1 Compressive Sensing

Compressive sensing is an idea proposed by Emmanuel Candès, Justin Romberg, Terence Tao [CRT06a, CRT06b] and David Donoho [Don06a, Don06b] around year 2005. In compressive sensing, we design a "sensing matrix" $A \in \mathbb{R}^{k\times d}(k < d)$, and try to recover a s-sparse vector x using $y = Ax$. We can't solve $y = Ax$ to find x directly because it is underdetermined, however we want A to preserve as much information as possible, and hopefully recover x by exploiting its sparsity.

Definition 1. *We define the quantity* $\|\mathbf{x}\|_0 := \{\#k : x_k \neq 0\}$*, i.e.,* $\|\boldsymbol{x}\|_0$ *counts the number of non-zero coefficients of vector* $\mathbf{x}$*.*

$\|x\|_0$ is not a norm, but we call it 0-norm anyway. This misnomer is partly justified by the fact that $\|x\|_p^p \to \|\mathbf{x}\|_0$ as $p \to 0$.

Suppose that $\mathbf{x} \in \mathbb{R}^d$, if $\|\mathbf{x}\|_0 = s$, we call x s-sparse. We are particularly interested in the case when $s \ll d$. If x is s-sparse, we may not know the locations of the non-zero entries *a priori*. Indeed, we usually assume that the non-zero locations of x are (initially) unknown.

Assume now we know that $x \in \mathbb{R}^d$ is s-sparse. We could compress x by first measuring all d coefficients of x then keeping the s non-zero ones and associated indices. Thus, we can compress $x \in \mathbb{R}^d$ into a vector of length $2s$ (without loss). But can we do almost as good without acquiring all d coefficients of $\boldsymbol{x}$ first.

Example: Let x be a vector of length $d = 2^n$ for some $n \in \mathbb{N} > 0$ and assume that $\boldsymbol{x}$ is 1-sparse and non-negative (the non-negativity assumption is just for convenience and will not be needed later):

$$\mathbf{x} = \left[0, \ldots 0, x_k, 0, \ldots, 0\right],$$

where $x_k > 0$ for some unknown index k. In order to locate the one non-zero entry x_k we could measure all d coefficients of x one by one. Or we could use the following more efficient strategy: The first measurement we take consists of suming up the first $d/2$ entries of x. If the sum is non-zero, we know that x_k is contained in the first $d/2$ entries of x and continue to step 2 using the first $d/2$ entries of x, otherwise we proceed to step 2 using the second $d/2$ entries of x. Let us assume w.l.o.g. that the first sum is zero. We now split the second half of x into two parts and sum up the first $d/4$ entries of the first part. If the sum is non-zero we proceed with that part, otherwise we take the other part. We can proceed in that way and within $\log_2 d$ steps we have located the non-zero coefficient x_k.

Admittedly, the procedure of taking measurements has changed now from point evaluations to summing (integrating) over parts of the signal. Depending on the application this may or may not be an issue. This adaptive strategy of measuring signals is a form of group testing, which was used for instance in the analysis of blood samples to test soldiers during World War II. However, we prefer a non-adaptive data acquisition scheme (which does not adapt the measurement strategy to a specific $\mathbf{x}$), as this is much more convenient in applications.

Difficulty: The space of all s-sparse vectors is not a linear subspace of $\mathbb{R}^d$ because the sum of two s-sparse signals can be 2s-sparse. Simply randomly sampling s-sparse signals does

not give useful compression because it would require $O(d)$ measurement. Hence, motivated by Johnson-Lindenstrauss, can we design a random sensing matrix $\mathbf{A}$ that allows us to take only k measurements of $\boldsymbol{x}$, where k is approximately proportional to s? More precisely, let $A = k \times d$ matrix, $x \in \mathbb{R}^d$, x is s-sparse, $k < d$ (ideally $k \sim s$). We measure y:

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

and hope that y preserves the information of x i.e. we can recover x from y.
Note that (1) is an under-determined system so there is no way of recovering an arbitrary x from y.

We want A to have the following properties:

1. If $x_1 \neq x_2$, we want to ensure that $y_1 \neq y_2$, where $y_1 = Ax_1$, $y_2 = Ax_2$;
2. We want to have an algorithm to solve x knowing y.

In general, it is not possible to satisfy (1), since A is underdetermined and has a non-empty null space. We could pick $h \in \text{null}(A)$, and have $y = Ax = A(x + h)$, so that both x and $x + h$ have the same y. However, if we can ensure that $x + h$ cannot be a s-sparse vector, we might be able to exclude this possibility, and thus find some good matrix A that satisfies property (1).

