Compressed Sensing

Huichao Xue

CS3750 Fall 2011

Table of Contents

Accidental Discovery

- In 2004, Candes accidentally discovered the fact that *L1-minimization* helps to fill in the blanks on an undersampled picture effectively.
- The recovered picture is not just slightly better than the original, rather, the picture looks sharp and perfect in every detail.

What will this technology bring to us

Being able to recover from incomplete data is very important:

- Less time spent on MRI or other sensing technologies
- Relieves storage requirement, because we only need incomplete data to recover all that we need
- Conserves energy

A very hot topic...

- I did a grep at http://dsp.rice.edu/cs, about 700 papers are published on CS during these 7 years.
- It is applied to many fields (of course including Machine Learning)
- Prof. Candes was rewarded with Waterman Prize[1].

[1] http://en.wikipedia.org/wiki/Alan_T._Waterman_Award

Table of Contents

Abstract Definition

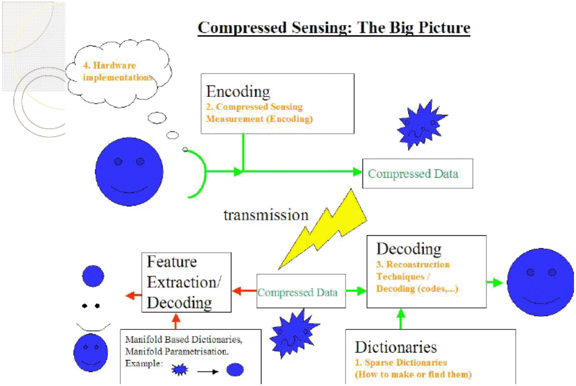
Definition

Compressed Sensing or Compressive Sensing(CS) is about acquiring and recovering a *sparse signal* in the most efficient way possible (subsampling) with the help of an *incoherent projecting basis*.[2]

1. The signal needs to be sparse
2. The technique acquires as few samples as possible
3. Later, the original sparse signal can be recovered
4. This done with the help of an *incoherent projecting basis*

[2] found it this definition here: https://sites.google.com/site/igorcarron2/cs

The big picture



I found it this here: https://sites.google.com/site/igorcarron2/cs

Note that there is no *compression* step in the framework. The compression is done when sensing, that why this technique got the name *Compressed Sensing*.

Table of Contents

L_p Norms

Compressed Sensing's algorithm makes use of L_1 norm's properties. So Let's have a review of it.

Definition

L_p norm of a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ is defined as:

$$||\mathbf{x}||_p = (|x_1|^p + |x_2|^p + \ldots + |x_n|^p)^{\frac{1}{p}} \tag{1}$$

L_p norms that are used in Compressed Sensing

In particular

Definition

L_0 norm of $\mathbf{x}$, $||\mathbf{x}||_0$, is the number of non-zero entries in $\mathbf{x}$.

Definition

L_1 norm of $\mathbf{x}$:

$$||\mathbf{x}|| = |x_1| + |x_2| + \ldots + |x_n| \quad (2)$$

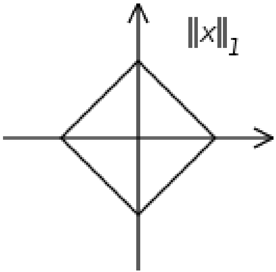
Definition

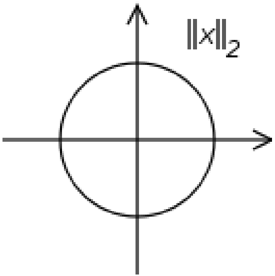
L_2 norm of $\mathbf{x}$:

$$||\mathbf{x}||_2 = \left(|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2\right)^{\frac{1}{2}} \quad (3)$$

L_p *balls*

Here are the illustrations of L_1 and L_2 balls in 2-D space:





Recovering f from an underdetermined linear system

Consider the scenario below:

$$\begin{bmatrix} y \end{bmatrix} = \begin{bmatrix} \quad & \Phi & \quad \end{bmatrix} \begin{bmatrix} \\ f \\ \\ \end{bmatrix}$$

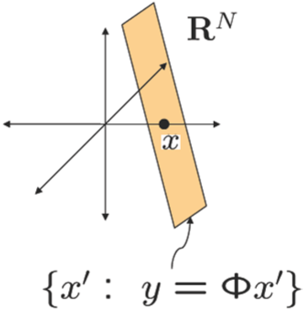
We want to recover f from the given y and Φ.

Is that even possible?

