Computational Intelligence

Samaneh Hosseini

Isfahan University of Technology

Outline

- Optimization Algorithms
  - Understanding Exponentially Weighted Averages
  - Bias Correction in Exponentially Weighted Averages
  - Gradient Descent with Momentum
  - RMSprop
  - Adam Optimization Algorithm
  - Learning Rate Decay
  - The Problem of Local Optima
- Hyperparameter Tuning
  - Tuning Process
  - Using an Appropriate Scale to pick Hyperparameters
  - Hyperparameters Tuning in Practice

Optimization Algorithms:
Adam Optimization Algorithm

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

$v_{dw} = \beta_1 v_{dw} + (1-\beta_1) dw$ $\quad v_{db} = \beta_1 v_{db} + (1-\beta_1) db$ ← momentum β_1

$S_{dw} = \beta_2 S_{dw} + (1-\beta_2) dw^2$ $\quad S_{db} = \beta_2 S_{db} + (1-\beta_2) db$ ← RMS prop

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

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

$$w := w - \alpha \frac{v_{dw}^{corrected}}{\sqrt{S_{dw}^{corrected}} + \varepsilon} , \quad b := b - \alpha \frac{v_{db}^{corrected}}{\sqrt{S_{db}^{corrected}} + \varepsilon}$$

Hyperparameters Choice

→ α : learning rate needs to be tune ←

β_1 : 0.9 → (dw)

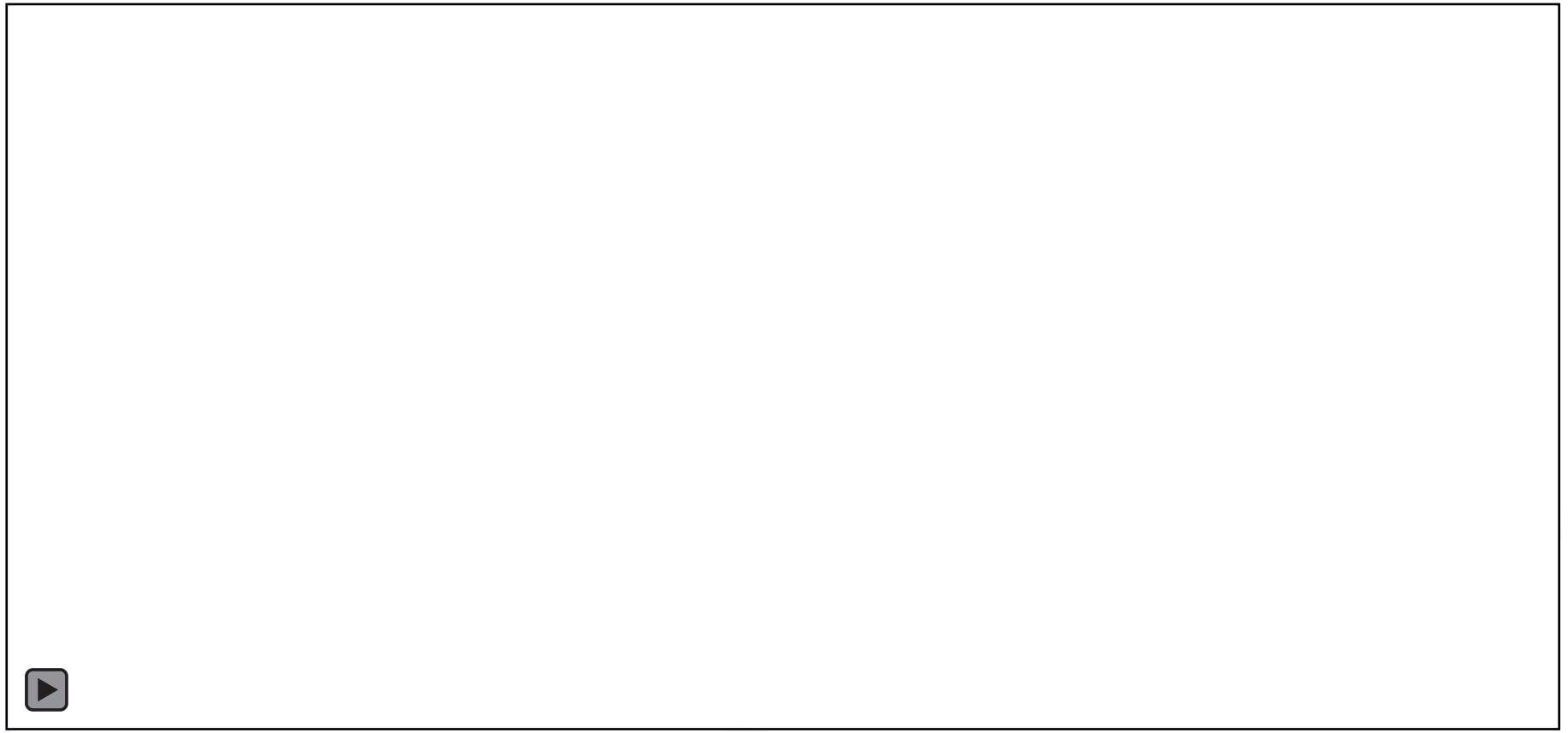
β_2 : 0.999 → (dw^2)

