IIIS 2014 Spring: ATCS - Selected Topics in Optimization
Lecture date: Apr 2nd, 2014
Instructor: Jian Li **Scribe: Hao Fu**

1 Introduction

Sparsity has gained a lot of interest in recent years in applied mathematics and in particular signal processing. This is due to the crucial observation that many types of signals can be well-represented by a small number of non-vanishing coefficients with respect to a suitable basis, that is, by a sparse representation. Indeed, this is the main reason why compression schemes such as JPEG or MPEG work so well.

Suppose that we observe

$$y = Ax, \tag{1}$$

where $x \in \mathbb{R}^n$ is an object we wish to reconstruct, $y \in \mathbb{R}^m$ are available measurements, and A is a known $m \times n$ matrix. Here, we are interested in the underdetermined case with fewer equations than unknowns, i.e. $m < n$, and ask whether it is possible to reconstruct x with good accuracy. As such, the problem is of course ill-posed but suppose now that x is known to be sparse or nearly sparse in the sense that it depends on a smaller number of unknown parameters. This premise radically changes the problem, making the search for solution feasible.

The following notation will be very useful. You can skin it and come to next section if you know them.

Definition 1 (Norm) *There are special norms which define by*

$$\|x\|_1 := \sum_{i=1}^{n} |x_i|, \quad \|x\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2} \quad \textit{and} \quad \|x\|_\infty := \max_i |x_i|,$$

where $x \in \mathbb{R}^n$.

We also write that

$$\|x\|_0 = \{\# \text{ nonzeros in } x\}.$$

Notice that it isn't a norm. We denote a set of vectors with constant K by

$$\Sigma_K = \{x : \|x\|_0 \le K\}.$$

Definition 2 (Nullspace) *Denote the null space of matrix* A *by*

$$\mathcal{N}(A) = \{x : Az = 0\}.$$

2 The spark

Definition 3 *The spark of a given matrix A is the smallest number of columns of A that are linearly dependent.*

This definition allows us to pose the following straightforward guarantee.

Theorem 4 *For any vector $y \in \mathbb{R}^m$, there exists at most one signal $x \in \Sigma_K$ such that $y = Ax$ if and only if $spark(A) > 2K$.*

Proof: We first assume that, for any $y \in \mathbb{R}^m$, there exists at most one signal $x \in \Sigma_K$ such that $y = Ax$. Now suppose for the sake of contradiction that $spark(A) \leq 2K$. This means that there exists some set of at most $2K$ columns that are linearly dependent, which implies that there exists an $h \in \mathcal{N}(A)$ such that $h \in \Sigma_{2K}$. In this case, we can write $h = x - x'$, where $x, x' \in \Sigma_K$. Thus, since $h \in \mathcal{N}(A)$ we have that $A(x - x') = 0$ and hence $Ax = Ax'$. But this contradicts our assumption when $y = 0$. Therefore, we must have that $spark(A) > 2K$.

Now suppose that $spark(A) > 2K$. Assume that for some y there exist $x, x' \in \Sigma_K$ such that $y = Ax = Ax'$. We therefore have that $A(x - x') = 0$. Let $h = x - x'$. Then we have $Ah = 0$. Since $spark(A) > 2K$, all sets of up to $2K$ columns of A are linearly independent, and therefore $h = 0$, which implies $x = x'$, proving the theorem. □

3 The null space property (NSP)

Definition 5 *For $\Lambda \subseteq [n]$, define $\Lambda^c = [n] \setminus \Lambda$. Denote x_Λ as $x_{\Lambda,i} = x_i$ if $i \in \Lambda$, and $x_{\Lambda,i} = 0$ if $i \notin \Lambda$.*

Definition 6 *[Null Space Property (NSP)] The matrix A satisfies the NSP of order K if there exists a constant $C > 0$ such that*

$$\|h_\Lambda\|_2 \leq C \frac{\|h_{\Lambda^c}\|_1}{\sqrt{K}} \tag{2}$$

for all $h \in \mathcal{N}(A)$ and for all Λ such that $|\Lambda| \leq K$.

The NSP quantifies the notion that vextors in the null space of A should not be too concentrated on a small subset of indices. For example, if a vector h is exactly K-sparse, then there exists a Λ such that $\|h_{\Lambda^c}\|_1 = 0$ and so (2) implies that $h_\Lambda = 0$ as well. Thus, if a matrix A satisfies the NSP then the only K-sparse vector in $\mathcal{N}(A)$ is $h = 0$.

