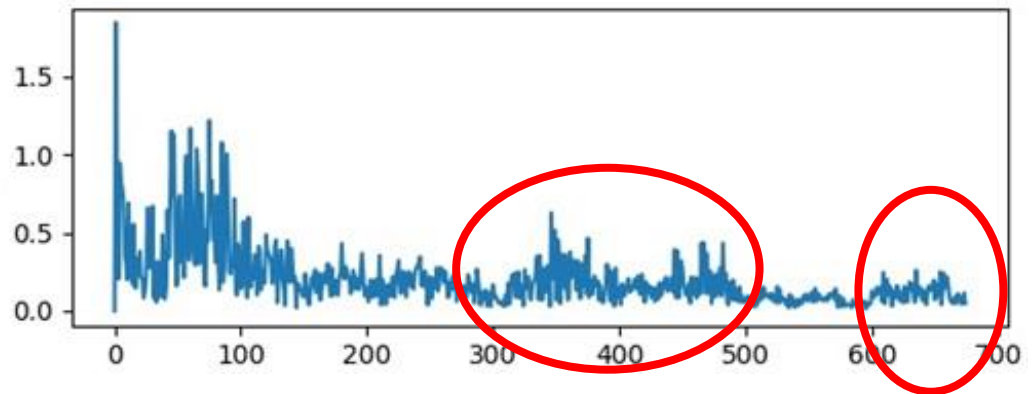
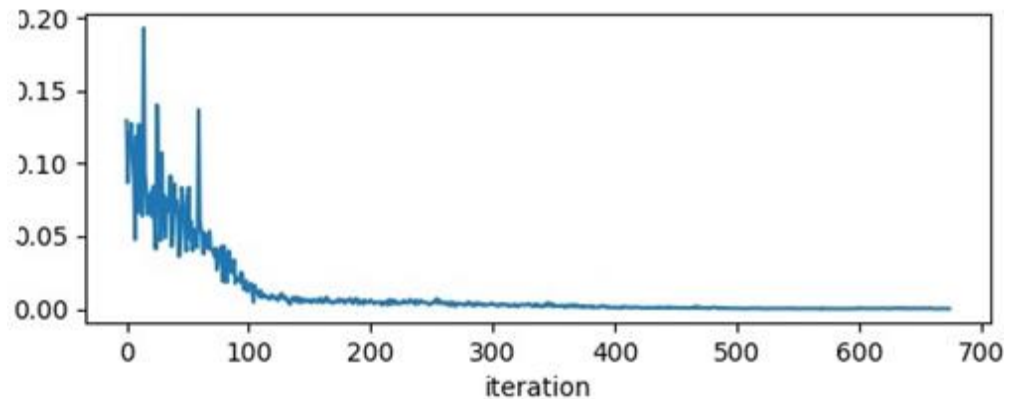
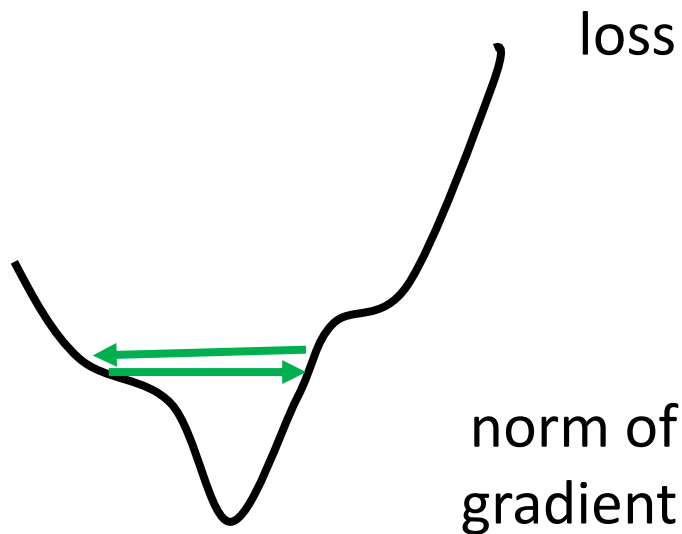


Error surface is rugged ...

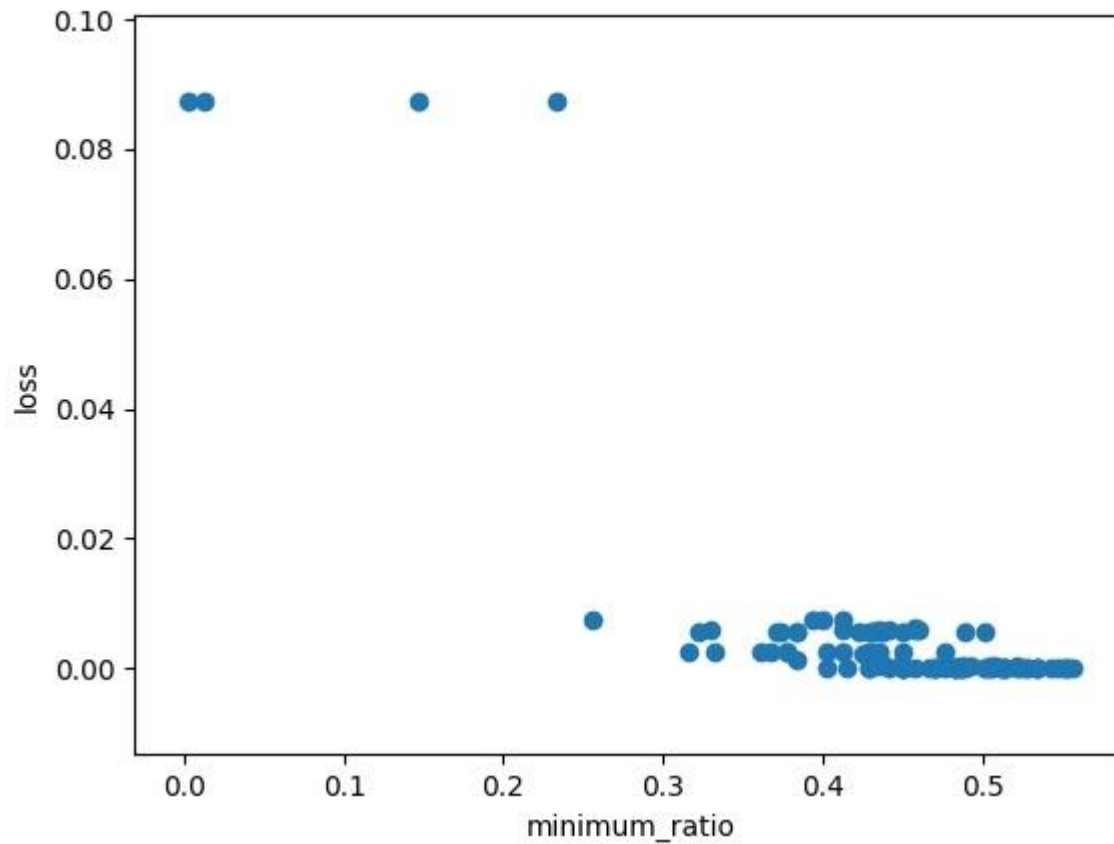
Tips for training: **Adaptive Learning Rate**

