Mathematics for Data Sciences **Date here**

Lecture 5. OMP, BP, and LASSO

Instructor: Yuan Yao, Peking University *Scribe: Zhan, Ruohan and Zhu, Weizhi*

1 Introduction to Compressed Sensing

Assuming $A \in \mathbb{R}^{n \times p}$, $b \in \mathbb{R}^p$, and $x^* \in \mathbb{R}^p$ be the solution of the following problem: $b = Ax$ (noise-free), or $b = Ax + \epsilon$ (noisy). We are interested in under which conditions we can recover x^*, and how we can do it when $k = |S| = |supp(x^*)| \ll n < p$.

A simple idea is to consider:

$$P_0 : \quad \min \|x\|_0, \quad s.t. \quad Ax = b. \tag{1}$$

which, in fact, is NP-hard.

Now we formally give some algorithms to solve the primal problem, without loss of generality, we assume each column of design matrix A has being standardized , that is , $\|A_j\|_2 = 1$, $j = 1, ..., p$.

Algorithm 1 Orthogonal Matching Pursuit,Mallet-Zhang,1993

initial $r_0 = b$, $x_0 = 0$, $S_0 = \emptyset$
repeat
$j_t = argmax_{1 \le j \le p} | < A_j, r_{t-1} > |$
$S_t = S_{t-1} \cup j_t$
$x_t = argmin_{x \in \mathbb{R}^p} \|b - A_{s_t} x\|$
$r_t = b - A x_t$
until $\|r_t\|_2 = 0$

Remarks:

1. OMP choose the column of maximal correlation with residue, in fact, it's also the one having the steepest decline in residue, which implies OMP is greedy.

2. In noisy case, we stop algorithm until $\|r_t\| \le \varepsilon$ for a given ε.

3. It's natural to ask how well OMP can recover x^* , the answer is yes under some conditions we will talk below.

Algorithm2 Basis Pursuit,Donoho,1999

$$P_1 : \quad \min \|x\|_1, \quad s.t. \quad Ax = b. \tag{2}$$

$$P_1' : \quad \min \|x\|_1, \quad s.t. \quad \|Ax - b\|_2 \le \lambda. \tag{3}$$

This is a convex relaxation of (P_0) , some equivalent conditions with (P_0) will be discussed later.

Algorithm3 LASSO, Tibshirani,1996

$$P_2 : \quad \min \frac{1}{2n} \|Ax - b\|_2 + \lambda \|x\|_1. \tag{4}$$

Some algorithms could be applied to (P_2) , for example, proximal gradient method.

Algorithm4 Dantzig Selector,Candes-Tao,2007

$$P_1: \quad \min \|x\|_1, \quad s.t. \quad \|A^*(b-Ax)\|_\infty \leq \lambda. \tag{5}$$

Now we turn to consider the conditions under which the algorithms before can recover x^*.

Uniqueness condition
$A_s^* A_s \geq rI$, for some $r > 0$.

Irrepresentable condition(Yu-Zhao, 2006**)**
$M =: \|A_{S^c}^* A_S (A_S^* A_S)^{-1}\|_\infty < 1$

Incoherence (Donoho-Huo, 2001**)**

Definition. Coherence: $\mu = \max_{i \neq j} | < A_i, A_j > |$.

Restricted-Isometry-Property(R.I.P.)(Candés-Recht-Tao, 2006**)**
All k-sparse $x \in \mathbb{R}^p$, $\exists \delta_k \in (0,1)$, s.t. $(1-\delta_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta_k)\|x\|_2^2$.

Remarks:

1. Uniqueness condition is a basic one, without which we can't even know which x^* we're going to recover.

2. Irrepresentable condition describe the relevance between A_S and A_{S^c} should be controlled. However, we may regard rows of $A_{S^c}^* A_S (A_S^* A_S)^{-1}$ to be the regression coefficient of $A_j = A_S \beta + \varepsilon$, for $j \in S^c$.

3. In fact, Irrepresentable condition could not be verified before we already have x^*, Incoherence condition cover the shortage of that by the following lemma.

Lemma 1.1. (Tropp, 2004)

$$\mu < \frac{1}{2k-1} \Rightarrow M \leq \frac{k\mu}{1-(k-1)\mu} < 1. \tag{6}$$

4. R.I.P condition is not easily to verified. But **Johnson-Lindestrauss Lemma** says some suitable random matrices will satisfy R.I.P. with high probability.

