AM 221: Advanced Optimization	Spring 2016
Prof. Yaron Singer	*Lecture 10 — February 29th*

1 Overview

In the previous lecture we introduced the gradient descent algorithm, and mentioned that it falls under a broader category of methods. In this lecture we describe this general approach called *steepest descent*. We will explain how gradient descent is an example of this method, and also introduce the *coordinate descent* algorithm which is another example of the steepest descent method. Lastly, we will present Newton's method. Newton's method is a general approach for solving systems of non-linear equations. Newton's method can conceptually be seen as a steepest descent method, and we will show how it can be applied for convex optimization.

2 Steepest Descent

As discussed in the previous lecture, one can consider a search for a stationary point as an iterative procedure of generating a point $\mathbf{x}^{(k+1)}$ which takes steps of certain length t_k at direction $\Delta\mathbf{x}^{(k)}$ from the previous point $\mathbf{x}^{(k)}$. The direction $\Delta\mathbf{x}^{(k)}$ decides which direction we search next, and the step size determines how far we go in that particular direction. We can write this update rule as:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t_k \Delta\mathbf{x}^{(k)}$$

A *steepest descent* algorithm would be an algorithm which follows the above update rule, where at each iteration, the direction $\Delta\mathbf{x}^{(k)}$ is the *steepest* direction we can take. That is, the algorithm continues its search in the direction which will minimize the value of function, given the current point. Or in other words, given a particular point $\mathbf{x}$, we would like to find the direction $\mathbf{d}$ s.t. $f(\mathbf{x}+\mathbf{d})$ is minimized.

Finding the steepest direction. In order to find the steepest direction, we can approximate the function via a first-order Taylor expansion:

$$f(\mathbf{x}+\mathbf{d}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal \mathbf{d}$$

The direction $\mathbf{d}$ that minimizes the function implies the following optimization problem[1]:

$$\min_{\mathbf{d}:\|v\|=1} \nabla f(\mathbf{x})^\intercal \mathbf{d}$$

In general, one may consider various norms for the minimization problem. As we will now see, the interpretation of steepest descent with different norms leads to different algorithms.

[1] Recall that a *direction* is a vector of unit length.