To fully illustrate the implications of the NSP in the context of sparse recovery, we now briefly dicuss how we will measure the performance of sparse recovery algorthems when dealing with general non-sparse x. Let $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ be the recovery algorthem. We will focus primarily on guarantees of the form

$$\|\Delta(Ax) - x\|_2 \leq C \frac{\delta_K(x)_1}{\sqrt{K}} \tag{3}$$

for all x, where we recall that

$$\delta_K(x)_p = \min_{\hat{x} \in \Sigma_K} \|x - \hat{x}\|_p.$$

We will show that the NSP of order $2K$ is sufficient to establish a guarantee of the form (3) for a pratical recovery algorithm (l_1 minimization). Morevoer

Theorem 7 *Let $A : \mathbb{R}^n \to \mathbb{R}^m$ denote a sensing matrix and $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ denote an arbitrary recovery algorithem. Then A satisfies NSP of order $2K$ if and only if there exists a pair (A, Δ) satisfies (3).*

Proof: We just prove the necessity. Suppose the pair (A, Δ) satisfies (3) and $h \in \mathcal{N}(A)$. Let Λ be the indices corresponding to the $2K$ larrest entries of h. W next split Λ into Λ_0 and Λ_1, where $|\Lambda_0| = |\Lambda_1| = K$. Set $x = h_{\Lambda_1} + h_{\Lambda^c}$ and $x' = -h_{\Lambda_0}$, so that $h = x - x'$. Since by construction $x' \in \Sigma_K$, we can apply (3) to obtain $x' = \Delta(Ax')$. Moreover, since $h \in \mathcal{N}(A)$, we have

$$Ah = A(x - x') = 0,$$

so that $Ax' = Ax$. Thus $x' = \Delta(Ax') = \Delta(Ax)$. Then, we have

$$\|h_\Lambda\|_2 \leq \|h\|_2 = \|x - x'\|_2 = \|x - \Delta(Ax)\|_2 \leq C\frac{\delta_K(x)_1}{\sqrt{K}} = \sqrt{2}C\frac{\|h_{\Lambda^c}\|}{\sqrt{2K}},$$

where the last inequality follows from (3). Last equation holds becuase

$$\delta_K(x)_1 = \min_{\hat{x}\in\Sigma_K} \|x - \hat{x}\|_1 = \min_{\hat{x}\in\Sigma_K} \|h_{\Lambda_1} + h_{\Lambda^c} - \hat{x}\|_1 = \|h_{\Lambda^c}\|_1,$$

where Λ is the indices corresponding to the $2K$ larrest entries of h. □

4 The restricted isometry property (RIP)

The NSP do not consider for *noise*. When the measurements are contaminated with noise or have been corrupted by some error such as quantization, it will be useful to consider somewhat stronger conditions. In [5], Candés and Tao introduced the following isometry condition on matrices A.

Definition 8 (Restricted Isometry Property (RIP)) *A matrix A satifies the RIP of order K if there exists a $\delta_K \in (0, 1)$ such that*

$$(1 - \delta_K)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_K)\|x\|_2^2, \tag{4}$$

for all $x \in \Sigma_K$.

As dissution at [4], we have following theorem which discovered by Candés, Romberg and Tao.

Theorem 9 *Suppose A satifies the RIP of order $4K$ with $\delta_{4K^ʿ} \leq \frac{1}{4}$. Denote l_0 minimization and l_1 minimization:*

$$x = \arg\min_z\{\|z\|_0 : Az = y\}$$
$$\hat{x} = \arg\min_z\{\|z\|_1 : Az = y\}.$$

If $\|x\|_0 \leq K$, then the solution of two problem are the same, i.e. $x = \hat{x}$.

