Probabilistic Reasoning in Compressed Sensing

Roman Vershynin

University of Michigan

Journées MAS
Clermont-Ferrand, France
August 31, 2012

Compressed Sensing

It is a young field on the crossroads of:

- Signal processing
- Probability
- Information theory
- Statistics
- Geometric functional analysis

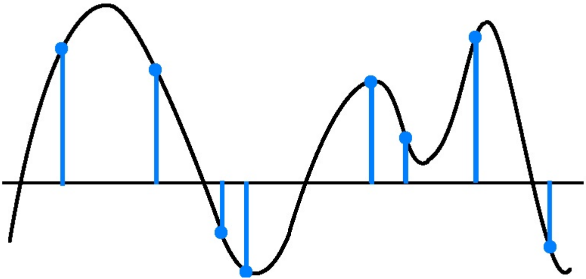
This talk: a very incomplete picture.
Emphasis on probabilistic, geometric insights.

Sampling

Problem: Recover a signal x from a sample of m linear measurements

$$f_1(x), \ldots, f_m(x).$$

Example: f_i are point evaluation functionals at random locations.

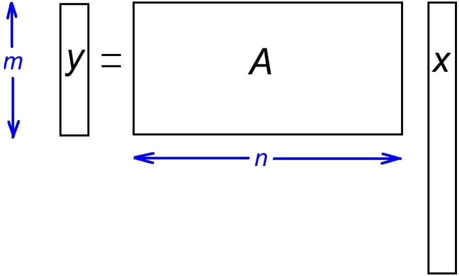


Sampling

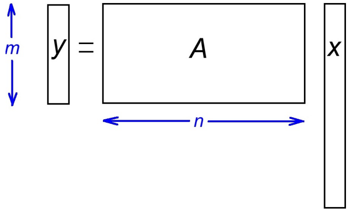
Unknown **signal** $x \in \mathbb{R}^n$.

Take m linear **samples**/measurements $y = Ax \in \mathbb{R}^m$.

Here A is a known measurement **matrix**, the sampling device.



Goal: recover x from y.



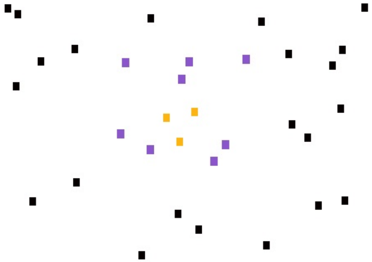
Goal: recover x from y.

- If $m \geq n$, the problem is well-posed, trivial: $x = A^{-1}y$.
- If $m < n$, the problem is **ill-posed**, recovery impossible due to $\ker(A)$.

Compressed sensing is seeking recovery strategies in the regime $m \ll n$.
[Donoho, Candes-Tao, ... 2004+]

Example:

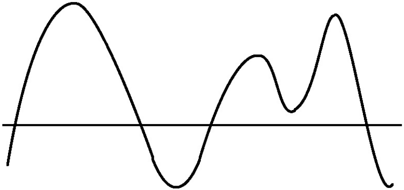
x = image, y = sample of m random pixels, $m \ll n$

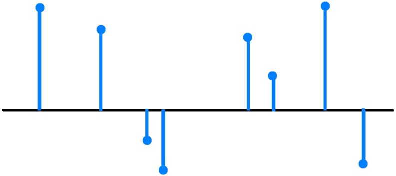


x = matrix, y = sample of m entries, $m \ll n$

More Examples:

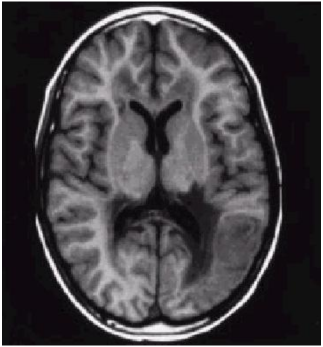
$x =$ audio signal, $y =$ sample of amplitudes at m random moments of time, $m \ll n = \infty$

