AM 221: Advanced Optimization	Spring 2018
Dr. Rasmus Kyng	*Problem Set 6 — Due Wednesday, Mar. 7th 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.
- For this homework, you will need to understand the section material from section on Mar. 2nd. You can find the notes here: `http://rasmuskyng.com/am221_spring18/sections/sec6.pdf`.

1. Subgradients. In this problem we consider a continuous convex function $f : \mathbb{R}^n \to \mathbb{R}$.

a. Show that the subdifferential $\partial f(\mathbf{x})$ of f at $\mathbf{x} \in \mathbb{R}^n$ is a closed convex set of $\mathbb{R}^n$.

b. Show that $\mathbf{x}^* \in \mathbb{R}^n$ is a minimizer of f ($i.e$ a solution to $\min_{\mathbf{x}\in\mathbb{R}^n} f(x)$) iff $0 \in \partial f(\mathbf{x})$.

2. Perceptron revisited. In this problem we will revisit the perceptron algorithm of Lecture 2 to find a separating hyperplane for a linearly separable dataset. The dataset $\mathcal{D}$ is a set of pairs $\mathcal{D} \stackrel{\text{def}}{=} \{(\mathbf{x}_i, y_i),\ 1 \le i \le n\}$ with $\mathbf{x}_i \in \mathbb{R}^{d-1}$ and $y_i \in \{0, 1\}$. We saw that after transformation of the data, finding a separating hyperplane amounts to finding $\mathbf{w} \in \mathbb{R}^d$ such that $\mathbf{w}^\intercal \mathbf{x}'_i > 0$ for all i, where the definition of $\mathbf{x}'_i$ using $\mathbf{x}_i$ and y_i is given in the lecture notes.

Let us define the following function:

$$f(\mathbf{w}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \max(0, -\mathbf{w}^\intercal \mathbf{x}'_i),\ \mathbf{w} \in \mathbb{R}^d$$

a. Show that f is convex and non-negative over $\mathbb{R}^d$. Is it differentiable?

b. Assume that the dataset $\mathcal{D}$ is linearly separable. Show that any $\mathbf{w}^* \in \mathbb{R}^d$ solution to:

$$\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

defines a separating hyperplane of $\mathcal{D}$. What is the value of f at $\mathbf{w}^*$?

c. Let us define:

$$f_i(\mathbf{w}) \stackrel{\text{def}}{=} \max(0, -\mathbf{w}^\intercal \mathbf{x}_i'),\ \mathbf{w} \in \mathbb{R}^d$$

and:

$$g_i(\mathbf{w}) = \begin{cases} 0 & \text{if } \mathbf{w}^\intercal \mathbf{x}_i' > 0 \\ -\mathbf{x}_i' & \text{otherwise} \end{cases}$$

show that for all $\mathbf{w} \in \mathbb{R}^d$, $g_i(\mathbf{w})$ is a subgradient of f_i at $\mathbf{w}$.

d. Describe the perceptron algorithm in the language of subgradients and the algorithms for convex optimization we saw in class. In particular, how would you describe the normalization by $\|\mathbf{x}_i'\|$ in the perceptron algorithm. Also note that each iteration of the perceptron focuses on one data point of the dataset at a time, can you draw an analogy with something we saw in class?

3. Gradient descent with weaker assumptions. In this problem we will analyze the gradient descent algorithm of Lecture 9 under weaker regularity assumptions. We consider a differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$ and only assume that f's gradient is L-Lipschitz continuous as introduced in the previous problem set:

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

We do **not** assume that f is twice-differentiable. Finally, we choose a constant step size $t = \frac{1}{L}$ instead of doing exact or backtracking line search.

a. Show that the L-Lipschitz continuous assumption on f's gradient implies the following quadratic upper bound on f:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2,\ \mathbf{x} \in \mathbb{R}^n,\ \mathbf{y} \in \mathbb{R}^n$$

Hint: use the fact you proved in Problem set 5 that $\frac{L}{2}\mathbf{x}^\intercal\mathbf{x} - f(\mathbf{x})$ is a convex function.

b. Let us denote by $\mathbf{x}^{(k)}$ the current solution at the kth iteration of the gradient descent algorithm. Show that:

$$f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^{(k)}) - \frac{1}{2L}\|\nabla f(\mathbf{x}^{(k)})\|^2,\ k \in \mathbb{N}$$

Show that this implies:

$$f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^{(k)})^\intercal(\mathbf{x}^{(k)} - \mathbf{x}^*) - \frac{1}{2L}\|\nabla f(\mathbf{x}^{(k)})\|^2,\ k \in \mathbb{N}$$

where $\mathbf{x}^*$ is a minimizer of f over $\mathbb{R}^n$.

c. Show that:

$$\nabla f(\mathbf{x}^{(k)})^\intercal(\mathbf{x}^{(k)} - \mathbf{x}^*) - \frac{1}{2L}\|\nabla f(\mathbf{x}^{(k)})\|^2 = \frac{L}{2}\left(\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|^2\right),\ k \in \mathbb{N}$$

hence, using b., we have:

$$f(\mathbf{x}^{(k+1)}) - f(\mathbf{x}^*) \leq \frac{L}{2}\left(\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|^2\right),\ k \in \mathbb{N}$$

d. Show that part c. implies:

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq \frac{L}{2k}\|\mathbf{x}^0 - \mathbf{x}^*\|^2,\ k \in \mathbb{N}$$

Given $\varepsilon > 0$, how many iterations of gradient descent are required to obtain $f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) < \varepsilon$? How does this compare to the strongly convex case?

4. Gradient descent, condition number, Newton's method. In this problem we will consider the following minimizing problem:

$$\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x}) \stackrel{\text{def}}{=} \min_{\mathbf{x}\in\mathbb{R}^2} \mathbf{x}^\intercal A\mathbf{x}$$

with:

$$A = \begin{pmatrix} 1+\lambda & 1-\lambda \\ 1-\lambda & 1+\lambda \end{pmatrix}$$

where λ is a real number with $\lambda \geq 1$.

a. Compute the gradient of f, its Hessian and its eigenvalues as a function of λ. What is the optimal solution of the above problem?

b. Implement the gradient descent algorithm for the above problem. Use backtracking line search as seen in section with $\alpha = 0.3$ and $\beta = 0.7$.

c. Run the gradient descent algorithm for several values of lambda between 1 and 10^5. For each value of λ, record the number of iterations required to reach a solution with error smaller than 10^{-10}. Choose $x^0 = [1.,\ 2.]$ as your starting point. Draw a plot of the number of iterations as a function of λ. How would you explain these results?

d. Implement Newton's method for the above problem. Use backtracking line search with $\alpha = 0.3$ and $\beta = 0.7$. For the same values of λ you used in c., show the number of the iterations required to reach the same error 10^{-10}. How would you explain those results?