Mathematics for Data Sciences **December 16th, 2014**

Lecture 13. Compressed Sensing and High Dimensional Statistics

Instructor: Yuan Yao, Peking University *Scribe: Xiaowei Wang, Yongyi Guo and Zhimei Ren*

1 Introduction

In this section we mainly discuss about compressed sensing and high dimensional statistics. The notion of *compressed sensing*, also known as *compressive sampling*, is introduced in 2006 by Emmanuel Candes [2] and David Donoho [1] who both set up the foundations of this field. The problem can be traced back to as early as the 1950s, when l_1 *norm* is taken into consideration to deal with sparse signal detection. In 1996, S.S. Chen, Donoho and M. Saunders proposed the problem of *Basis Pursuit* [3] as well as its denoting version. In this paper, the following problem is stated:

Let $x^* \in R^p$ be sparse but unknown. Here being sparse means $\#\{x_i^* \neq 0\} = k < p$. We can recover x^* by solving the following problem with linear measurements $b = Ax^*$ which is noise-free and underdetermined ($n \leq p$):

P_0:

$$\begin{aligned} min\ &\|x\|_0 := \#\{x_i \neq 0\} \\ s.t.\quad &Ax = b \end{aligned}$$

which is NP-hard. Donoho et al. then turns this problem into a "convex relaxation":

P_1:

$$\begin{aligned} min\ &\|x\|_1 \\ s.t.\quad &Ax = b \end{aligned}$$

and asked when the solution of P_0 uniquely meets x^* and when the solution of P_1 meets that of P_0. In [1] they present a sufficient condition based on mutual incoherence defined as the maximal correlation between columns of A. In 2004, Joel Tropp [7] gives the exact recovery condition for $\hat{x} = x^*$, which are necessary and sufficient simultaneously for both BP and OMP (Orthogonal Matching Pursuit), another greedy sparse recovery algorithm. Such a condition, is later called as *incoherence* or *irrepresentable condition* especially in noisy settings. 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

1. Input: A, b, active set $S_0 = \emptyset$, $x_0 = 0$, and $r_0 = b$
2. Output: x_t
3. Let $r_t = b - Ax_t$;
4. $i_t = \arg\max_{i \in S_t^c} \langle A_i, r_t \rangle$
5. $S_t = S_{t-1} \cup \{i_t\}$
6. $x_t = \arg\min_x \|A_{S_t} x - b\|^2$ (or equivalently, $x_t = (A_{S_t}^T A_{S_t})^{-1} A_{S_t}^T r_t$)

When noisy measurements are taken, $b = Ax^* + \varepsilon$, Donoho et al. suggests the following *Basis Pursuit DeNoising* (BPDN) problem as constraint optimization:

$P_{1,\delta}$:

$$\begin{aligned} & min \ \|x\|_1 \\ s.t. \quad & \|Ax - b\|_2^2 \leq \delta \end{aligned}$$

where δ is a regularization parameter decided by noise etc.

Such a constraint optimization is inconvenient to solve in practice, though welcomes theoretical analysis. Simultaneously in 1996, Tibshirani introduced the method of LASSO (Least Absolute Shrinkage Selection Operator) [4] as an unconstraint optimization problem to deal with noisy measurements:

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

where λ is a regularization parameter. The solution of LASSO $\hat{x}$ embodies sparsity. Three questions concerning the recovery are raised then:

- $l_2 - norm$ consistency: $\|\hat{x} - x^*\|_2 \leq ?$
- Model-selection consistency:supp($\hat{x}$)=supp(x^*)?
- Sign-consistency:sign($\hat{x}$)=sign(x^*)?

In 2007, Candes and Tao proposed the *Dantzig Selector* [6] in the following format:

$$\begin{aligned} & min \ \|x\|_1 \\ s.t. \quad & \|A^T(Ax - b)\|_\infty \leq \delta \end{aligned}$$

In 2009, P.J.Bickel, Y Ritov and AB Psybakov solved the l_2 consistency problem in his paper [5], which simultaneously analyzes LASSO and Dantzig Selector.