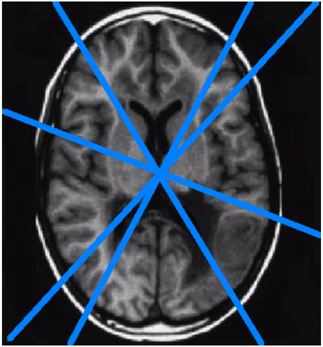




More Examples:

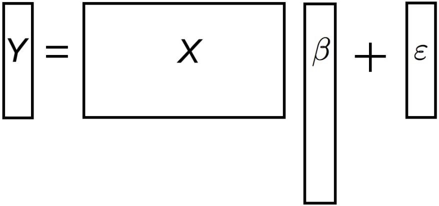
x = brain, y = MRI scan in m random directions. $m \ll n = \infty$

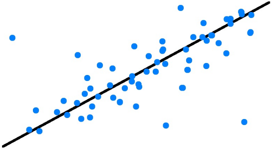




More Examples:

Linear Regression $\qquad Y = X\beta + \varepsilon$





$\beta \in \mathbb{R}^p$: unknown coefficient vector ($\sim$ signal x)
$X \in \mathbb{R}^{n \times p}$: sample of n i.i.d. predictor variables ($\sim$ matrix A)
Y = sample of n i.i.d. response variables ($\sim$ measurement vector y)

$n \ll p$: small sample, large number of parameters

Recall: problem ill-posed. Recovery **impossible** in general, due to $\ker(A)$.

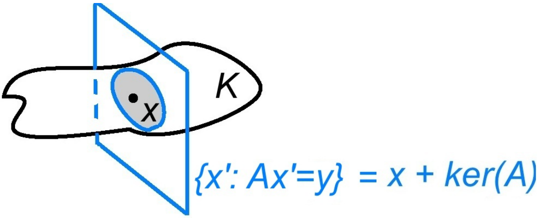
However, signal x may not be completely arbitrary.

Model: $x \in K$, a known signal set in $\mathbb{R}^n$.

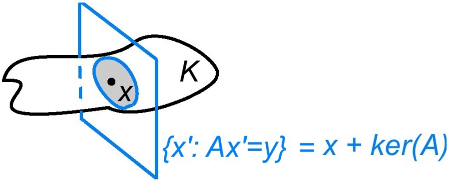
Can recover x up to $K \cap \ker(A)$. So, if

$$\operatorname{diam}(K \cap \ker(A)) \le \varepsilon$$

then we can recover x with error ε.



If $\text{diam}(K \cap \ker(A)) \le \varepsilon$ then we can recover x with error ε.



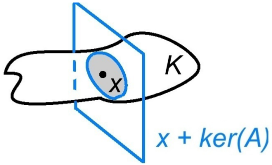
Recovery is achieved by solving the program:

$$\text{Find } x' \in K \text{ such that } Ax' = y.$$

In words: "Find a signal consistent with the model (K) and with the measurements (y)."

How to solve in practice?

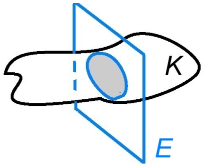
- If K is convex, this is a **convex program**. Many solvers exist.
- If not, convexity: replace K by $\text{conv}(K)$.



The recovery problem reduces to a *geometric question*:

Question. For what convex sets $K \subset \mathbb{R}^n$ and what matrices $A \in \mathbb{R}^{m \times n}$ is $\mathrm{diam}(K \cap \ker(A))$ small?

A is a **random matrix**.
Thus $E = \ker(A)$ is a **random subspace** in $\mathbb{R}^n$ of codimension m.



Question. For what convex sets $K \subset \mathbb{R}^n$ is $\mathrm{diam}(K \cap E)$ small, where E is a random subspace of given codimension m?

Geometric Functional Analysis.
[Pajor-Tomczak '85, Mendelson-Pajor-Tomczak '07]

Trivial answer: for *small* sets K.