ε : 10^{-8}

Adam : Adaptive momentam estimation

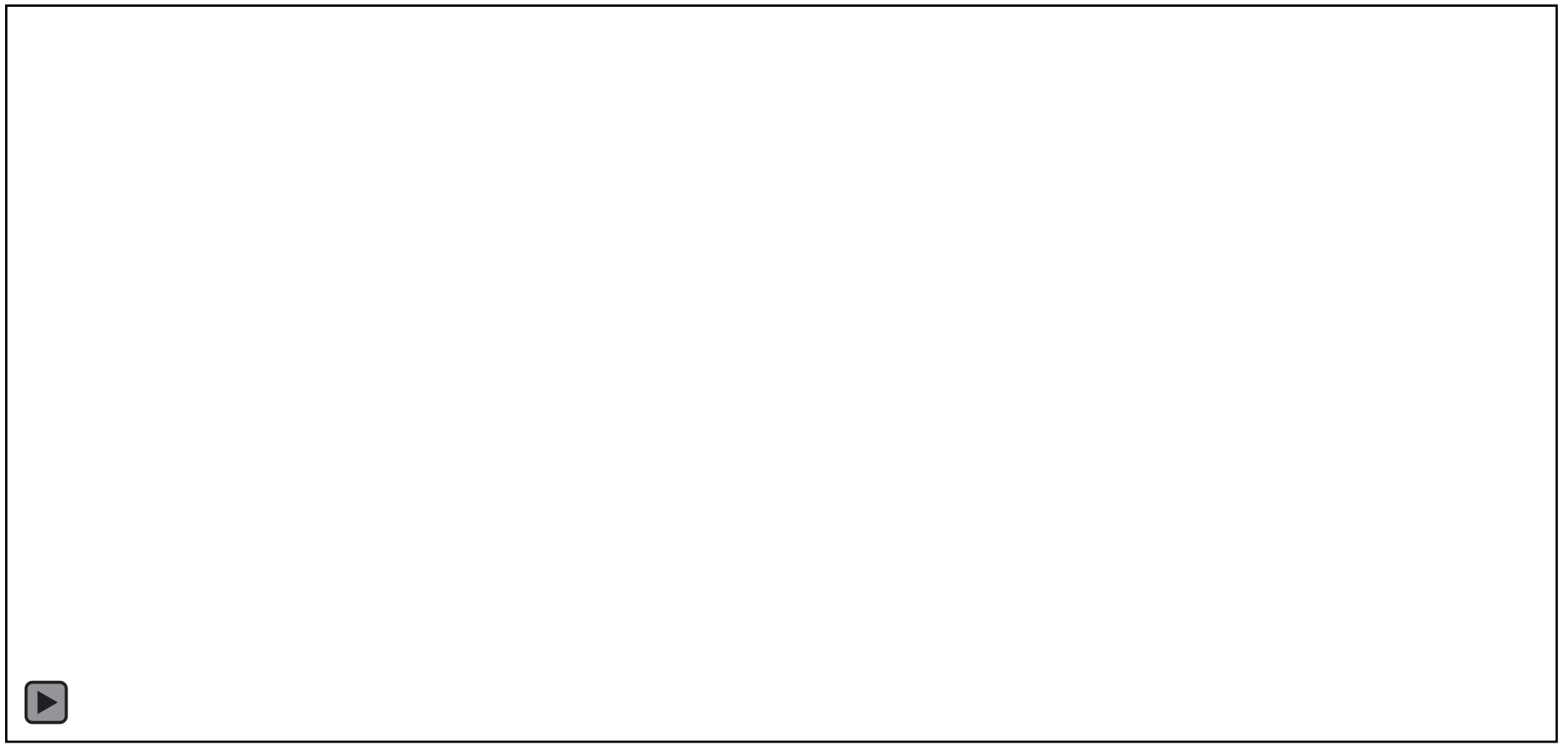
Setting the Learning Rate

Small learning rate *converges slowly and gets stuck in false local minima*



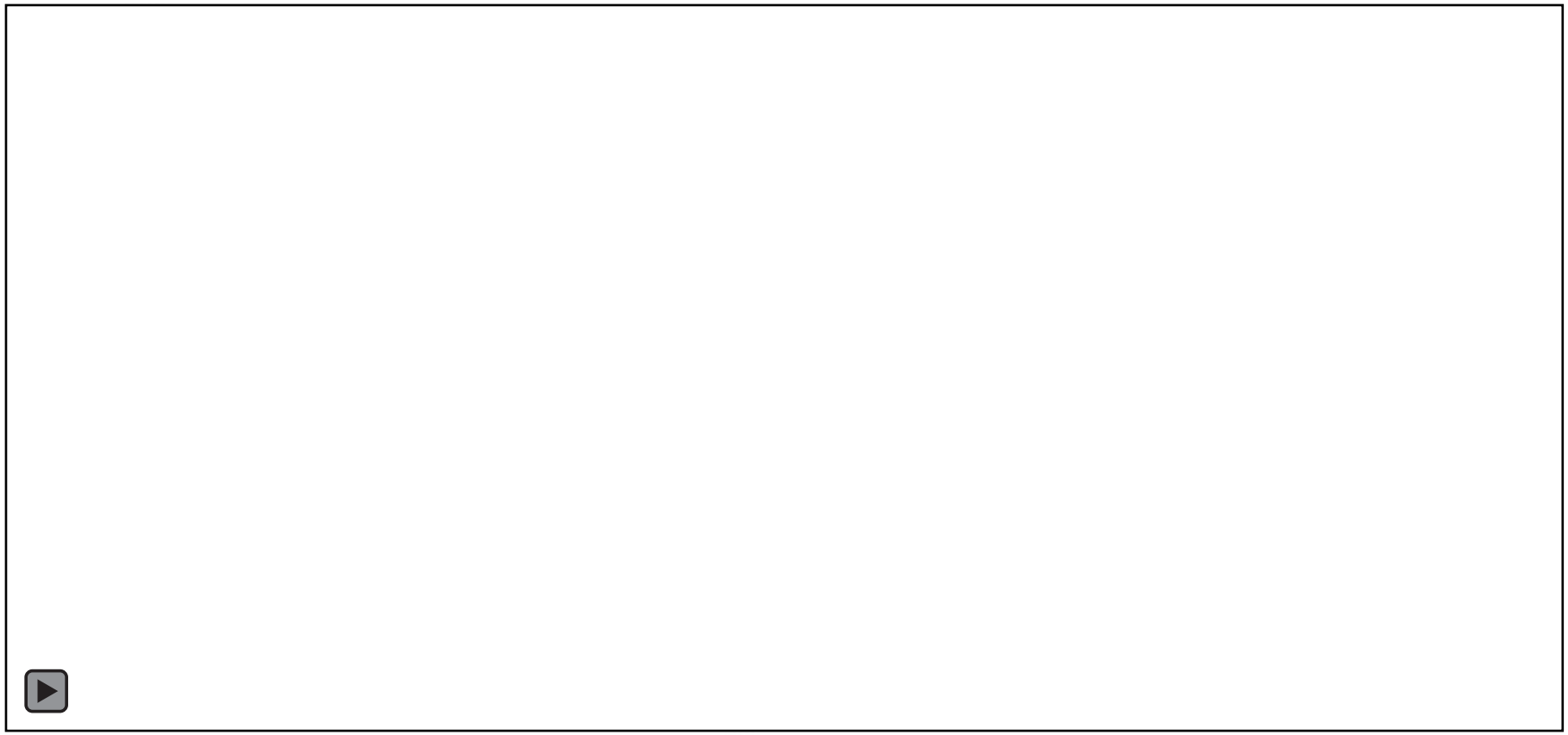
Setting the Learning Rate

Large learning rates *overshoot, become unstable and diverge*



Setting the Learning Rate

Stable learning rates *converge smoothly and avoid local minima*



How to deal with this?

Idea 1:

Try lots of different learning rates and see what works "just right"

Adoptive Learning Rates

Idea 1:

Try lots of different learning rates and see what works "just right"

Idea 2:

Do something smarter!

