AM 221: Advanced Optimization Spring 2018

Kojin Oshiba *Problem Set 5 — Due Wednesday, Feb. 28th at 23:59*

Instructions:

- All your solutions should be prepared in LaTeX and the PDF and .tex should be submitted to Canvas. Please submit all your files as ONE archive of filetype zip, tgz, or tar.gz.
- Name the file [your-first-name]_[your-last-name].[filetype]. For example, I would call my submission rasmus_kyng.zip.
- INCLUDE your name in the submisson pdf and any files with code.
- If the TFs cannot easily deduce your identity from your files alone, they may decide not to grade your submission.
- For each question, a well-written and correct answer will be selected a sample solution 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.

1. Least-squares regression. In lecture 1, we introduced the problem of least-squares regression. Given a dataset of n data points $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \le i \le n$, 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.

a. Let

$$Y = \frac{1}{\sqrt{n}} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \frac{1}{\sqrt{n}} \begin{pmatrix} \mathbf{x}_1^\intercal & 1 \\ \mathbf{x}_2^\intercal & 1 \\ \vdots & \vdots \\ \mathbf{x}_n^\intercal & 1 \end{pmatrix}, \mathbf{d} = \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix}$$

Then can rewrite the given optimization problem as:

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

b. Since $f(\mathbf{d}) = ||X\mathbf{d} - Y||^2 = ||Y||^2 - 2X^\intercal Y\mathbf{d} + ||X\mathbf{d}||^2$,

$$\nabla f(\mathbf{d}) = -2X^\intercal Y + 2X^\intercal(X\mathbf{d}) = -2X^\intercal(X\mathbf{d} - Y)$$

$$H_f(\mathbf{d}) = 2X^\intercal X$$

Because for $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v}^\intercal X^\intercal X\mathbf{v} = ||X\mathbf{v}||^2 > 0$, $H_f(\mathbf{d})$ is positive semi-definite. Hence we have that f is convex.

c. From the definition of strong convexity, we would like to have, for some $m < M \in \mathbb{R}_{\geq 0}$,

$$mI \leq H_f(\mathbf{x}) \leq MI$$

.

Hence, the sufficient and necessary condition for f to be strongly convex is

$$mI \leq X^\intercal X \leq MI$$

d.

$$\begin{aligned} \nabla f(\mathbf{x}) = 0 &\Leftrightarrow 2X^\intercal(X\mathbf{d} - Y) \\ &\Leftrightarrow X^\intercal X\mathbf{d} = X^\intercal Y \\ &\Leftrightarrow \mathbf{d} = (X^\intercal X)^{-1}X^\intercal Y \end{aligned}$$

e. Since f is convex, the stationary point is the global minimum. Hence,

$$\begin{aligned}\nabla f(\mathbf{d} + \lambda\delta) = 0 &\Leftrightarrow 2(X\delta)^\intercal(X(\mathbf{d} + \lambda\delta) - Y) = 0\\ &\Leftrightarrow (X\delta)^\intercal X(\mathbf{d} + \lambda\delta) = (X\delta)^\intercal Y\\ &\Leftrightarrow ||X\delta||^2\lambda = (X\delta)^\intercal Y - (X\delta)^\intercal X\mathbf{d}\\ &\Leftrightarrow \lambda = \frac{(X\delta)^\intercal(Y - X\mathbf{d})}{||X\delta||^2}\end{aligned}$$

2. Wine quality revisited. In this problem, we will re-use the dataset from Homework 3 available at `http://rasmuskyng.com/am221_spring18/psets/hw3/wines.csv`. Please refer to Homework 3 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 1 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.

a. See hw5.py.

b. a: [2.49905528e-02 -1.08359026e+00 -1.82563950e-01 1.63312698e-02 -1.87422516e+00 4.36133331e-03 -3.26457970e-03 -1.78811638e+01 -4.13653144e-01 9.16334413e-01 2.76197699e-01]

   b: 21.9652084243

   val: 0.416767167221

   See hw5.py for code.

c. a: [0.05153577 -1.16512617 -0.26828834 0.00558491 -0.44400488 0.00370356 -0.00251642 0.43634826 0.08281569 0.78552489 0.3120705]

   b: 1.49761264935

   val: 0.423019391955

The value is very similar to the optimal solution, although it's not entirely the same. The weights and bias are different from the values in b.