- There could be an infinite number of solutions for f
- But what if we already know that f is sparse [3]?

[3] being sparse means having only a few non-zero values among all f's dimensions

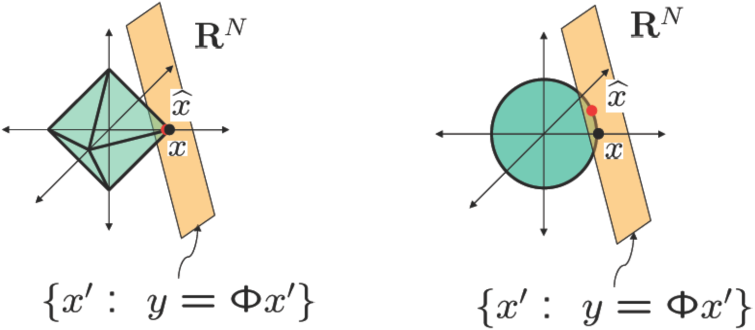
Consider recovering x from projection from the given y and Φ



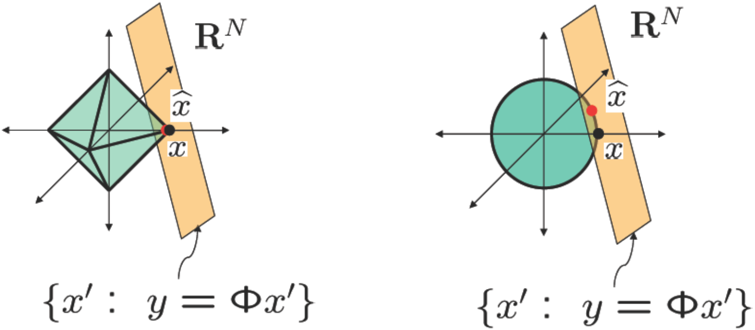
- The possible solutions for x lie in the yellow colored hyperplane.
- To limit the solution to be just one single point, we want to pick the sparsest x from that region.
- How do we define sparsity?

Comparison between L_1 *and* L_2

- Norms will help us here. We hope: smaller norm $\Rightarrow$ sparser
- But which norm should we choose?

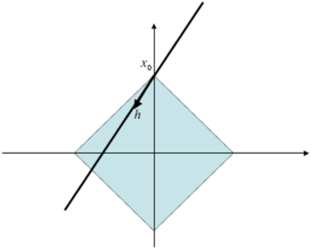


L_1 *prefers sparsity*



- Here minimizing L_1 provides a better result because in its solution $\hat{x}$, most of the dimensions are zero.
- Minimizing L_2 results in small values in some dimensions, but not necessarily zero.

But L_1 isn't always better



- Consider the graph on the left, when we try to find a solution for $y = \Phi x$ for the given y and Φ
- The original sparse vector x_0 which generates y from the linear transformation Φ is shown in the graph
- When we solve the equation $y = \Phi x$, we get the hyperplane indicated by h.
- If we choose to minimize the L_1-norm on h, then we will get a totally wrong result, which lies on a different axis than x_0's.

In Compressed Sensing, people develop conditions to ensure that this never happens.

Table of Contents

The algorithm's Context

$$\begin{bmatrix} y \end{bmatrix} = \begin{bmatrix} & \Phi & \end{bmatrix} \begin{bmatrix} \\ f \\ \\ \end{bmatrix}$$

- The linear system is underdetermined
- We want f to be sparse

The algorithm to find the proper f

- When no noise:

$$\min_{f \in R^n} ||f||_1 \quad \text{s.t.} \;\; y = \Phi f$$

- When there is noise:

$$\min_{f \in R^n} ||f||_1 \quad \text{s.t.} \;\; ||y - \Phi f||_2 \leq \epsilon$$

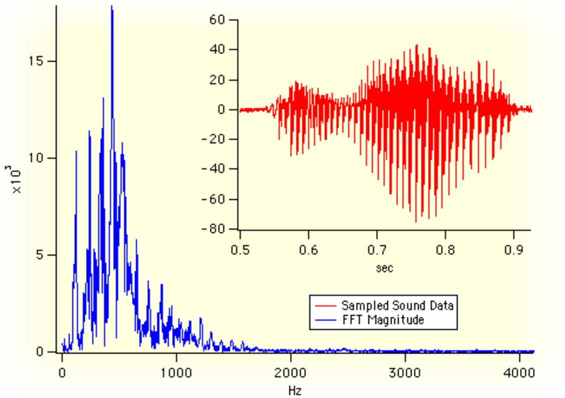
The whole literature is trying to show that: in most cases, this is going to find a very good solution.

