Reading Notes for Week 2

1.1 ~ 1.2 Supervised Learning
Just introduce some notations and framework knowledge. Not very much insights here.

8.2.1 Steepest Gradient Descent
For an unconstrained optimization problem, if the function is not convex, we cannot rely on a traditional framework to solve for the optimal. Instead, we opt for gradient descent. The gradient descent method utilizes what we have learned in multivariable calculus about the directional derivative, where for any point on the function, we want to find a direction where, if we move along that direction, the function's **directional derivative** hits the minimum. Mathematically, at a point x, the directional derivative is $\nabla f(\vec{x})^T \vec{v}$, where v is the descent direction. The book proposes steepest gradient descent method and later on stochastic gradient descent to determine such direction.

8.2.2 Search for step size
Once we have our descent direction, we now decide the step size. It could be constant, but should be bounded in order for the gradient descent to converge, the detail of convergence analysis is based on eigenvalues and control theory, which is not difficult. **The most interesting** part is the search for step size. In the book pp284 where Murphy proposes an inexact line search algorithm called **Armijo backtracking method**. According to NW06's book numerical optimization, typically for a line search algorithm to find a good step size, **sufficient decrease condition**(upper bound) and **curvature condition**(lower bound) should be met. But for Armijo's method, since it prevents the step size from getting too small, so it only needs to satisfy the sufficient decrease condition.

8.4.1 Finite sum problem
This chapter states that the batched GD is slower than SGD since SGD compute gradient using different portion of dataset, which won't repeatedly iterate on the same data.

8.4.2 SGD for fitting linear regression
SGD may require multiple pass through the data to find the optimum.

8.4.3 Choosing the step size
This section introduces lots of scheme of choosing learning rate for SGD to converge. **Surprisingly**, the choice of learning rate can be accomplished by line search if the variance of the gradient noise goes to zero.

8.4.4 Iterative Averaging
This section introduces a way to reduce variance of the estimate and thus the gradient noise. Thus i guess with this scheme we can use line search to find a good learning rate.