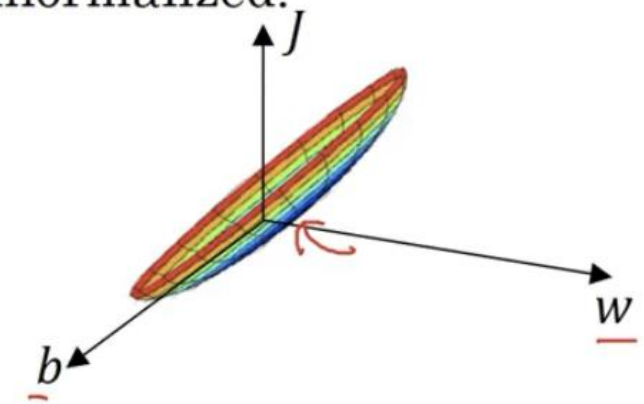
Why normalize inputs?

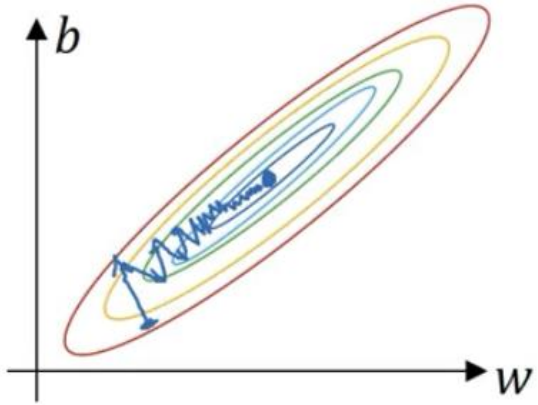
$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right)$$

w_1 x_1: 1....1000

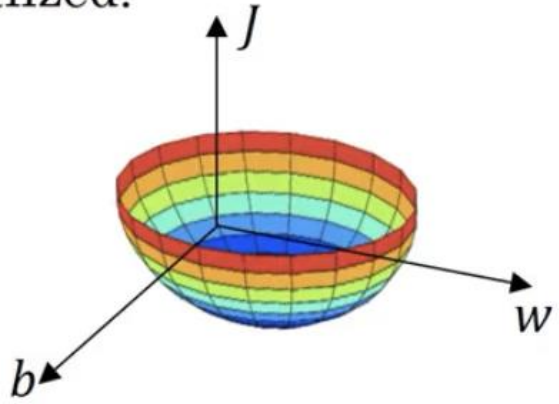
w_2 x_2: 0...1

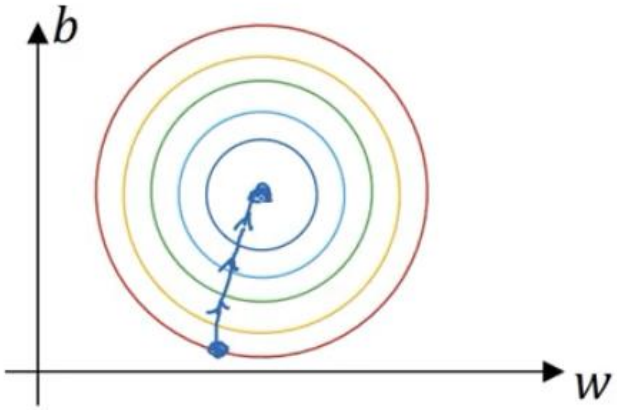
Unnormalized:





Normalized:





