Tuning Learning Rate(Optimizer)

Momentum

Let v_k be the variable that stores previous move, i.e. the momentum. In the beginning, $v_0 = 0$.

$$
\begin{aligned}
&\text{Initialize } \theta_0 \text{ and let } v_0 = 0. \\
&\qquad \theta_1 = \theta_0 + v_1, \quad v_1 = \lambda v_0 - \alpha \nabla L(\theta_0) = -\alpha \nabla L(\theta_0), \\
&\qquad \theta_2 = \theta_1 + v_2, \quad v_2 = \lambda v_1 - \alpha \nabla L(\theta_1) = -\lambda \alpha \nabla L(\theta_0) - \alpha \nabla L(\theta_1), \\
&\qquad\quad \vdots \\
&\qquad \theta_{t+1} = \theta_t + v_{t+1}, \quad v_{t+1} = \lambda v_t - \alpha \nabla L(\theta_t)
\end{aligned}
$$

Briefly, momentum method perturb current gradient by previous gradient(momentum).

Nesterov Accelerated Gradient(NAG)

Similar to momentum

$$
\begin{aligned}
&\text{Initialize } \theta_0 \text{ and let } v_0 = 0. \\
&\qquad \theta_1 = \theta_0 + v_1, \quad v_1 = \lambda v_0 - \alpha \nabla L(\theta_0 + \lambda v_0) = -\alpha \nabla L(\theta_0), \\
&\qquad \theta_2 = \theta_1 + v_2, \quad v_2 = \lambda v_1 - \alpha \nabla L(\theta_1 + \lambda v_1) = -\lambda \alpha \nabla L(\theta_0) - \alpha \nabla L(\theta_1 + \lambda v_1), \\
&\qquad\quad \vdots \\
&\qquad \theta_{t+1} = \theta_t + v_{t+1}, \quad v_{t+1} = \lambda v_t - \alpha \nabla L(\theta_t + \lambda v_t)
\end{aligned}
$$

Here, instead of perturbing current gradient, we perturb current parameter by previous gradient.

Adagrad(Adaptive Gradient)

Use first derivative to estimate second derivative.

$$
\begin{aligned}
\alpha &\leftarrow \frac{\alpha}{\sqrt{\sum_{i=0}^{t} (\nabla L(\theta_i))^2}}, \\
\theta_t &\leftarrow \theta_{t-1} - \frac{\alpha}{\sqrt{\sum_{i=0}^{t} (\nabla L(\theta_i))^2}} \nabla L(\theta^{t-1})
\end{aligned}
$$

In practice we will add ϵ in the denominator to avoid dividing by zero

$$
\theta_t \leftarrow \theta_{t-1} - \frac{\alpha}{\sqrt{\sum_{i=0}^{t} (\nabla L(\theta_i))^2 + \epsilon}} \nabla L(\theta^{t-1})
$$

Adadelta

This method needs the average of gradients $\mathrm{E}\left[\nabla L(\theta)^2\right]_t$ at step t and the average of parameter update $\mathrm{E}[\Delta\theta^2]_t$. Initialize $\mathrm{E}\left[\nabla L(\theta)^2\right]_0 = 0$, $\mathrm{E}[\Delta\theta^2]_0 = 0$ and choose a decay rate ρ, learning rate α and small ϵ.

$$\text{I. } \theta_1 = \theta_0 + \Delta\theta_0, \Delta\theta_0 = -\frac{\alpha}{\sqrt{\mathrm{E}\left[\nabla L(\theta)^2\right]_0 + \epsilon}} \nabla L(\theta_0)$$

$$\text{II. } \mathrm{E}[\Delta\theta^2]_0 = 0$$

$$\text{II. } \mathrm{E}\left[\nabla L(\theta)^2\right]_1 = \rho\mathrm{E}\left[\nabla L(\theta)^2\right]_0 + (1-\rho)\nabla L(\theta_1)^2$$

$$\text{II. } \theta_2 = \theta_1 + \Delta\theta_1, \Delta\theta_1 = -\frac{\sqrt{\mathrm{E}[\Delta\theta^2]_0 + \epsilon}}{\sqrt{\mathrm{E}\left[\nabla L(\theta)^2\right]_1 + \epsilon}} \nabla L(\theta_1) \equiv -\frac{RMS[\Delta\theta]_0}{RMS[\nabla L(\theta)]_1} \nabla L(\theta_1)$$

