10-605 Recitation 4

Daniel, Zach

Carnegie Mellon University

Today's Recitation

- SGD Recap
- Optimize SGD
- Learning Rate Tuning
- Coding Example

SGD Recap

Stochastic Gradient Descent

$$for\ i\ in\ range(m):$$

$$w_j := w_j - \alpha \frac{\partial J_i}{\partial w_j}$$

J_i is the cost of ith training example

Gradient Descent

$$w_j := w_j - \alpha \frac{\partial J}{\partial w_j}$$

J is the cost over all the training data points

Carnegie Mellon University

SGD Recap

Stochastic Gradient Descent

- Computationally cheap
- High variance
- More step to converge
- Prone to find local minima

Gradient Descent

- Computationally expensive
- Low variance
- Less step to converge
- Prone to find global minima

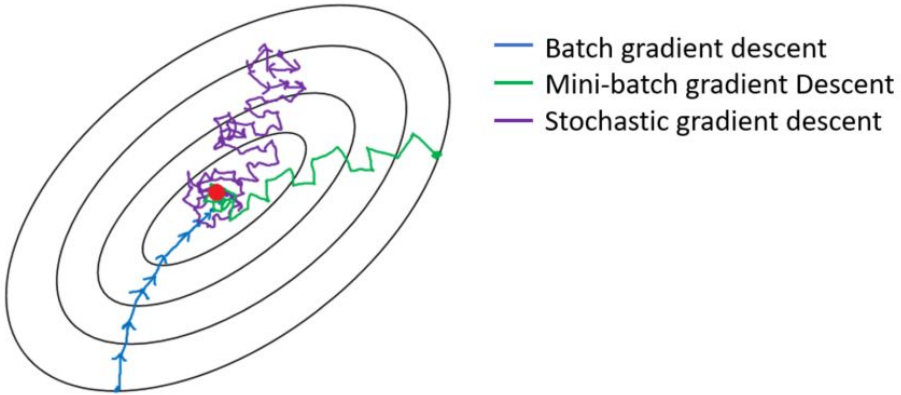
SGD Recap

Mini Batch Gradient Descent

$$for\ b\ in\ \textbf{batches}:$$

$$w_j := w_j - \alpha \frac{\partial J_b}{\partial w_j}$$

J_b is the cost of bth batch



Carnegie Mellon University

SGD with Momentum

"Noisy" derivatives for SGD.

Only estimate on a small batch, which might not be the optimal direction.

Momentum:

Define a way to get the "moving" average of some sequence, which will change along with data.

$$V_t = \beta V_{t-1} + \alpha \nabla_w L(W, X, y)$$

$$W = W - V_t$$

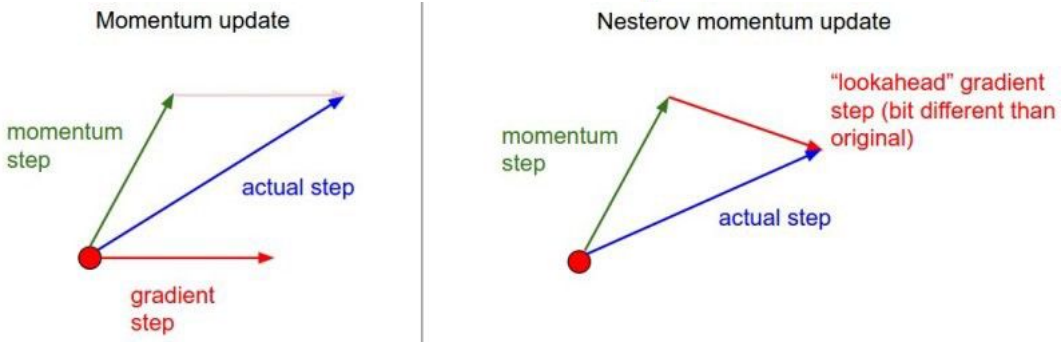
SGD with Nesterov

Slightly different from Momentum.

Calculate gradient at “look ahead” position, instead of current position.

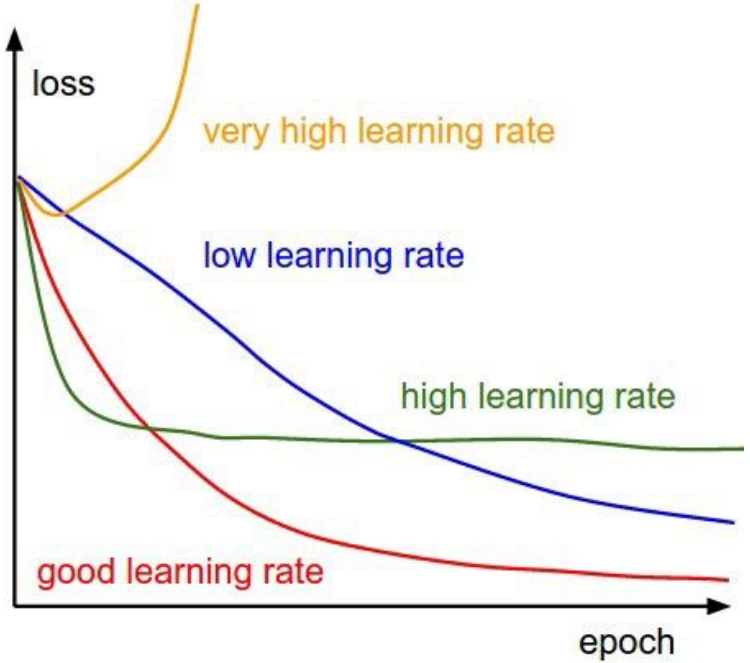
$$V_t = \beta V_{t-1} + \alpha \nabla_w L(W - \beta V_{t-1}, X, y)$$

