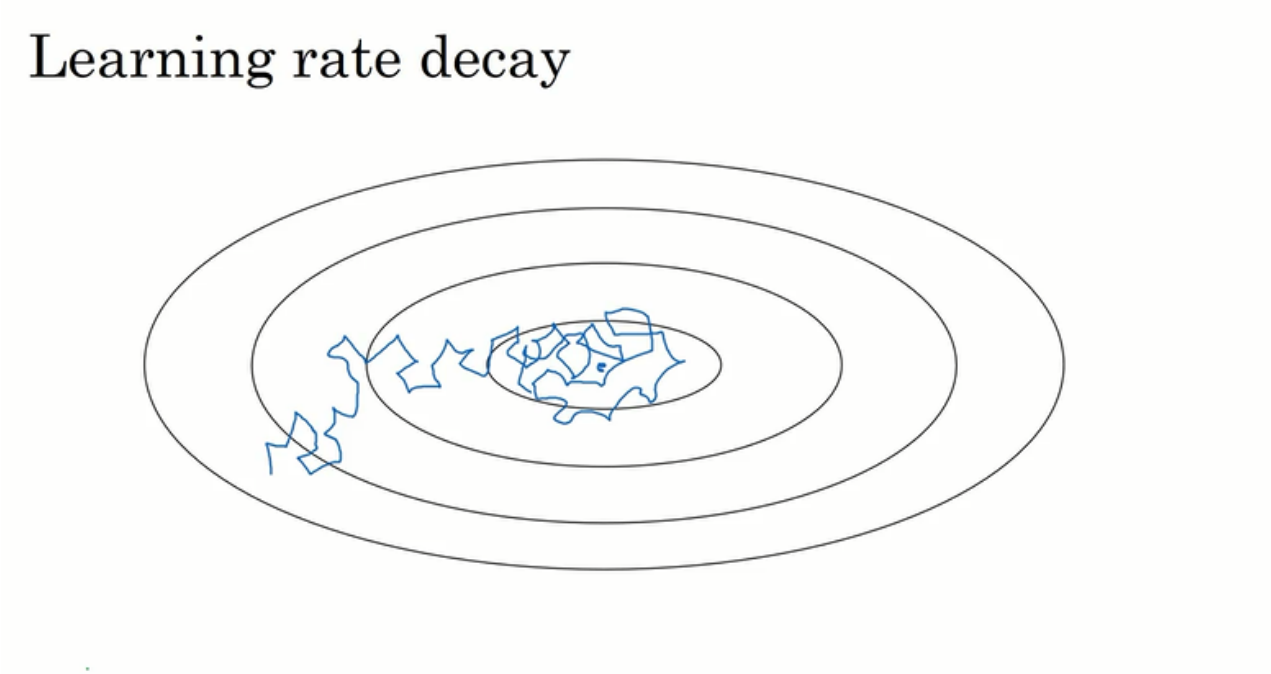
Learning rate decay

- fixed learning rate, and noise in mini-batches lead to wandering around the optimal point



- slowly reduce learning rate
  - at first it is ok to take bigger steps
  - when come close to the optimal point, take smaller step

Implement learning decay

- 1 **epoch** = 1 pass through the network
- $\alpha = \frac{1}{1+decay\_rate*epoch\_num}\alpha_0$
- $Hyperparameters: \ decay\_rate \ \& \ \alpha_0$

Other learning rate decay methods

- $\alpha = 0.95^{epoch\_num}\alpha_0$
- $\alpha = \frac{k}{\sqrt{epoch\_num}}\alpha_0 \ or \frac{k}{\sqrt{t}}\alpha_0$
- discrete staircase