Suppose we have two s-sparse vector x_1 and x_2, and we want $Ax_1 \neq Ax_2$ when $x_1 \neq x_2$. When $Ax_1 = Ax_2$, we have $A(x_1 - x_2) = 0$, where $x_1 - x_2$ is up to $2s$-sparse. We want to make sure that $x_1 - x_2$ is not in A's null space to exclude this situation. In other words, we want A's null space to have no $2s$-sparse vectors. To make our idea more precise, we introduce the restricted isometry property (RIP):

Definition 2 (Restricted Isometry Property)**.** *We say that A satifies the Restricted Isometry Property (s, δ_s)-RIP when for each $s = 1, 2, 3, \cdots$, define the isometry constraints δ_s of matrix A as the smallest constants such that*

$$(1 - \delta_s)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_s)\|x\|_2^2$$

holds for all s-sparse vector x.

RIP quantifies how well A ensures that measurements of $2s$-sparse vectors are different. If $\delta_{2s} < 1$ then $x_1 \neq x_2 \implies Ax_1 \neq Ax_2$.

Now let us look at property (2). We can show that, finding the unique solution x to $Ax = y$, where x is a s-sparse vector, is equivalent to the following optimization problem:

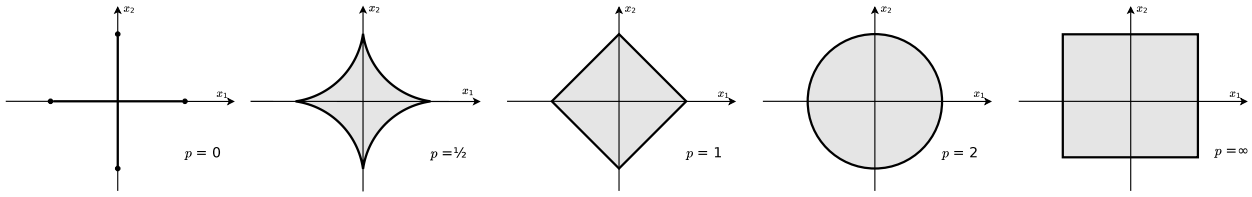
Definition 3 (The L_0 minimization problem)**.**

$$\min_z \|z\|_0, \quad \text{subject to } Az = y$$

The L_0 norm $\|\cdot\|_0$ reflects the number of non-zero elements in a vector. Unfortunately, solving the L_0 minimization problem has been proven to be NP-hard, and there are no effective algorithms when s is relatively large. However, we recall that

$$\lim_{p\to 0} \|z\|_p^p \to \|z\|_0$$

we can probably use some other small p other than 0 to solve the minimization problem. Figure 1 shows the unit ball for p-norm $\|x\|_p \leq 1$ with different p. In the two dimension case, we can clearly see the how the shape of the unit ball gradually changes, from 0 to ∞.



(a) $p = 0$ (b) $p = \frac{1}{2}$ (c) $p = 1$ (d) $p = 2$ (e) $p = \infty$

Figure 1: Unit balls of L_p norm with different p

Now we consider the L_p minimization problem:

$$\min_z \|z\|_p, \quad \text{subject to } Az = y, \quad \text{for some } p > 0$$

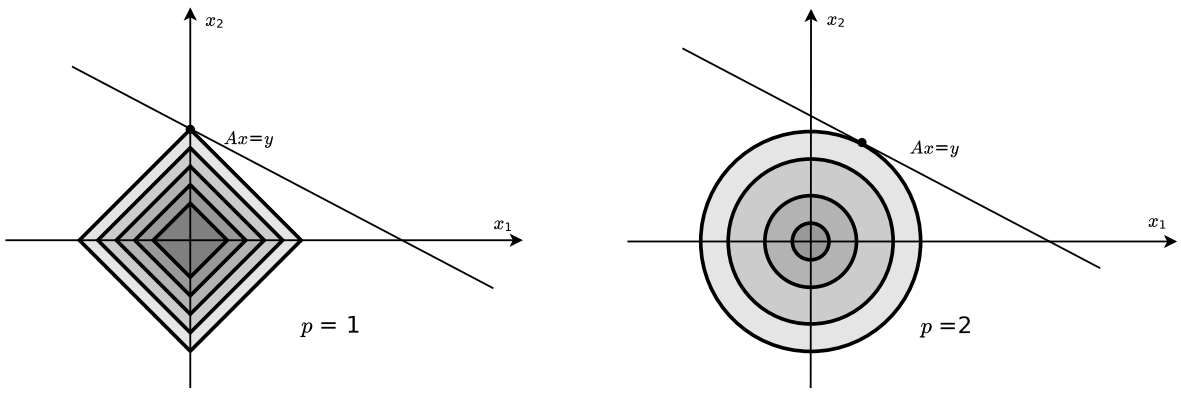
When $0 < p < 1$, this optimization problem is non-convex and is hard to solve. When $p = 1$, this becomes a convex optimization problem and good algorithms (for example, gradient descent) exist to solve it efficiently. We also observe that, L_1 minimization tends to give us a sparse solution. In figure 2, the constraint $Ax = y$ is visualized as a line, and we gradually increase the radius of the ball $\|x\|_p \leq r$. We stop increasing the radius when the ball just touches the line $Ax = b$, and the intersecting point is the optimal solution of the L_p minimization problem.