See hw5.py for code.

d. The upperbound of the convergence rate is:

$$f(\mathbf{d}^{(k+1)}) - \alpha^*) \leq (1 - \frac{m}{M})^k (f(\mathbf{x}^{(0)}) - \alpha^*)$$

With $k = 100000$, the upperbound is, 21.7.

The strength of method b is that the solution is guranteed to be optimal. The weakness of method b is that weight matrix must be invertible, and even if it is, inverting it might take depending on the dataset size.

The strength of method b is that the weight matrices don't have to be invertible, and if the number of iterations is not too high, the algorithm runs fast. The weakness of method d is that the solution is not guranteed to be optimal. And as we just saw, the theoretical gurantee of convergence is actually quite weak. So we are not certain if the solution reached is optimal.

3. 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:

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

In class, we saw a definition of L-Lipschitz-continuous for the special case of *twice* differentiable functions. In this problem, you will show (among other things) that the definition state above implies the condition stated in class whenever the function is twice differentiable.

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.

a. We would like to show (i) $g : \mathbb{R}^n \to \mathbb{R}$ is convex $\Leftrightarrow$ (ii) $(\nabla g(\mathbf{x}) - \nabla g(\mathbf{y}))^\intercal(\mathbf{x} - \mathbf{y}) \geq 0, (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n$.

   Proof of (i) $\to$ (ii)

   Since g is convex,

$$g(\mathbf{y}) \geq g(\mathbf{x}) + \nabla g(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x})$$
$$g(\mathbf{x}) \geq g(\mathbf{y}) + \nabla g(\mathbf{y})^\intercal(\mathbf{x} - \mathbf{y})$$

   Adding these two inequalities, we have

$$g(\mathbf{y}) + g(\mathbf{x}) \geq g(\mathbf{x}) + g(\mathbf{y})0(\nabla g(\mathbf{x}) - \nabla g(\mathbf{y}))^\intercal(\mathbf{y} - \mathbf{x})$$

   Hence,

$$(\nabla g(\mathbf{x}) - \nabla g(\mathbf{y}))^\intercal(\mathbf{x} - \mathbf{y}) \geq 0$$

   Proof of (ii) $\to$ (i)

   (ii) $\to \nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is monotone. Let

$$h(t) = g(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$$

   Then,

$$h'(t) = \nabla g(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^\intercal(\mathbf{y} - \mathbf{x})$$

   Because ∇f is monotone, we have that $h'(t) \geq h'(0)$ for $t \geq 0$. Hence,

$$\begin{aligned} g(\mathbf{y}) &= h(1) \\ &= h(0) + \int_0^1 h'(t)dt \\ &\geq h(0) + h'(0) \\ &= g(\mathbf{x}) + \nabla g(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}) \end{aligned}$$

   Hence, g is convex.

b. $\nabla g(\mathbf{x}) = L\mathbf{x} - \nabla f(\mathbf{x})$. Since $\mathbb{R}^n$ is a convex set, all we need to show is that $\nabla g(\mathbf{x})$ is monotone.

$$\begin{aligned} &||\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})|| \leq L||\mathbf{x} - \mathbf{y}|| \\ \Rightarrow &||\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})||||\mathbf{x} - \mathbf{y}|| \leq L||\mathbf{x} - \mathbf{y}||^2 \\ \Rightarrow &(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\intercal(\mathbf{x} - \mathbf{y}) \leq L||\mathbf{x} - \mathbf{y}||^2 \ (\because \text{Cauchy-Shwartz}) \\ \Rightarrow &(L(\mathbf{x} - \mathbf{y}) - (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})))^\intercal(\mathbf{x} - \mathbf{y}) \geq 0 \\ \Rightarrow &(\nabla g(\mathbf{x}) - \nabla g(\mathbf{y}))^\intercal(\mathbf{x} - \mathbf{y}) \geq 0 \end{aligned}$$

   Hence, $\nabla g(\mathbf{x})$ is monotone. Thus, $g(\mathbf{x})$ is convex.

c. Since f is twice differentiable,

$$\nabla g(\mathbf{x}) = L\mathbf{x} - \nabla f(\mathbf{x})$$
$$H_g(\mathbf{x}) = LI_n - H_f(\mathbf{x})$$

   Since we know that $g(\mathbf{x})$ is convex from (b), $H_g(\mathbf{x}) \geq 0$. Hence

$$H_f(\mathbf{x}) \leq LI_n$$