Design an adaptive learning rate that "adapts" to the landscape

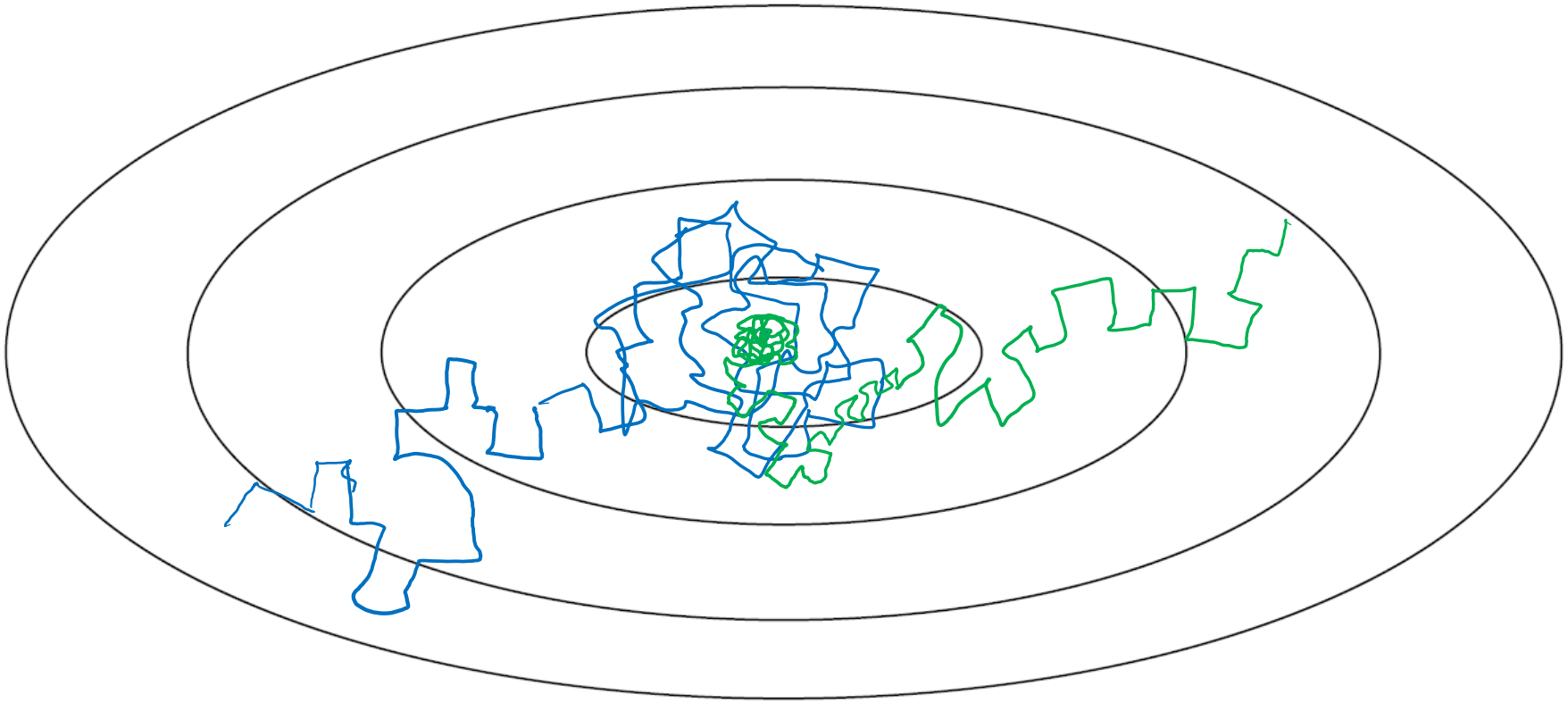
Adoptive Learning Rates

- Learning rates are no longer fixed
- Can be made larger or smaller depending on:
  - how large gradient is
  - how fast learning is happening
  - size of particular weights
  - etc...

Optimization Algorithms: Learning Rate Decay

Learning Rate Decay

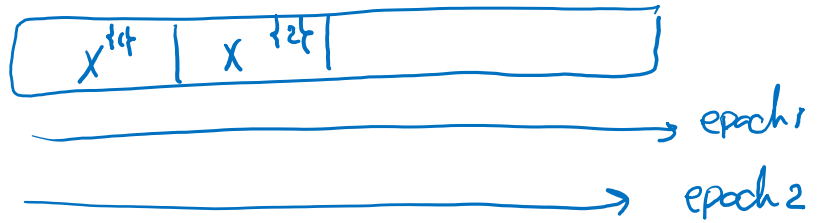
slowly reduce α



Learning Rate Decay

1 epoch : 1 pass through data

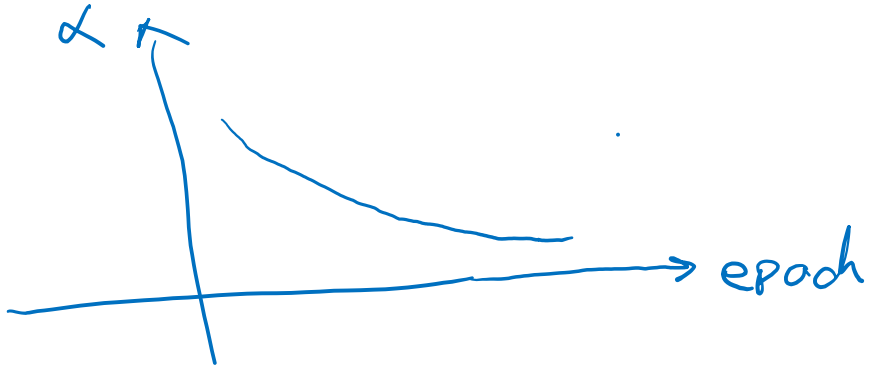
$$\alpha = \frac{1}{1 + \text{decay\_rate} * \text{epoch\_num}} \alpha_0$$