Andrew Ng

Hyperparameters **To Tune**

α

β 0.9

$\beta_1, \beta_2, \varepsilon$

layers

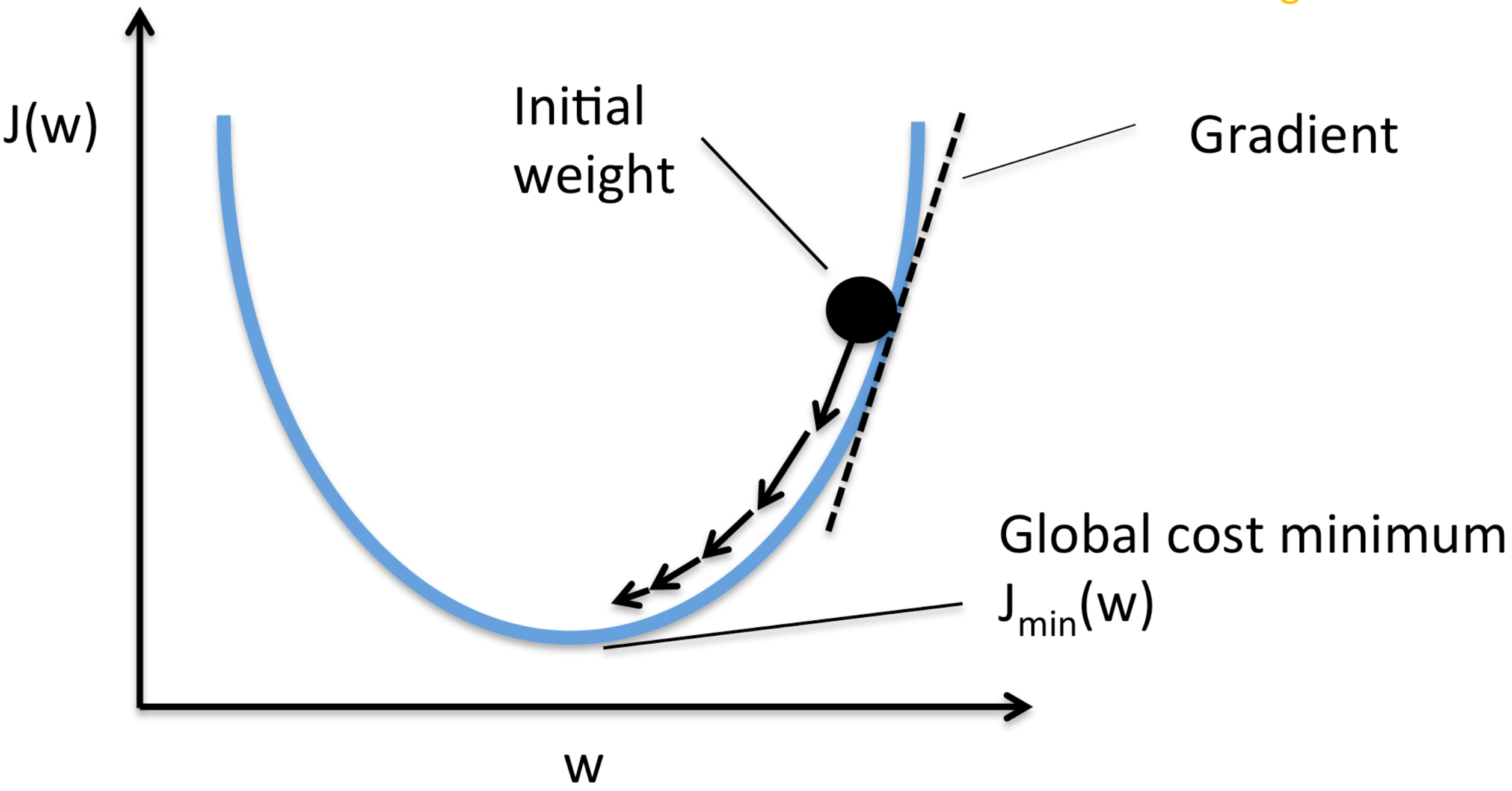
hidden units

learning rate decay

mini-batch size

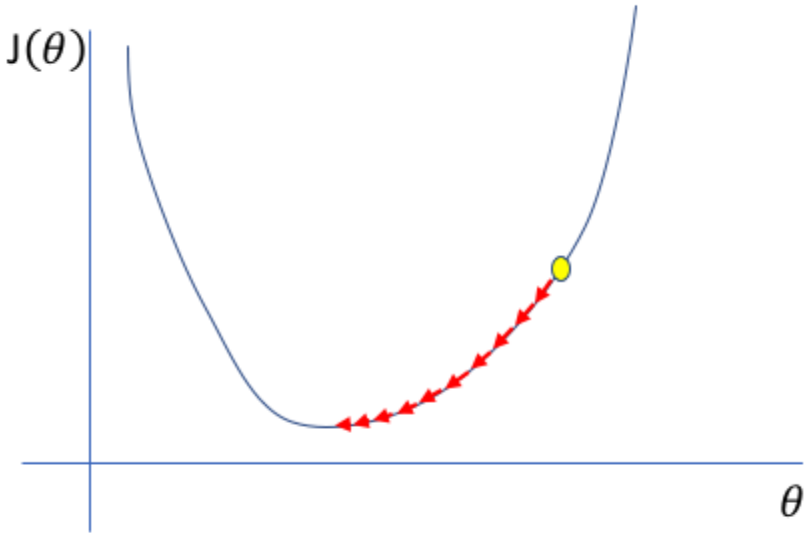
Andrew Ng

Learning Rate



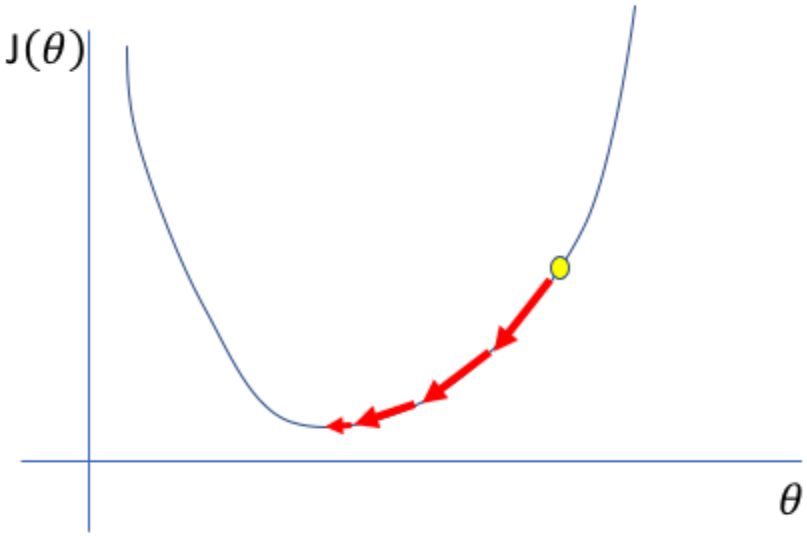