Proof. (of Lemma1.1) First, we have

$$M = \|A_{S^c}^* A_S (A_S^* A_S)^{-1}\|_\infty \leq \|(A_S^* A_S)^{-1}\|_\infty \|A_{S^c}^* A_S\|_\infty. \tag{7}$$

It's easy to verify that

$$\|A_{S^c}^* A_S\|_\infty \leq k\mu. \tag{8}$$

Then we consider $\|(A_S^* A_S)^{-1}\|_\infty$.

Decompose $A_S^* A_S = I_k + \Delta$, then

$$
\begin{aligned}
& \max|\Delta_{i,j}| \leq \mu, \quad \mathrm{diag}(\Delta) = \mathbf{0}; \\
\Rightarrow \quad & \|\Delta\|_\infty \leq \frac{k-1}{2k-1} < 1; \\
\Rightarrow \quad & (A_S^* A_S)^{-1} = (I_k + \Delta)^{-1} = \sum_{j=0}^{\infty}(-\Delta)^j; \\
\Rightarrow \quad & \|(A_S^* A_S)^{-1}\|_\infty = \|\sum_{j=0}^{\infty}(-\Delta)^j\|_\infty \leq \sum_{j=0}^{\infty}\|\Delta\|_\infty^j = \frac{1}{1-\|\Delta\|_\infty} \leq \frac{1}{1-(k-1)\mu}.
\end{aligned} \tag{9}
$$

Thus, we reach our conclusion

$$M \leq \frac{k\mu}{1-(k-1)\mu}. \tag{10}$$

□

Theorems

Theorem 1.2. (Tropp, 2014) Under uniqueness and Irrepresentable conditions , OMP and BP recovers x^*.

Proof. (I) OMP recovers x^*.

The key to the proof is to show that at each step $t \leq k$, OMP selects atom from S rather than S^c. Then we only need to examine

$$\rho(r_t) = \frac{\|A_{S^c}^* r_t\|_\infty}{\|A_S^T r_t\|_\infty} < 1. \tag{11}$$

In noise-free case,

$$\left.\begin{aligned} b = Ax^* \in im(A_S) \\ r_t = b - Ax_t \in im(A_S) \end{aligned}\right\} \Rightarrow r_t \in im(A_S). \tag{12}$$

$P_S = A_S(A_S^* A_S)^{-1}A_S^*$ is the projection operator onto $im(A_S)$, thus we have $r_t = P_S r_t$. Hence,

$$\rho(r_t) = \frac{\|A_{S^c}^*(P_S r_t)\|_\infty}{\|A_S^* r_t\|_\infty} = \frac{\|A_{S^c}^* A_S(A_S^* A_S)^{-1}A_S^* r_t\|_\infty}{\|A_S^* r_t\|_\infty} \leq \|A_{S^c}^* A_S(A_S^* A_S)^{-1}\|_\infty < 1. \tag{13}$$

(II) BP recovers x^*.

Assume $\hat{x} \neq x^*$ solves

$$P_1: \quad \min\|x\|_1, \quad s.t. \quad Ax = b. \tag{14}$$

Denote $\hat{S} = \mathrm{supp}(\hat{x})$ and $\hat{S} \setminus S \neq \emptyset$. We have

$$
\begin{aligned}
\|x^*\|_1 &= \|(A_S^* A_S)^{-1}A_S^* b\|_1 \\
&= \|(A_S^* A_S)^{-1}A_S^* A_{\hat{S}}\hat{x}_{\hat{S}}\|_1 \quad (A\hat{x} = b) \\
&= \|(A_S^* A_S)^{-1}A_S^* A_S\hat{x}_S + (A_S^* A_S)^{-1}A_S^* A_{\hat{S}\setminus S}\hat{x}_{\hat{S}\setminus S}\|_1 \quad (\hat{x}_{\hat{S}} = \hat{x}_S + \hat{x}_{\hat{S}\setminus S}) \\
&< \|\hat{x}_S\|_1 + \|\hat{x}_{\hat{S}\setminus S}\|_1 = \|\hat{x}_{\hat{S}}\|_1,
\end{aligned} \tag{15}
$$

which is a contradictory. □

Theorem 1.3. The following conclusions hold,

1. $\delta_{2k} < 1 \Rightarrow$ BP can recover has x^* which is a unique solution of P_0;
2. $\delta_{2k} < \sqrt{2} - 1 \Rightarrow$ BP can recover x^* which is a unique solution of P_1.

