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

EE381V: Adv. Convex Optimization — Spring 2019

Problem Set Two

Constantine Caramanis Due: Tuesday, February 5, 2019.

Reading Assignments

1. (?) Read Chapter 3 in S. Bubeck's notes. Also read Chapter 3 in Nocedal and Wright, about backtracking line search. Note that an e-version of this book is available for free from UT Austin Libraries.

Computational Problems

Save your completed code in a file names `hw2.py`. Don't use stock optimization code, you should develop the core part of this assignment yourself. All plots should have titles, axis labels, and lines with different colors, markers, and legend labels.

1. **Solving the Lasso**

   Consider a least squares problem with ℓ^1 regularization:

   $$\min_{\boldsymbol{x}} \left[f(\boldsymbol{x}) = \frac{1}{2} \left|\left| A\boldsymbol{x} - \boldsymbol{b} \right|\right|_2^2 + \lambda \left|\left| \boldsymbol{x} \right|\right|_1 \right]$$

   This problem is often called LASSO (least absolute shrinkage and selection operator) and is known to induce *sparse* solutions with few nonzero elements in x, which can have advantages in terms of computation and interpretability. This problem is nonsmooth due to the regularization term. It is also not strongly convex when A has more columns than rows. We (i.e., you) will solve this problem using two different algorithms. The dataset represented in the matrices provided in the `numpy` binary files `A.npy` and `b.npy` are from a diabetes dataset [1] with 10 features that has been corrupted with an additional 90 noisy features. Thus a sparse solution should be very effective. See `hw2_lasso.py` for skeleton code to help with loading the data, running the algorithms and plotting the results, and save your completed code using the same file name. Again, please don't use stock optimization code, you should develop the core part of this assignment yourself.

   (a) Minimize $f(\boldsymbol{x})$ using 10^4 iterations of the subgradient method starting with $t = 0$ and $\boldsymbol{x}_0 = \boldsymbol{0}$. Use a decreasing step size of $\eta_t = c/\sqrt{t+1}$ with values for c that (roughly) optimize the empirical performance. Separately record the (unsquared) error $||A\boldsymbol{x}_t - \boldsymbol{b}||$ and the regularization term $||\boldsymbol{x}||_1$.

   (b) The Frank-Wolfe (or conditional gradient) algorithm minimizes a smooth function $f(x)$ subject to a convex constraint $x \in \mathcal{X}$ and is defined as follows:

[1] `scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html`

$$\begin{aligned}\boldsymbol{s}_t &= \arg\min_{\boldsymbol{s}\in\mathcal{X}} \langle \boldsymbol{s}, \nabla f(\boldsymbol{x}_t)\rangle \\ \boldsymbol{x}_{t+1} &= \boldsymbol{x}_t + \eta_t(\boldsymbol{s}_t - \boldsymbol{x}_t), \eta_t \in [0,1]\end{aligned}$$

When it is easy to minimize a linear function over $\mathcal{X}$, Frank-Wolfe (FW) has several advantages including that it often produces sparse iterates and does not require a projection step to stay within $\mathcal{X}$. In order to apply FW to the LASSO problem above, we can reformulate the problem as

$$\begin{array}{ll}\min\limits_{\boldsymbol{x}} : & \frac{1}{2}\left|\left|A\boldsymbol{x} - \boldsymbol{b}\right|\right|_2^2, \\ \text{s.t.} & \left|\left|\boldsymbol{x}\right|\right|_1 \leq \gamma.\end{array}$$

These problems are equivalent for a suitable pair of λ and γ. Run the FW algorithm for 10^4 steps using the stepsize schedule $\eta_t = 2/(t+2)$. Find a γ that performs well empirically (you might consider the ℓ^1 norm of your solution from part a). Again, separately record the (unsquared) error $||A\boldsymbol{x}_t - \boldsymbol{b}||$ and the regularization term $||\boldsymbol{x}||_1$.

(c) Repeat the experiment on both algorithms using some form of backtracking line search (BTLS) instead of a fixed stepsize schedule (see, e.g., Chapter 3, Algorithm 3.1 of Nocedal and Wright). Plot the error and ℓ^1 norm of both algorithms with both stepsize procedures on the same plot. There should be two plots (with titles) each having four lines with different colors, markers, and legend labels.

Written Problems

1. **Convex functions**

(a) If f_i are convex functions, show that $f(x) := \sup_i f_i(x)$ is also convex.

(b) Show that the largest eigenvalue of a symmetric matrix is a convex function of the matrix (i.e. $\lambda_{\max}(M)$ is a convex function of M). Is the same true for the eigenvalue of largest magnitude ?

(c) Consider a weighted graph with edge weight vector w. Fix two nodes a and b. The *weighted shortest path* from a to b is the path whose sum of edge weights is the minimum, among all paths with one endpoint at a and another at b. Let $f(w)$ be the weight of this path. Show that f is a concave function of w.

2. **Projection**.

(a) Suppose that $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed and bounded convex set. Let $\mathbf{y} \in \mathbb{R}^2$ be any point. The projection of $\mathbf{y}$ on $\mathcal{X}$ is defined by

$$\Pi_{\mathcal{X}}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{y} - \mathbf{x}\|_2^2.$$

Show that the solution to the optimization problem is unique.

(b) Show that for $\mathcal{X}$ as above, if $\mathbf{y} \notin \mathcal{X}$, there exists a hyperplane with $\mathcal{X}$ on one side, and $\mathbf{y}$ strictly on the other side. That is, show that there is a vector $\mathbf{s}$ and a scalar b, such $\langle \mathbf{s}, \mathbf{x}\rangle \leq b$ for all $\mathbf{x} \in \mathcal{X}$, and $\langle \mathbf{s}, \mathbf{y}\rangle > b$.

(c) Projected gradient descent is used for constrained optimization problems. Instead of just taking gradient steps, we follow up each gradient step with a projection onto the feasible set. That is, the update is given by:

$$x^{(k+1)} = \mathrm{Proj}_{\mathcal{X}}(x^{(k)} - t_k \nabla f(x^{(k)})).$$

Show that this is equivalent to the update:

$$x^{(k+1)} = \arg\min_{x \in \mathcal{X}} \left\{ \langle x, \nabla f(x^{(k)}) \rangle + \frac{1}{2t_k} \|x - x^{(k)}\|_2^2 \right\}.$$

3. **Computing Projections**. For the given convex set $\mathcal{X}$, compute the projection of a point z.

   (a) $\mathcal{X} = \mathbb{R}^n_+$.

   (b) Euclidean ball: $\{x : \|x\|_2 \leq 1\}$.

   (c) Positive semidefinite cone: $\mathbb{S}^n_+ = \{M \in \mathbb{S}^n : x^\top M x \geq 0, \ \forall x \in \mathbb{R}^n\}$. Assume for this problem that $z \in \mathbb{S}^n$.

   (d) $\mathcal{X}$ is a rectangle defined by vectors L and U that satisfy $U_i \geq L_i$. Thus, $\mathcal{X} = \{x : L_i \leq x_i \leq U_i, \ i = 1, \ldots, n\}$.

4. **Computing More Projections**. Now for two more tricky ones.

   (a) 1-norm ball: $\{x : \sum_i |x_i| \leq 1\}$.

   (b) Probability simplex: $\mathcal{X} = \{x : \sum_i x_i = 1, \ x_i \geq 0, \ i = 1, \ldots, n\}$.

5. Show that sub gradients have the following properties:[2] Note that the "$\subseteq$" inclusion can be much more tricky than the reverse inclusion to show. Do your best for this direction.

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

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

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

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

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

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

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.

[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.