If we draw several different lines $Ax = y$, we find that in the L_1 minimization problem the solution is more likely to be on an axis, indicating a sparse solution. In L_2 minimization, a sparse solution can only be found when $Ax = y$ is either vertical or horizontal. L_1 promotes sparsity much better then L_2 does. From what we learned in lecture 1, in high dimensions, the polytope of L_1 unit ball will be more "peaky", thus it is a good approximation of the L_0 minimization problem.

To precisely describe our idea of replacing L_0 minimization with L_1 minimization, we state the following theorem:

Theorem 4. *Let $y = Ax_0$ where x_0 is a s-sparse vector. Assume that A satisfies the RIP property with $\delta_{2s} < \frac{1}{3}$, then the solution x_* to the L_1 minimization problem*

$$\min_z \|z\|_1, \text{subject to } Az = y = Ax_0$$



(a) $p = 1$ (b) $p = 2$

Figure 2: Sparsity of the solution of L_p minimization

becomes x_0 exactly, i.e., $x_* = x_0$

The theorem basically says that under certain conditions, the optimizer of the L_1 minimization problem $Az = Ax_0$, where x_0 is a s-sparse vector, will be exactly x_0. Note that in general we don't know x_0, but we have the measurements Ax_0. If A satisfies the RIP property as in Theorem 4, we can recover x_0 by finding the L_1 minimizer, which can be solved efficiently.

2 Proof of the Theorem

Lemma 5. *There holds*

$$|\langle Ax, Ax' \rangle| \leq \delta_{s+s'} \|x\|_2 \|x'\|_2$$

for all x, x' supported [1] on disjoint subsets $S, S' \subseteq [1, \cdots, d]$, $x, x' \in \mathbb{R}^d$, and $|S| \leq s$, $|S'| \leq s'$

Proof. Without loss of generality, we can assume $\|x\|_2 = \|x'\|_2 = 1$, so that the right hand size of the inequality becomes just $\delta_{s+s'}$. From A's RIP property,

$$(1 - \delta_{s+s'})\|x \pm x'\|_2^2 \leq \|A(x \pm x')\|_2^2 \leq (1 + \delta_{s+s'})\|x \pm x'\|_2^2$$

Since the support of x and x' is disjoint, $\|x \pm x'\|_2^2 = \|x\|_2^2 + \|x\|_2^2 = 2$; the RIP property then becomes

$$2(1 - \delta_{s+s'}) \leq \|Ax \pm Ax'\|_2^2 \leq 2(1 + \delta_{s+s'})$$

The polarization identity then implies:

[1] The support of a vector, $\text{supp}(x)$, is the set of indices i such that $x_i = 0$. The size of the support of x, $|\text{supp}(x)|$, is denoted $\|x\|_0$

$$
\begin{aligned}
|\langle Ax, Ax' \rangle| &= \frac{1}{4}\Big| \|Ax + Ax'\|_2^2 - \|Ax - Ax'\|_2^2 \Big| \\
&\leq \frac{1}{4}\Big| 2(1+\delta_{s+s'}) - 2(1-\delta_{s+s'}) \Big| \\
&= \delta_{s+s'}
\end{aligned}
$$

□

Lemma 6 (Null-space property)**.** *Given $A \in \mathbb{R}^{k\times d}$. Every vector $x \in \mathbb{R}^d$ supported on a set S is the unique solution to the L_1 optimization problem*

$$\min_z \|z\|_1, \quad \text{subject to } Az = Ax$$

if and only if A satisfies the following null-space property (NP) w.r.t S

$$\|h_S\|_1 < \|h_{S^C}\|_1, \quad \text{for all } h \in \text{null}(A) \setminus 0$$

for all S of cardinality $|S| \leq s$.