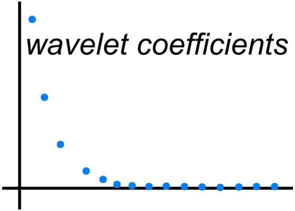
But why are common signal sets small?

Common signal sets are "small"

$K = \{\textbf{common images}\}$.

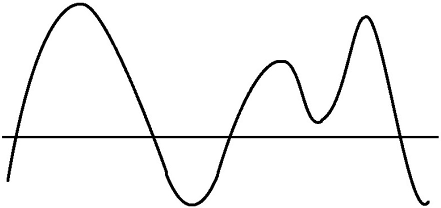
Few wavelet coefficients are large.

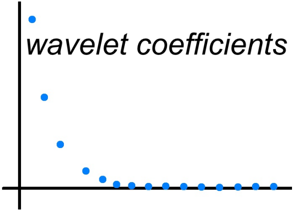
Thus images are **sparse in the wavelet domain**.



Common signal sets are "small"

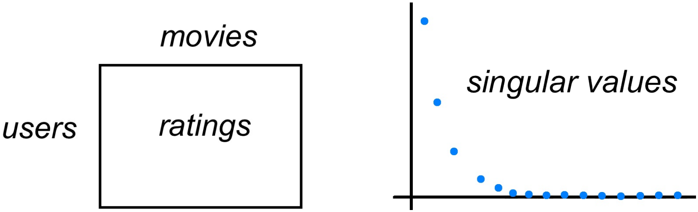
$K = \{$**common audio signals**$\}$.
Band-limited. Few leading frequencies (Fourier coefficients) are large. So these signals are **sparse in the Fourier domain**.





Common signal sets are "small"

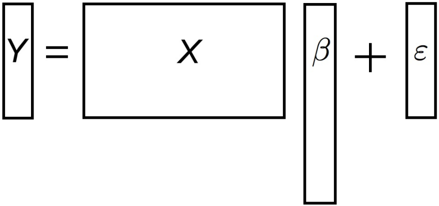
$K = \{$**common matrices**$\}$.
For example, the matrix of Netflix preferences. Nearly **low-rank.**

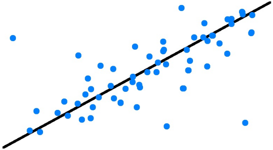


[Candes-Recht '08, ...]: matrix completion.

Common signal sets are “small”

Regression $Y = X\beta + \varepsilon$



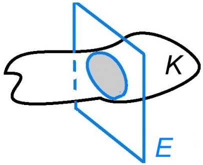


Only few of the predictor variables have significant influence.
Thus β has only few large coefficients, hence is **sparse**.

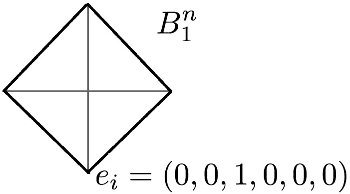
Lasso [Tibshirani ’96]; Danzig Selector [Candes-Tao ’05, . . .]

Back to our geometric question:

Question. Consider a "small" set $K \subset \mathbb{R}^n$, and a random subspace E of given codimension m. Is $\mathrm{diam}(K \cap E)$ small?



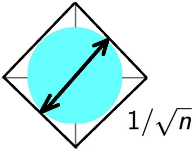
Example: $K = \text{conv}(\pm e_i) = \{x : \|x\|_1 \le 1\} = B_1^n$, the ℓ_1 ball.

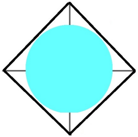


Theorem [Kashin '77]. If $\text{codim}(E) = m = \varepsilon n$, then

$$\text{diam}(B_1^n \cap E) \le \frac{C(\varepsilon)}{\sqrt{n}} \qquad \text{with high probability.}$$

Hence $B_1^n \cap E \sim$ inscribed *round ball*!





