CS 189 Introduction to Machine Learning
Fall 2019 Jennifer Listgarten & Stella Yu DIS6

1 Newton's Method

Newton's Method is an alternative to gradient descent that takes into account the second-order terms of the Taylor expansion. In other words, instead of optimizing

$$\min_{\mathbf{w}} \bar{f}(\mathbf{w}) = f(\mathbf{w}^{(t)}) + \nabla f(\mathbf{w}^{(t)})^\top(\mathbf{w} - \mathbf{w}^{(t)}),$$

we optimize

$$\min_{\mathbf{w}} \bar{f}(\mathbf{w}) = f(\mathbf{w}^{(t)}) + \nabla f(\mathbf{w}^{(t)})^\top(\mathbf{w} - \mathbf{w}^{(t)}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^{(t)})^\top \nabla^2 f(\mathbf{w}^{(t)})(\mathbf{w} - \mathbf{w}^{(t)}).$$

The reason we might want to do this is to reach convergence faster. In fact, we will soon see that Newton's method can be viewed as a generalized gradient descent algorithm that decides the step size based on the Hessian. We will also see other methods later on in the course such as applying momentum to the gradient step to achieve a similar effect.

1. First, let us derive the update step for Newton's method. Using gradient methods, solve for the optimal w. This will be our next weight vector in our sequence of weights, since this is an iterative algorithm, so this will actually be our new value for w, i.e. $w^{(t+1)}$.

   Now, let's try using Newton's Method to solve a problem we've studied before, ridge regression. We'll warm up to it with some Hessian exercises, and then derive the Newton update step afterwards.

2. Compute the Hessian of $f(x, y, z) = x^2 + y^2 + z^2$.

3. Compute the Hessian of $f(x, y) = (x^2 + y^2) \cdot e^{-y}$, and approximate the function at $(x, y) = (0, 0)$

4. Recall that Ridge Regression has the loss function:

   $$L(w) = ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \lambda||\mathbf{w}||_2^2$$

   for $\mathbf{X} \in \mathbb{R}^{nxd}$, $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^n$. Derive the gradient and Hessian for this loss function.

5. What is the Newton's method update rule for this problem?

2 Connections between OLS, Ridge Regression, TLS, PCA, and CCA

We will review several topics we have learned so far: ordinary least-squares, ridge regression, total least squares, principle component analysis, and canonical correlation analysis. We emphasize their basic attributes, including the objective functions and the explicit form of their solutions. We will also discuss the connections and distinctions between these methods.

1. What are the objective functions and closed-form solutions to OLS, ridge regression, and TLS? How do the probabilistic interpretations vary?

2. Consider the matrix inversion in the solution to OLS, ridge regression, and TLS. How do the eigenvalues compare to those of the matrix $\mathbf{X}^\top\mathbf{X}$?

3. Suppose you have a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and output $\mathbf{y} \in \mathbb{R}^{n \times 1}$. Use PCA to compute the first k principal components of $[\mathbf{X} \quad \mathbf{y}]$. Describe how this solution would relate to performing TLS on the problem.