Training stuck \neq Small Gradient

- People believe training stuck because the parameters are around a critical point ...

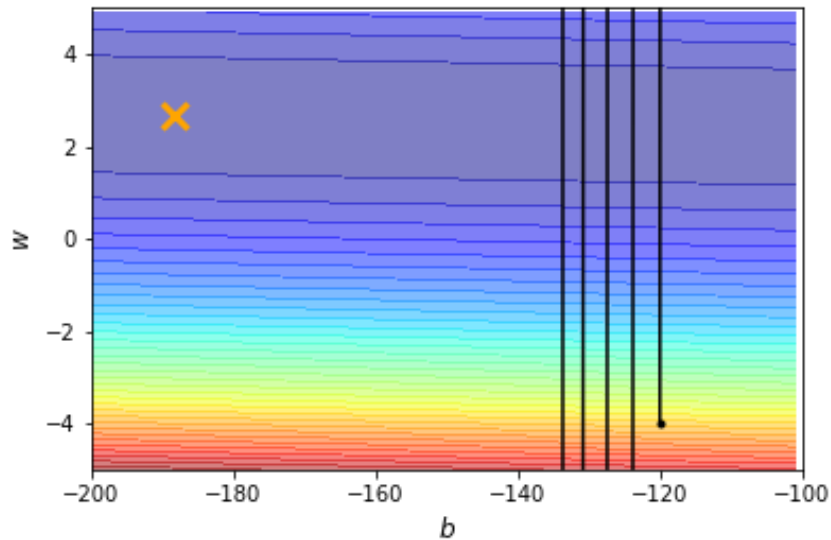
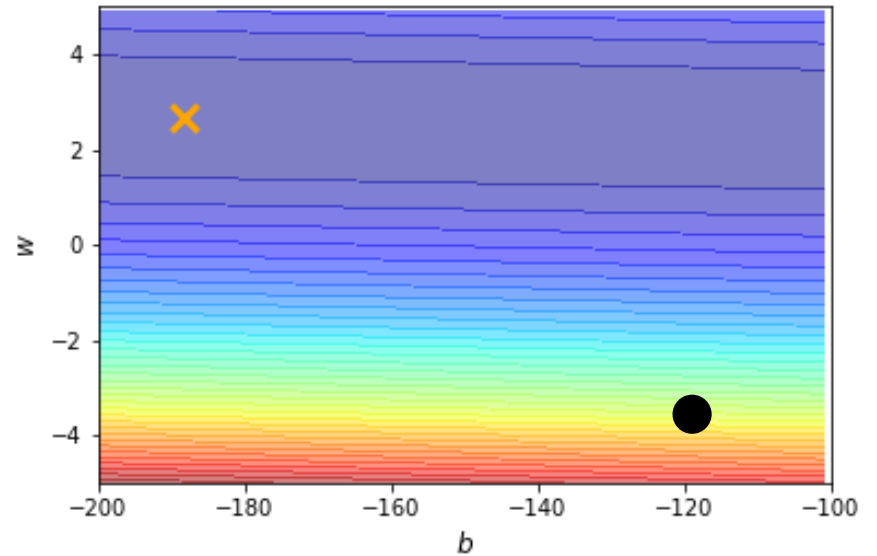


Wait a minute ...