Similar result for arbitrary m (not just proportional to n):

Theorem [Garnaev-Gluskin '84]. If $\text{codim}(E) = m$, then

$$\text{diam}(B_1^n \cap E) \leq C\sqrt{\frac{\log n/m}{m}} \qquad \text{with high probability.}$$

In particular: if $m \gg \log n$ then the diameter is small, $o(1)$.

Corollary. One can accurately recover any signal $x \in B_1^n$ from $m = O(\log n)$ random linear measurements $y = Ax \in \mathbb{R}^m$.

Very few measurements! Indeed, one needs $\log n$ bits to specify a vertex $x = e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$.

General signal sets K.

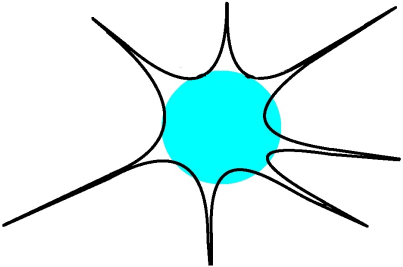
Question. What does a general convex set look like?

Concentration insight (recall Olivier Guedon's talk):

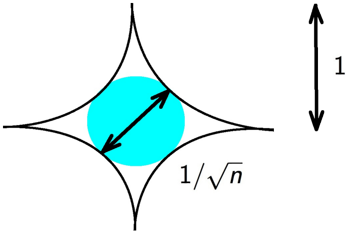
$$K \approx \text{bulk} + \text{outliers}.$$

Bulk = round ball, makes up most volume of K.
Outliers = few faraway tentacles, contain little volume.

V. Milman's heuristic picture of a general convex body:



Example: $K = B_1^n$**.** Heuristic picture:

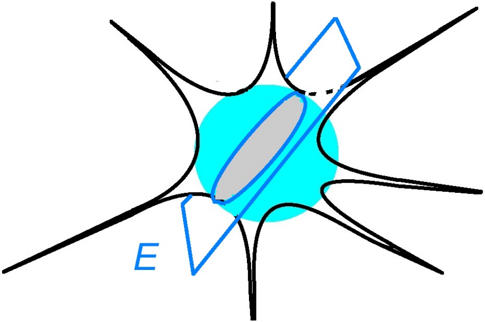


Concentration of volume:

$$\mathrm{Vol}(K)^{1/n} \sim \mathrm{Vol}(\bullet)^{1/n} \sim \frac{1}{n}.$$

For **general sets** K - recall Oliver Guedon's talk.

Heuristic consequences.



A random subspace E should tend to **miss the outliers**, pass through the bulk of K.

If so,

$$\text{diam}(K \cap E) \approx \text{diam}(\textit{bulk}) \quad \text{is small.}$$

As we desired!

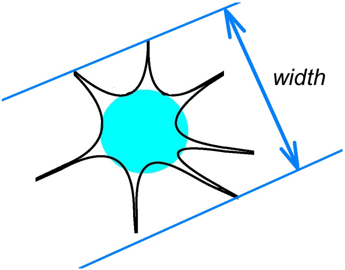
Rigorous results.

Theorem [Pajor-Tomczak '85]. Consider a convex set K in $\mathbb{R}^n$, and a random subspace E of codimension m. Then

$$\operatorname{diam}(K \cap E) \le \frac{C\, w(K)}{\sqrt{m}} \qquad \text{with high probability.}$$

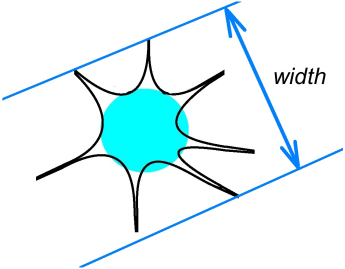
Here $w(K)$ is the **mean width** of K.