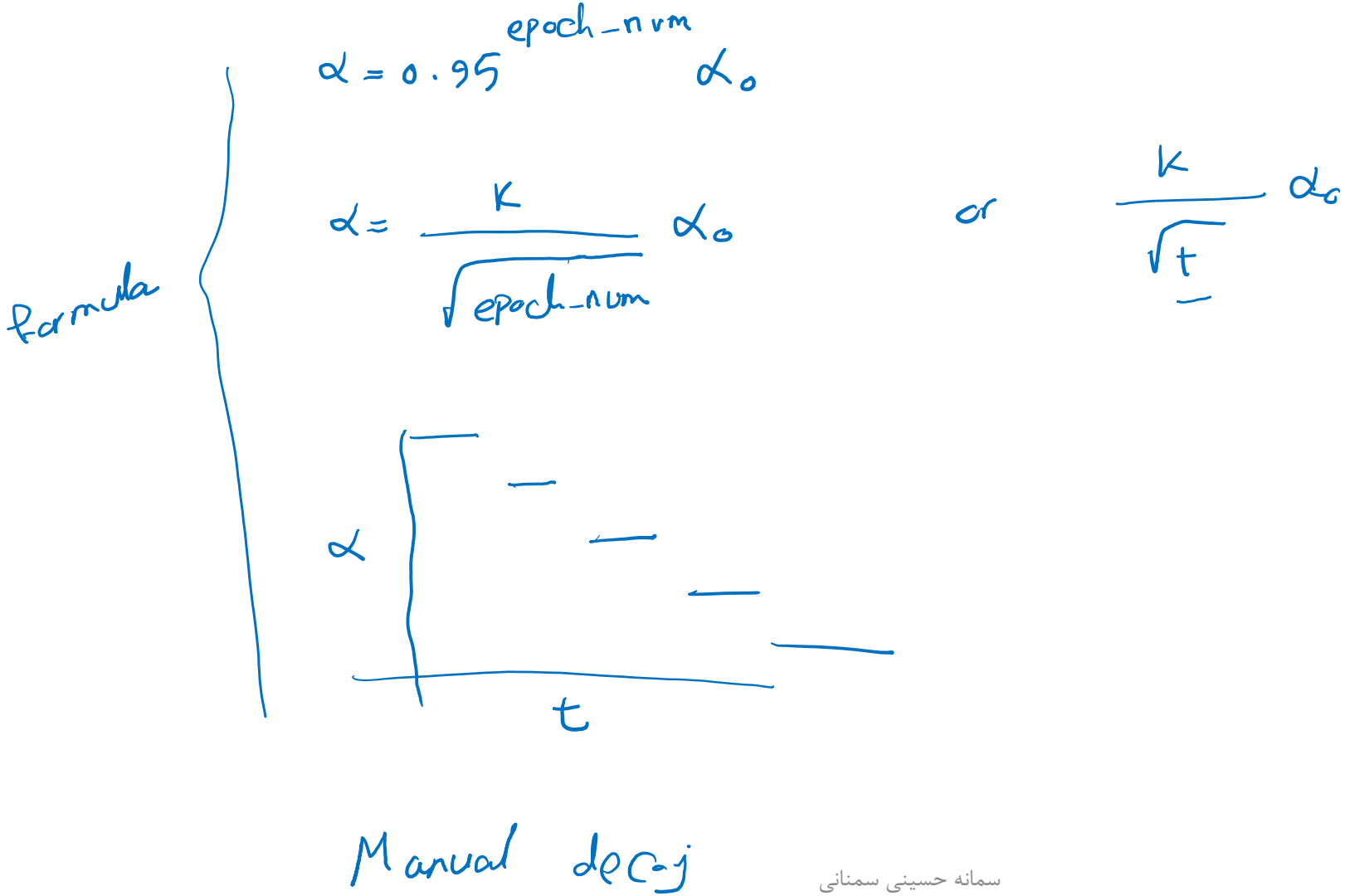
$\alpha_0 = 0.2$

decay_rate = 1

Epoch	α
1	0.1
2	0.067
3	0.05
4	0.04
⋮	⋮

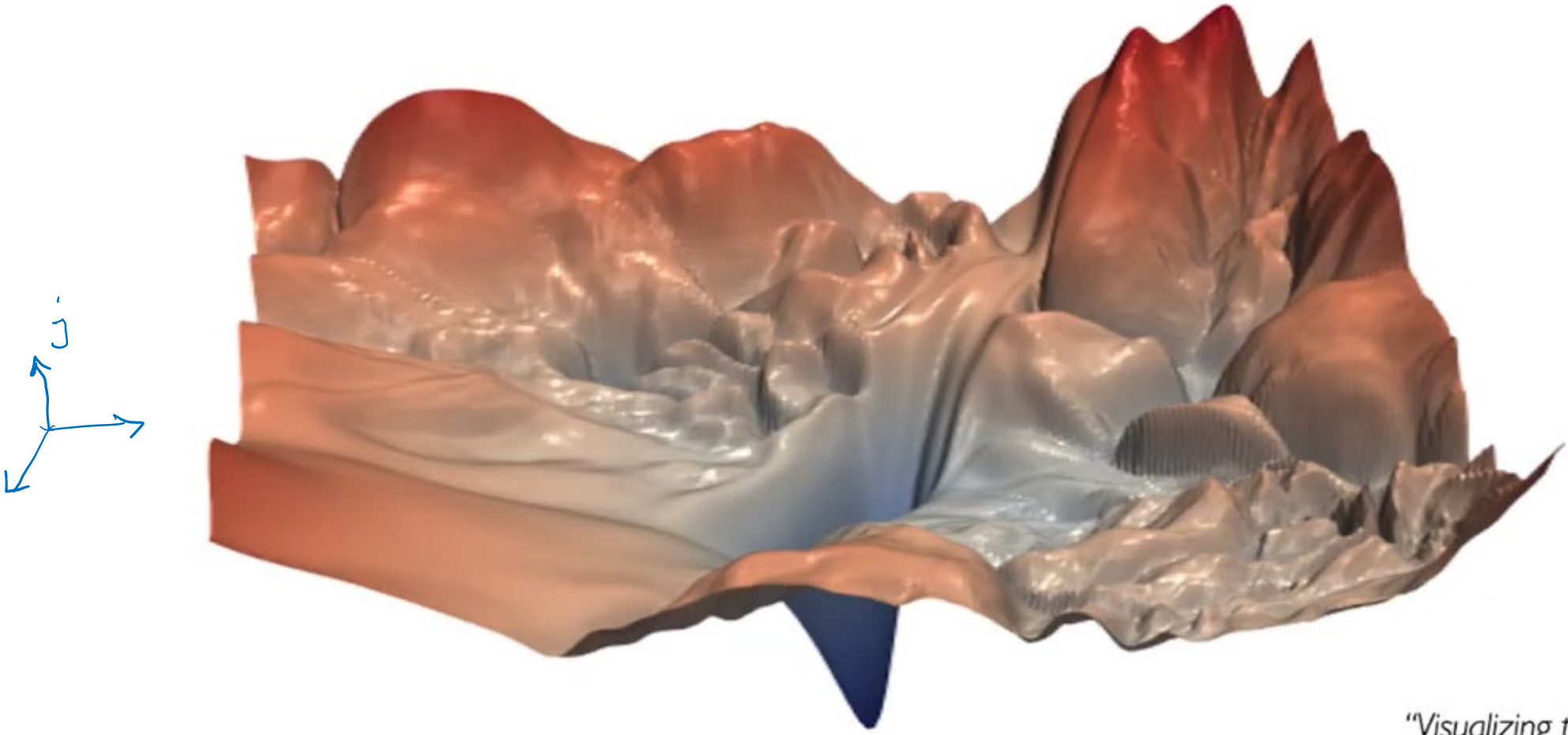


Other Learning Rate Decay Methods



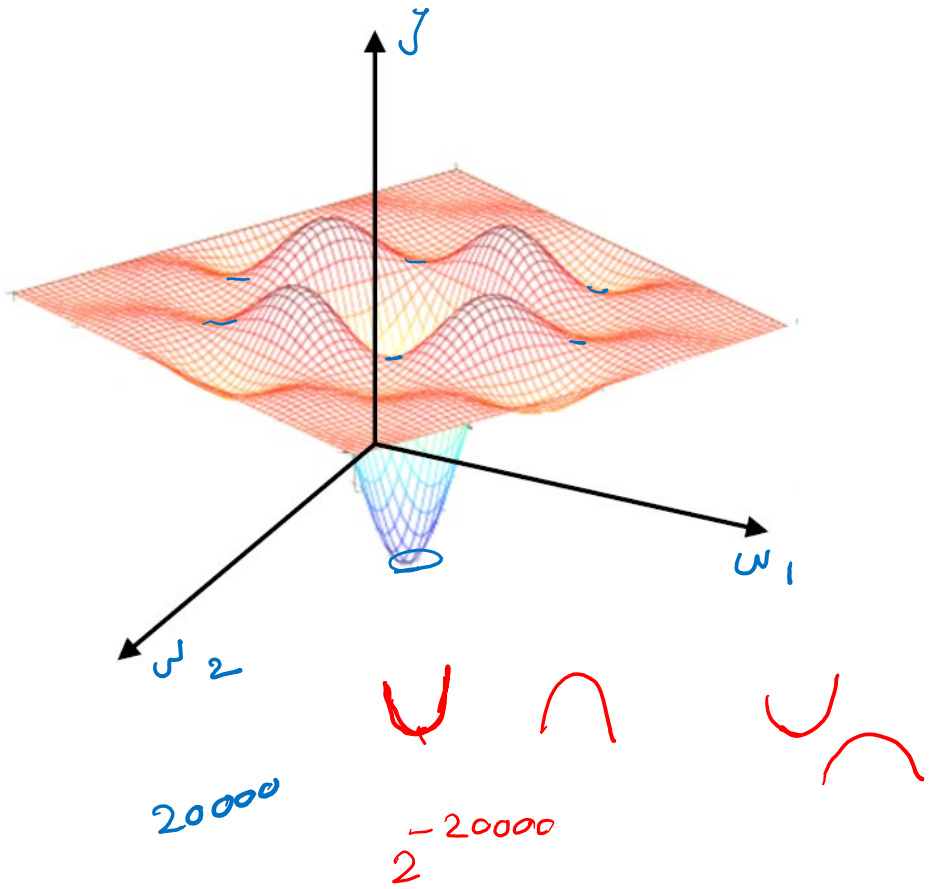