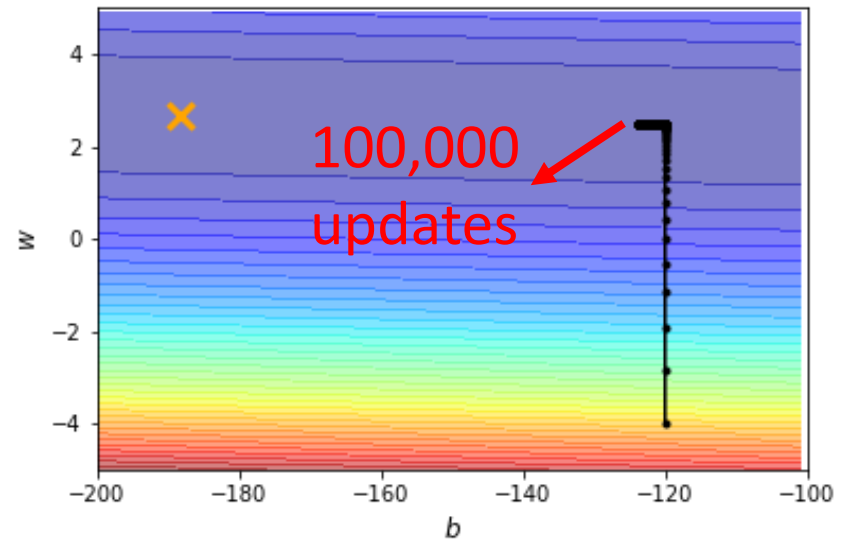


Training stuck without critical points

Learning rate **cannot** be
one-size-fits-all

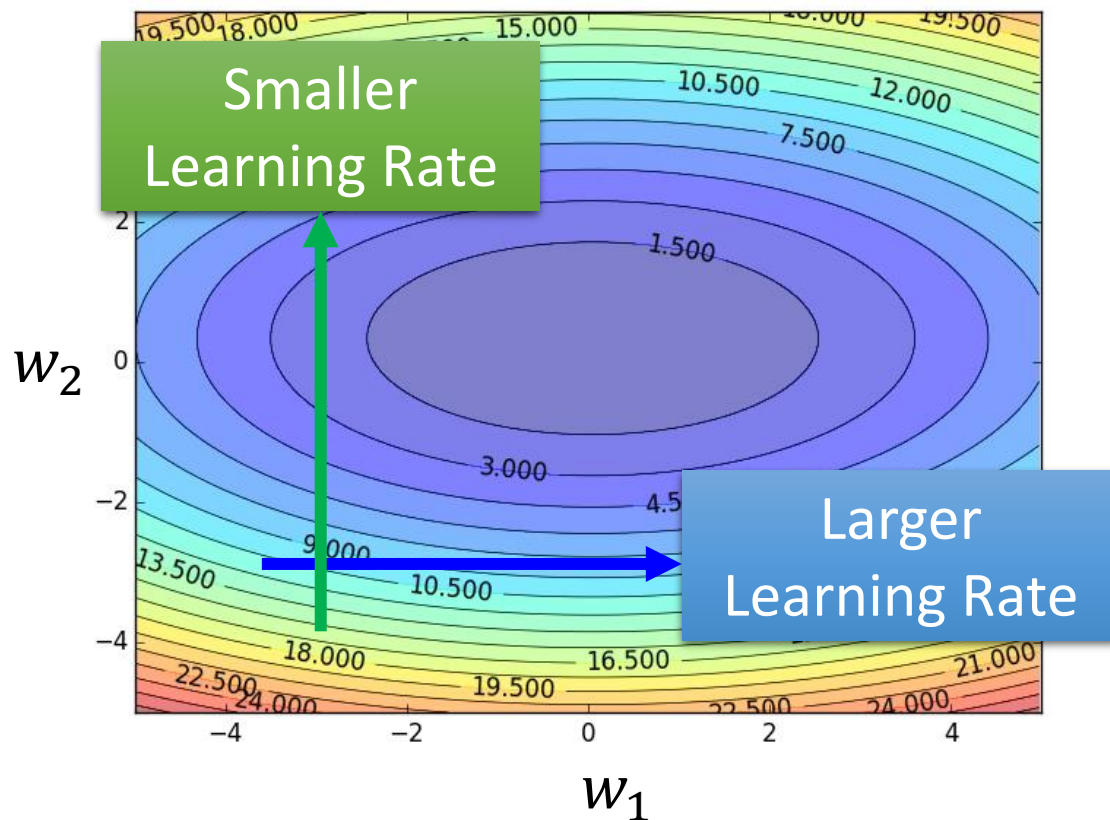


$$\eta = 10^{-2}$$



$$\eta = 10^{-7}$$

Different parameters need different learning rate



Update one parameter:

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

$$g_i^t = \frac{\partial L}{\partial \theta_i} \bigg|_{\theta = \theta^t}$$

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

Parameter
dependent

Root Mean Square

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

$$\boldsymbol{\theta}_i^1 \leftarrow \boldsymbol{\theta}_i^0 - \frac{\eta}{\sigma_i^0} \mathbf{g}_i^0 \quad \sigma_i^0 = \sqrt{(\mathbf{g}_i^0)^2}$$

$$\boldsymbol{\theta}_i^2 \leftarrow \boldsymbol{\theta}_i^1 - \frac{\eta}{\sigma_i^1} \mathbf{g}_i^1 \quad \sigma_i^1 = \sqrt{\frac{1}{2} [(\mathbf{g}_i^0)^2 + (\mathbf{g}_i^1)^2]}$$

$$\boldsymbol{\theta}_i^3 \leftarrow \boldsymbol{\theta}_i^2 - \frac{\eta}{\sigma_i^2} \mathbf{g}_i^2 \quad \sigma_i^2 = \sqrt{\frac{1}{3} [(\mathbf{g}_i^0)^2 + (\mathbf{g}_i^1)^2 + (\mathbf{g}_i^2)^2]}$$

\vdots

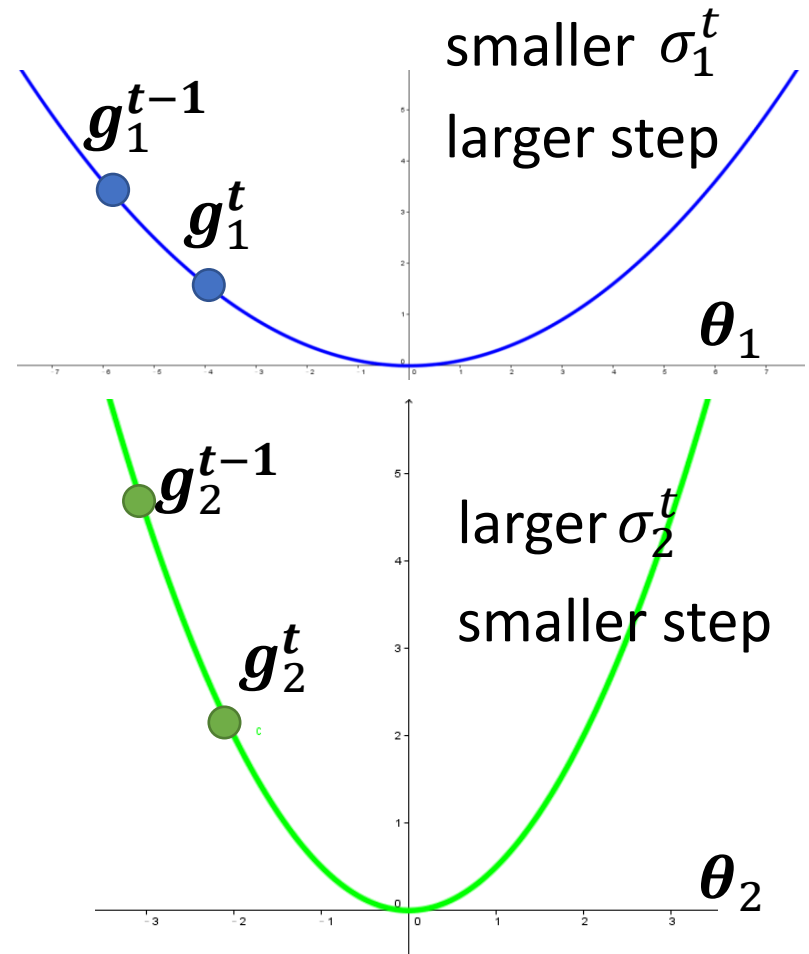
$$\boldsymbol{\theta}_i^{t+1} \leftarrow \boldsymbol{\theta}_i^t - \frac{\eta}{\sigma_i^t} \mathbf{g}_i^t \quad \sigma_i^t = \sqrt{\frac{1}{t+1} \sum_{i=0}^t (\mathbf{g}_i^t)^2}$$