Table of Contents

Questions at this point

- When do we need to solve such underdetermined linear system problems? Is that really important?
- If it is important, how did people deal with this before CS was discovered?
- Why does CS always find a good solution?

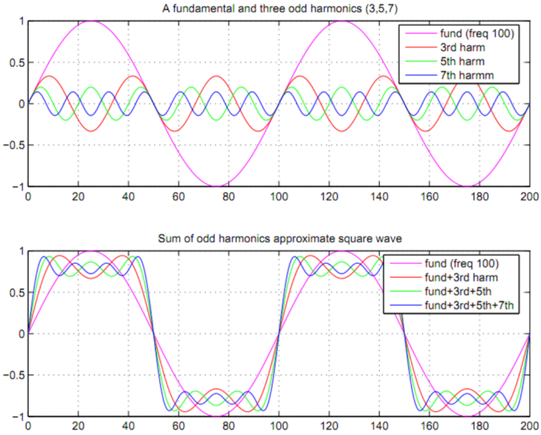
When? And how did people handle that?



- The sound data, colored in red, is quite complicated. It is a time domain representation because the x-axis is time.
- Luckily, it also has another representation in frequency domain, colored in blue. This representation has the benefit

A review of Fourier Transform

This is a demonstration of how data in time domain(lower graph) also can be constructed using a superposition of periodic signals(upper graph), each of which has a different frequency.



Formulas for Fourier Transform

To go between the time domain and the frequency domain, we use Fourier Transforms:

$$H_n = \sum_{k=0}^{N-1} h_k e^{\frac{2\pi ikn}{N}} \quad (4)$$

$$h_n = \frac{1}{N} \sum_{n=0}^{N-1} H_n e^{-\frac{2\pi ikn}{N}} \quad (5)$$

Here H is the frequency domain representation, and h is the time domain signal.
Note that the transformations are linear.

Shannon-Nyquist Sampling Theorem

Theorem

If a function $x(t)$ contains no frequencies higher than B hertz, it is completely determined by giving its ordinates at a series of points spaced $\frac{1}{2B}$ seconds apart.

This basically says:

- x's frequency domain representation is sparse in the sense that all dimensions higher than B are zero.
- No information loss if we sample at 2 times the highest frequency.
- To do this, use Fourier transform.

The mapping to the underdetermined linear system

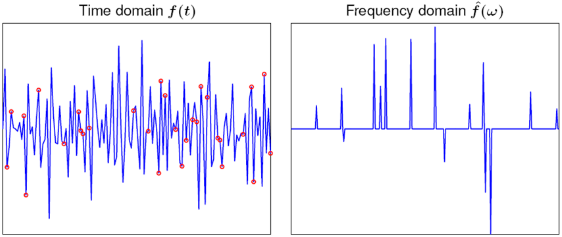
$$\begin{bmatrix} y \end{bmatrix} = \begin{bmatrix} & \Phi & \end{bmatrix} \begin{bmatrix} \\ f \\ \\ \end{bmatrix}$$

Here is the mapping between the equation above and the Shannon-Nyquist scenario:

- f is the low frequency signal. Higher dimensions are all zero.
- Φ is the inverse Fourier Transform
- y is our samples in the time basis

What's different in CS

- Rather than trying to recover all information on low **frequencies**, CS recovers those with high **amplitudes**.
- With the assumption that *only a few* frequencies have high amplitudes, CS requires much less samples to recover them



Measure M samples
(red circles = samples)

K nonzero components
$\#\{\omega : \hat{f}(\omega) \neq 0\} = K$

How is that possible?

It sounds quite appealing. But how do we do it?

- How do we pick the measurements so that the peaks' information is preserved?
- Don't we need to know how the data look like beforehand?

The big findings in CS:

- We only need the measurements to be incoherent to the sparse basis.
- Several randomly generated measurements are incoherent to every basis.

Table of Contents

The Uniform Uncertainty Principle

Definition

Φ obeys a UUP for sets of size K if

$$0.8 \cdot \frac{M}{N} \cdot ||f||_2^2 \leq ||\Phi f||_2^2 \leq 1.2 \cdot \frac{M}{N} \cdot ||f||_2^2$$

for every K-sparse vector f. Here M and N are the numbers of dimensions for x and f, correspondingly.