Here, h_S is defined as h restricted to set S, and S^C is the complement of S.

Proof. We are going to show both directions are true:

$\Rightarrow$):

For any $h \in \text{null}(A) \setminus 0$, h_S is supported on S and thus the unique minimizer of

$$\min_z \|z\|_1, \quad \text{s.t. } Az = Ah_S$$

Because $h = h_S + h_{S^C}$ and $Ah = 0$, we have $A(-h_{S^C}) = Ah_S$. Now we consider $-h_{S^C}$ which is also a feasible solution of this minimization problem. Because S and S^C are complementary, we have $-h_{S^C} \neq h_S$. Also since $h_S \neq 0$ is the unique minimizer, $\| - h_S^C\|_1$ must be greater than the minimizer:

$$\|h_S\|_1 < \| - h_{S^C}\|_1 = \|h_{S^C}\|_1$$

$\Leftarrow$):

Now assume null-space property holds. Then, given an s-sparse x and any $z \neq x$ both satisfies the constraint $Az = Ax$. Consider $h = x - z \in \text{null}(A)$. Using the null-space property, we get:

$$
\begin{aligned}
\|x\|_1 &= \|x - z_S + z_S\|_1 \\
&\leq \|x - z_S\|_1 + \|z_S\|_1 \\
&= \|h_S\|_1 + \|z_S\|_1 && \text{(by definition } h = x - z\text{)} \\
&< \|h_{S^C}\|_1 + \|z_S\|_1 && \text{(by null-space property)} \\
&= \| - z_{S^C}\|_1 + \|z_S\|_1 && \text{(because } x_{S^C} = 0\text{)} \\
&= \|z\|_1 && \text{(by additive of disjoint support)}
\end{aligned}
$$

$\|x\|_1$ is less than any other $\|z\|_1$ where $z \neq x$, thus it is the unique minimizer.

□

Now we are ready to prove Theorem 4. We simply need to show that Null-space Property holds for the given conditions.

Proof. Take $h \in \text{null}(A) \setminus 0$. Let index set S_0 be the set of indices of s largest entries (by modulus) of h. Let index sets $S_1, S_2, \cdots$ be index sets corresponding to the next s to $2s$, $2s$ to $3s$, $\cdots$ largest entries of h.

Because of the RIP property of A, we have

$$
\begin{aligned}
\|h_{S_0}\|_2^2 &\leq \frac{1}{1-\delta_s}\|Ah_{S_0}\|_2 && && (2) \\
&= \frac{1}{1-\delta_s}\sum_{j\geq 1}\langle Ah_{S_0}, A(-h_{S_j})\rangle && \text{Because } h_{S_0} = \sum_{j\geq 1}(-h_{S_j}) && (3) \\
&= \frac{1}{1-\delta_s}\sum_{j\geq 1}\delta_{2s}\|h_{S_0}\|_2\|h_{S_j}\|_2 && \text{By Lemma 6} && (4) \\
&\leq \frac{\delta_{2s}}{1-\delta_s}\|h_{S_0}\|_2\sum_{j\geq 1}\|h_{S_j}\|_2 && && (5) \\
\|h_{S_0}\|_2 &\leq \frac{\delta_{2s}}{1-\delta_s}\sum_{j\geq 1}\|h_{S_j}\|_2 && && (6)
\end{aligned}
$$

Note that

$$\|h_{S_j}\|_2 \leq s^{\frac{1}{2}}\|h_{S_j}\|_\infty \leq s^{-\frac{1}{2}}\|h_{S_{j-1}}\|_1$$

We can rewrite (6) as

$$\|h_{S_0}\|_2 \leq \frac{\delta_{2s}}{1-\delta_s} s^{-\frac{1}{2}} \sum_{j\geq 1} \|h_{S_{j-1}}\|_1 \tag{7}$$

$$= \frac{\delta_{2s}}{1-\delta_s} s^{-\frac{1}{2}} \|h\|_1 \tag{8}$$

Also, by Cauchy-Schwartz inequality,

$$\|h_{S_0}\|_1 = \sum_{i\in S_0} 1 \times |h_i| \leq \sqrt{\sum_{i\in S_0} 1^2} \sqrt{\sum_{i\in S_0} h_i^2} = \sqrt{s}\|h_{S_0}\|_2 \tag{9}$$

We have $\delta_{2s} < \frac{1}{3}$ as a condition, so

$$\frac{\delta_{2s}}{1-\delta_s} < \frac{\delta_{2s}}{1-\delta_{2s}} < \frac{1}{2} \quad \text{for } \delta_{2s} < \frac{1}{3} \tag{10}$$

