma v_{t-1} + \eta_t g_t$$

$$x_t \leftarrow x_{t-1} - v_t$$

hyperparameter $0 < \gamma \le 1$, when $\gamma = 0$ it's normal SGD

related to: exponentially weighted moving average (EWMA)

by writing: $v_t \leftarrow \gamma v_{t-1} + (1-\gamma)\left(\frac{\eta_t}{1-\gamma} g_t\right)$

we see v_t is EWMA on the series $\left\{\frac{\eta_t}{1-\gamma} g_t\right\}_{t=0,1,\dots}$

focus on recent $\frac{1}{1-\gamma}$ steps result.

- AdaGrad

adjustable learning rates on each dimension for flexibility

$$t=0, \quad \text{set } S_0 = 0$$

$$t>0, \quad S_t \leftarrow S_{t-1} + g_t \odot g_t$$

$\odot$ element-wise multiply

$$x_t \leftarrow x_{t-1} + \frac{\eta}{\sqrt{S_t + \varepsilon}} \odot g_t$$

Here η is learning rate, ε is constant $\approx 10^{-6}$ for stability

By accumulating the square of gradient in S_t, if one dimension's gradient is always big, it's learning rate drop quickly.

If one dimension's gradient is always small, then its gradient drop slowly.

- ⚠ the learning rate for AdaGrad is always dropping.
  so it might not find optimal solution in later stage due to too small learning rate.

- RMSProp

to solve the "later-too-low-learning-rate" in AdaGrad

$$t=0, \quad S_0 = 0$$

$$t>0, \quad S_t \leftarrow \gamma S_t + (1-\gamma)\cdot g_t \odot g_t$$

$$X_t \leftarrow X_{t-1} - \frac{\eta}{\sqrt{S_t+\varepsilon}} \odot g_t$$

Hyperparameter $0 < \gamma \leq 1$, ε stability constant

notice RMSProp is EWMA on $\{g_t \odot g_t\}_{t=0,1,\dots}$ series, so that the learning rate does not always drop.

- AdaDelta

similarly to solve the low learning rate problem in AdaGrad

$$t=0: \quad S_0 = 0$$

$$\Delta X_0 = 0$$

$$t>0: \quad S_t \leftarrow \gamma S_{t-1} + (1-\gamma)\cdot g_t \odot g_t$$

$$g_t' \leftarrow \sqrt{\frac{\Delta X_{t-1}+\varepsilon}{S_t+\varepsilon}} \odot g_t$$

$$X_t \leftarrow X_{t-1} - g_t'$$

$$\Delta X_t \leftarrow \gamma \Delta X_{t-1} + (1-\gamma)\cdot g_t' \odot g_t'$$

Hyperparameter $0 \leq \gamma < 1$, ε stability constant

The difference between AdaDelta and RMSProp is use $\sqrt{\Delta X_{t-1}}$ to replace learning rate η_t.

- Adam

a combination of RMSProp and Momentum

$$t=0: \quad V_0 = 0, \quad S_0 = 0$$

$$t>0: \quad V_t \leftarrow \gamma_1 V_{t-1} + (1-\gamma_1)\cdot g_t$$

$$S_t \leftarrow \gamma_2 \cdot S_{t-1} + (1-\gamma_2)\cdot g_t \odot g_t$$

$$\hat{v}_t \leftarrow \frac{v_t}{1-\gamma_1^t}$$

$$\hat{s}_t \leftarrow \frac{s_t}{1-\gamma_2^t}$$

$$g_t' \leftarrow \frac{\eta \cdot \hat{v}_t}{\sqrt{\hat{s}_t} + \varepsilon}$$

$$x_t \leftarrow x_{t-1} - g_t'$$

Hyperparameter $0 \leq \gamma_1 < 1$ (suggested 0.9)

$0 \leq \gamma_2 < 1$ (suggested 0.999)

ε stability constant

notice by divide $1-\gamma_1^t$ and $1-\gamma_2^t$, we can the sum of weights of past time-stamp gradients equal 1.