Machine Learning	**Summer 2017**

Topic 1: Gradient Descent

Author: Andy Lee

1.1 Introduction

This set of notes provides intuition that gradient descent and Newton's method can be interpreted as minimizing quadratic approximations to a loss function.

1.2 General Form

Let the kth solution iterate be denoted x^k, the loss function be denoted f. D^k is a positive semi-definite matrix. Then, in general, gradient descent type algorithms can be formulated as such

$$x^{k+1} = x^k + \alpha^k D^k \nabla f(x^k)$$

1.3 Gradient Descent

Gradient descent takes the form

$$x^{k+1} = x^k + \alpha^k I \nabla f(x^k) \text{ , } (I \text{ is the } n \times n \text{ identity matrix})$$

We will now show that gradient descent can be thought of as minimizing a particular quadratic function, namely the following quadratic approximation to our loss function, $\hat{f}$. Note α can be thought of either as a regularization parameter or step-size parameter.

$$\hat{f}(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2\alpha}||y-x||^2$$

Now, we will solve for the optimal solution iterate y by setting the derivative $\frac{\partial}{\partial y}(\hat{f}(y))$ equal to 0 and solving for y.

$$\frac{\partial}{\partial y}(\hat{f}(y)) = \frac{\partial}{\partial y}(f(x)) + \frac{\partial}{\partial y}\left(\nabla f(x)^T(y-x)\right) + \frac{\partial}{\partial y}\left(\frac{1}{2\alpha}||y-x||^2\right) \tag{1.1}$$

$$= 0 + \frac{\partial}{\partial y}\left(\nabla f(x)^T(y-x)\right) + \frac{1}{2\alpha}\frac{\partial}{\partial y}\left((y-x)(y-x)\right) \tag{1.2}$$

$$= \text{ ... algebra} \tag{1.3}$$

$$= \nabla f(x)^T + \frac{1}{\alpha}(y-x) \tag{1.4}$$

By solving this, we obtain the solution

$$y^* = x - \alpha \nabla f(x)$$

1.4 Newton's Method

There derivation of Newton's method relies on using a better quadratic approximation to the loss function. In particular, we take the full 2nd order Taylor approximation to our loss function.

$$\hat{f}(y) = f(x) + \nabla(f(x)^T(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x)$$

Note the derivation.

$$\frac{\partial}{\partial y}(\hat{f}(y)) = \frac{\partial}{\partial y}(f(x)) + \frac{\partial}{\partial y}\Big(\nabla f(x)^T(y - x)\Big) + \frac{\partial}{\partial y}\left(\frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x)\right) \tag{1.5}$$

$$= \nabla f(x) + \frac{1}{2}\frac{\partial}{\partial y}\Big((y - x)^T \nabla^2 f(x)(y - x)\Big) \tag{1.6}$$

$$= \text{ ... algebra} \tag{1.7}$$

$$= \nabla f(x) + \frac{1}{2}\Big((y - x)^T \nabla^2 f(x) + \nabla^2 f(x)(y - x)\Big) \tag{1.8}$$

$$= \nabla f(x) + \nabla^2 f(x)(y - x) \tag{1.9}$$

By solving this, we obtain the solution

$$y^* = x - (\nabla^2 f(x))^{-1} \nabla f(x)$$