Model selection consistency is typically achieved via sign-consistency, which however requires much stronger condition than l_2 consistency. There are several independent papers dealing with sign-consistency or model-selection consistency in noisy settings, which are all based on a tightening of the exact recovery condition in [7]. For example, Zhao and Yu (2006) [8] called it as irrepresentative condition, among others such as Yuan and Lin (2006) [9], Zou (2007), and Wainwright's paper in 2009 [10]. Incidentally, the l_2 *-consistency* can be deduced from the *Model-selection consistency*.

2 Noise Free Condition

We first discuss when $\hat{x} = x^*$ under noise free condition. Recall that x^* is the solutio of P_0 and $\hat{x}$ is the solution of the P_1, we have:

$$\begin{aligned} & Ax^* = b = A\hat{x} \qquad \Rightarrow \qquad A(x^* - \hat{x}) \\ & \|\hat{x}\|_1 \leq \|x^*\|_1 \end{aligned} \tag{1}$$

Let $\Delta = \hat{x} - x^*$ and $S = supp(x^*), S^c = \{1, ..., p\} - S$. Then we have:

$$\|x^*\|_1 = \|x_s^*\|_1 \geq \|\hat{x}_s\|_1 + \|\hat{x}_{S^c}\|_1$$

Meanwhile,

$$\|\Delta_{S^c}\|_1 = \|\hat{x}_{S^c}\|_1 \leq \|x^*_S\|_1 - \|\hat{X}_s\|_1 \leq \|\Delta_s\|_1$$

That is to say:

$$\|\Delta_{S^c}\|_1 \leq \|\Delta_s\|_1 \tag{2}$$

Notice that this equation represents a cone. With (1) and (2) we have the following:

$$\hat{x} = x^* \Rightarrow Ker(S) \cap C_{x^*} = \{0\}$$

Here, C_{x^*} denotes the cone:$\|\Delta_{S^c}\|_1 \leq \|\Delta_s\|_1$. In fact, this is the *necessary and sufficient* condition for P_1 to exactly recover x^*.

2.1 RIP(Restrict Isometry Property)

How to ensure the condition above? Candès and Tao proposed the restricted isometry property (RIP) as a sufficient condition. First we regard $A \approx I_p$ and consider $\frac{\|Ax\|_2}{\|x\|_2}$, where $A \in \mathbb{R}^{n\times p}, n \leq p$ and we have:

$$1 - \delta_k \leq \frac{\|Ax_{[k]}\|_2}{\|x_{[k]}\|_2} \leq 1 + \delta_k$$

Here $x_{[k]}$ means x is sparse, with at most k nonzero entries. As can be deduced, P_0 is NP-hard and under RIP, x^* can be exactly recovered if and only if $\delta_{2k} < 1$. On the contrary, P_1 is an linear programming program and with RIP, x^* can be recovered iff $\delta_{2k} \leq \sqrt{2} - 1$ (Tony Cai et al.).

How to construct matrices A satisfying RIP? Most deterministic matrices fail for this, but *random matrix A* typically works – for instance, elements of A are independent and subject to normal or sub-Gaussian distribution. A summary concerning this fact is written by Vershynin.

But joint normal distributed random matrices might not satisfy RIP. A typical example is constructed by Raskutti et al.

$$A \sim \mathcal{N}(0, \Sigma)$$

with

$$\Sigma = (1-\mu)I + \mu 1 \cdot 1^T.$$

In this case, with high probability

$$\left\| \frac{A^T A}{n} - I_d \right\| \geq \mu(k-1) + (1-\epsilon).$$

Therefore as k increases, δ_k might go unbounded. So RIP fails here.