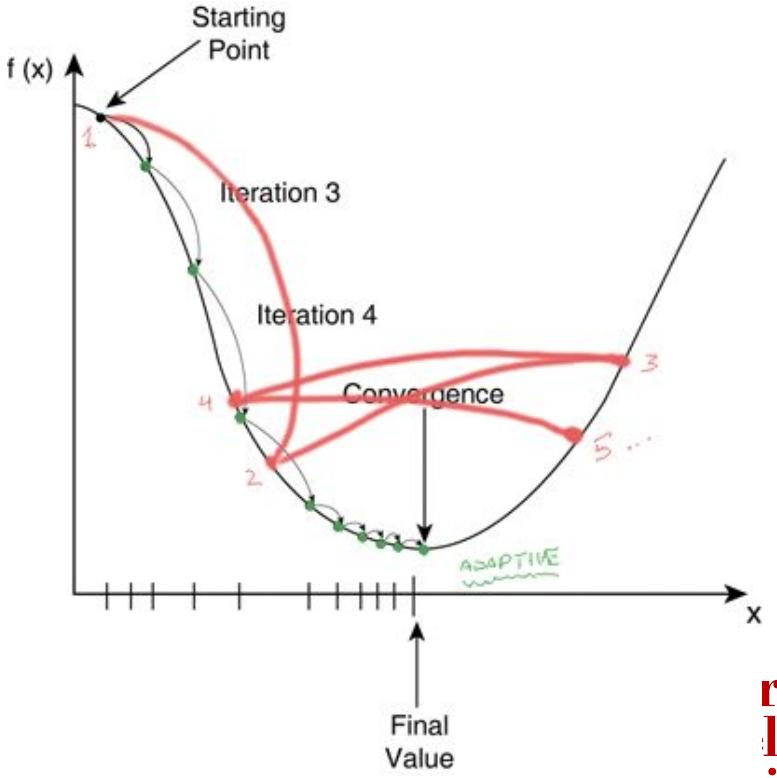
$$W = W - V_t$$



Carnegie Mellon University

Learning Rate Tuning





Learning Rate Tuning

Adaptive learning rate methods

- Time-based decay

$$lr = lr \times \frac{1}{decay \times num\_epoch}$$

- Step decay

$$lr = lr_{initial} \times drop^{\frac{num\_epoch}{epochs\_drop}}$$

- Exponential decay

$$lr = lr_{initial} \times e^{k \times num\_iter}$$

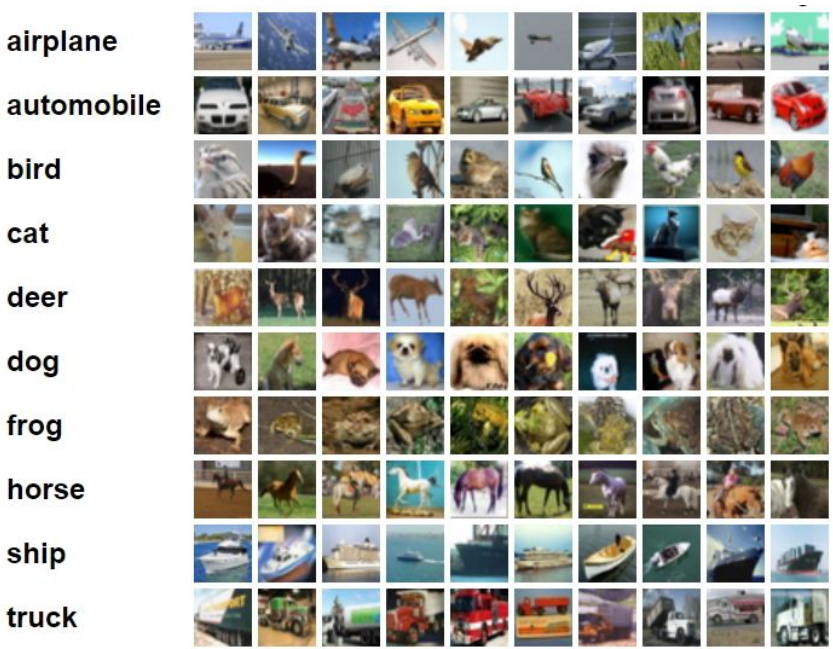
Learning Rate Tuning

Other adaptive learning rate methods

- Multi-step scheduler
- Reduce lr on plateau
- Newbob strategy

Coding Example

- Task: Image Classification
- Data: CIFAR10
- Model: CNN



Carnegie Mellon University

Reference

- https://towardsdatascience.com/https-towardsdatascience-com-why-stochastic-gradient-descent-works-9af5b9de09b8
- https://towardsdatascience.com/stochastic-gradient-descent-with-momentum-a84097641a5d
- https://deepnotes.io/sgd-momentum-adaptive#momentum