Note. How LASSO works?

LASSO:

$$\min_x \frac{1}{2}\|b - Ax\|_2^2 + \lambda\|x\|_1 \tag{16}$$

Sufficient Conditions:

- $M \leq 1 - \eta$;
- $\min_{i\in S} |x_i^*| \geq \frac{\|b\|_\infty}{\eta\gamma}(1+\eta)\sqrt{k}$.

Proof. Suppose $\hat{x}$ solve the LASSO problem 16. According to KKT condition, we have

$$\lambda p(\hat{x}) = A^T(b - A\hat{x}), \quad p(\hat{x}) \in \partial\|\hat{x}\|_1. \tag{17}$$

The sign consistency at $(\hat{\lambda}, \hat{x})$ implies

$$\hat{\lambda}\text{sign}(x_S^*) = A_S^T(b - A_S\hat{x}) \tag{18}$$

$$\|A_{S^C}^T(b - A_S\hat{x})\|_\infty \leq \hat{\lambda} \tag{19}$$

Combine equation 18 with $b = A_S x^* + \epsilon$, we have

$$\hat{x} = x^* - \hat{\lambda}(A_S^T A_S)^{-1}\text{sign}(x_S^*) + (A_S^T A_S)^{-1}A_S^T\epsilon. \tag{20}$$

Replace $\hat{x}$ in 19 with 20 , we get

$$\begin{aligned} &\|A_{S^c}^T A_S x^* + A_{S^c}^T\epsilon - A_{S_c}^T A_S x^* - A_{S^c}^T A_S(A_S^T A_S)^{-1}A_S^T\epsilon + \lambda A_{S^c}^T A_S(A_S^T A_S)^{-1}\text{sign}(x_S^*)\|_\infty \leq \hat{\lambda} \\ \Leftrightarrow \quad &\|A_{S^c}^T(I - P_S)\epsilon + \hat{\lambda}A_{S^C}^T A_S(A_S^T A_S)^{-1}\text{sign}(x_S^*)\|_\infty \leq \hat{\lambda} \end{aligned} \tag{21}$$

Since $M \leq 1 - \eta$, we have

$$\|A_{S^c}^T A_S(A_S^T A_S)^{-1}\text{sign}(x_S^*)\|_\infty \leq 1 - \eta \quad (M \leq 1 - \eta). \tag{22}$$

Then it suffices

$$\|A_{S^c}^T(I - P_S)\epsilon\|_\infty < \hat{\lambda}\eta. \tag{23}$$

Since ϵ is Gaussian noise, usually,

$$\|A_S^T\epsilon\|_\infty < \|b\|_\infty, \quad \text{w.h.p.} \tag{24}$$

so $\hat{\lambda} = \frac{\|b\|_\infty}{\eta}$, where $\|b\|_\infty = c\sigma\sqrt{\log p}$. So 19 is satisfied.

from 18, $\text{sign}(x_S^*) = \text{sign}(\hat{x}_S)$. So we have

$$\text{sign}(\hat{x}_S) = \text{sign}(x^* - \hat{\lambda}(A_S^T A_S)^{-1}\text{sign}(x_S^*) + (A_S^T A_S)^{-1}A_S^T\epsilon),$$

which requires

$$\min_{i\in S} |x_i^*| > \|\hat{\lambda}(A_S^T A_S)^{-1}\text{sign}(x_S^*) - (A_S^T A_S)^{-1}A_S^T\epsilon\|_\infty, \tag{25}$$

We now have:

$$\|\hat{\lambda}(A_S^T A_S)^{-1}\text{sign}(x_S^*)\|_\infty \leq \hat{\lambda}\frac{\sqrt{k}}{\gamma}, \quad \|(A_S^T A_S)^{-1}A_S^T\epsilon\|_\infty \leq \frac{\hat{\lambda}\eta\sqrt{k}}{\gamma}. \tag{26}$$

Therefore, we only need

$$\min |x_i^*| > \frac{\hat{\lambda}(1+\eta)\sqrt{k}}{\gamma}. \tag{27}$$

□