Learning Rate

Too low



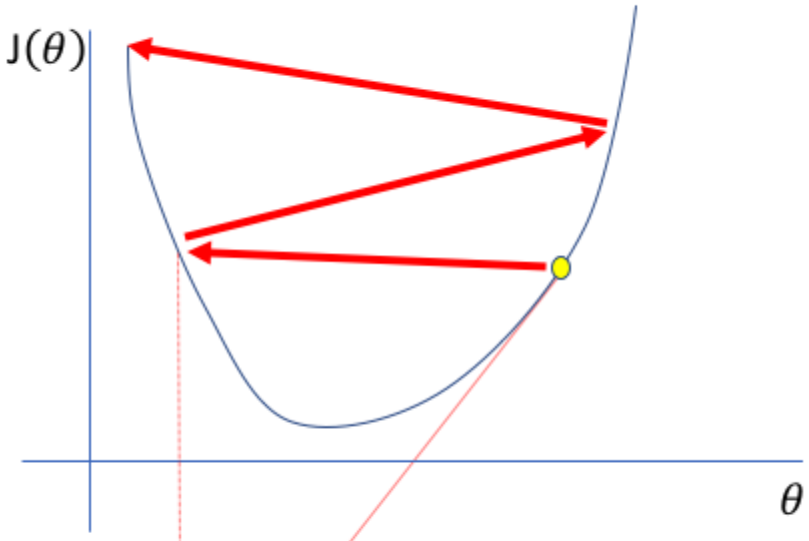
A small learning rate requires many updates before reaching the minimum point

Just right



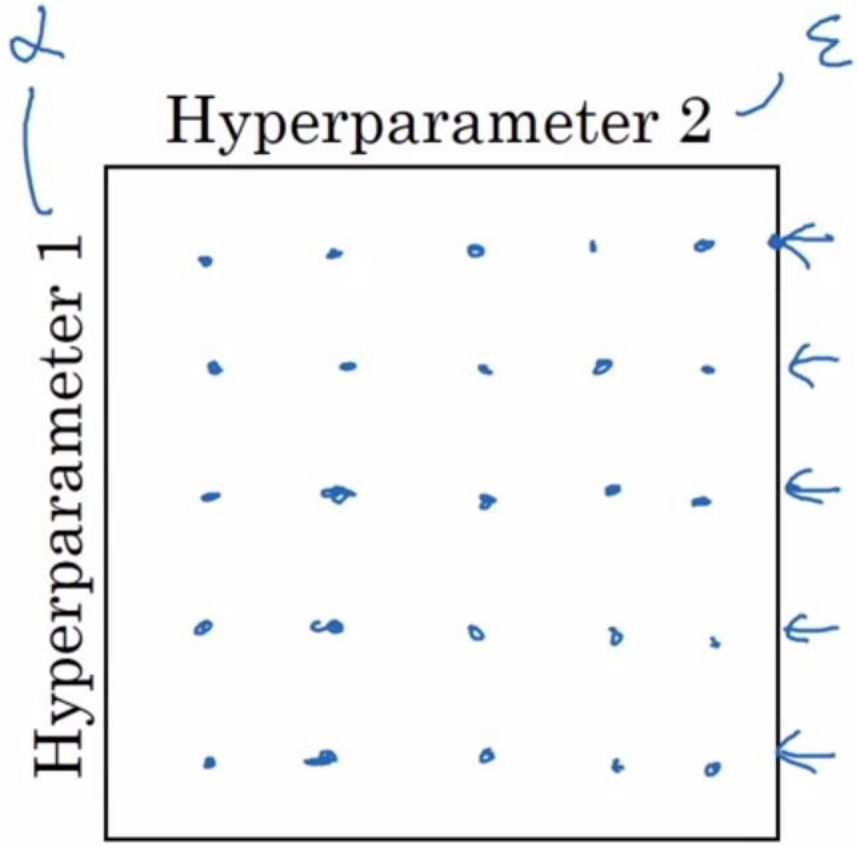
The optimal learning rate swiftly reaches the minimum point

