EE 381V: Large Scale Learning **Spring 2013**

Lecture 20 — March 28

Lecturer: Caramanis & Sanghavi *Scribe: Zheng Lu*

20.1 Reviews of last lecture

In Lecture 18, we begin to study the methods for sparse recovery. Recall the basic problem is $y = W \cdot x$, where x is an $n \times 1$ high dimensional sparse data vector, W is an $m \times n$ measurement matrix with $m << n$, and y is an $m \times 1$ observation vector. The objective of sparse recovery can be described as: given (y, W), find x such that $||x||_0 \leq S$, where $||x||_0$ is the number of non-zero elements in x.

We have briefly introduced three types of algorithms to perform sparse recovery:

- Recovery via l_1-norm minimization also known as Compressed Sensing;
- A greedy algorithm also known as Orthogonal Matching Pursuit (OMP);
- Iterative methods including iterative soft thresholding and iterative hard thrsholding.

We will discuss theses algorithms in more details in the next few lectures. Our plan is to study l_1-norm minimization in this lecture, OMP in the next lecture and then iterative methods.

20.2 Recovery via l_1-norm minimization

The problem of sparse recovery via l_1-norm minimization is often referred to as compressed sensing. A useful reference is [1] which can be found on the course website. Like we discussed earlier, the original sparse recovery problem can be described as the following Problem (20.1):

$$\text{Given } (y, W), \text{ find } x, \text{ such that } ||x||_0 \leq S \text{ and } y = Wx. \tag{20.1}$$

Since this problem is not easy to solve in general, compressed sensing instead tries to solve an optimization problem and obtain an estimation $\hat{x}$ of the true solution x. The optimization is defined as the following Problem (20.2):

$$\begin{aligned} \underset{\tilde{x}}{\text{minimize}} \quad & ||\tilde{x}||_1 \\ \text{subject to} \quad & y = W\tilde{x}. \end{aligned} \tag{20.2}$$

Note in the remaining of this lecture we will always refer to x as the (true) solution to Problem (20.1), and $\hat{x}$ as the solution to Problem (20.2).

Our hope is that we can solve the sparse recovery problem by solving compressed sensing, i.e., $\hat{x} = x$. However we quickly realize that this is not true in general, since Problem (20.2) always has a feasible solution (actually it can be written as a linear programming) while Problem (20.1) can be infeasible to solve or it can have multiple solutions. Thus in order for $\hat{x} = x$ to hold, we need to add some conditions on x and W.

Before we describe the exact conditions we need, let us see some intuitions first. Basically what we want here are the uniqueness of the sparse solution x and the hope of sparse recovery by compressed sensing.

- **Uniqueness**. If x_1 and x_2 are two solutions to Problem (20.1), then we need $x_1 = x_2$ so that we are guaranteed to have only one solution x, which is the true solution.
- **Hope of recovery**. Note that if W is such that $Wx = 0$, i.e., x is in the null space of W, then there is no hope of finding a non-zero solution using compressed sensing, since the solution of Problem (20.2) will always be $\hat{x} = 0$. Thus for sparse x not falling in the null space of W, we need some conditions that "preserves" sparse vectors, for example,

$$1 - \epsilon \leq \frac{||W\tilde{x}||_2}{||\tilde{x}||_2}, \quad \forall \tilde{x} \text{ such that } ||\tilde{x}||_0 \leq S.$$

Now let us define **Restricted Isometry Property (RIP)** which is a sufficient condition needed for both uniqueness of sparse solution and sparse recovery by compressed sensing.

Definition 1. *A matrix W satisfies (ϵ, S)-RIP if*

$$1 - \epsilon \leq \frac{||W\tilde{x}||_2}{||\tilde{x}||_2} \leq 1 + \epsilon, \quad \forall \tilde{x} \textit{ such that } ||\tilde{x}||_0 \leq S.$$

Here are some remarks on the definition of RIP.