To get exact recovery under more general conditions than $\delta_{2k} \leq \sqrt{2} - 1$, one natural idea is to transform P_0 into a non-convex problem

$$P_q : \min \|x\|_q^q := \sum_i |x_i|^q \quad (0 < q < 1)$$
$$subject\ to\ Ax = b$$

A good thing about this transform is that x^* can be exactly recovered iff $\delta_{2k} < C$, where the constant $C > \sqrt{2} - 1$ (but so far we do not know the exact value of C yet). However, an evident shortcoming of this idea lies in that when $q < 1$, P_q is *strongly* NP-hard [13] and has been proved to have no polynomial approximating algorithms. In fact, if the penalty function is strongly concave at point 0 (with non-trivial curvature, as is shown in figure 1), then similar troubles will always exist. Thus we hope to find every local optimal for solution.

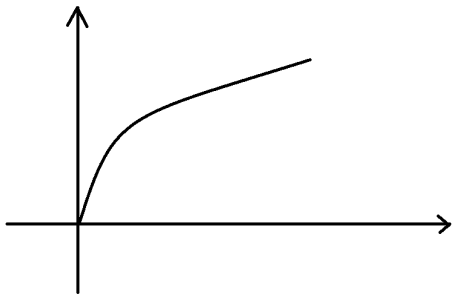

Figure 1: example of a function with non-trivial curvature

2.2 RE condition

Now let's turn to the noise induced case:

$$b = Ax^* + \epsilon.$$

It is impossible to get exact recovery in this setting, so our goal becomes that

$$\|\hat{x} - x^*\|_2$$

be bounded.

Compared with RIP, we are going to introduce Restricted Eigenvalue condition here. Firstly, construct a matrix A_S consisting columns of A whose indices are in signal support set S. Assume the following condition holds,

$$\frac{\|A\Delta\|^2}{2n} \geq \gamma\|\Delta\|_2^2 \tag{3}$$

when Δ ranges in the cone

$$\|\Delta_{S^c}\|_1 \leq \alpha\|\Delta_S\|_1. \tag{4}$$

This is equivalent to say that the Hessian of $\frac{1}{2n}\|Ax - b\|^2$ becomes

$$[\frac{\partial^2 L}{\partial x_i \partial x_j}] = \frac{A^T A}{n}$$

which is positive definite when restricting on the cone $\|\Delta_{S^c}\|_1 \leq \alpha\|\Delta_S\|_1$. When such a condition holds, we shall see that the LASSO solution $\hat{x}$, satisfies

$$\|\hat{x} - x^*\|_2 \leq \frac{C\lambda\sqrt{k}}{r}, \tag{5}$$

when the correlation of noise and measuring matrix are small enough such that

$$\|\frac{A^T}{n} \cdot \epsilon\|_\infty \leq \frac{\lambda}{2}.$$

Similarly,

$$\|\hat{x} - x^*\|_1 \leq \frac{Ck\lambda}{\gamma}.$$

An intuitive understanding of RE is that γ described the curvature of quadratic loss in the restricted cone. The larger is the curvature, the smaller is the error (see figure 2).

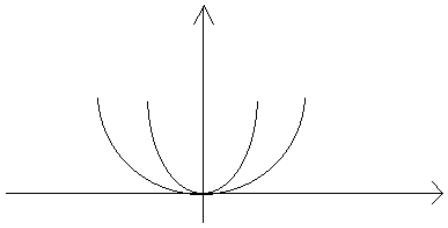

Figure 2: an illustration

For example, under Gaussian noise in application, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, the bound can be controlled within the bound

$$\|\frac{A^T}{n} \cdot \epsilon\|_\infty \leq C\delta \sqrt{\frac{\log p}{n}}.$$

How to derive the l_2 consistency under RE condition? The following reasoning gives you a fast glimpse. Since $\hat{x}$ is the optimal value to minimize the target value in LASSO, noticing $b = Ax^* + \epsilon$, we have

$$\begin{aligned}
0 &\geq (\frac{1}{2n}\|A\hat{x} - b\|_2^2 + \lambda\|\hat{x}\|_1) - (\frac{1}{2n}\|Ax^* - b\|_2^2 + \lambda\|x^*\|_1) \\
&= \frac{1}{2n}\|A(\hat{x} - x^*) - \epsilon\|_2^2 + \lambda\|\hat{x}\|_1 - \frac{1}{2n}\|\epsilon\|_2^2 - \lambda\|x^*\|_1 \\
&= \frac{1}{2n}\|A\Delta\|_2^2 + \left\langle \frac{A^T\epsilon}{n}, \Delta \right\rangle + \lambda\|\hat{x}\|_1 - \lambda\|x^*\|_1 \\
&\geq \frac{1}{2n}\|A\Delta\|_2^2 + \left\langle \frac{A^T\epsilon}{n}, \Delta \right\rangle - \lambda\|\Delta_S\|_1 + \lambda\|\Delta_{S^C}\|_1
\end{aligned}$$