Optimization Algorithms:
The Problem of Local Optima

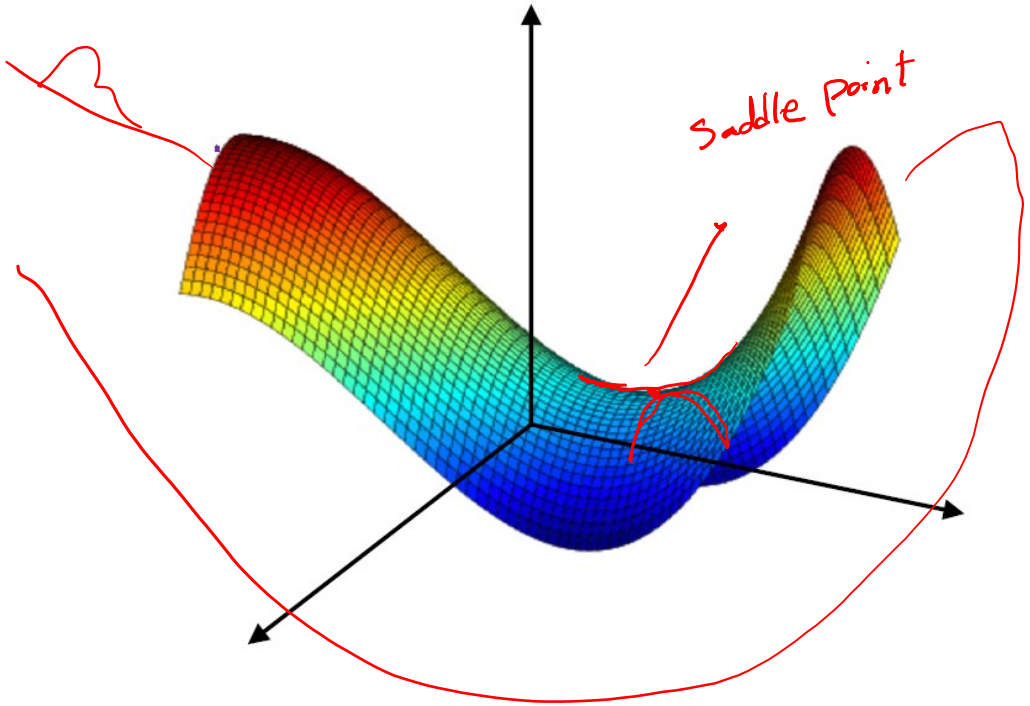
Training Neural Network is Difficult



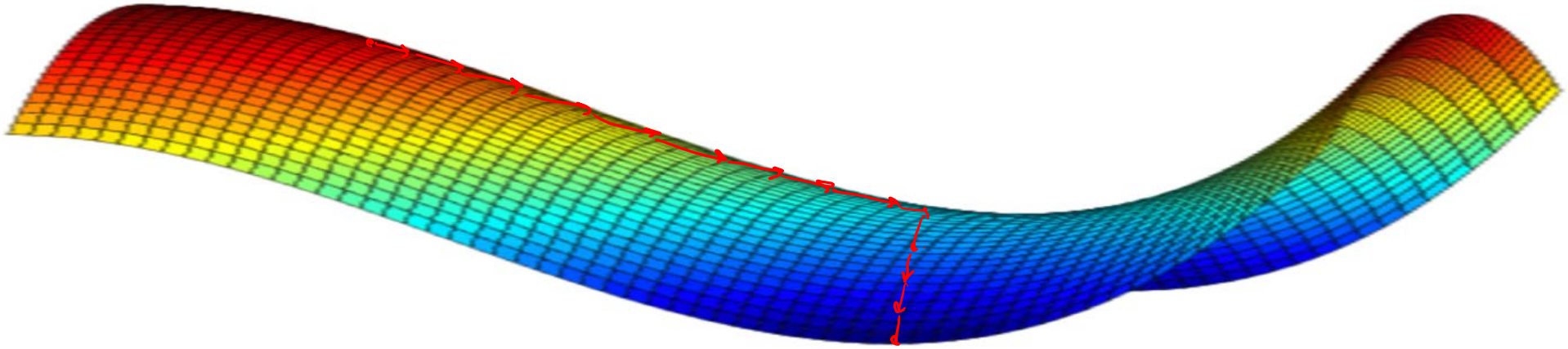
"Visualizing the loss landscape of neural nets". Dec 2017.