- The upper bound $1 + \epsilon$ is required in analysis, but we are not sure whether it is fundamentally required.
- Under special conditions $\epsilon = 0$, $S = n$, the above definition implies that W is an isometry on $\mathbb{R}^n$. This is why we use the term "restricted isometry".

We will first show uniqueness of sparse solution under RIP.

Lemma 20.1. *If W satisfies $(\epsilon, 2S)$-RIP for any $\epsilon < 1$, then $y = Wx$ cannot have two S-sparse solutions.*

Proof: Suppose there exist two S-sparse solutions x_1 and x_2. Let $\tilde{x} = x_1 - x_2$ and note that $||W\tilde{x}||_2 = 0$. Also note that $\tilde{x}$ is a $2S$-sparse vector, thus by RIP, we have $\frac{||W\tilde{x}||_2}{||\tilde{x}||_2} \geq 1 - \epsilon > 0$, which implies $\frac{0}{||\tilde{x}||_2} > 0$. Thus there must be $\tilde{x} = 0$, i.e., $x_1 = x_2$. □

Note that Lemma 20.1 guarantees that under the conditions stated in the lemma, Problem (20.1) has a unique solution x, which is the true solution. Then we can state the following theorem.

Theorem 20.2. *If $||x||_0 \leq S$ and W satisfies $(\epsilon, 2S)$-RIP for $\epsilon < \sqrt{2}-1$, then $\hat{x} = x$.*

We will not prove Theorem 20.2, but instead we will prove a stronger theorem which is describe as follows:

Theorem 20.3. *If W satisfies $(\epsilon, 2S)$-RIP with $\epsilon < \sqrt{2}-1$, then*

$$||x - \hat{x}||_2 \leq \frac{2}{\sqrt{S}(1-\rho)}||x - x_S||_1,$$

where $\rho = \frac{\sqrt{2}\epsilon}{1-\epsilon}$, and x_S is the largest S entries of x.

Proof: Let $\tilde{x}$ be any vector, and $h = \tilde{x} - x$. Define disjoint index sets $T_i \in [n]$ as follows:

$T_0 =$ largest (in magnitude) S elements of vector x;
$T_1 =$ largest (in magnitude) S elements of vector $h_{T_0^C}$;
$T_2 =$ largest (in magnitude) S elements of vector $h_{T_{0,1}^C}$;
......

Note that T_0 is an index set on vector x, and by its definition we have $x_{T_0} = x_S$. But for $i \geq 1$, T_i is an index set on vector h. Under these definitions, we can state two claims which will be proved in Lemma 20.4 and Lemma 20.5 respectively.

- **Claim 1**. For any $\tilde{x}$ such that $||\tilde{x}||_1 \leq ||x||_1$, we have

$$||h_{T_{0,1}^C}||_2 \leq ||h_{T_0}||_2 + \frac{2}{\sqrt{S}}||x - x_{T_0}||_1.$$

- **Claim 2**. For any $\tilde{x}$ such that $W\tilde{x} = Wx$, and W satisfies $(\epsilon, 2S)$-RIP with $\epsilon < \sqrt{2}-1$, then

$$||h_{T_{0,1}}||_2 \leq \frac{\rho}{1-\rho}\frac{1}{\sqrt{S}}||x - x_{T_0}||_1,$$

  where $\rho = \frac{\sqrt{2}\epsilon}{1-\epsilon}$.