$$\text{III. } \mathrm{E}[\Delta\theta^2]_1 = \rho\mathrm{E}[\Delta\theta^2]_0 + (1-\rho)\Delta\theta_1^2$$

$$\text{III. } \mathrm{E}\left[\nabla L(\theta)^2\right]_2 = \rho\mathrm{E}\left[\nabla L(\theta)^2\right]_1 + (1-\rho)\nabla L(\theta_2)^2$$

$$\text{III. } \theta_3 = \theta_2 + \Delta\theta_2, \Delta\theta_2 = -\frac{RMS[\Delta\theta]_1}{RMS[\nabla L(\theta)]_2} \nabla L(\theta_2)$$

$$\#.\ \mathrm{E}[\Delta\theta^2]_{t-2} = \rho\mathrm{E}[\Delta\theta^2]_{t-3} + (1-\rho)\Delta\theta_{t-2}^2$$

$$\#.\ \mathrm{E}\left[\nabla L(\theta)^2\right]_t = \rho\mathrm{E}\left[\nabla L(\theta)^2\right]_{t-1} + (1-\rho)\nabla L(\theta_t)^2$$

$$\#.\ \theta_t = \theta_{t-1} + \Delta\theta_{t-1}, \Delta\theta_{t-1} = -\frac{RMS[\Delta\theta]_{t-2}}{RMS[\nabla L(\theta)]_{t-1}} \nabla L(\theta_{t-1})$$

RMSprop(Root Mean Square Propagation)

Manually determine a weight β.

$$\begin{aligned}
\theta_1 &\leftarrow \theta_0 - \frac{\alpha}{\sigma_0}\nabla L(\theta_0), \quad \sigma_0 = \nabla L(\theta_0), \\
\theta_2 &\leftarrow \theta_1 - \frac{\alpha}{\sigma_2}\nabla L(\theta_1), \quad \sigma_1 = \sqrt{\beta(\sigma_0)^2 + (1-\beta)\ (\ \nabla L(\theta_1))^2 + \epsilon}, \\
&\vdots \\
\theta_{t+1} &\leftarrow \theta_t - \frac{\alpha}{\sigma_t}\nabla L(\theta_t), \quad \sigma_t = \sqrt{\beta(\sigma_{t-1})^2 + (1-\beta)\ (\ \nabla L(\theta_t))^2 + \epsilon}
\end{aligned}$$

Adam(RMSprop+Momentum)

Two weight numbers β_1 and β_2. Two moment vectors v_k and σ_k. In the beginning $v_0 = 0$ and $\sigma_0 = 0$.

$$\begin{aligned}
\theta_1 &= \theta_0 - \alpha\frac{\sigma_1}{\sqrt{v_1} + \epsilon}, \quad \sigma_1 = \frac{\beta_1\sigma_0 + (1-\beta_1)\nabla L(\theta_0)}{1-\beta_1}, \quad v_1 = \frac{\beta_2 v_0 + (1-\beta_2)\ (\ \nabla L(\theta_0))^2}{1-\beta_2}, \\
\theta_2 &= \theta_1 - \alpha\frac{\sigma_2}{\sqrt{v_2} + \epsilon}, \quad \sigma_2 = \frac{\beta_1\sigma_1 + (1-\beta_1)\nabla L(\theta_1)}{1-\beta_1^2}, \quad v_2 = \frac{\beta_2 v_1 + (1-\beta_2)\ (\ \nabla L(\theta_1))^2}{1-\beta_2^2}, \\
&\vdots \\
\theta_{t+1} &= \theta_t - \alpha\frac{\sigma_{t+1}}{\sqrt{v_{t+1}} + \epsilon}, \quad \sigma_{t+1} = \frac{\beta_1\sigma_t + (1-\beta_1)\nabla L(\theta_t)}{1-\beta_1^t}, \quad v_{t+1} = \frac{\beta_2 v_t + (1-\beta_2)\ (\ \nabla L(\theta_t))^2}{1-\beta_2^t}
\end{aligned}$$