Root Mean Square

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

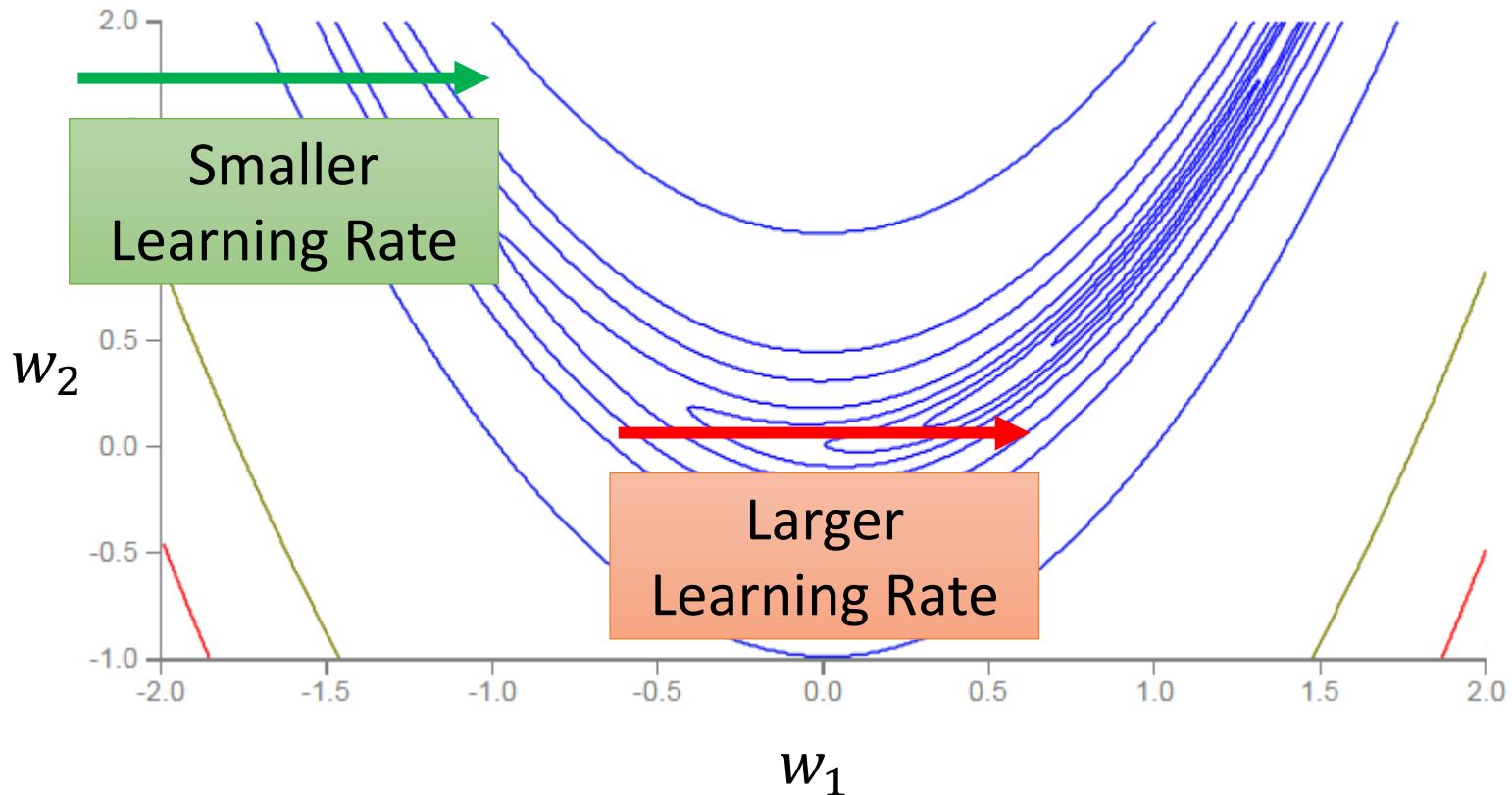
$$\sigma_i^t = \sqrt{\frac{1}{t+1} \sum_{i=0}^t (g_i^t)^2}$$

Used in **Adagrad**



Learning rate adapts dynamically

Error Surface can be very complex.



RMSProp

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

$$\boldsymbol{\theta}_i^1 \leftarrow \boldsymbol{\theta}_i^0 - \frac{\eta}{\sigma_i^0} \mathbf{g}_i^0 \quad \sigma_i^0 = \sqrt{(\mathbf{g}_i^0)^2} \quad 0 < \alpha < 1$$

$$\boldsymbol{\theta}_i^2 \leftarrow \boldsymbol{\theta}_i^1 - \frac{\eta}{\sigma_i^1} \mathbf{g}_i^1 \quad \sigma_i^1 = \sqrt{\alpha(\sigma_i^0)^2 + (1 - \alpha)(\mathbf{g}_i^1)^2}$$

$$\boldsymbol{\theta}_i^3 \leftarrow \boldsymbol{\theta}_i^2 - \frac{\eta}{\sigma_i^2} \mathbf{g}_i^2 \quad \sigma_i^2 = \sqrt{\alpha(\sigma_i^1)^2 + (1 - \alpha)(\mathbf{g}_i^2)^2}$$