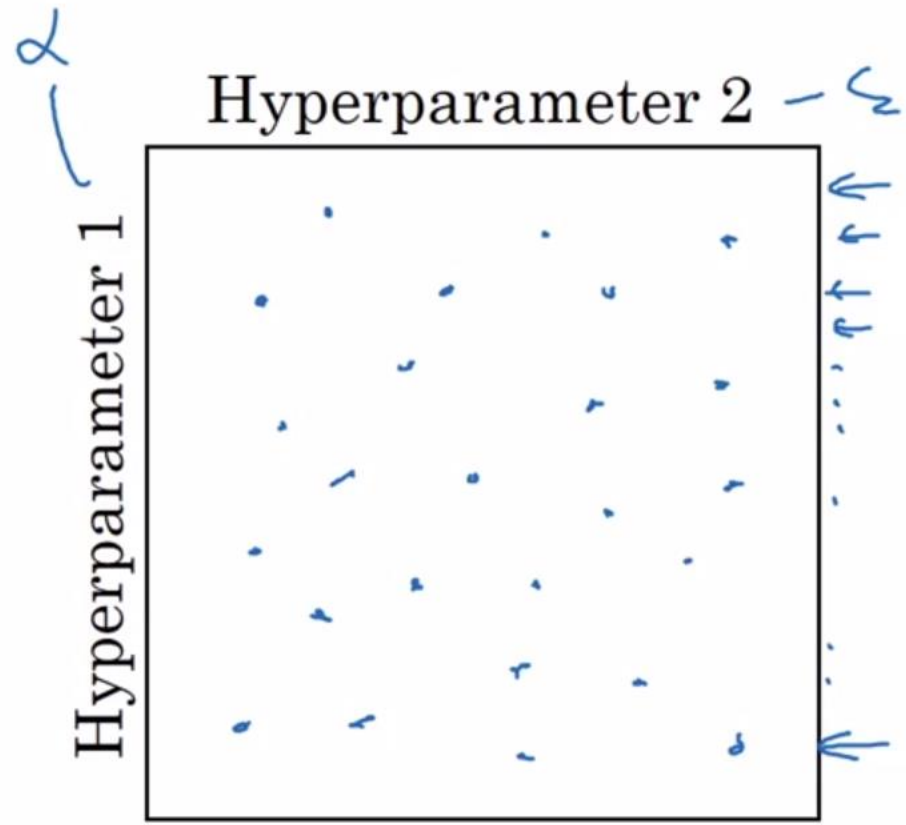
Too high



Too large of a learning rate causes drastic updates which lead to divergent behaviors