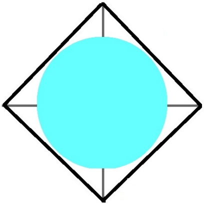
$$w(K) := \mathbb{E} \sup_{x \in K-K} \langle g, x \rangle = \sqrt{n} \cdot \mathbb{E}\left[\text{width of } K \text{ in random direction}\right].$$



Mean width.

$$w(K) := \mathbb{E} \sup_{x \in K-K} \langle g, x \rangle = \sqrt{n} \cdot \mathbb{E} \left[\text{width of } K \text{ in random direction} \right].$$





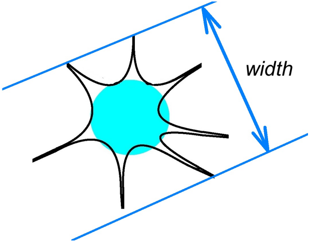
Remark: $w(K) = w(\text{conv}(K))$. Survives convexification.

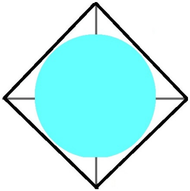
Example 1. $K = B_1^n$ or just the vertices $\{\pm e_i\}$.
Here $w(K) \sim \sqrt{\log n}$. Almost the same as $w(\bullet) = 1$.

Hence: the mean width **sees the bulk**, ignores the outliers.

Mean width.

$$w(K) := \mathbb{E} \sup_{x \in K-K} \langle g, x \rangle = \sqrt{n} \cdot \mathbb{E}\left[\text{width of } K \text{ in random direction}\right].$$





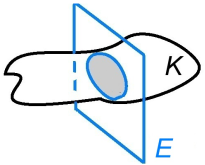
Example 2. $K = \{s\text{-sparse vectors in } \mathbb{R}^n\}$. Here $w(K) \sim \sqrt{s \log n}$.

Intuition: $w(K)^2$ is an **effective dimension** of K.
The amount of information in K.

Examples: Effective dim. of $\{\pm e_i\}$ is $\log n = \#$ bits to specify the signal.
Effective dim. of $\{s\text{-sparse vectors in } \mathbb{R}^n\}$ is $s \log n$.
(Intuition: need $\log \binom{n}{s} \sim s \log n$ bits to specify the sparsity pattern $+s$ bits to specify magnitudes of coefficients.)

Pajor-Tomczak's Thm: $\text{diam}(K \cap E) \lesssim w(K)/\sqrt{m}$ for random E of codimension m.

Consequence of Pajor-Tomczak's Theorem:
if $m \gg w(K)^2$ then diameter is small, $o(1)$.



Corollary. One can accurately recover any signal $x \in K$ from $m = w(K)^2$ random linear measurements $y = Ax \in \mathbb{R}^m$.

The sample size $m \sim$ **effective dimension** of K.

Surprisingly, **non-linear measurements** are also possible.

$$y = \theta(Ax)$$

where a function $\theta : \mathbb{R} \to \mathbb{R}$ is applied to each coordinate of Ax.

Examples:

1. Generalized Linear Models (GLM) in Statistics.
In particular, **logistic regression**. [Plan-V '12]

2. For $\theta(\cdot) = \mathrm{sign}(\cdot)$, **one-bit compressed sensing** [Plan-V '11]:

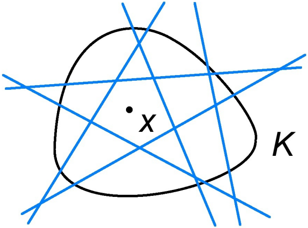
One-bit compressed sensing

$$y = \text{sign}(Ax) \quad \in \{-1, 1\}^m.$$

(Writing in coordinates, $y_i = \text{sign}(\langle A_i, x \rangle)$, $i = 1, \ldots, m$.)

Extreme quantization: one bit per measurement.

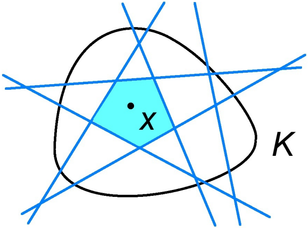
Geometric interpretation:
$y =$ vector of orientations of x with respect to m random hyperplanes (with normals A_i).