⋮

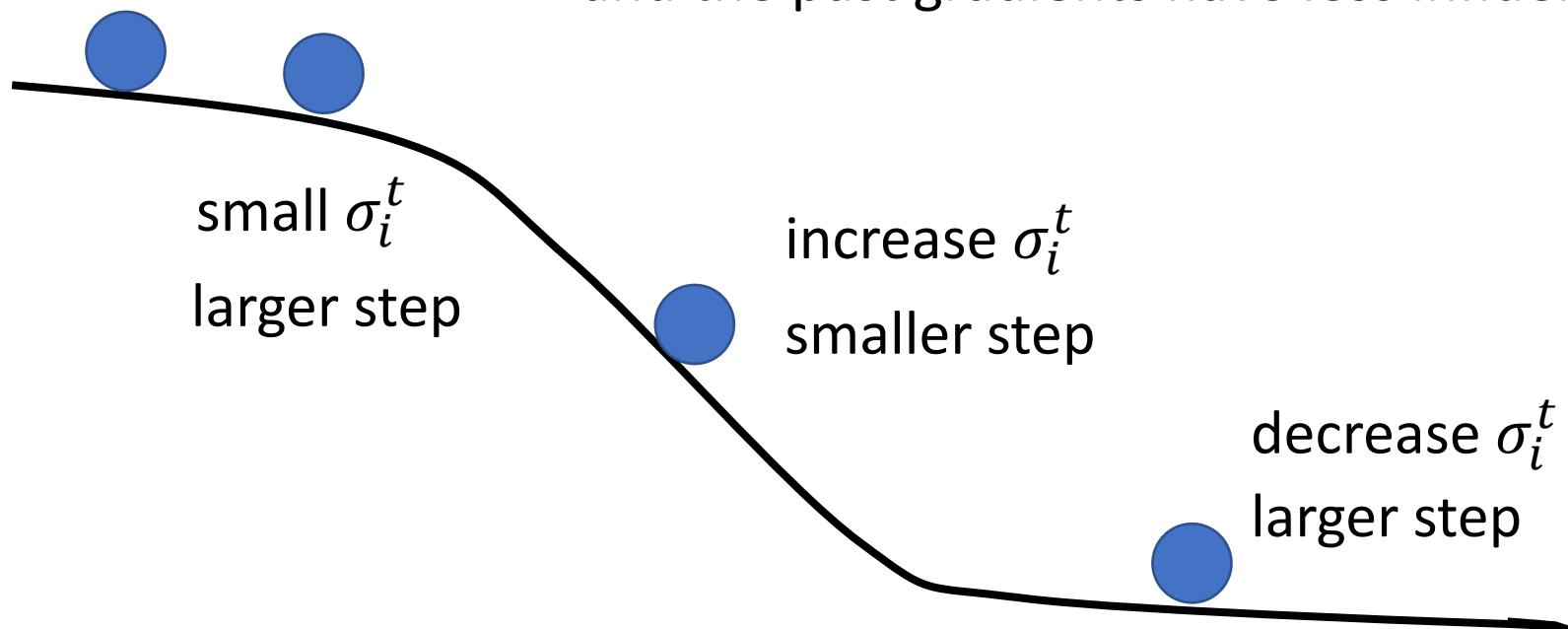
$$\boldsymbol{\theta}_i^{t+1} \leftarrow \boldsymbol{\theta}_i^t - \frac{\eta}{\sigma_i^t} \mathbf{g}_i^t \quad \sigma_i^t = \sqrt{\alpha(\sigma_i^{t-1})^2 + (1 - \alpha)(\mathbf{g}_i^t)^2}$$

RMSProp

$$0 < \alpha < 1$$

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

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



Adam: RMSProp + Momentum

Algorithm 1: *Adam*, our proposed algorithm for stochastic optimization. See section 2 for details, and for a slightly more efficient (but less clear) order of computation. g_t^2 indicates the elementwise square $g_t \odot g_t$. Good default settings for the tested machine learning problems are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. All operations on vectors are element-wise. With β_1^t and β_2^t we denote β_1 and β_2 to the power t .

Require: α : Stepsize

Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates

Require: $f(\theta)$: Stochastic objective function with parameters θ

Require: θ_0 : Initial parameter vector

$m_0 \leftarrow 0$ (Initialize 1st moment vector) \rightarrow for momentum

$v_0 \leftarrow 0$ (Initialize 2nd moment vector) \rightarrow for RMSprop

$t \leftarrow 0$ (Initialize timestep)

while θ_t not converged **do**

$t \leftarrow t + 1$

$g_t \leftarrow \nabla_{\theta} f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)

$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)

$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)

$\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)

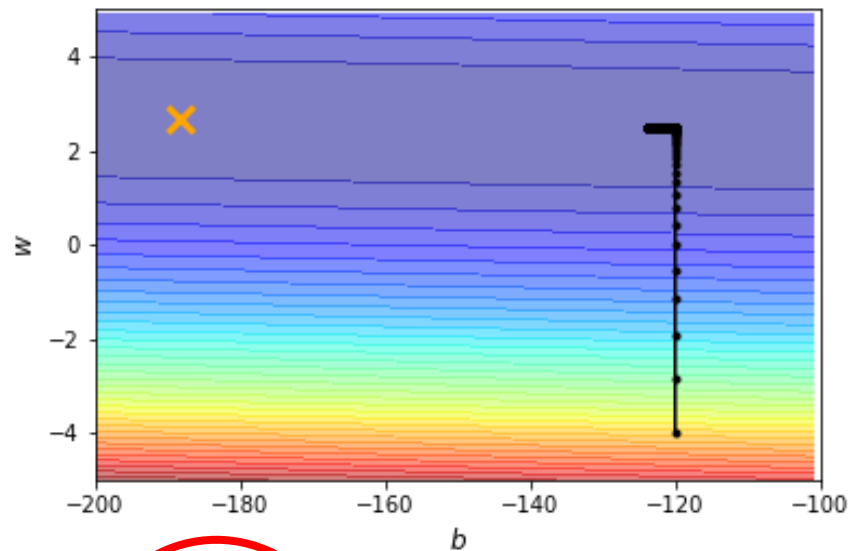
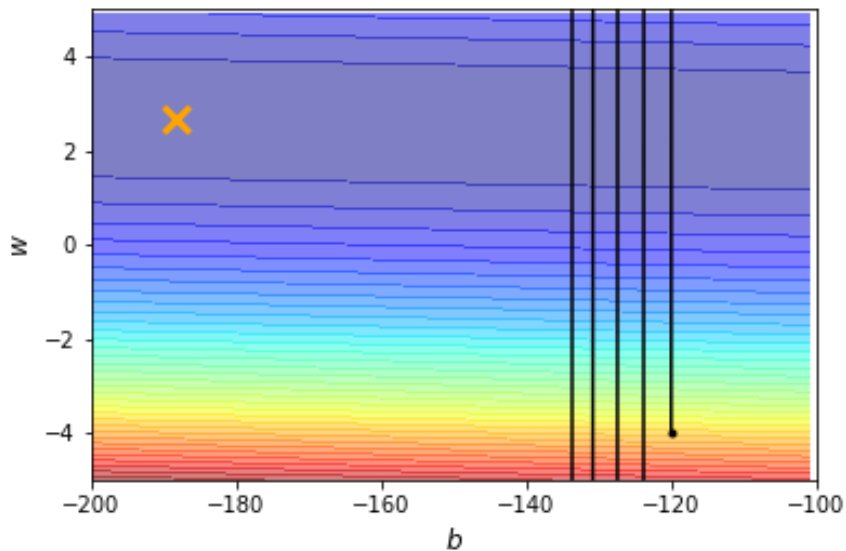
$\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)