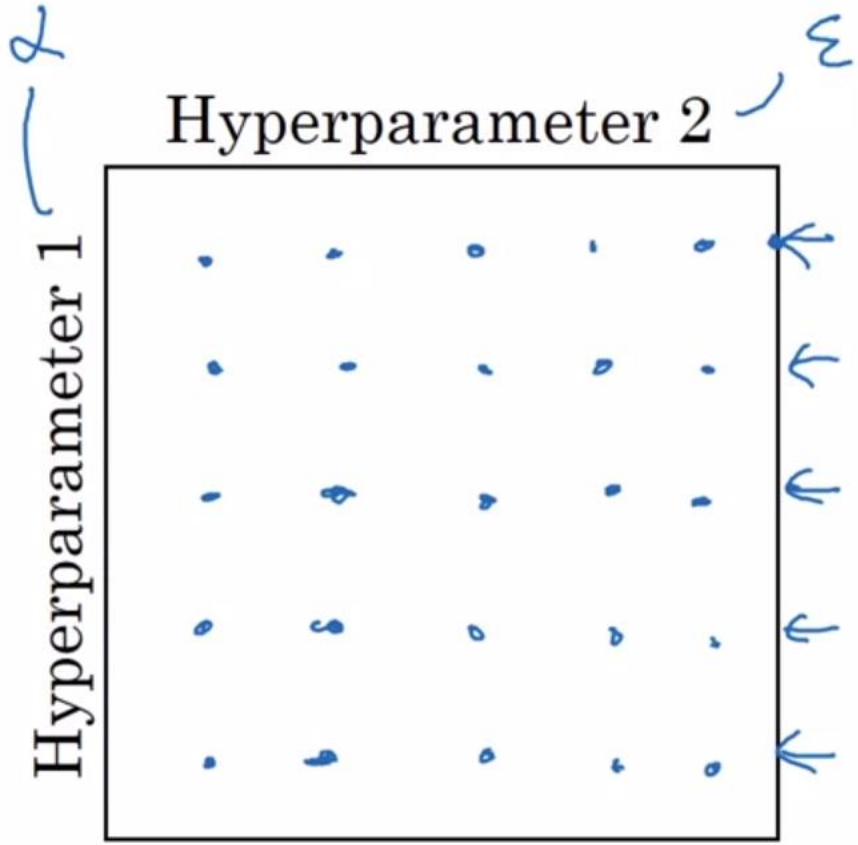
Try random values: Don't use a grid

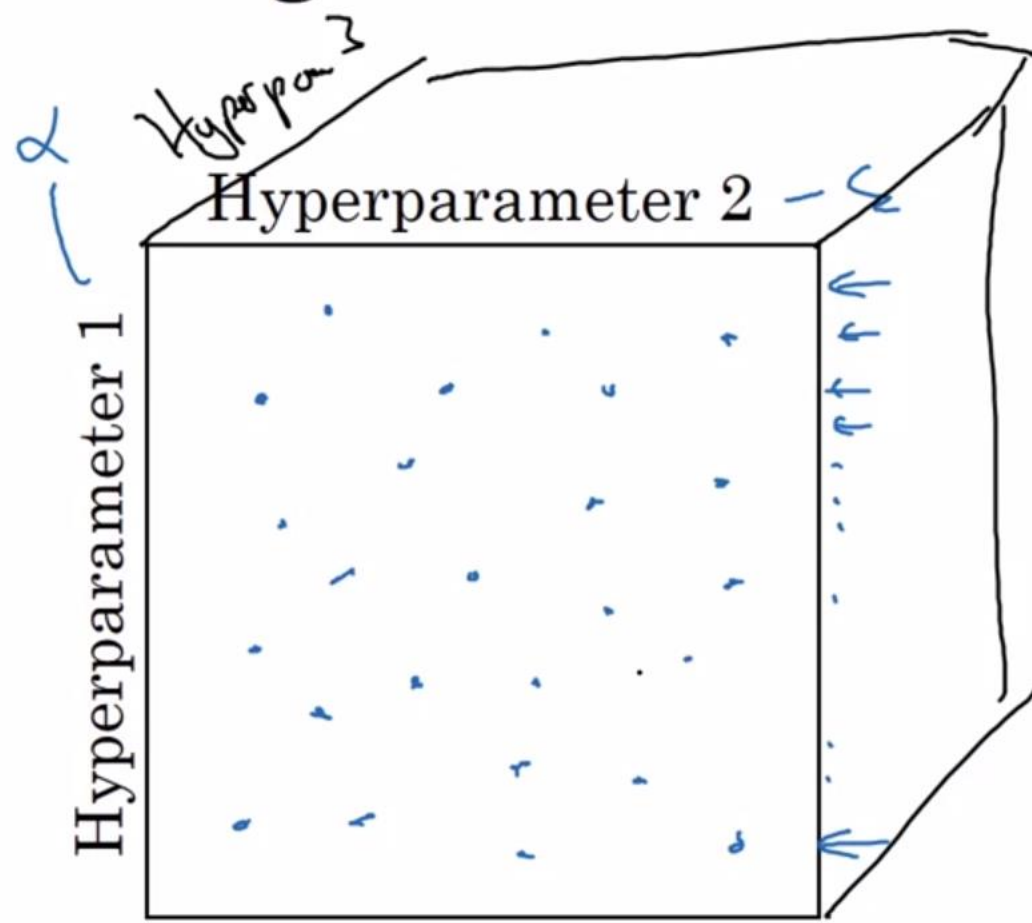




Andrew Ng

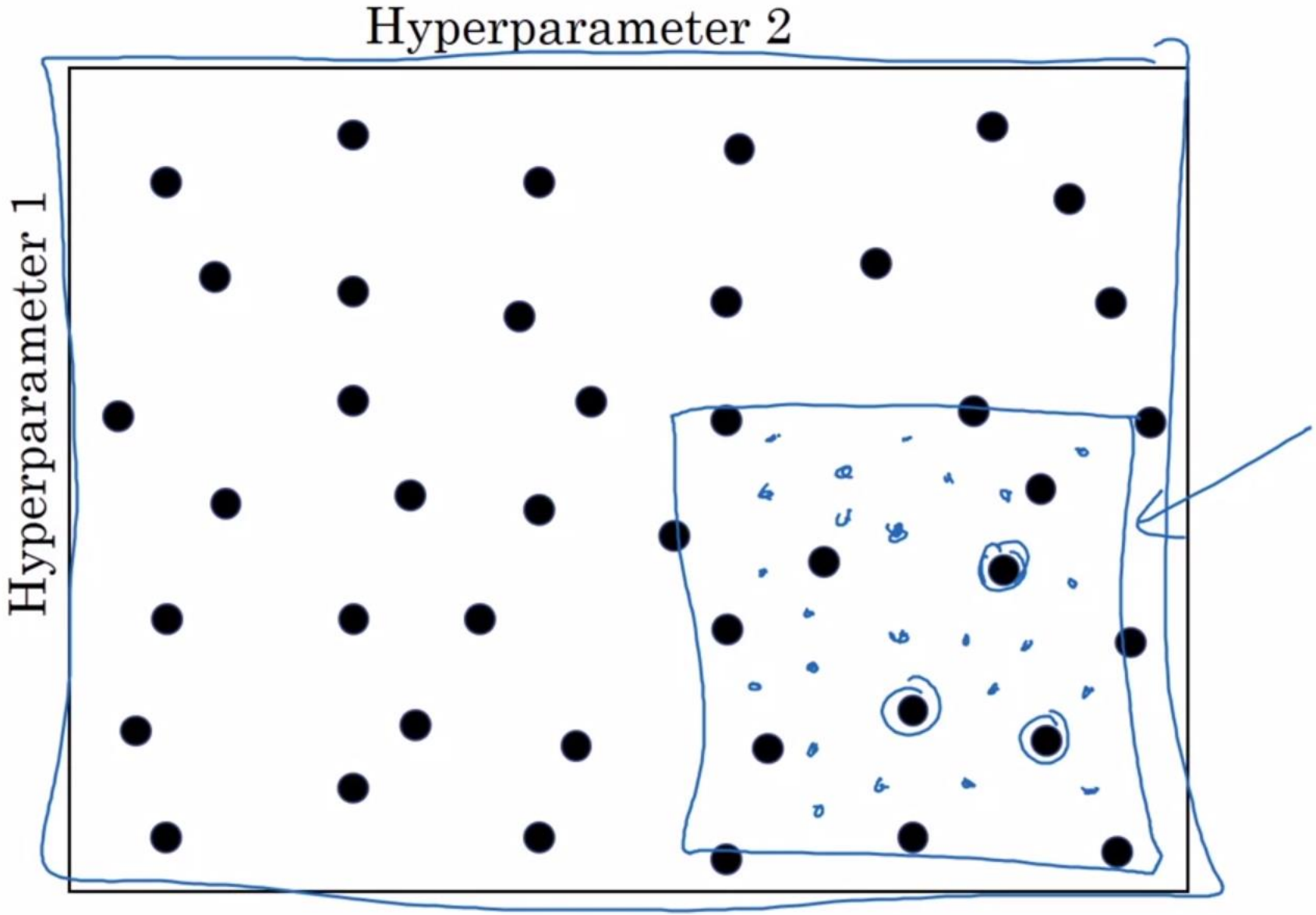
Try random values: Don't use a grid





Andrew Ng

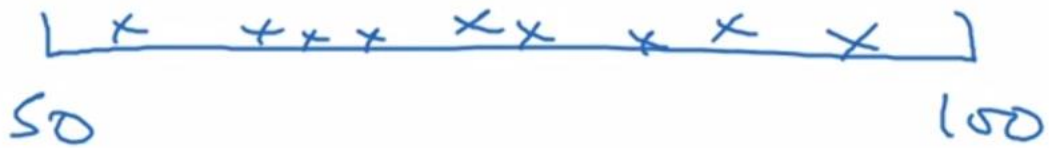
Coarse to fine