Random hyperplane tessellation (cutting) of K.

One-bit compressed sensing: $y = \text{sign}(Ax) \in \{-1, 1\}^m$

y = vector of orientations of x with respect to m random hyperplanes.



Knowing $y \iff$ knowing the **cell** $\ni x$.

If diam(every cell) $\leq \varepsilon$ then we can recover x with error ε.

Recovery is achieved by solving the program:

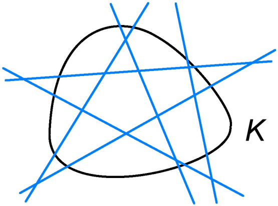
$$\text{Find } x' \in K \text{ such that } \text{sign}(Ax') = y.$$

Again, if K is convex, this is a **convex program**. (Many algorithms.)

It remains to answer the geometric question on the diameter:

Random hyperplane tessellations

Question. Given a set $K \subset \mathbb{R}^n$, how many random hyperplanes does it take in order to cut K in pieces of diameter $\leq \varepsilon$?



Non-trivial even for $K = S^{n-1}$.

Stochastic geometry: mostly focuses on the shape of a *fixed* cell.

Kendall's Conjecture. Let $m \to \infty$. If diameter of the zero cell $\to 0$, then its shape $\to$ round ball.
Proofs by [Kovalenko '97] ($n = 2$), [Hug, Reitzner, Schneider '04] ($n \geq 2$).

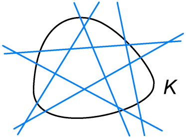
Irrelevant: we need to control *all* cells.

Random hyperplane tessellations

Theorem [Plan-V '12]. Consider a convex set $K \subset S^{n-1}$ and m random hyperplanes. Then, with high probability,

$$\text{diam(every cell)} \leq \left[\frac{C\, w(K)}{\sqrt{m}} \right]^{1/3}.$$

Here, as before, $w(K)$ is the **mean width** of K.



Very similar to Pajor-Tomczak's bound on $\text{diam}(K \cap \text{random subspace})$, except for the exponent $1/3$. It is probably not optimal.

Like before, a consequence for one-bit compressed sensing:

Corollary. One can accurately recover any signal $x \in K$ from $m = w(K)^2$ random one-bit linear measurements $y = Ax \in \{-1, 1\}^m$.

Can replace sign$(\cdot)$ by **general function** $\theta(\cdot)$:

$$y = \theta(Ax)$$

Recovery of x is achieved by solving the program

$$\max\langle y, Ax'\rangle \quad \text{subject to} \quad x' \in K.$$

In words, "maximize the correlation with measurements (y), while staying consistent with model (K)".

K convex $\implies$ convex program.

Surprise: the solver **does not need to know** θ; it may be unknown or unspecified.

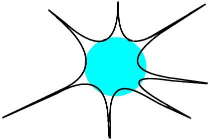
[Plan-V '12]

Summary

Compressed Sensing problem:
Recover signal $x \in \mathbb{R}^n$ from $m \ll n$ random measurements/samples

$$y = Ax \quad \text{(linear)}, \qquad\qquad y = \theta(Ax) \quad \text{(non-linear)}.$$

Model: $x \in K$, where K is a known signal set.
Convex set $\approx$ bulk $+$ outliers:



If the "bulk" of K is small, accurate recovery is possible. Precisely,

$$m \sim w(K)^2 = \text{the } \textbf{effective dimension} \text{ of } K.$$

Here $w(K)$ is the mean width of K, a computable quantity.

Compressed Sensing

- **Signal processing** (sampling)
- **Probability** (random matrices, stochastic geometry)
- **Information theory** (effective dimension $\sim$ information in K)
- **Statistics** (regression)
- **Geometric functional analysis** (sections of convex sets)

Where to find literature:

- Compressed Sensing Webpage at Rice University
  `http://dsp.rice.edu/cs`
- My webpage at Michigan: recent papers with Yaniv Plan
  `www.umich.edu/~romanv`

Thank you!