Note that Claim 1 involves the l_1-norm minimization feature, but it does not have any RIP requirements on W. Claim 2 does not involve any l_1-norm minimization feature, but it requires RIP on W. Now if we take $\tilde{x}$ to be the solution $\hat{x}$ to the compressed sensing

Problem (20.2), then the conditions of Claim 1 and Claim 2 are both satisfied, and thus we have:

$$\begin{aligned}||x-\hat{x}||_2 = ||h||_2 &\leq ||h_{T_{0,1}}||_2 + ||h_{T_{0,1}^C}||_2 \text{ (by triangle inequality)},\\ &\leq 2||h_{T_{0,1}}||_2 + \frac{2}{\sqrt{S}}||x-x_S||_1 \text{ (by Claim 1 and the fact that } ||h_{T_0}||_2 \leq ||h_{T_{0,1}}||_2),\\ &\leq 2(\frac{\rho}{1-\rho}+1)\frac{1}{\sqrt{S}}||x-x_S||_1 \text{ (by Claim 2)},\\ &\leq \frac{2}{\sqrt{S}(1-\rho)}||x-x_S||_1.\end{aligned}$$

□

Here are the two lemmas for the two claims in the proof above.

Lemma 20.4. *With the definitions in the proof of Theorem 20.3, for any $\tilde{x}$ such that $||\tilde{x}||_1 \leq ||x||_1$, we have*

$$||h_{T_{0,1}^C}||_2 \leq ||h_{T_0}||_2 + \frac{2}{\sqrt{S}}||x - x_{T_0}||_1.$$

Proof: Take $j > 1$, for any $i \in T_j$ and $i' \in T_{j-1}$ we have $|h_i| \leq |h_{i'}|$. Therefore, $||h_{T_j}||_\infty \leq \frac{||h_{T_{j-1}}||_1}{S}$. Thus we have:

$$||h_{T_j}||_2 \leq \sqrt{S}||h_{T_j}||_\infty \leq \frac{1}{\sqrt{S}}||h_{T_{j-1}}||_1.$$

Summing the above over $j = 2, 3, ...$ and using the triangle inequality we obtain:

$$||h_{T_{0,1}^C}||_2 \leq \sum_{j\leq 2}||h_{T_j}||_2 \leq \frac{1}{\sqrt{S}}||h_{T_0^C}||_1. \tag{20.3}$$

Since $||\tilde{x}||_1 \leq ||x||_1$, using triangle inequality we have:

$$||x||_1 \geq ||x+h||_1 = \sum_{i\in T_0}|x_i+h_i| + \sum_{i\in T_0^C}|x_i+h_i| \geq ||x_{T_0}||_1 - ||h_{T_0}||_1 + ||h_{T_0^C}||_1 - ||x_{T_0^C}||_1, \tag{20.4}$$

and since $||x_{T_0^C}||_1 = ||x-x_S||_1 = ||x||_1 - ||x_{T_0}||_1$, Equation (20.4) can be further written as:

$$||h_{T_0^C}||_1 \leq ||h_{T_0}||_1 + 2||x_{T_0^C}||_1. \tag{20.5}$$

Combining Equation (20.3) and (20.5), we obtain:

$$||h_{T_{0,1}^C}||_2 \leq \frac{1}{\sqrt{S}}(||h_{T_0}||_1 + 2||x_{T_0^C}||_1) \leq ||h_{T_0}||_2 + \frac{2}{\sqrt{S}}||x_{T_0^C}||_1.$$

□

Lemma 20.5. *With the definitions in the proof of Theorem 20.3, for any $\tilde{x}$ such that $W\tilde{x} = Wx$, and W satisfies $(\epsilon, 2S)$-RIP with $\epsilon < \sqrt{2} - 1$, then*

$$||h_{T_{0,1}}||_2 \leq \frac{\rho}{1-\rho}\frac{1}{\sqrt{S}}||x - x_{T_0}||_1,$$

where $\rho = \frac{\sqrt{2}\epsilon}{1-\epsilon}$.