$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$ (Update parameters)

end while

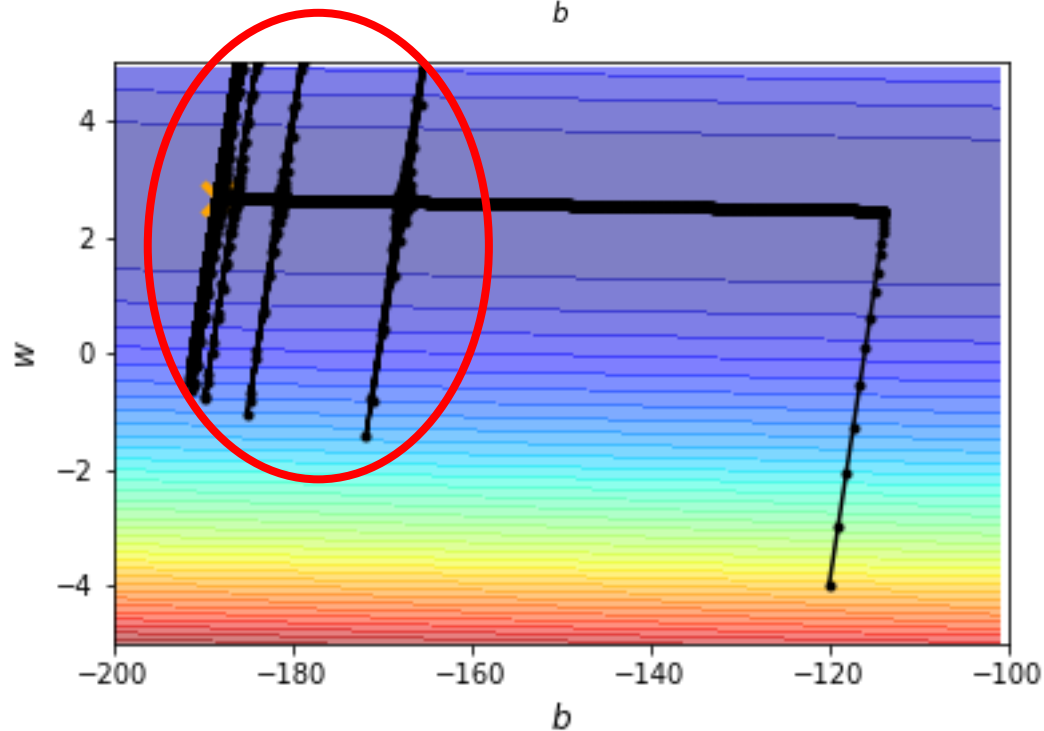
return θ_t (Resulting parameters)

Without Adaptive Learning Rate



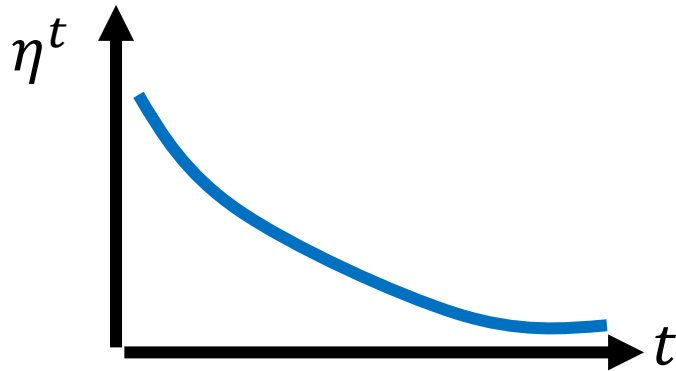
$$\theta_i^{t+1} \leftarrow \theta_i^t - \frac{\eta}{\sigma_i^t} g_i^t$$