Example

Φ obeys UUP for $K \cdot M / \log N$ when

- $\Phi =$ random Gaussian
- $\Phi =$ random binary
- $\Phi =$ randomly selected Fourier samples (extra log factors apply)

We call these types of measurements incoherent

Sparse Recovery

- UUP basically means preserving the L_2 norms.
- UUP for sets of size $2K \Rightarrow^4$ there is only one K-sparse explanation for y.
- Therefore, say f_0 is K-sparse, and we measure $y = \Phi f_0$: If we search for the sparsest vector that explains y, we will find f_0

$$\min_f \#\{t : f(t) \neq 0\} \quad \text{s.t. } \Phi f = y$$

Note that here we need to minimize L_0-norm, which is hard. Can we make it a convex optimization problem?

[4] This basically means preserving L_2 distances.

Using L_1 norm

UUP for sets of size $4K \Rightarrow$

$$\min_f ||f||_1 \quad \text{s.t. } \Phi f = y$$

will recover f_0 exactly

Table of Contents

Coherence

Definition

The coherence between the sensing basis Φ and the representation basis Ψ is

$$\mu(\Phi, \Psi) = \sqrt{n} \times \max_{1 \leq k,j \leq n} |\langle \phi_k, \psi_j \rangle|$$

Here sensing basis is used for sensing the object f, and the representation basis is used to represent f.

Note that: $\mu(\Phi, \Psi) \in [1, \sqrt{n}]$

Example

- Time-frequency pair: $\mu(\Phi, \Psi) = 1$
- When $\Phi = \Psi$, $\mu(\Phi, \Psi) = \sqrt{n}$

Sample Size VS Coherence[5]

$$\min_{\tilde{x} \in R^n} ||\tilde{x}||_1 \quad \text{s.t. } y_k = \langle \phi_k, \Psi_{\tilde{x}}, \forall k \in M \rangle \tag{6}$$

Theorem

Fix $f \in R^n$ and suppose that the coefficient sequence x of x in the basis Ψ is S-sparse. Select m measurements in the Φ domain uniformly at random. Then if

$$m \geq C \cdot \mu^2(\Phi, \Psi) \cdot S \cdot \log n$$

for some positive constant C, the solution to (6) is exact with overwhelming probability.

- In the randomly generated matrices, if we choose the sensing basis uniformly at random, the coherence is likely to be $\sqrt{2 \log n}$
- This means: $m \approx \log^2 n \times S$

[5] This was developed by Candes and Romberg in 2007.

Table of Contents

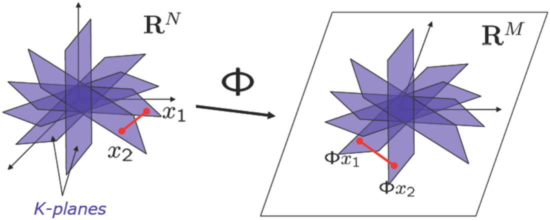
RIP(Restricted Isometry Property) aka UUP

Definition

For each integer $S = 1, 2, \ldots$, define the isometry constant δ_S of a matrix A as the smallest number 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 implies accurate reconstruction

If the RIP holds, then the following linear program gives an accurate reconstruction:

$$\min_{\widetilde{x} \in R^n} ||\widetilde{x}||_1 \quad \text{s.t.} \;\; A\widetilde{x} = y (= Ax) \tag{7}$$

Theorem

Assume that $\delta_{2S} < \sqrt{2} - 1$. Then the solution x^* to (7) obeys

$$||x^* - x||_2 \leq \frac{C_0}{\sqrt{S}} ||x - x_S||_1$$

and

$$||x^* - x||_1 \leq C_0 \, ||x - x_S||_1$$

for some constant C_0, where x_S is the vector x with all but the largest S components set to 0.

RIP implies robustness

$$\min_{\tilde{x} \in R^n} ||\tilde{x}||_1 \quad \text{s.t.} \quad ||A\tilde{x} - y||_2 \leq \epsilon \tag{8}$$

Theorem

Assume that $\delta_{2S} < \sqrt{2} - 1$. Then the solution x^* to (8) obeys

$$||x^* - x||_2 \leq \frac{C_0}{\sqrt{S}} ||x - x_S||_1 + C_1\epsilon$$

for some constants C_0 and C_1.

How do we find such A's

- The relations between m and S are missing in theorems (13) and (14).
- δ_{2S} provides the notion of incoherency. What kind of A and m support such a property? The answer is:
  - A can be m rows of random numbers, where $m \approx C \times S \log(n/S)$
  - You can't do much better than this.