The University of Texas at Austin
Department of Electrical and Computer Engineering

EE381K: Large Scale Optimization — Fall 2015

Problem Set Seven

Constantine Caramanis — Due: Thursday, November 12, 2015.

Related Reading. I find the notes by Sebastien Bubeck to be very well written. Chapter 3 in those notes covers much of the discussion of the last several lectures. You can find these here: `http://www.princeton.edu/~sbubeck/Bubeck14.pdf`.

Computational Problems

1. **Dual proximal gradient method for total variation de-noising**[1].

   In this problem we apply the proximal gradient method to the one-dimensional total variation de-noising problem

   $$\min_x \ \frac{1}{2}\,||x-\hat{x}||^2 + \mu \sum_{i=1}^{n-1} |x_{i+1}-x_i|,$$

   where $\hat{x}$ is a given noisy signal and $\mu \geq 0$. By introducing a new variable $y_i = x_{i+1} - x_i$ and defining $B_i = \mu\,[0,\ldots,0,-1,1,0,\ldots,0] \in \mathbb{R}^{1\times n}$ (B_i has an entry $-\mu$ in position i, an entry μ in position $i+1$, and zeros elsewhere), we can rewrite the problem as

   $$\begin{aligned} \min_{x,y}: \quad & \frac{1}{2}\|x-\hat{x}\|^2 + \mu \sum_{i=1}^{n-1} |y_i| \\ \text{s.t.}: \quad & Bx = y. \end{aligned}$$

   Denoise the corrupted one-dimensional signal `x_hat` provided in `tv_denoising.mat` using the following approach:

   (a) Derive the dual of this problem.

   (b) Solve the dual using the proximal gradient algorithm with a fixed step size.

   (c) With the dual solution, recover the solution of the primal problem. Use duality theory to formulate a stopping criterion that guarantees the returned value of x satisfies

   $$\frac{f(x)-f^*}{f^*} \leq 10^{-4}.$$

   (d) Plot the signal before and after denoising using $\mu = 0.2$. Examine the influence of μ on the convergence rate and on the quality of the computed x. Quality can be evaluated by inspection, the true underlying signal should be evident from the plot.

[1] This problem is adapted from an assignment in L. Vandenberghe's optimization course

2. **Accelerated Gradient**

   Compare the performance of gradient descent and Nesterov's accelerated gradient method on the logistic regression model, now with ℓ_2 regularization, using the data from `logistic-news.mat` (see Assignment 6 for details on logistic regression and the dataset).

$$\min_{\boldsymbol{\beta}} = \frac{1}{N}\sum_{i=1}^{N}\left(\beta_{y_i}^\top x_i + \log\sum_{j=1}^{C} e^{-\beta_j^\top x_i}\right) + \mu\,||\boldsymbol{\beta}||^2\,.$$

   (a) Find the value of μ that gives you (approximately) the best generalization performance (error on test set). What value do you get for the test loss after convergence?

   (b) Plot the loss against iterations for both the test and training data using the value of μ from part (a).

   (c) How do the two algorithms differ in performance, and how does this change as you decrease μ?

   (d) Explain the difference in convergence in terms of the condition number of the problem (note that the loss is μ-strongly convex).

3. **Proximal Gradient Descent**. Consider the third computational problem from Problem Set 1. Here you considered various optimization formulations for sparse regression, but you used CVX to solve them. Consider the LASSO formulation described in the third bullet of computational exercise 3. You now have two algorithms that can solve this formulation: the sub gradient method, and then also proximal gradient descent. Implement both and compare their convergence times.

Written Problems

1. Show that sub gradients have the following properties:[2]

   (a) $\partial(\alpha f(x)) = \alpha\partial f(x)$.

   (b) $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$.

   (c) If $g(x) = f(Ax + b)$, then $\partial g(x) = A^\top \partial f(Ax + b)$.

   (d) If $f(x) = \max_{1\le i\le m} f_i(x)$, then

$$\partial f(x) = \text{conv}\bigcup_i \{\partial f_i(x),\ f_i(x) = f(x)\}.$$

2. Compute the sub gradient of the $\|\cdot\|_{2,1}$ norm on matrices: For M a matrix with columns M_i, this is defined as:

$$\|M\|_{2,1} = \sum_i \|M_i\|_2.$$

3. (more tricky) Suppose $A_0, A_1, \ldots, A_m$ are symmetric matrices. Consider the function

$$f(x) = \lambda_{\max}(A(x)),$$

[2] We don't need convexity to define sub gradients, but we did use it in the definitions we gave in class. Therefore, in the problems below, you can assume that all functions are convex, coefficients nonnegative, etc.

where

$$A(x) = A_0 + x_1 A_1 + \cdots + x_m A_m.$$

Compute the sub gradient of $f(x)$. Hint: use the fact that

$$f(x) = \sup_{\|y\|_2=1} y^\top A(x) y,$$

and the last property you proved from the first problem.

4. The indicator function of a set $\mathcal{X}$ is defined as:

$$I_{\mathcal{X}}(x) = \begin{cases} 0 & x \in \mathcal{X} \\ +\infty & \text{otherwise.} \end{cases}$$

   (a) Show that the sub differential of $I_{\mathcal{X}}$ is the *normal cone* to $\mathcal{X}$ at the point x.

   (b) Now consider the constrained optimization problem:

$$\begin{aligned} \min: &\quad f(x) \\ \text{s.t.}: &\quad x \in \mathcal{X}, \end{aligned}$$

   for f and $\mathcal{X}$ convex. This can be rewritten as the equivalent unconstrained problem:

$$\min: \ f(x) + I_{\mathcal{X}}(x).$$

   This is again a convex function. Write down conditions for a point x^* to be optimal to the unconstrained problem.

5. (Monotonicity). Show that the subdifferential of a convex function $f(\cdot)$ is an example of what is known as a *monotone operator*. That is, show that

$$\langle u - v, x - y \rangle \geq 0, \quad \forall u \in \partial f(x), \ v \in \partial f(y).$$

6. Consider the ℓ_1-regularized regression problem

$$\min_x \quad \frac{1}{2}\|y - Ax\|_2^2 \ + \ \lambda \|x\|_1 \tag{1}$$

   Where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$. Show that a point $\bar{x}$ is an optimum of this problem if and only if there exists a $z \in \mathbb{R}^n$ such that both the following hold:
   (a) $-A^\top(y - A\bar{x}) + \lambda z = 0$
   (b) For every $i \in [n]$, $z_i = \text{sign}(\bar{x}_i)$ if $\bar{x}_i \neq 0$, and $|z_i| \leq 1$ if $\bar{x}_i = 0$.

7. In class we saw that the convergence of sub gradient descent is given by

$$f_{k,\text{best}} - f^* \ \leq \ \frac{R^2 + G^2 \sum_{i \leq k} h_i^2}{\sum_{i \leq k} h_i}$$

   where h_i were the step sizes. We also saw that, for a *fixed* k, the lowest this bound could be is $\frac{RG}{\sqrt{k+1}}$; this is achieved by choosing every $h_i = \frac{R/G}{\sqrt{k+1}}$. Thus, each k needs a different sequence to achieve the best lower bound.

   Suggest a *single* sequence of h_i's that makes the above bound decay as $O(\frac{\log k}{\sqrt{k}})$, for *every* k. Prove your result.