Proof: Since vector $h_{T_{0,1}}$ is $2S$-sparse, we can use the RIP condition to get:

$$(1-\epsilon)||h_{T_{0,1}}||_2^2 \leq ||Wh_{T_{0,1}}||_2^2. \tag{20.6}$$

Note that $Wh_{T_{0,1}} = Wh - \sum_{j\geq 2} Wh_{T_j} = -\sum_{j\geq 2} Wh_{T_j}$, thus we have:

$$||Wh_{T_{0,1}}||_2^2 = -\sum_{j\geq 2}\langle Wh_{T_{0,1}}, Wh_{T_j}\rangle = -\sum_{j\geq 2}\langle Wh_{T_0} + Wh_{T_1}, Wh_{T_j}\rangle. \tag{20.7}$$

Note that for all i and j, $i \neq j$, $\frac{h_{T_i}}{||h_{T_i}||_2} + \frac{h_{T_j}}{||h_{T_j}||_2}$ and $\frac{h_{T_i}}{||h_{T_i}||_2} - \frac{h_{T_j}}{||h_{T_j}||_2}$ are $2S$-sparse, and then by RIP we have:

$$||\frac{Wh_{T_i}}{||h_{T_i}||_2} + \frac{Wh_{T_j}}{||h_{T_j}||_2}||_2^2 \leq (1+\epsilon)(||\frac{h_{T_i}}{||h_{T_i}||_2}||_2^2 + ||\frac{h_{T_j}}{||h_{T_j}||_2}||_2^2) = 2(1+\epsilon),$$

and

$$-||\frac{Wh_{T_i}}{||h_{T_i}||_2} - \frac{Wh_{T_j}}{||h_{T_j}||_2}||_2^2 \leq -(1-\epsilon)(||\frac{h_{T_i}}{||h_{T_i}||_2}||_2^2 + ||\frac{h_{T_j}}{||h_{T_j}||_2}||_2^2) = -2(1-\epsilon).$$

Thus we have:

$$|\langle\frac{Wh_{T_i}}{||h_{T_i}||_2}, \frac{Wh_{T_j}}{||h_{T_j}||_2}\rangle| = |\frac{||\frac{Wh_{T_i}}{||h_{T_i}||_2} + \frac{Wh_{T_j}}{||h_{T_j}||_2}||_2^2 - ||\frac{Wh_{T_i}}{||h_{T_i}||_2} - \frac{Wh_{T_j}}{||h_{T_j}||_2}||_2^2}{4}| \leq \epsilon,$$

which implies:

$$|\langle Wh_{T_i}, Wh_{T_j}\rangle| \leq \epsilon||h_{T_i}||_2||h_{T_j}||_2.$$

Take the above equation into Equation (20.7), we obtain:

$$||Wh_{T_{0,1}}||_2^2 \leq \epsilon(||h_{T_0}||_2 + ||h_{T_1}||_2)\sum_{j\geq 2}||h_{T_j}||_2 \leq \sqrt{2}\epsilon||h_{T_{0,1}}||_2\sum_{j\geq 2}||h_{T_j}||_2. \tag{20.8}$$

Combining Equation (20.3) (20.6) and (20.8), we obtain:

$$(1-\epsilon)||h_{T_{0,1}}||_2^2 \leq \sqrt{2}\epsilon||h_{T_{0,1}}||_2\frac{1}{\sqrt{S}}||h_{T_0^C}||_1.$$

Using Equation (20.5) and rearranging, we get:

$$||h_{T_{0,1}}||_2 \leq \frac{\rho}{\sqrt{S}}(||h_{T_0}||_1 + 2||x_{T_0^C}||_1) \leq \rho||h_{T_0}||_2 + \frac{2\rho}{\sqrt{S}}||x_{T_0^C}||_1,$$

but since $||h_{T_0}||_2 \leq ||h_{T_{0,1}}||_2$, we have:

$$||h_{T_{0,1}}||_2 \leq \frac{\rho}{1-\rho}\frac{1}{\sqrt{S}}||x - x_{T_0}||_1.$$

$\square$

Bibliography

[1] Shai Shalev-Shwartz. Compressed sensing: Basic results and self contained proofs. in manuscript, 2009.