$$\sigma_i^t = \sqrt{\frac{1}{t+1} \sum_{i=0}^t (g_i^t)^2}$$



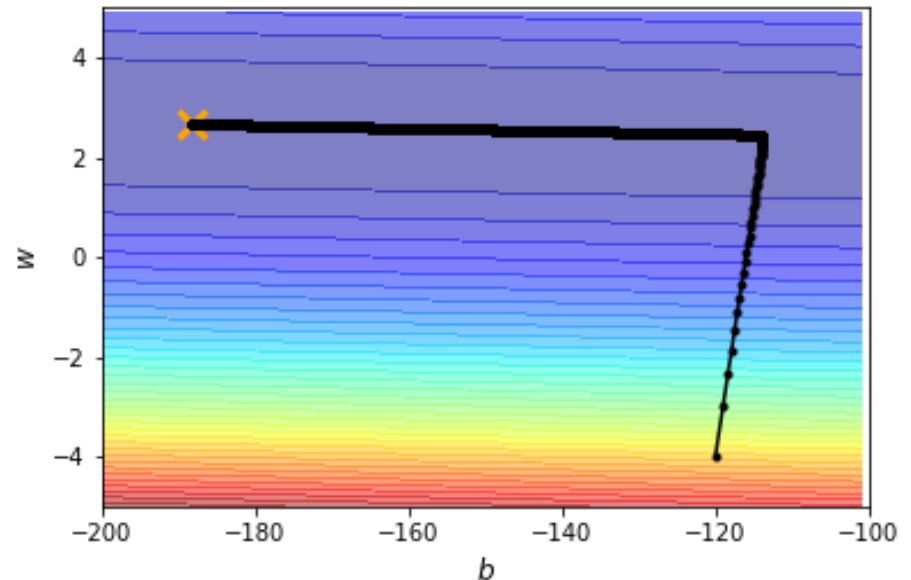
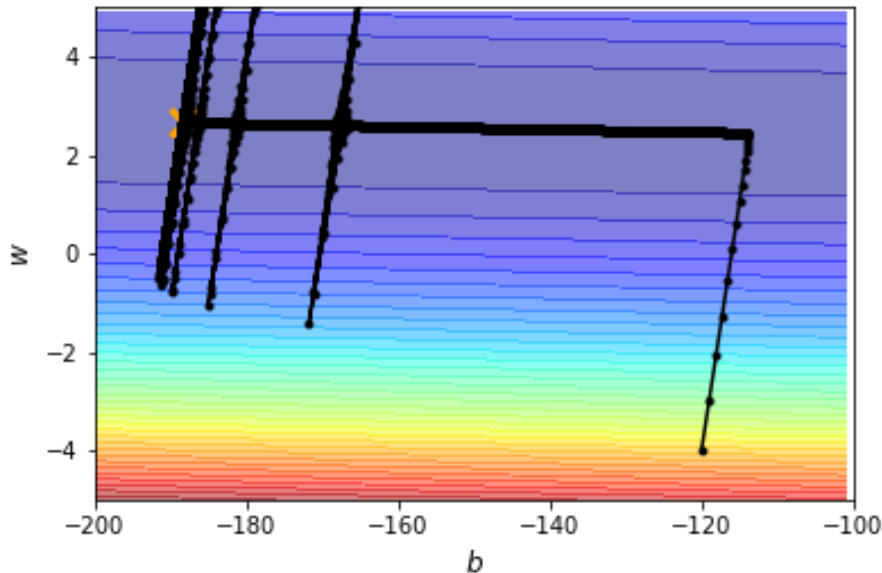
Learning Rate Scheduling

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



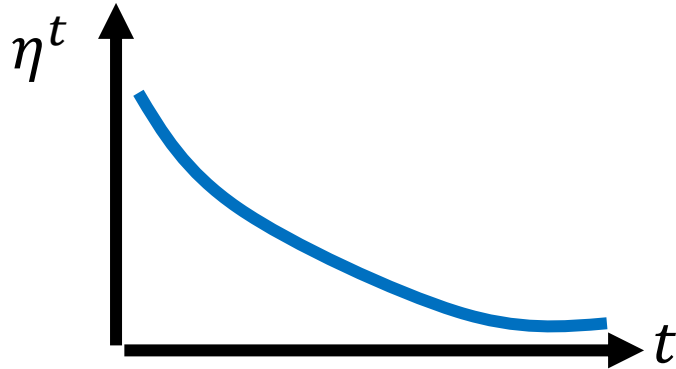
Learning Rate Decay

After the training goes, we are close to the destination, so we reduce the learning rate.



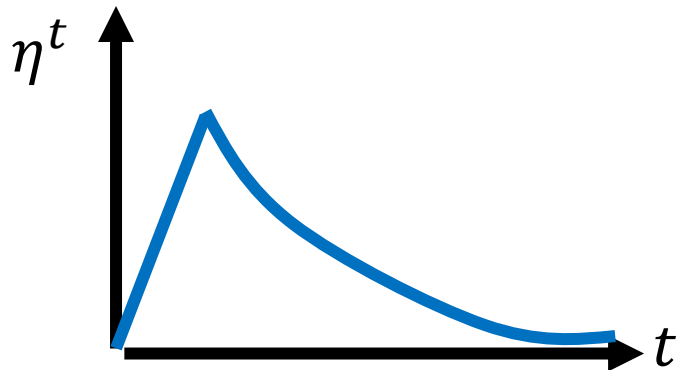
Learning Rate Scheduling

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



Learning Rate Decay

After the training goes, we are close to the destination, so we reduce the learning rate.



Warm Up

Increase and then decrease?

We further explore $n = 18$ that leads to a 110-layer ResNet. In this case, we find that the initial learning rate of 0.1 is slightly too large to start converging⁵. So we use 0.01 to warm up the training until the training error is below 80% (about 400 iterations), and then go back to 0.1 and continue training. The rest of the learning schedule is as done previously. This 110-layer network converges well (Fig. 6, middle). It has *fewer* parameters than other deep and thin

⁵With an initial learning rate of 0.1, it starts converging (<90% error) after several epochs, but still reaches similar accuracy.

5.3 Optimizer

We used the Adam optimizer [17] with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. We varied the learning rate over the course of training, according to the formula:

$$lrate = d_{\text{model}}^{-0.5} \cdot \min(\text{step_num}^{-0.5}, \text{step_num} \cdot \text{warmup_steps}^{-1.5}) \quad (3)$$

This corresponds to increasing the learning rate linearly for the first *warmup_steps* training steps, and decreasing it thereafter proportionally to the inverse square root of the step number. We used *warmup_steps* = 4000.

Residual Network

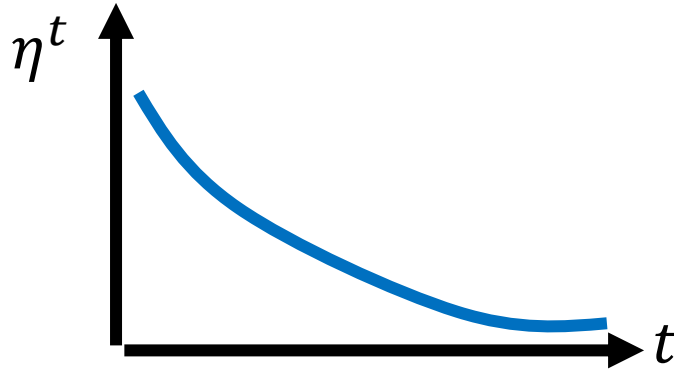
<https://arxiv.org/abs/1512.03385>

Transformer

<https://arxiv.org/abs/1706.03762>

Learning Rate Scheduling

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



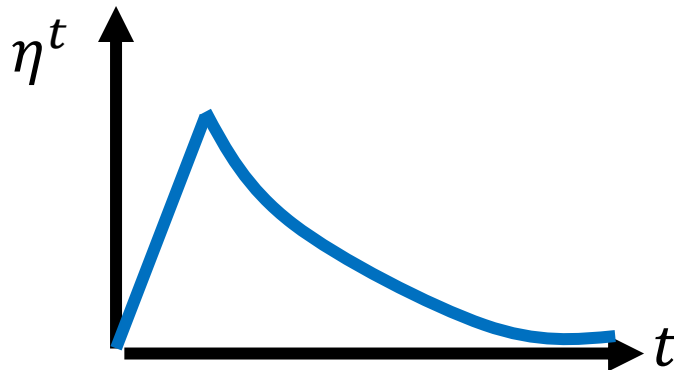
Learning Rate Decay

After the training goes, we are close to the destination, so we reduce the learning rate.

