Gradient Descent Tips

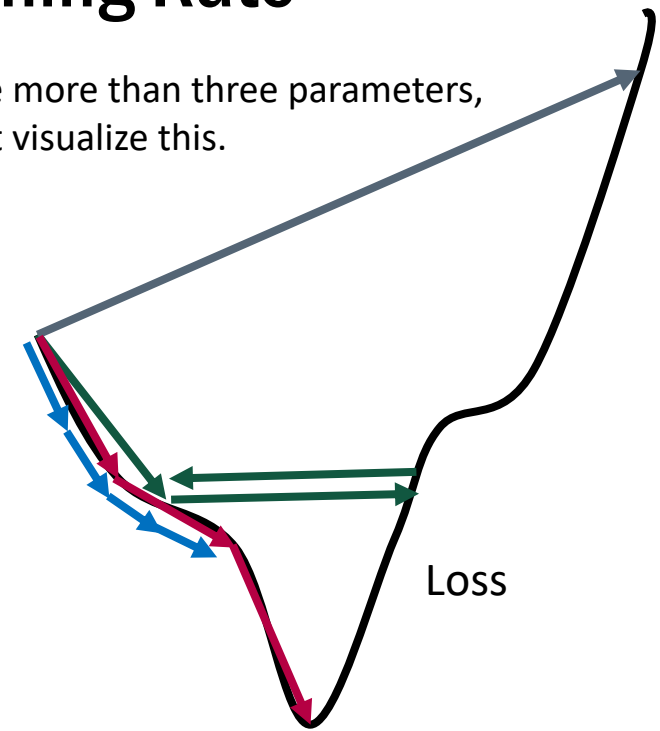
Tip One: Tuning Your Learning Rates

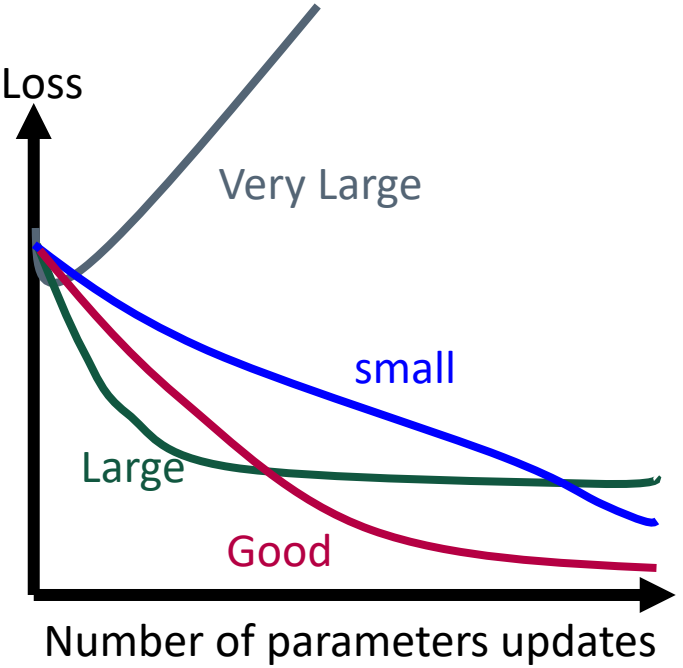
Learning Rate

$$\theta^i = \theta^{i-1} - \eta \nabla L\left(\theta^{i-1}\right)$$

But you can always visualize this.

If there are more than three parameters, you cannot visualize this.





To choose learning rate, try the following:
Declare convergence if loss decreases by less than 0.001 (tolerance) in one iteration.

Adaptive Learning Rates

- Popular and simple idea: Reduce the learning rate by some factor every few epochs.
  - At the beginning, we are far from the destination. So, we use the larger learning rate.
  - After several epochs, we are close to the destination. So, we reduce the learning rate.
  - E.g. 1/t decay: $\eta^t = \eta/\sqrt{t+1}$
- Learning rate cannot be one-size-fits-all.
  - You should assign different parameters to different learning rates.