Combining (8) (9) (10) we get

$$\|h_{S_0}\|_1 < \frac{1}{2}\|h\|_1 \tag{11}$$

Now we show (11) is equivalent to $\|h_S\|_1 < \|h_{S^C}\|_1$:

$$\begin{aligned} & \|h_S\|_1 < \|h_{S^C}\|_1 \\ \Leftrightarrow \quad & 2\|h_S\|_1 < \|h_{S^C}\|_1 + \|h_S\|_1 \\ \Leftrightarrow \quad & 2\|h_S\|_1 < \|h\|_1 \\ \Leftrightarrow \quad & \|h_S\|_1 < \frac{1}{2}\|h\|_1 \end{aligned}$$

Thus, we have shown that $\|h_{S_0}\|_1 < \|h_{S^C}\|_1$, which is the null-space property. Then by Lemma 6, we know that every vector $x \in \mathbb{R}^d$ supported on the set S_0 is the unique solution of the L_1 minimization problem; in other words, we can recover the s-sparse vector x by solving the L_1 minimization problem.

□

3 Constructing RIP matrices

We start by presenting two ways of constructing the compressive sensing matrix A that satisfies the Restricted Isometry Property.

Theorem 7. *Let $A = (a_{ij})_{k\times d}$, where $a_{ij} \sim N(0,1)$, be a $k \times d$ Gaussian random matrix. Then there exists a constant C such that if $k \leq Cs\log(\frac{d}{s})$, then $\frac{1}{\sqrt{k}}A$ satisfies the $(s, \frac{1}{3})$-RIP with probability at least $1 - e^{-d}$.*

Theorem 8. *Let $F_d = \{e^{-\frac{2\pi i}{d}kl}\}_{k,l=0,1,\dots d-1}$ be the $d \times d$ Discrete Fourier Transform (DFT) matrix. Let A be the $k \times d$ matrix attained by selecting k rows from F_d uniformly randomly. If $k \leq Cs(\log d)^4$, then A satisfies the $(s, \frac{1}{3})$-RIP with probability $1 - e^{-k}$.*

Remarks:

- Theorem 1 is theoretical and it does not appear too much in practical application compared to Theorem 2 which constructs RIP matrix by DFT.
- One may read [1] for how to use the latest CVX toolbox in Matlab.
- Note that
$$\delta_S = \max_{|s|\leq S} \|A_s^* A_s - I_s\|_{op}$$
where A_s consists of s columns of A with $|S| = s$. Also $A_x = A_s x_s$ where x_s is x restricted to its s nonzero entries.
- The above two theorems are based on probabilistic arguments. However, one should note that they are the probabilities that one matrix A satisfies the RIP and that once A satisfies RIP, it is true for all s-sparse vectors x.

4 RIP versus JL

In the previous lectures, we discussed linear dimension reduction technique via Johnson-Lindenstrauss (JL) projection. Here, we make several comments on the features of JL projection and RIP.

Johnson-Lindenstrauss:

- In JL projection, the number of vectors are supposed to be fixed and finite. The vectors do not have to be sparse.
- JL projection preserves (up to ϵ) pairwise distances between vectors.
- JL projection does not allow one to recover x from $y = Ax$.

Restricted Isometry Property:

- RIP works for infinitely many vectors. However, it requires the vectors to be s-sparse in d-dimension. As a result, if one does not have sparsity on vectors, one should use JL.

- RIP allows one to recover x from $y = Ax$ numerically efficiently.

Connection between JL and RIP:

Question: If a matrix A satisfies RIP, can we use it for dimension reduction matrix in JL? The answer is NO.

Example: randomly subsampled DFT matrix (or Hadamard matrix). Then the matrix we get satisfies RIP with high probability but it does not fit in JL. To get JL, we need to pre-multiply A by a random diagonal sign matrix $D = (d_{ij})$, where $d_{ij} = 0$ if $i \neq j$ and $d_{ii} = \pm 1$ with probability $\frac{1}{2}$, respectively. As a result, if A is a RIP matrix, then AD is a JL matrix with high probability.

To solve L_1 problems in Matlab, it is convenient to use the software package *cvx*, see [1].

References

[1] Michael C. Grant and Stephen P. Boyd. *CVX: Matlab Software for Disciplined Convex Programming.* http://cvxr.com/cvx/.

[CRT06a] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[CRT06b] 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.

[Don06a] David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[Don06b] David L Donoho. For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.