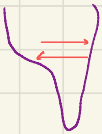
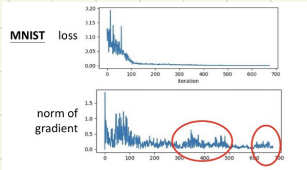
Error surface is rugged ... 误差表面凹凸不平

Adaptive Learning Rate

Training stuck ≠ Small Gradient

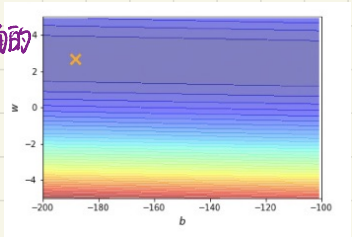


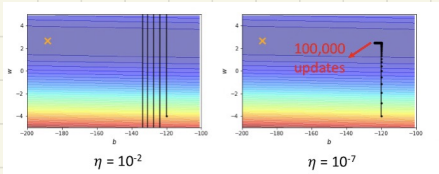
来回振荡



Training stuck without critical points

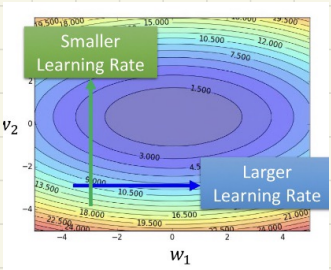
Convex 凸面的





无法靠近 critical points

Different parameters needs different learning rate



Update one parameter

$$\theta_i^{t+1} \leftarrow \theta_i^t - \eta g_i^t$$

$$g_i^t = \left.\frac{\partial L}{\partial \theta_i}\right|_{\theta=\theta^t}$$

$$\theta_i^{t+1} \leftarrow \theta_i^t - \frac{\eta}{\sigma_i^t} g_i^t$$

a. Root Mean Square (均方根) —— Adagrad

$$\theta_i^1 \leftarrow \theta_i^0 - \frac{\eta}{\sigma_i^0} g_i^0 \qquad \sigma_i^0 = \sqrt{(g_i^0)^2} = |g_i^0|$$

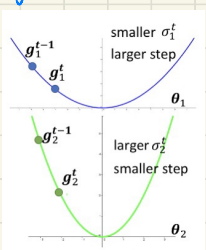
$$\theta_i^2 \leftarrow \theta_i^1 - \frac{\eta}{\sigma_i^1} g_i^1 \qquad \sigma_i^1 = \sqrt{\frac{1}{2}\left[(g_i^0)^2 + (g_i^1)^2\right]}$$

$$\theta_i^3 \leftarrow \theta_i^2 - \frac{\eta}{\sigma_i^2} g_i^2 \qquad \sigma_i^2 = \sqrt{\frac{1}{3}\left[(g_i^0)^2 + (g_i^1)^2 + (g_i^2)^2\right]}$$

$$\vdots$$

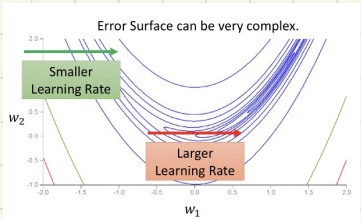
$$\theta_i^{t+1} \leftarrow \theta_i^t - \frac{\eta}{\sigma_i^t} g_i^t \qquad \sigma_i^t = \sqrt{\frac{1}{t+1}\sum_{i=0}^{t}(g_i^t)^2}$$

Used in Adagrad



b. RMSProp

Learning rate adapt dynamically



在绿色箭头上 需要 small LR

在红色箭头上 需要 Larger LR

需要在同一参数同一方向，LR也能动态调整 ⇒ RMSprop

↓

RMSprop

$$\theta_i^1 \leftarrow \theta_i^0 - \frac{\eta}{\sigma_i^0} g_i^0 \qquad \sigma_i^0 = \sqrt{(g_i^0)^2}$$

$$0 < \alpha < 1$$

$$\theta_i^2 \leftarrow \theta_i^1 - \frac{\eta}{\sigma_i^1} g_i^1 \qquad \sigma_i^1 = \sqrt{\alpha(\sigma_i^0)^2 + (1-\alpha)(g_i^1)^2}$$

$$\theta_i^3 \leftarrow \theta_i^2 - \frac{\eta}{\sigma_i^2} g_i^2 \qquad \sigma_i^2 = \sqrt{\alpha(\sigma_i^1)^2 + (1-\alpha)(g_i^2)^2}$$

$$\vdots$$

$$\theta_i^{t+1} \leftarrow \theta_i^t - \frac{\eta}{\sigma_i^t} g_i^t \qquad \sigma_i^t = \sqrt{\alpha(\sigma_i^{t-1})^2 + (1-\alpha)(g_i^t)^2}$$

$\alpha \to 0$ g_i^t 相较于之前计算的 gradient 比较重要

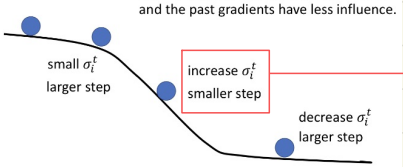
$\alpha \to 1$ 现在算出的 g_i^t 不重要，之前的重要

RMSProp

$$\theta_i^{t+1} \leftarrow \theta_i^t - \frac{\eta}{\sigma_i^t} g_i^t \qquad \sigma_i^t = \sqrt{\alpha(\sigma_i^{t-1})^2 + (1-\alpha)(g_i^t)^2}$$

$$0 < \alpha < 1$$

The recent gradient has larger influence, and the past gradients have less influence.

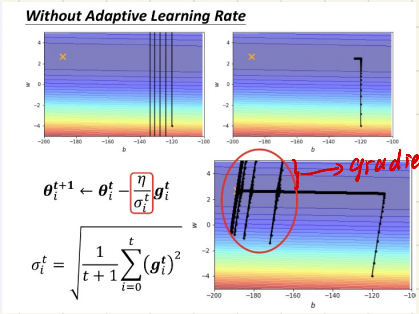


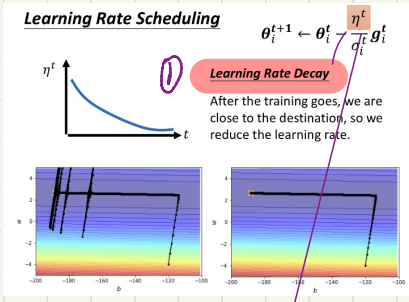
走到第三个球，按照 Adagrad 的话 LR会很大，导到第三个球飞出；所以得控制，α 可以用来控制 α

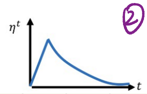
↓

Adam： RMSprop + Momentum

learning Rate Scheduling







② **Warm Up**

Increase and then decrease?

At the beginning, the estimate of σ_i^t has large variance.

Summary of optimization

(Vanilla) Gradient Descent

$$\boldsymbol{\theta}_i^{t+1} \leftarrow \boldsymbol{\theta}_i^{t} - \eta \boldsymbol{g}_i^{t}$$

Various Improvements

$$\boldsymbol{\theta}_i^{t+1} \leftarrow \boldsymbol{\theta}_i^{t} - \frac{\eta^t}{\sigma_i^t} \boldsymbol{m}_i^{t}$$

- η^t → Learning rate scheduling
- $\boldsymbol{m}_i^t$ → Momentum: weighted sum of the previous gradients 考虑大小和方向
- σ_i^t → root mean square of the gradients 只考虑 gradient 大小