Under the assumption $\|\frac{A^T\epsilon}{n}\|_\infty \leq \frac{\lambda}{2}$, we have

$$\left\langle \frac{A^T\epsilon}{n}, \Delta \right\rangle \geq -\frac{\lambda}{2}\|\Delta\|_1 = -\frac{\lambda}{2}(\|\Delta_S\|_1 + \|\Delta_{S^C}\|_1)$$

Using the RE condition $\frac{\|A^T\Delta\|^2}{2n} \geq r\|\Delta\|_2^2$, we have

$$\begin{aligned}
r\|\Delta\|_2^2 \leq \frac{1}{2n}\|A\Delta\|_2^2 &\leq \frac{3}{2}\lambda\|\Delta_S\|_1 - \frac{1}{2}\lambda\|\Delta_{S^C}\|_1 \\
&\leq \frac{3}{2}\lambda\|\Delta_S\|_1 \leq \frac{3}{2}\lambda\sqrt{|s|}\|\Delta_S\|_2 = \frac{3\lambda\sqrt{k}}{2}\|\Delta_S\|_2
\end{aligned}$$

Thus, the equations above indicates

$$\|\Delta\|_2 \leq \frac{3\lambda}{2r}\sqrt{k}$$

Under Gaussian noise, this bound holds with high probability.

Furthermore, researchers have proved that whenever the RIP condition holds, the RE condition shall hold. This can be found in the S.Vander Geer's paper. Therefore, whenever the recovery problem can be solved under noise-free condition, it shall be solvable under noisy condition.

2.3 RSC condition

There exists certain situations where LASSO analysis makes sense even when RIP condition doesn't hold. For instance, in the previous example

$$A \sim N(0, \Sigma), \Sigma = (1-\mu)I + \mu \mathbf{1} \cdot \mathbf{1}^T$$

RIP doesn't hold, however, we can still use LASSO to find a good estimator.

Here we give a condition under which the LASSO analysis should make sense. It is called the RSC (Restricted Strongly Convex) condition, proposed by Raskutti, Wainwright and Bin Yu [11].

When row vector $A_i \sim N(0, \Sigma)$, we have

$$\frac{\|A\Delta\|_2}{\sqrt{n}} \geq \lambda_{min}(\Sigma)\|\Delta\|_2 - ck_\Sigma \sqrt{\frac{\log p}{n}}\|\Delta\|_1$$

where $k := \max_i |\Sigma_{ii}|$.

In this situation, if $\|\frac{A^T\epsilon}{n}\|_\infty \leq \frac{\lambda}{2}$ holds,we can ensure that

$$\|\Delta\|_2 \leq \frac{c\lambda\sqrt{k}}{2}$$

This result can be extended into other situations, for instance, sub-Gaussian case and non-convex case are prospective candidates for extension.

3 The Bias of LASSO and Non-convex Regularization

LASSO estimator is biased and to remove bias, it is necessary to introduce non-convex regularization [12].

