AM 221: Advanced Optimization Spring 2016

Prof. Yaron Singer *Problem Set 5 — Due Wed, March. 2nd at 23:59*

Instructions: All your solutions should be prepared in LaTeX and the PDF and .tex should be submitted to canvas. For each question, the best and correct answers will be selected as sample solutions for the entire class to enjoy. If you prefer that we do not use your solutions, please indicate this clearly on the first page of your assignment.

The programming parts can be written in the programming language of your choice and the code should be submitted alongside your solutions.

Throughout this problem set, $\|\mathbf{x}\|$ will denote the Euclidean norm of $\mathbf{x} \in \mathbb{R}^n$, that is $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\intercal \mathbf{x}}$. For A and B to matrices in $\mathbb{R}^{n \times n}$, $A \preceq B$ denotes the partial order defined by:

$$A \preceq B \Leftrightarrow B - A \text{ is positive semi-definite}$$

1. Optimality criteria. Let f be a differentiable convex function from $\mathbb{R}^n$ to $\mathbb{R}$. We consider the following optimization problem:

$$\min_{\mathbf{x} \in C} f(\mathbf{x})$$

where $C \subseteq \mathbb{R}^n$ is a closed convex set.

a. Show that $\mathbf{x}^* \in C$ is an optimal solution to the above problem if and only if:

$$\forall \mathbf{y} \in C,\ \nabla f(\mathbf{x}^*)^\intercal (\mathbf{y} - \mathbf{x}^*) \geq 0 \tag{P}$$

b. Show that when $\nabla f(\mathbf{x}^*) \neq 0$ this implies that $\mathbf{x}^*$ lies on the boundary of C and that $\nabla f(\mathbf{x}^*)$ defines a supporting hyperplane of C at $\mathbf{x}^*$. Remember that the boundary $\delta(C)$ of a closed set C is defined by:

$$\delta(C) \stackrel{\text{def}}{=} \{\mathbf{x} \in C \,|\, \forall r > 0, B(\mathbf{x}, r) \nsubseteq C\}$$

where $B(\mathbf{x}, r)$ denotes the ℓ_2-ball of center $\mathbf{x}$ and radius r.

c. Show that when $C = \mathbb{R}^n$, $\mathbf{x}^* \in \mathbb{R}^n$ satisfies condition (P) of part a. iff:

$$\nabla f(\mathbf{x}^*) = 0$$

Then observe (you don't have to prove it) that by combining part a. and c. we obtain the following theorem.

Theorem 1. *$\mathbf{x}^* \in \mathbb{R}^n$ is an optimal solution to the problem:*

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$$

if and only if $\nabla f(\mathbf{x}^*) = 0$.

2. Lipschitz-continuous Gradient. A common smoothness assumption made to show convergence of optimization algorithms for convex functions is to assume that the gradient is Lipschitz-continuous. We say that a differentiable function f from $\mathbb{R}^n$ to $\mathbb{R}$ has a gradient which is L-Lipschitz-continuous iff:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|,\ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n$$

a. Show that a differentiable function $g : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if:

$$\big(\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\big)^\intercal(\mathbf{x} - \mathbf{y}) \geq 0,\ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n$$

b. Assume that f's gradient is L-Lipschitz continuous (f is not necessarily convex), then show that the function g defined by:

$$g(\mathbf{x}) = \frac{L}{2}\|\mathbf{x}\|^2 - f(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^n$$

is convex.

c. Assume that f twice differentiable and that its gradient is L-Lipschitz-continuous, then show that:

$$H_f(\mathbf{x}) \preceq LI_n,\ \mathbf{x} \in \mathbb{R}^n$$

where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix.

Remark. It is possible to show that when f is convex, the reverse statement is true: if $H_f(\mathbf{x}) \preceq LI_n$ for all $\mathbf{x}$, then f's gradient is L-Lipschitz-continuous.

3. Least-squares regression. In lecture 1, we introduced the problem of least-squares regression. Given a dataset of n data points $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \leq i \leq n$, the goal is to find $\mathbf{a} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ so as to minimize:

$$RSS(\mathbf{a}, b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\intercal\mathbf{a} - b)^2$$

In other words, we are trying to approximate $y_i \simeq \mathbf{x}_i^\intercal\mathbf{a} - b$ and the approximatation error is measured by the function RSS above.

a. Rewrite the least-squares regression problem in matrix form, that is find $X \in \mathbb{R}^{n \times (d+1)}$ and $Y \in \mathbb{R}^n$ such that the problem above takes the form:

$$\min_{\mathbf{d} \in \mathbb{R}^{d+1}} \|X\mathbf{d} - Y\|^2$$

and express X and Y in terms of the data points $(\mathbf{x}_i, y_i)$.

b. Define $f : \mathbb{R}^{d+1} \to \mathbb{R}$ by $f(\mathbf{d}) = \|X\mathbf{d} - Y\|^2$ for all $\mathbf{d} \in \mathbb{R}^{d+1}$. Compute the gradient and Hessian of f and show that f is convex.

c. Give a sufficient and necessary condition for f to be strongly convex.

In all the following questions, we will assume that the condition of part c. is satisfied.

d. Solve the equation $\nabla f(\mathbf{x}) = 0$ and explain how you would use this to find an optimal solution to the least-squares regression problem.

e. An alternative approach to d. is to use gradient descent. For this, we need to solve for any direction $\delta \in \mathbb{R}^{d+1}$ the following line search problem:

$$\min_{\lambda \in \mathbb{R}} f(\mathbf{d} + \lambda \delta)$$

Give a closed-form formula for the optimal solution to the line-search problem. The solution should be expressed in termes of $X, Y, \mathbf{d}$ and δ.

4. Wine quality revisited. In this problem, we will re-use the dataset from last week available at `http://thibaut.horel.org/wines.csv`. Please refer to Problem Set 4 for a description of the dataset. We will again fit a linear model to predict wine quality as a function of the chemical measurements. However we will use least-squares regression (as presented in Problem 3 above) instead of ℓ_1-regression.

a. Verify that for this dataset, matrix X as defined in Problem 3 satisfies the condition of Problem 3, part c.

b. Write code to compute the optimal solution to the least-squares regression problem using the method derived in Problem 3, part d. Report your code, the linear model ($\mathbf{a}$ and b) and the value of function RSS for this model.

c. Implement the gradient descent algorithm for the least-squares regression problem. You are not allowed to use already existing implementations of gradient descent (but you can of course use libraries for matrix computation). You should use exact line search as derived in Problem 3, part e. Report your code, the linear model and the value of RSS for this model. How does this compare to the result found in part b.?

d. Compute from matrix X an upper-bound on the convergence rate of the gradient descent algorithm. Discuss the relative strengths and weaknesses of method b. and method c.