Andrew Ng

Picking hyperparameters at random

→ $n^{[l]} = 50, \ldots, 100$



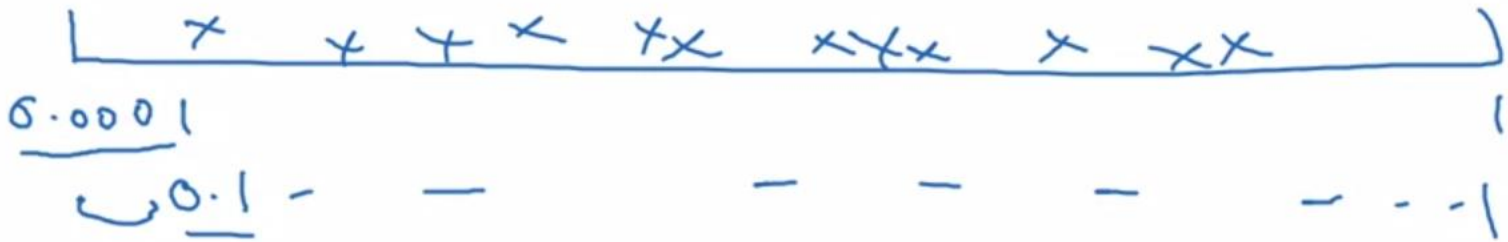
→ #layers L: 2 – 4

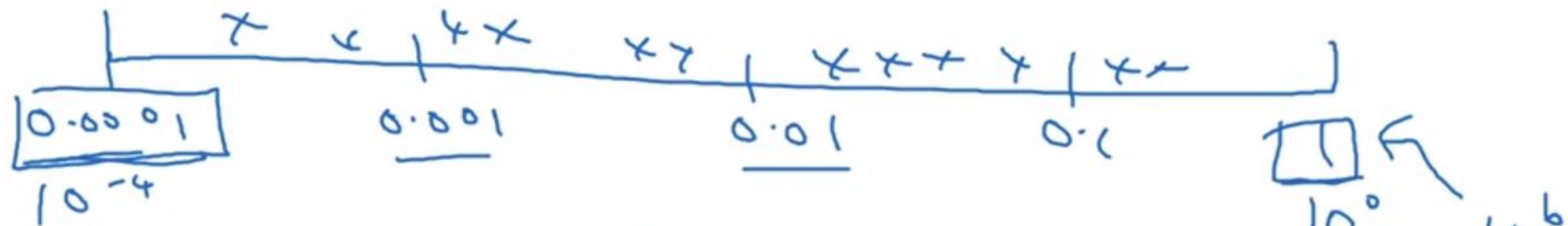
2, 3, 4

Andrew Ng

Appropriate scale for hyperparameters

$\alpha = 0.0001, \ldots, 1$





10^a