為什麼需要warm up至今依然是個謎

Warm Up

不過有個猜想是認為可以幫助sigma的估計
Increase and then decrease?



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

Please refer to **RAdam**

<https://arxiv.org/abs/1908.03265>

Summary of Optimization

(Vanilla) Gradient Descent

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

Various Improvements

movement與gradient成正相關

sigma與gradient負相關

為什麼？這樣不是會有抵銷作用嗎？

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

Learning rate scheduling

Momentum: weighted sum of the previous gradients

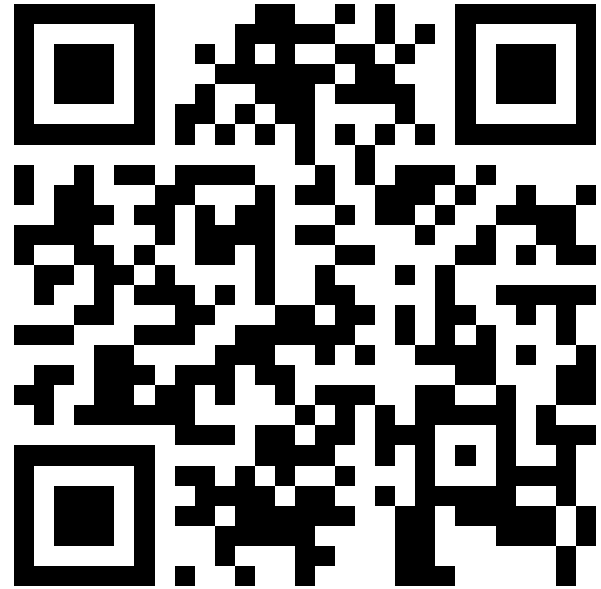
root mean square of the gradients

To Learn More



<https://youtu.be/4pUmZ8hXlHM>

(in Mandarin)

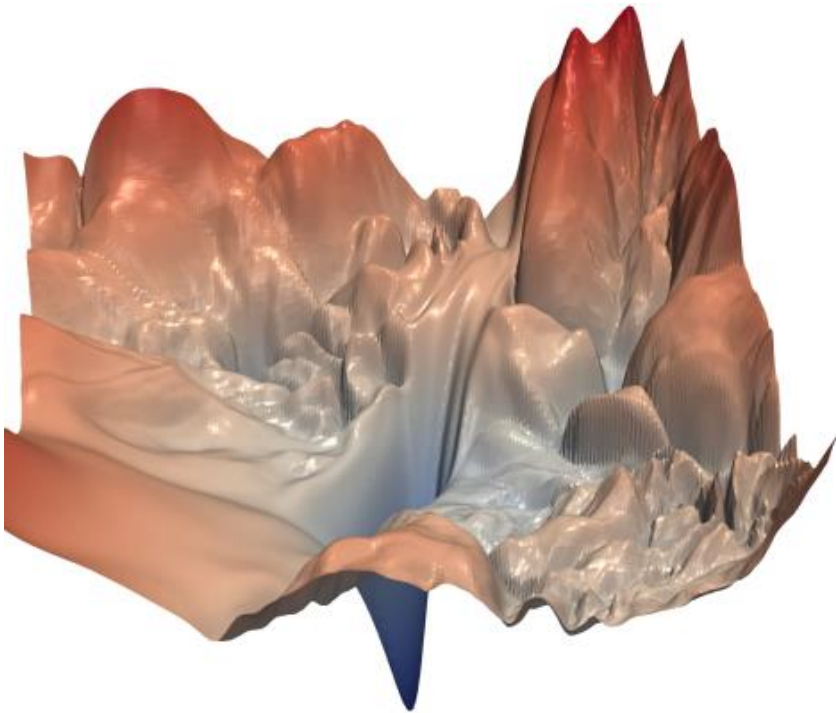


<https://youtu.be/e03YKGHXnL8>

(in Mandarin)

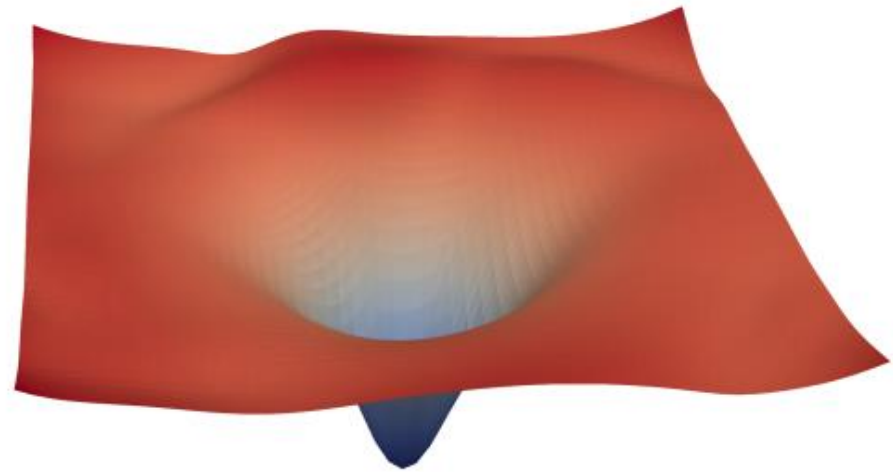
Next Time

Source of image: <https://arxiv.org/abs/1712.09913>



Better optimization strategies:
If the mountain won't move,
build a road around it.

Next time



Can we change the error
surface?
Directly move the mountain!