Proof: Let $h = \hat{x} - x$, then $h \in \mathcal{N}(A)$. Let Λ be the indices corresponding to nonzero entries of x, i.e. Λ =Supp$(x) = \{i|x_i \neq 0\}$ (which deduces $|\Lambda| \leq K$). Then we have

$$\|x\|_1 \geq \|\hat{x}\|_1 = \|h_{x+\Lambda}\|_1 + \|h_{\Lambda^c}\|_1 \geq \|x\|_1 - \|h_\Lambda\|_1 + \|h_{\Lambda^c}\|_1,$$

which implied

$$\|h_{\Lambda^c}\|_1 \leq \|h_\Lambda\|_1. \tag{5}$$

We begin by dividing Λ^c into subsets of size $3K$ and enumerate Λ^c as $a_1, a_2, \ldots, a_{n-|\Lambda|}$ in decreading order of magnitude of h_{Λ^c}. Set $\Lambda_j = \{a_i, 3K(j-1)+1 \leq i \leq 3Kj\}$. That is, Λ_1 contains the indices of the $3K$ largest of coeffients of h_{Λ^c}, Λ_2 contains the indices of the next $3K$ largest coeffients, and so on. Then we have following claim.

Claim 10

$$\sum_{j\geq 2} \|h_{\Lambda_j}\| \leq \frac{1}{\sqrt{3K}} \|h_{\Lambda^c}\|_1 \tag{6}$$

Then, by the (5) and the (6), we have

$$\sum_{j\geq 2} \|h_{\Lambda_j}\|_2 \leq \frac{1}{\sqrt{3K}} \|h_{\Lambda^c}\|_1 \leq \frac{1}{\sqrt{3K}} \|h_\Lambda\|_1 \leq \frac{\sqrt{|\Lambda|}}{\sqrt{3K}} \|h_\Lambda\|_2 \leq \frac{1}{\sqrt{3}} \|h_\Lambda\|_2.$$

Hence, we have

$$\begin{aligned}
0 = \|A\hat{x} - Ax\| = \|Ah\|_2 &= \|A(h_\Lambda + h_{\Lambda_1} + \sum_{j\geq 2} h_{\Lambda_j})\|_2 \\
&\geq \|A(h_\Lambda + h_{\Lambda_1}\|_2 - \sum_{j\geq 2} \|Ah_{\Lambda_j}\|_2 \\
&\geq (1-\delta_{4k})\|h_\Lambda + h_{\Lambda_1}\|_2 - (1+\delta_{4K}) \sum_{j\geq 2} \|h_{\Lambda_j}\|_2 \\
&\geq (1-\delta_{4k})\|h_\Lambda\|_2 - \frac{1}{\sqrt{3}}(1+\delta_{4k})\|h_\Lambda\|_2 \\
&\geq 0,
\end{aligned}$$

where we use the assumption that $\delta_{4k} \leq \frac{1}{4}$ in the last inequality.

The both sides of inequality are equal, which decuce we can get all equality instead of inequality. This implies that $\|h_\Lambda\|_2 = 0$. So $x_\Lambda = x'_\Lambda$, then $x = x'$. □

Proof:[The proof of claim 10] By the construction of Λ_j, we have

$$\|h_{\Lambda_{j+1}}\|_2 \leq \sqrt{3K} \|h_{\Lambda_{j+1}}\|_\infty \leq \frac{1}{\sqrt{3K}} \|h_{\Lambda_j}\|_1,$$

when $j = 1, 2, \ldots$

To sum all the inequality, we have

$$\sum_{j\geq 2} \|h_{\Lambda_j}\|_2 \leq \sum_{j\geq 1} \frac{1}{\sqrt{3K}} \|h_{\Lambda_j}\|_1 = \frac{1}{\sqrt{3K}} \|h_{\Lambda^c}\|_1.$$

□

We now turn to the question of how to construct matrices that satisfy the RIP. It is possible to deterministically construct matrices of size m, n that satisfy the RIP of order K, but such constructions also require m to be relatively large. Fortunately, these limitations can be overcome by randomizing the matrix construction. We will construct our random matrices as follows: given m and n, generate random matrices A by choosing the entries A_ij as independent realizations from some probability distribution.

5 Moreover

Definition 11 (Mutual Coherence) *The Mutual Coherence of matrix A is defined as*

$$\mu(A) = \max_{1 \le i \ne j \le n} \frac{|\langle a_i, a_j \rangle|}{\|a_i\| \|a_j\|},$$

where $a_1, a_2, \ldots, a_n \in \mathcal{R}^m$ are the colunms of matrix A.

Then we have following lemma without proof.