$a = \log_{10} 0.0001 = -4$

$r = -4 * np.random.rand() \leftarrow r \in [-4, 0]$

$\alpha = 10^r \leftarrow 10^{-4} \ldots 10^0$

10^b

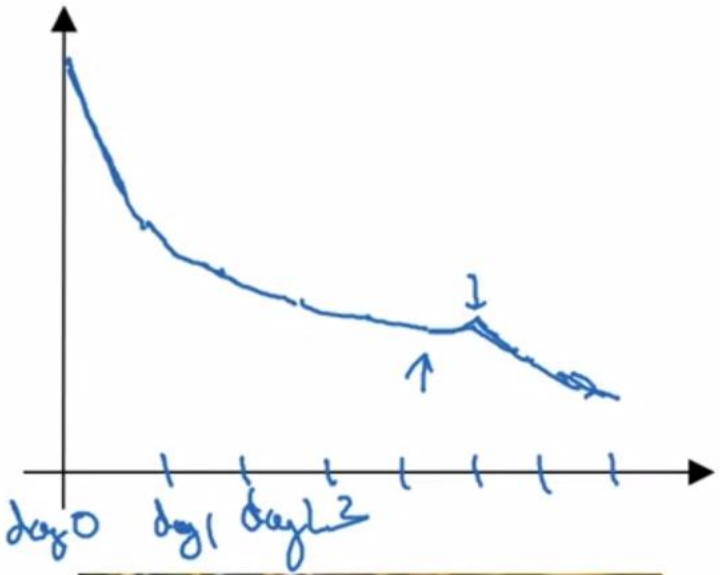
$b = \log_{10} 1 = 0$

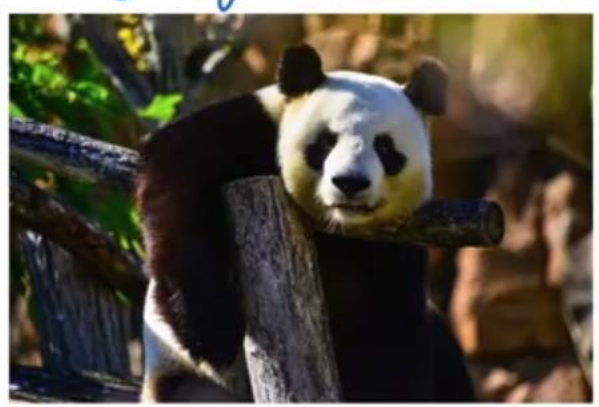
$10^a \ldots 10^b \qquad r \in [a, b] \qquad \alpha = 10^r$

