Mathematical Introduction to Data Science **Dec 16, 2013**

Homework 9. Regularized M-estimators in High Dimensional Statistics

Instructor: Yuan Yao *Due: Tuesday Dec 23, 2013*

The questions below marked by $\star$ will be optional.

1. *Variable Selection by LASSO:* Use the following matlab codes to simulate a n-by-p measurement equation:

$$b = Ax + \varepsilon$$

```
%% dimensions and sparsity
n = 10; % # of rows of A
p = 20; % # of columns of A
s = 4; % sparsity

%% generate random measurement matrix
A = randn(n,p);
A = zscore(A)/sqrt(n-1);
u_ref = zeros(p,1);
u_ref(randsample(p,s)) = round(10*rand(s,1)+1); % true sparse signal
supp_ref = find(u_ref); % true support

sigma = 1; % noise-level
e = sigma*randn(n,1)/sqrt(n); % error
b = A*u_ref + e; % measurements
```

(a) Try to find an estimator using LASSO with $\lambda = c\sigma\sqrt{\log p/n}$ with $c \in (0, 1)$.

(b) Try Least Angle Regression (LARs, R-package 'glmnet') to find regularization paths.

(c) Compare solution paths of Orthogonal Matching Pursuit (OMP) and LASSO.

The idea of OMP is very simple, which recursively adds the column of A_i which has maximal correlation with the residue, i.e. the following procedure

(a) Input: A, b, active set $S_0 = \emptyset$, $x_0 = 0$, and $r_0 = b$

(b) Output: x_t

(c) Let $r_t = b - Ax_t$;

(d) $i_t = \arg\max_{i \in S_t^c} \langle A_i, r_t \rangle / \|A_i\|_2$

(e) $S_t = S_{t-1} \cup \{i_t\}$

(f) $x_t = \arg\min_x \|A_{S_t} x - b\|^2$ (or equivalently, $x_t = (A_{S_t}^T A_{S_t})^{-1} A_{S_t}^T r_t$)

2. * *RSC and l_2-consistency*: Consider the regularized empirical risk minimization

$$\min_{\theta\in\Theta} \mathcal{L}(\theta, z_1^n) + \lambda_n R(\theta)$$

whose minimizer is $\hat{\theta}_{\lambda_n}$. Let the minimizer of population risk be $\theta^* \in \arg\min_{\theta\in\mathbb{R}^p} \mathbb{E}_{z_1^n}\mathcal{L}(\theta, z_1^n) \in \mathcal{M}$, for some subspace $\mathcal{M} \in \mathbb{R}^p$. Assume that

(A1) (Decomposable Regularizer) there is a subspace $\overline{\mathcal{M}} \subseteq \mathbb{R}^p$ which contains $\mathcal{M}$, such that $R(\alpha+\beta) = R(\alpha) + R(\beta)$ for $\alpha \in \mathcal{M}$ and $\beta \in \overline{\mathcal{M}}^\perp$;

(A2) (Restricted Strongly Convex Loss) for all $\Delta \in C(\mathcal{M}, \overline{\mathcal{M}}^\perp; \theta^*) := \{\Delta \in \mathbb{R}^p : R(\Delta_{\overline{\mathcal{M}}^\perp}) \le 3R(\Delta_{\overline{\mathcal{M}}})\}$,

$$\mathcal{L}(\theta^*+\Delta) - \mathcal{L}(\theta^*) - \langle \nabla\mathcal{L}(\theta^*), \Delta\rangle \ge \gamma\|\Delta\|_2^2 - \tau_n^2 R^2(\Delta), \quad \gamma > 0, \tau_n \ge 0$$

(A3) (Lipschitz Regularizer) for any subspace $\overline{\mathcal{M}} \subseteq \mathbb{R}^p$,

$$\Psi(\overline{\mathcal{M}}) := \sup_{\theta\in\overline{\mathcal{M}}\setminus\{0\}} \frac{R(\theta)}{\|\theta\|_2} < \infty$$

Show that

(a) If (A1) holds and the dual regularizer satisfies $\lambda_n \ge 2R^*(\nabla\mathcal{L}(\theta^*))$, then $\hat{\Delta} := \hat{\theta}_{\lambda_n} - \theta^* \in C(\mathcal{M}, \overline{\mathcal{M}}^\perp; \theta^*)$, i.e.

$$R(\hat{\Delta}_{\overline{\mathcal{M}}^\perp}) \le 3R(\hat{\Delta}_{\overline{\mathcal{M}}})$$

(note: use $|\langle u, v\rangle| \le R(u)\cdot R^*(v)$, decomposability and triangle inequality for $R(\theta)$.)

(b) Moreover if (A2-A3) hold, and for large enough n such that $16\tau_n^2\Psi^2(\overline{M})) \le \gamma/2$, then

$$\|\Delta\|_2 \le \frac{3\lambda_n}{\gamma}\Psi(\overline{\mathcal{M}})$$

(note: A2 is reduced to $\mathcal{L}(\theta^*+\Delta) - \mathcal{L}(\theta^*) - \langle\nabla\mathcal{L}(\theta^*), \Delta\rangle \ge \frac{\gamma}{2}\|\Delta\|_2^2$ for $\Delta \in C(\mathcal{M}, \overline{\mathcal{M}}^\perp; \theta^*)$)

(c) Moreover if for all $\Delta \in C(\mathcal{M}, \overline{\mathcal{M}}^\perp; \theta^*)$,

$$R^*(\nabla\mathcal{L}(\theta^*+\hat{\Delta}) - \nabla\mathcal{L}(\theta^*)) \ge \gamma R^*(\Delta)$$

then

$$R^*(\hat{\theta}_{\lambda_n} - \theta^*) \le \frac{3}{2\gamma}\lambda_n$$

(note: Consider the first order KKT condition and triangle inequality for $R^*(u)$)

(d) apply the results above to standard LASSO

$$\min_\theta \frac{1}{2n}\|Y - X\theta\|_2^2 + \lambda_n\|\theta\|_1$$

with $\tau_n = 0$, $\lambda_n = 2\sigma\sqrt{\frac{\log p}{n}}$ and $\Psi(\overline{\mathcal{M}}) = \sqrt{s}$, and derive the bounds

$$\|\hat{\theta}_{\lambda_n} - \theta^*\|_2 \le O(\sqrt{\frac{s\log p}{n}})$$

and

$$\|\hat{\theta}_{\lambda_n} - \theta^*\|_\infty \leq O(\sqrt{\frac{\log p}{n}})$$

(Reference: A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, Negahban, Ravikumar, Wainwright, Yu, 2010)