Lemma 12 *When $n >> m$, by welch bould[2], we have*

$$\mu(A) \ge \frac{1}{\sqrt{m}}.$$

Lemma 13 *By Gershgorin circle theorem, we have*

$$spark(A) \ge 1 + \frac{1}{\mu(A)}.$$

Lemma 14 *If A has unit-norm columns, then A satifies RIP of order $K(K < \frac{1}{\mu(A)})$ with $\delta_K = (K-1)\mu(A)$.*

So far we have discussed various properties that might be useful in the context of sparse recovery, including the null space property(NSP6) and restricted isometry property(RIP8). We have also mentioned how to get a matrix that satisfies the RIP by using randomness. We will now turn to the problem of how to use these properties to establish performance guarantees for practical sparse recovery algorithms. We will begin by considering l_1-minimization and the RIP. We also observe that the RIP actually implies the NSP, adn hence is somewhat stronger condition.

The canonical l_1-minimization problem that we will analyze is given by:

$$x^* = \arg\min_{\tilde{x} \in \mathbb{R}^n} \|\tilde{x}\|_1 \qquad \text{subject to} \qquad \tilde{x} \in \mathcal{B}(y), \tag{7}$$

We will leave $\mathcal{B}$ general for now, but examples could include choices such as $\mathcal{B}(y) = \{x : Ax = y\}$ or $\mathcal{B}(y) = \{x : \|Ax - y\|_2 \le \epsilon\}$.

The key ideas in the proof of this result follow from [3]. To state the result, we need a bit of notation. Let $h = x^* - x$ denote the recovery error vector. Then we let Λ_0 denote the index set corresponding to the K entries of x with largest magniute, Λ_1 denote the index set corresponding to the K largest entries of $h_{\Lambda_0^c}$(largest magniute), Λ_2 to the next K largest entries of $h_{\Lambda_0^c}$ and so on.

Lemma 15 *Suppose that A satifies the RIP of order $2K$ with $\delta_{2K} < \sqrt{2} - 1$. If $\|x^*\|_1 \le \|x\|_1$, then*

$$\|h\|_2 \le C_0 \frac{\delta_K(x)_1}{\sqrt{K}} + C_1 \frac{|\langle Ah_\Lambda, Ah \rangle|}{\|h_\Lambda\|_2},$$

where the error $h = x^* - x$, $\Lambda = \Lambda_0 \cup \Lambda_1$.

Theorem 16 (Noiseless recovery) *Suppose that A satifies the RIP of order $2K$ with $\delta_{2K} < \sqrt{2}-1$.Then the solution x^* to*

$$\min_{\tilde{x}\in\mathbb{R}^n} \|\tilde{x}\|_1 \qquad \textit{subject to} \qquad A\tilde{x} = y, \tag{8}$$

obeys

$$\|x^* - x\|_2 \le C_0 \frac{\delta_K(x)_1}{\sqrt{K}}.$$

Proof: we have that $y = Ax = Ax^*$, and hence $Ah = A(x^* - x) = 0$. Applying Lemma 15, we can obtain the desired result. □

Theorem 17 (Noisy recovery) *Suppose that A satifies the RIP of order $2K$ with $\delta_{2K} < \sqrt{2}-1$.Then the solution x^* to*

$$\min_{\tilde{x}\in\mathbb{R}^n} \|\tilde{x}\|_1 \qquad \textit{subject to} \qquad \|A\tilde{x} - y\| \le \epsilon, \tag{9}$$

obeys

$$\|x^* - x\|_2 \le C_0 \frac{\delta_K(x)_1}{\sqrt{K}} + C_2\epsilon.$$

Proof: We have that

$$\begin{aligned} & \frac{|\langle Ah_\Lambda, Ah\rangle|}{\|h_\Lambda\|_2} \\ \le\ & \frac{\|Ah_\Lambda\|_2\|Ah\|_2}{\|h_\Lambda\|_2} \\ \le\ & \sqrt{1+\delta_{2K}}\|Ah\|_2 \\ =\ & \sqrt{1+\delta_{2K}}\|A(x^*-x)\|_2 \\ \le\ & \sqrt{1+\delta_{2K}}(\|Ax^*-y\|_2 + \|y-Ax\|_2) \\ \le\ & 2\epsilon\sqrt{1+\delta_{2K}}. \end{aligned}$$

Applying Lemma 15, we can obtain the desired result. □

More detail can see the notes[1].

References

[1] Notes. `http://users.ece.gatech.edu/~mdavenport/ece-8823a-spring2013/notes/`.

[2] Wikipedia. `http://en.wikipedia.org/wiki/Welch_bounds`.

[3] Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

[4] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[5] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.