Local Optima in Neural Networks





Problem of plateaus

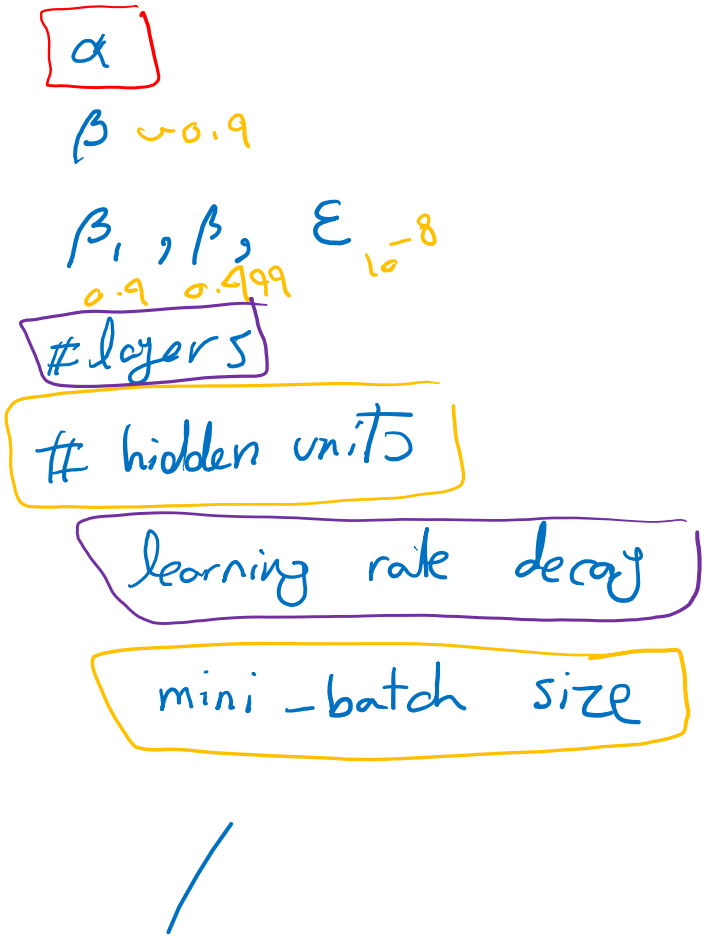


- Unlikely to get stuck in a bad local optima
- Plateaus can make learning slow

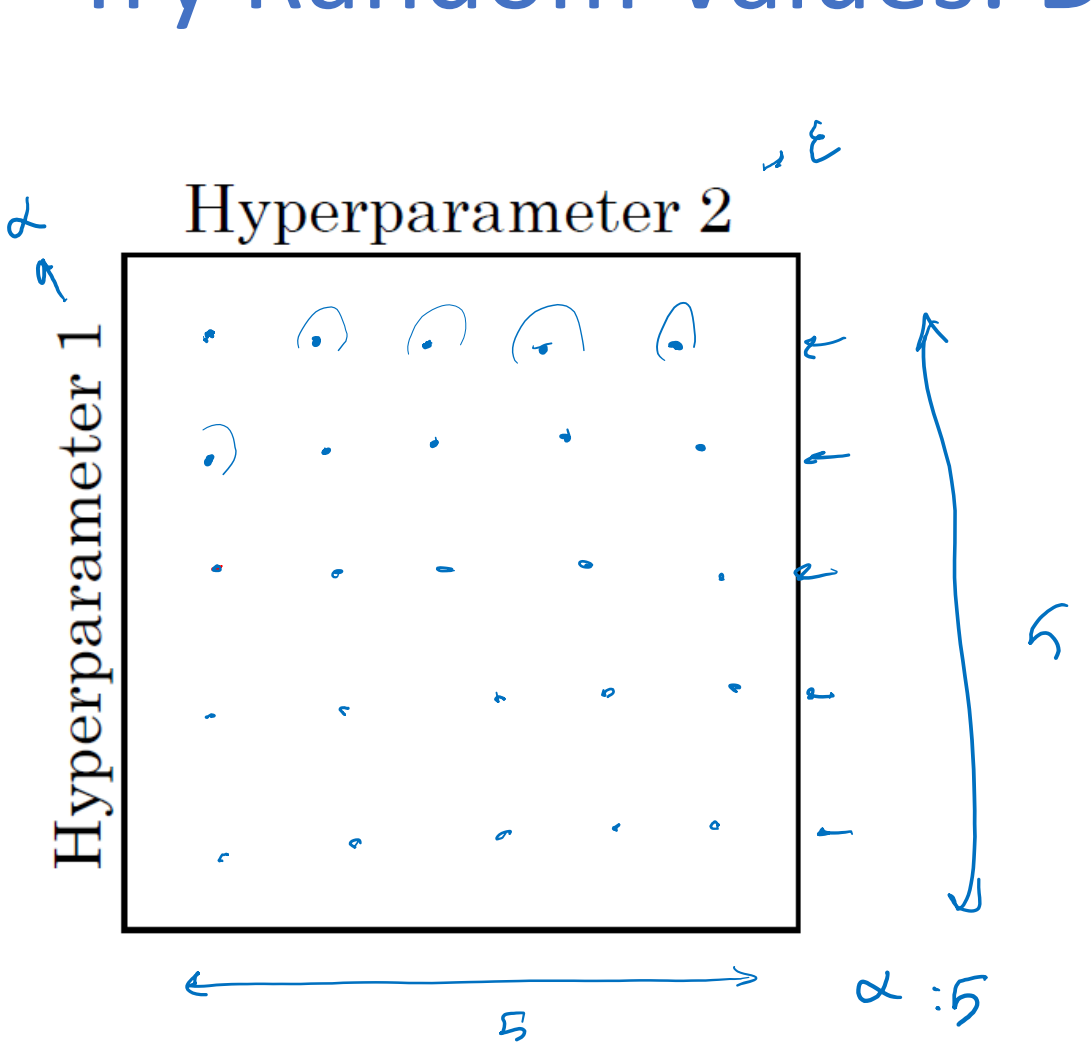
Hyperparameter Tuning: Tuning Process

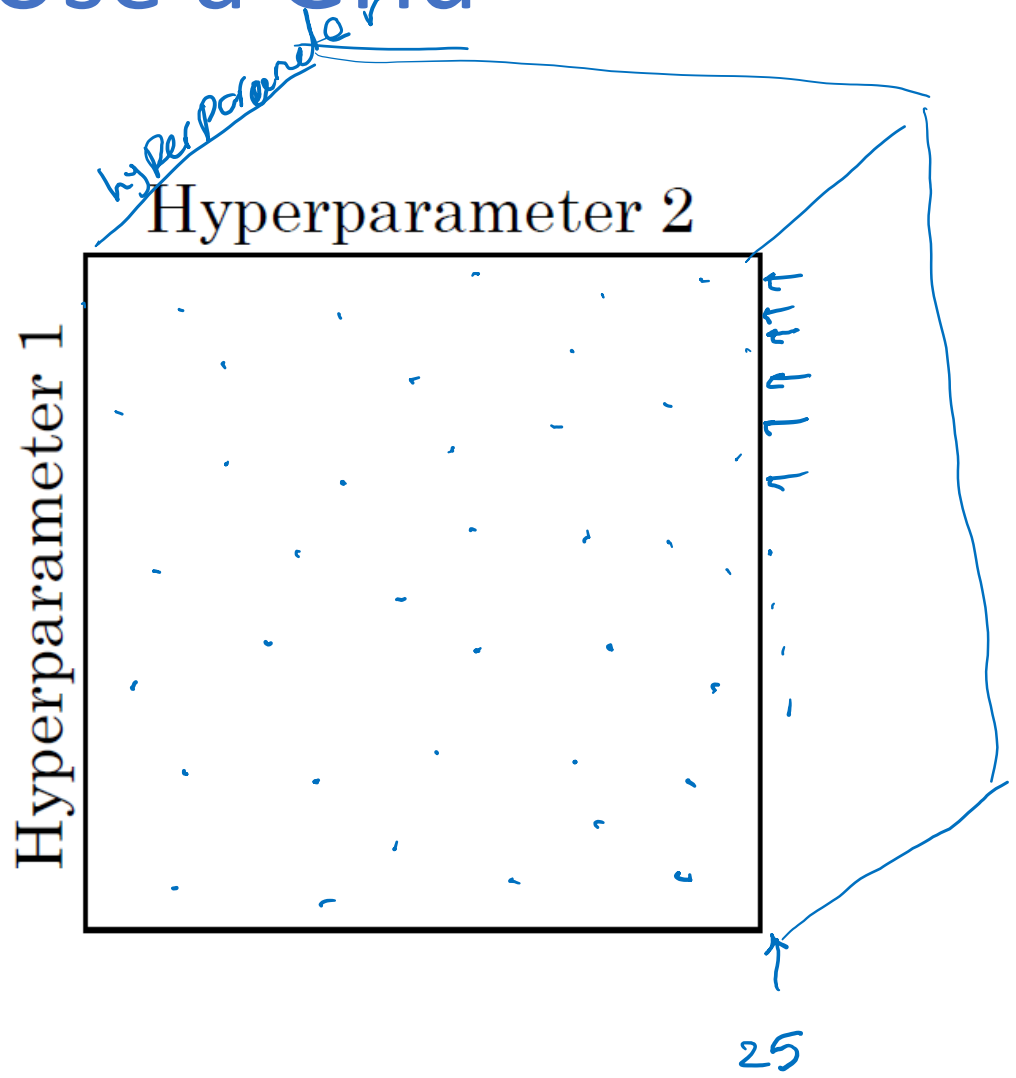
Hyperparameters

- Hyperparameter Tuning

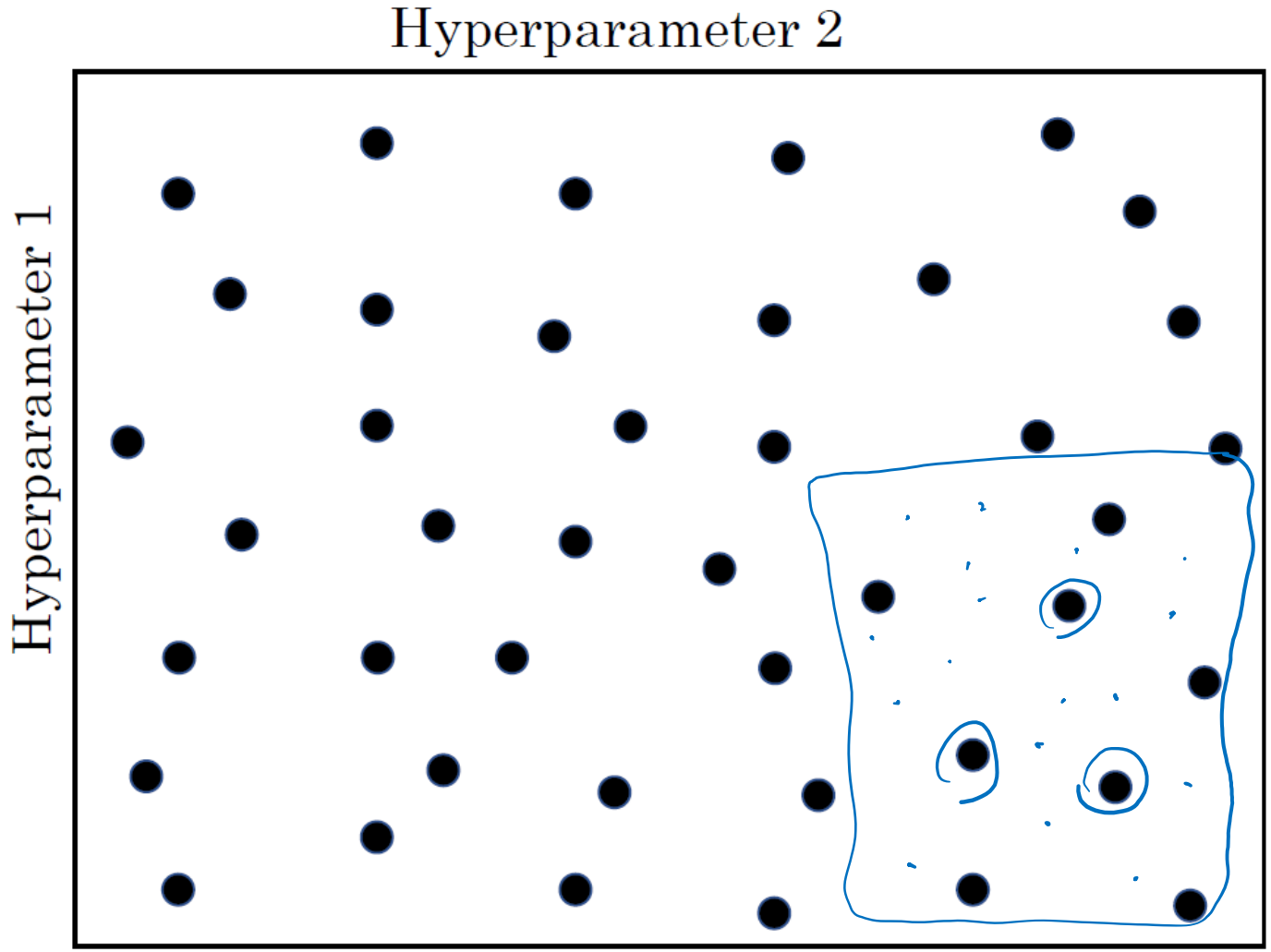


Try Random Values: Don't Use a Grid





Coarse to fine



Hyperparameter Tuning:
Using an appropriate scale to pick hyperparameters

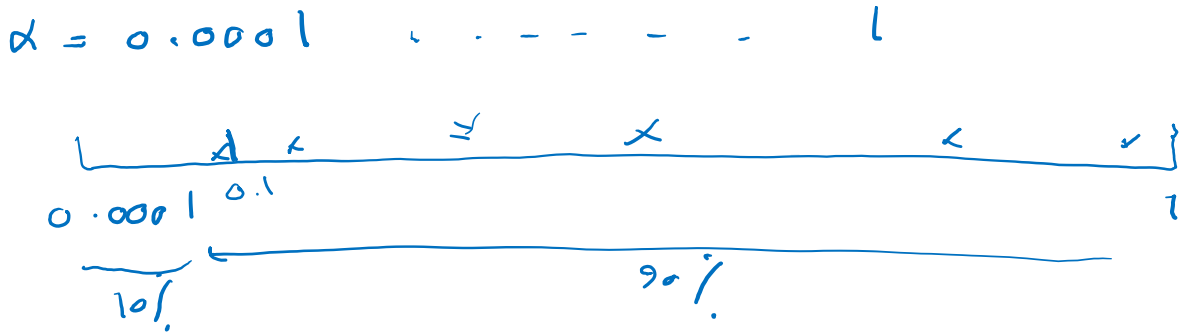
