STATS 608A, Fall 2015
Homework 2

Ambuj Tewari

Assigned: Nov 11, 2015. Due: Nov 23, 2015.

1 Properties of subdifferential (5 points)

Prove these properties directly using the definition of subdifferential.

1. $\partial f(x)$ is always a convex set
2. $\partial(f+g)(x) \supseteq \partial f(x) + \partial g(x)$ (+1 extra credit for proving the reverse inclusion perhaps under some additional mild conditions)
   (Note: If A, B are sets of vectors, we define $A + B = \{a + b \,:\, a \in A, b \in B\}$.)
3. If A is square and invertible, show that $g(x) = f(Ax)$ then $\partial g(x) = A^\top \partial f(Ax)$ (+1 extra credit for showing this for general, rectangular A)
4. $\partial(af)(x) = a\partial f(x)$ for $a \geq 0$
5. $\partial(\max_{i=1}^k f_i)(x) \supseteq \text{conv}\left(\cup_{i:f_i(x)=f(x)} \partial f_i(x)\right)$ (+1 extra credit for proving the reverse inclusion too)

2 Convergence analysis for inexact subgradient method (5 points)

Consider g to be an ϵ-subgradient of f at x, if for all y,

$$f(y) \geq f(x) + g^\top (y - x) - \epsilon.$$

Note that the standard subgradient is the same as an ϵ-subgradient with $\epsilon = 0$. Consider an inexact version of subgradient method

$$x^+ = x - tg$$

where g is an ϵ-subgradient of f at x. Show the following convergence guarantee:

$$f(x_{\text{best}}^{(k)}) - f^\star \leq \frac{R^2}{2kt} + \frac{(G+\epsilon)^2 t}{2} + \epsilon$$

where $R = \|x^{(0)} - x^\star\|_2$, G is the Lipschitz constant for f and $f(x_{\text{best}}^{(k)}) = \min_{i=0}^k f(x^{(i)})$.

3 Lipschitz constant for a regularized hinge loss minimization problem (5 points)

Consider the following (non-differentiable) objective function:

$$f(\beta) = \lambda \|\beta\|_2 + \frac{1}{n} \sum_{i=1}^n \max\{0, 1 - y_i \beta^\top x_i\}.$$

Establish the best possible upper bound you can for the Lipschitz constant of the function $f(\beta)$ in terms of properties of the design matrix $X \in \mathbb{R}^{n \times p}$ that has the n p-dimensional vectors x_i's as its rows. Note that this problem is in a binary classification context, i.e., the label $y_i \in \{-1, +1\}$.

4 Prox mappings (5 points)

Recall the prox mapping for a given function h is defined as

$$\text{prox}_t(x) = \underset{z}{\text{argmin}} \ \frac{1}{2t}\|x - z\|_2^2 + h(z).$$

Compute the prox mapping explicitly for the following cases.

1. $h(x) = \sum_{i=1}^{p} \lambda_i |x_i|$ (weighted lasso penalty)
2. $h(x) = \lambda\|x\|_2^2 + \mu\|x\|_1$ (elastic net penalty)
3. $h(x) = I_C(x), C = \{x : \|x\|_\infty \leq 1\}$ (indicator of ℓ_∞ ball)
4. $h(x) = I_C(x), C = \{x : \|x\|_2 \leq 1\}$ (indicator of ℓ_2 ball)
5. Let $x = (x_1, \ldots, x_k) \in \mathbb{R}^{kp}$ be partitioned into k blocks of dimension p each and let $h(x) = \lambda \sum_{b=1}^{k} \|x_b\|_2$ (group lasso penalty)

5 Computational Problem (5 points)

We will test the subgradient method, proximal gradient method, and accelerated proximal gradient method on the trace-norm regularized approach to matrix completion

$$f(B) = \frac{1}{2} \sum_{(i,j) \in \Omega} (Y_{i,j} - B_{i,j})^2 + \lambda\|B\|_{\text{tr}}$$

Generate the "true low rank" matrix $M \in \mathbb{R}^{n \times n}$ and the set Ω of observed entries as follows. Set $n = 200$ and rank $r = 2$ and generate two rectangular matrices $M_L, M_R \in \mathbb{R}^{n \times r}$ each with iid entries drawn from $\mathcal{N}(0, \sigma_n^2)$ with $\sigma_n^2 = 20/\sqrt{n}$. Set $M = M_L M_R^\top$. Let Ω be chosen uniformly at random from among all subsets of size $0.2n^2$ of the set of all n^2 entries, i.e., 20% entries are observed. Moreover, the observed entries are observed with noise as follows:

$$\forall (i,j) \in \Omega, \quad Y_{i,j} = M_{i,j} + Z_{i,j}$$

where $Z_{i,j}$ are iid draws from $\mathcal{N}(0, 1)$. Set $\lambda = \sqrt{0.4n}$.

Methods to test: (i) subgradient method with step size $t_k = c_0/\sqrt{k}$ for $c_0 = 0.1, 0.5$ and 1; (ii) proximal gradient method with $t_k = t = 1$ and (iii) accelerated proximal gradient method with $t_k = t = 1$.

Stopping rule: stop your optimization once the relative change $\|B^{(k+1)} - B^{(k)}\|_F / \|B^{(k)}\|_F$ falls below 10^{-3} in successive iterations.

Submit 2 plots. In plot 1, show the objective function vs. number of iterations for all 5 methods. In plot 2, show the objective function vs. time taken for all 5 methods.

Answer the following questions:

1. What is the starting value of the objective function (start every method at $B^{(0)} = \mathbf{0}_{n \times n}$)?
2. What is the ending value of the objective function for each method?
3. For each of the 5 methods, what was the rank of the final B matrix at the end of optimization?
4. For each of the 5 methods, what was the RMSE error $\sqrt{\frac{\|B - M\|_F^2}{n^2}}$?