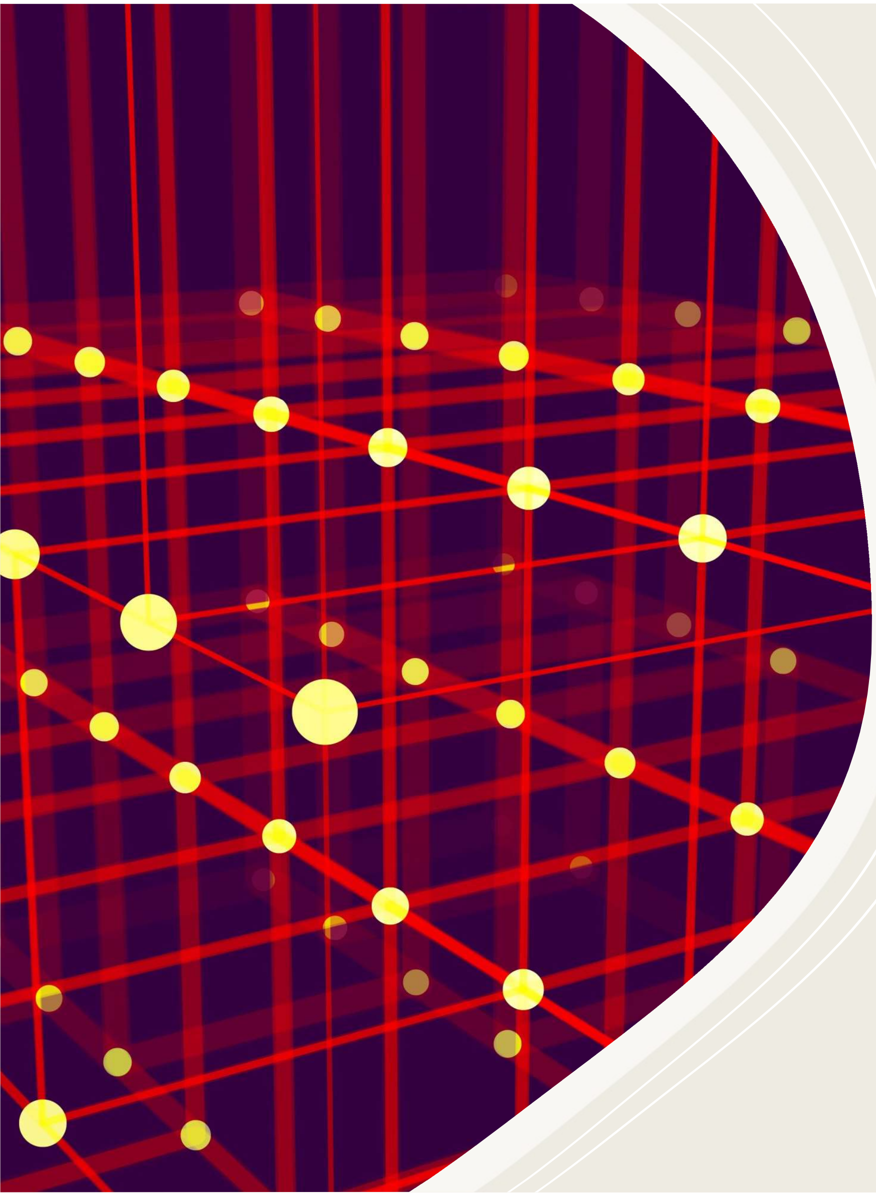


CPSC 340 – Tutorial 5

Michael Liu
mfliu@students.cs.ubc.ca

Slides courtesy of Shahriar Shayesteh

Template from Lironne Kurzman

University of British Columbia

October 13th, 2021

Agenda

- Gradient Descent

Motivation

- Recall the loss function for linear least squares

$$f(w) = \|Xw - y\|_2^2 = \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

- The gradient of this function is

$$\nabla f(w) = X^T X w - X^T y$$

- By setting the gradient to 0, we arrive at the normal equations

$$X^T X w = X^T y$$

- Solving this gives us the value(s) of w that minimize the loss

Motivation

$$X^T X w = X^T y$$

- How long does it take to solve this linear system of d equations?
  - Answer: $O(nd^2 + d^3)$
- What if d is large (e.g. $d > n$)?
  - Solving the normal equations might take a long time!

Gradient Descent for Finding a Local Minimum

- Start with some initial guess w^0
- Generate new guess by moving in the **negative gradient direction**

$$w^1 = w^0 - \alpha^0 \nabla \mathrm{f}(\mathrm{w}^0)$$

  - This decreases f if the "step size" α^0 is small enough
  - Usually, we decrease α^0 if it increases f
- Repeat to **successively refine the guess**

$$w^{t+1} = w^t - \alpha^t \nabla f(w^t)$$

- Stop if not making progress (when we are very close to a local minimum)
  - E.g. $\|\nabla f(w^t)\| \leq \epsilon$ for some small ϵ

Initializing Gradient Descent

- Pick a random initial value for the parameters

$$w^0 = \left[w_1^0, w_2^0, \dots, w_d^0\right]^T$$

- Recall the definition of the gradient

$$\nabla_{\mathrm{w}} f(w) = \left[\frac{\partial f(w)}{\partial w_1}, \frac{\partial f(w)}{\partial w_2}, \dots, \frac{\partial f(w)}{\partial w_d}\right]$$

- Initialize the gradient function by plugging in w^0

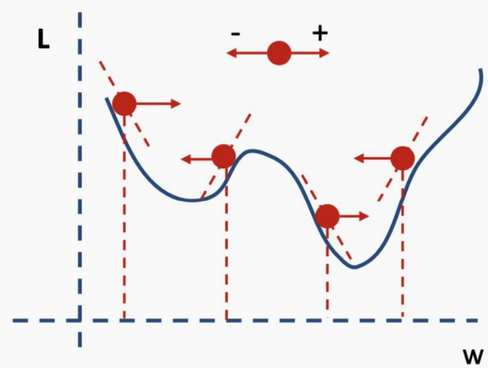
$$\nabla_{\mathrm{w}} f(w^0) = \left[\frac{\partial f(w^0)}{\partial w_1}, \frac{\partial f(w^0)}{\partial w_2}, \dots, \frac{\partial f(w^0)}{\partial w_d}\right]^T$$

Running Gradient Descent

- Calculate the step for each feature as:
  step = gradient * learning rate
- Calculate new parameters as:
  new params = old params – step
- Repeat **gradient update**, **step**, and **parameter update** until gradient is almost 0

One Step of Gradient Descent, Graphically

$$w^{(i+1)} = w^{(i)} - \lambda \frac{d\mathcal{L}}{dw}$$



Matthew Stewart, PhD Researcher: https://towardsdatascience.com/simple-introduction-to-neural-networks-ac1d7c3d7a2c