$[-4, 0]$

Andrew Ng

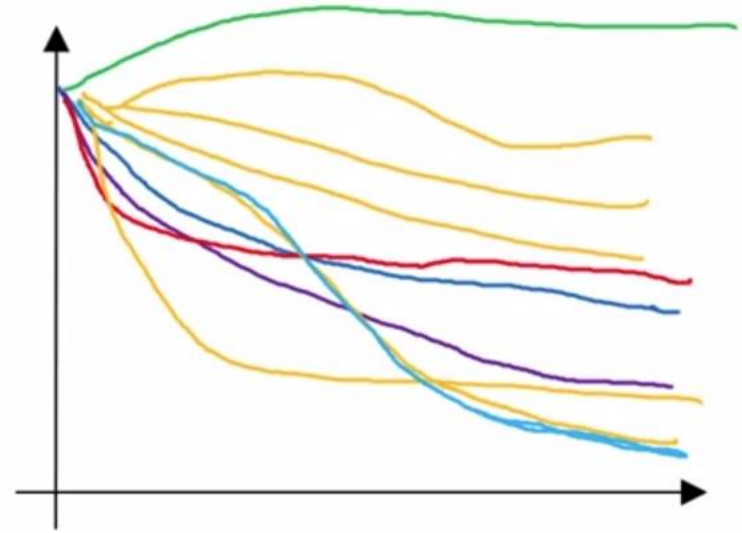
Babysitting one model





Panda

Training many models in parallel



Caviar

Andrew Ng

Adam optimization algorithm

$V_{dW} = 0$, $S_{dW} = 0$. $V_{db} = 0$, $S_{db} = 0$

On iteration t:

Compute dW, db using current mini-batch

$V_{dW} = \beta_1 V_{dW} + (1-\beta_1) dW$, $V_{db} = \beta_1 V_{db} + (1-\beta_1) db$ ← "momentum" β_1

$S_{dW} = \beta_2 S_{dW} + (1-\beta_2) dW^2$, $S_{db} = \beta_2 S_{db} + (1-\beta_2) db$ ← "RMSprop" β_2

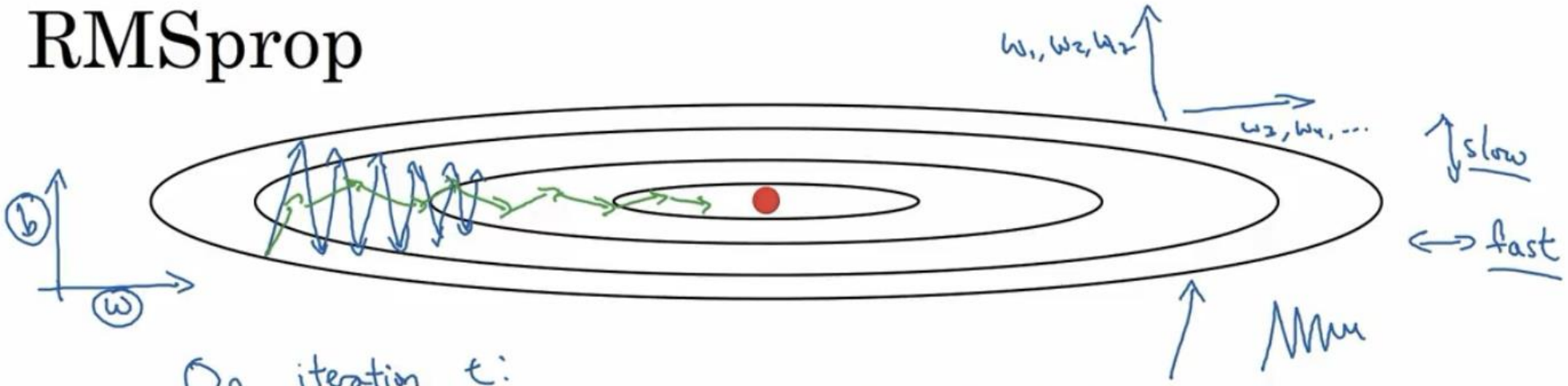
$V_{dW}^{corrected} = V_{dW} / (1-\beta_1^t)$, $V_{db}^{corrected} = V_{db} / (1-\beta_1^t)$

$S_{dW}^{corrected} = S_{dW} / (1-\beta_2^t)$, $S_{db}^{corrected} = S_{db} / (1-\beta_2^t)$

$$W := W - \alpha \frac{V_{dW}^{corrected}}{\sqrt{S_{dW}^{corrected}} + \varepsilon} \qquad b := b - \alpha \frac{V_{db}^{corrected}}{\sqrt{S_{db}^{corrected}} + \varepsilon}$$

Andrew Ng

RMSprop



On iteration t:

Compute dW, db on current mini-batch

element-wise

$$S_{dW} = \beta_2 S_{dW} + (1-\beta_2) dW^2 \leftarrow \text{small}$$

$$\rightarrow S_{db} = \beta_2 S_{db} + (1-\beta_2) db^2 \leftarrow \text{large}$$

$$W := W - \alpha \frac{dW}{\sqrt{S_{dW}}} \qquad b := b - \alpha \frac{db}{\sqrt{S_{db}}}$$

Andrew Ng

Implementation details

On iteration t:

Compute dW, db on the current mini-batch

$$v_{dW} = \beta v_{dW} + (1-\beta)dW$$

$$v_{db} = \beta v_{db} + (1-\beta)db$$

$$W = W - \alpha v_{dW}, \quad b = b - \alpha v_{db}$$

Hyperparameters: α, β $\qquad$ $\beta = 0.9$

Andrew Ng