To see this, for simplicity assume that $n = p$,

$$A = I$$
$$b = x^* + \epsilon, \qquad \epsilon \sim N(0, I)$$

In this case, Johnstone and Donoho showed that

$$E\|\hat{x} - x^*\|_2^2 = o(\sqrt{\frac{k \log \frac{p}{k}}{n}})^2$$

which is minimax optimal. However, such an error bound hides a fact that the estimator $\hat{x}$ is biased. Bias might not be bad in terms of mean square error, such as what we have seen in Stein's phenomenon before. To see the bias, let $n = 1, p = 1$ and $b > 0$

$$L(x) = \frac{1}{2}\|b - x\|_2^2 + \lambda\|x\|_1$$

$\hat{x}$ is a optimum if it satisfies the first order KKT condition $0 \in \partial L(\hat{x})$, i.e.

$$0 = -(b - \hat{x}) + \lambda sign(\hat{x})$$

Solving this gives

$$\hat{x} = \begin{cases} x^* - \lambda sign(x^*) + \epsilon, x^* > b - \lambda > 0 \\ 0, \quad x^* \leq b - \lambda \end{cases}$$

Here $E(\hat{x}) = x^* - \lambda sign(x^*) \neq x^*$. Therefore the estimator is biased. To remove the bias, a general non-convex regularization is proposed

$$L(x) = \frac{1}{2}||b - x||_2^2 + \lambda p(x)$$

whose KKT condition gives

$$0 = -(b - \hat{x}) + \lambda \partial p(\hat{x})$$

One can see that when $|\hat{x}| \sim b >> 0$, one can remove bias by setting $\partial p(x) = 0$. This implies that $p(x)$ must level-off on approaching infinity, hence is non-convex. For example, the q-norm: $p(x) = ||x||_q^q$, where $q \in (0, 1)$. This is a strongly NP-hard problem.

Recently, Loh and Wainwright (2014) shows that l_2 consistency can be achieved at RSC when $p(x)$ satisfies the following condition among others:

$$p(x) + \mu||x||_2^2 \text{ is convex for some } \gamma > \mu > 0,$$

Sign-consistency and model-consistency conditions are much stronger than RSC.

4 Extensions to Matrix Estimation (Completion)

Compressed sensing schemes above can be extended to matrix completion (estimation) when x is a matrix and each row $A_i^T x$ represents a matrix-matrix inner product

$$<< A_i, X >>= b_i$$

and l_1-norm of x is replaced by matrix nuclear norm or others (max-norm) etc.

For example, in matrix completion problem, one can arbitrarily select any element in X. Tony Cai and Anru Zhang recently studies rank-one measurements A_i.

References

[1] Donoho, David L. "Compressed sensing." Information Theory, IEEE Transactions on 52.4 (2006): 1289-1306.

[2] E. J. Candès. Compressive sampling. *Proceedings of the International Congress of Mathematicians*, Madrid, Spain, 2006.

[3] Chen, Shaobing, and David Donoho. "Basis pursuit." Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on. Vol. 1. IEEE, 1994.

[4] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological) (1996): 267-288.

[5] Bickel, Peter J., Ya'acov Ritov, and Alexandre B. Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector." The Annals of Statistics (2009): 1705-1732.

[6] Candes, Emmanuel, and Terence Tao. "The Dantzig selector: Statistical estimation when p is much larger than n." The Annals of Statistics (2007): 2313-2351.

[7] Tropp, Joel A. "Greed is good: Algorithmic results for sparse approximation." Information Theory, IEEE Transactions on 50.10 (2004): 2231-2242.

[8] Zhao, Peng, and Bin Yu. "On model selection consistency of Lasso." The Journal of Machine Learning Research 7 (2006): 2541-2563.

[9] Yuan, Ming, and Yi Lin. "Model selection and estimation in regression with grouped variables." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68.1 (2006): 49-67.

[10] Wainwright, Martin J. "Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (Lasso)." Information Theory, IEEE Transactions on 55.5 (2009): 2183-2202.

[11] G. Raskutti, M. J. Wainwright and B. Yu (2010). Restricted nullspace and eigenvalue properties for correlated Gaussian designs. Journal of Machine Learning Research, 11:2241-2259, August 2010.

[12] Jianqing Fan and Runze Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of American Statistical Association (2001), 1348–1360.

[13] X. Chen, D. Ge, Z. Wang and Y. Ye, Complexity of Unconstrained $L_2 - L_p$ Minimization, Math. Programming, 143 (2014), 371-383.