Picking Hyperparameters at Random

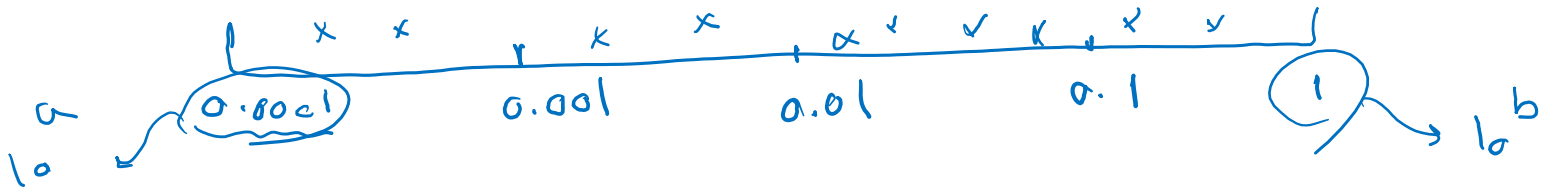
$n^{[l]} = 50 \ldots 100$

#layer $L: 2 - 4$

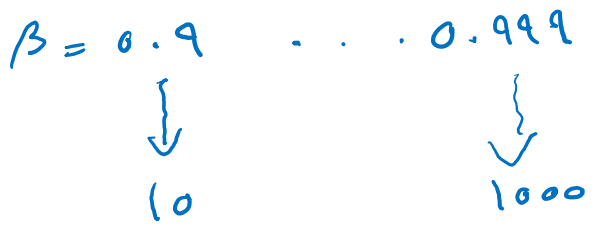
2, 3, 4

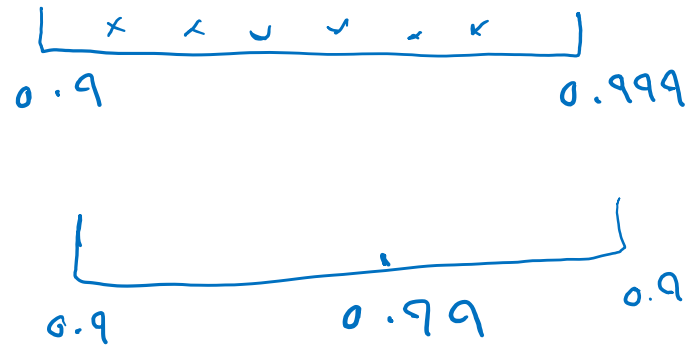
Appropriate Scale for Hyperparameters



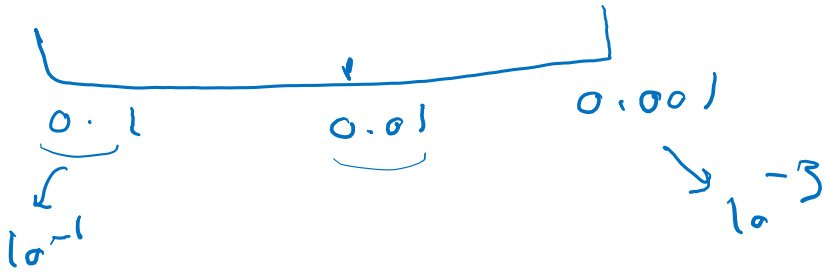


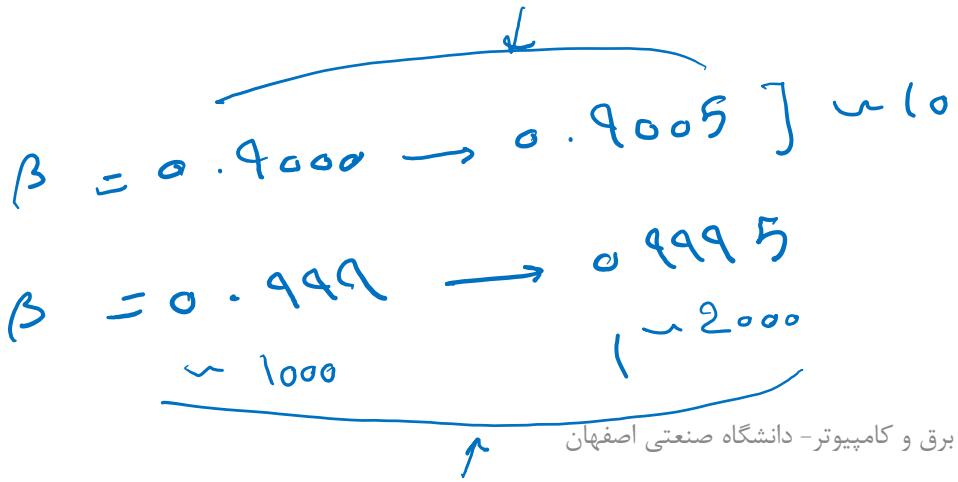
Hyperparameters for Exponentially Weighted Averages





$1-\beta = 0.1 \quad \dots \quad 0.001$





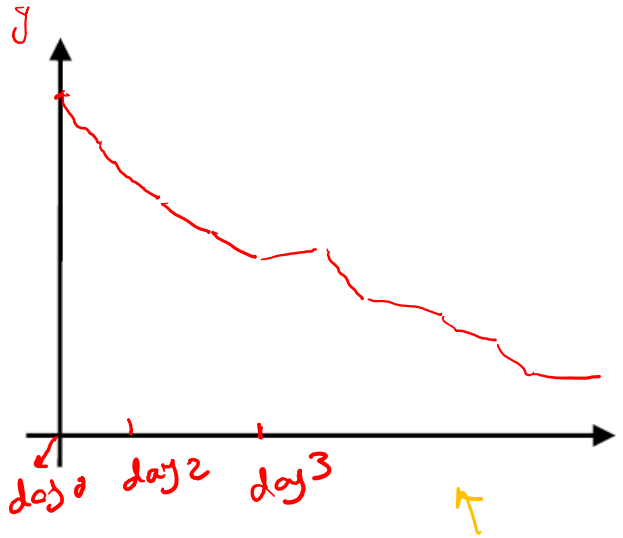
$r = [-3, -1]$

$1-\beta = 10^{r}$

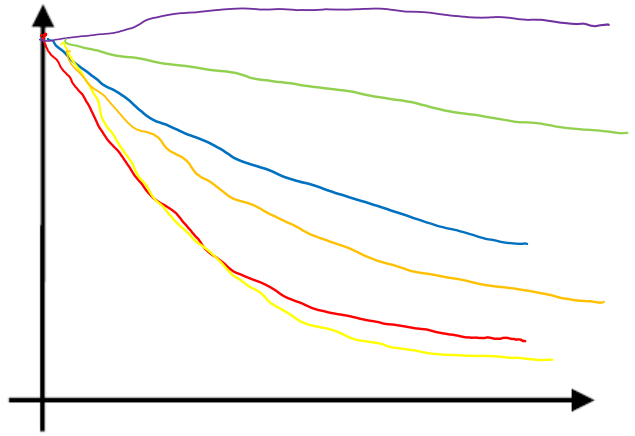
$\beta = 1 - 10^{r}$

Hyperparameters tuning in practice

Babysitting one model



Training many models in parallel



Outline

- Optimization Algorithms
  - Understanding Exponentially Weighted Averages
  - Bias Correction in Exponentially Weighted Averages
  - Gradient Descent with Momentum
  - RMSprop
  - Adam Optimization Algorithm
  - Learning Rate Decay
  - The Problem of Local Optima
- Hyperparameter Tuning
  - Tuning Process
  - Using an Appropriate Scale to pick Hyperparameters
  - Hyperparameters Tuning in Practice