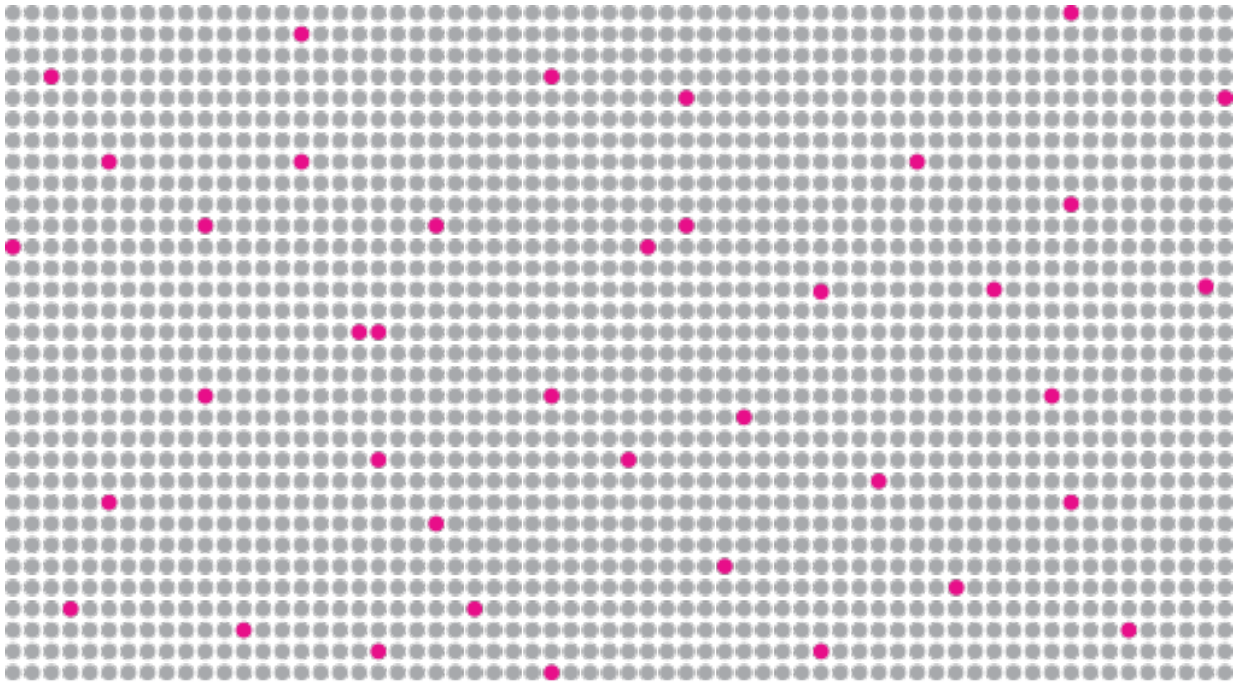
Recovering sparse high dimensional data:

how to do it in the cheapest possible way?



Ivan Nourdin (Université du Luxembourg)

Let $x_0 \in \mathbb{R}^d$, where d is meant to be large (100, 1000, 10^6, etc.)

Our problem: We want to *acquire* x_0 with the *smallest* possible number of *linear* measurements.

Mathematically speaking

- One considers $x_0 \in \mathbb{R}^d$ (unknown)

- One makes m linear measurements of x_0: $\langle a_1, x_0 \rangle, \dots, \langle a_m, x_0 \rangle$.

That is, one observes $Ax_0 \in \mathbb{R}^m$, where $A = \begin{pmatrix} \overline{\quad a_1 \quad} \\ \vdots \\ \overline{\quad a_m \quad} \end{pmatrix} \in \mathbb{R}^{m \times d}$.

Notations that will be used throughout the talk:

- d is the ambient dimension

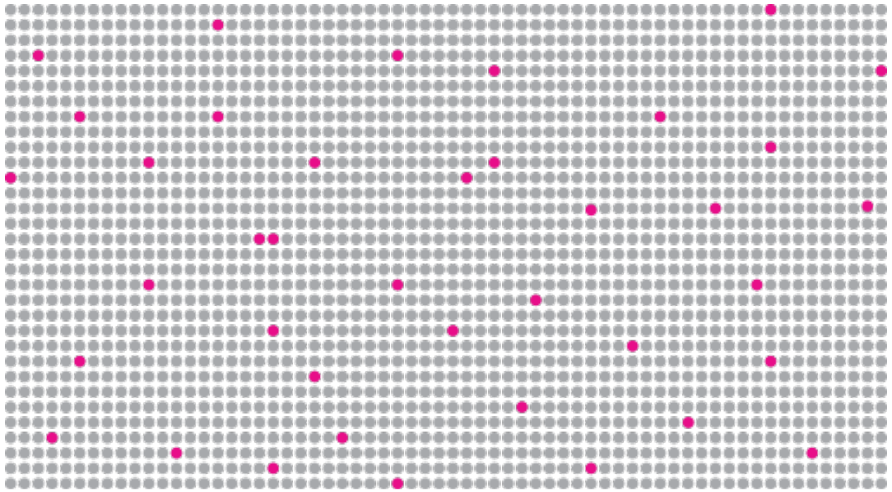
- m is the number of measurements

- $A \in \mathbb{R}^{m \times d}$ is the measurement matrix

Problem: "I give you A and Ax_0. Are you able to recover x_0?".

If $m < d$, we are dealing with an *undetermined* system, so there is no hope to provide a positive answer.

Extra assumption: x_0 is s-sparse, that is, at most s of its entries are nonzero.



Questions:

- is such an assumption realistic in practice? (which practice?)

- what is the gain of doing such an assumption?

Example 1: pictures taken with smartphones

- Assume x_0 encodes a picture of size $n \times m$, e.g. $n = 3456$, $m = 4608 \Rightarrow d \simeq 1,6 \times 10^7$

- Each entry of x_0 has a value between 0 (black) and 15 (white), depending on the luminosity at the corresponding pixel

- x_0 itself is not sparse. But $\widetilde{x}_0 \in \mathbb{R}^d$ defined as $\widetilde{x}_0(k) = x_0(k) - x_0(k-1)$ $(2 \leq k \leq d)$ is. And if you know $\widetilde{x}_0$, you know x_0.

Example 2: group testing.

- Consider a population of d persons (d large), of which a *small* proportion (representing, say, $s << d$ persons) is sick.

- Suppose one can determine whether a given person is sick or healthy by means of a blood test. Taking blood is easy, but testing it is costly (in time or in money).

- We want to find all the sick persons.

- The first (bad!) idea is to test each person *individually*. It would lead to d blood tests to only find s sick persons.

- There is in fact a much better strategy to apply!

- Mathematically speaking, we can model the situation by means of an (unknown) string $x_0 \in \{0,1\}^d$ with at most s ones in it.

- We are allowed to test any subset $S \subset [d]$ of the indices. The answer to the test tells whether $x_0(i) = 0$ for <u>all</u> $i \in S$ or not.

- Subdividing into two groups of 'same' cardinality at each step, we can find all the ones in about $\boxed{s \log_2 d}$ steps.

Example 3: Medical Resonance Imagery 

Example 4: Seismology 

Example 5: High-resolution radar 

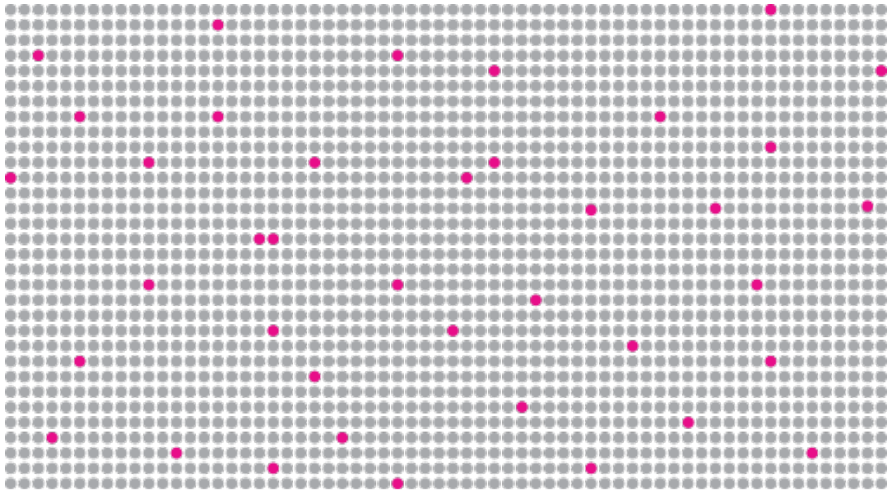
Example 6: Analog-to-digital converters 

.....

....

..

Extra assumption: x_0 is s-sparse, that is, at most s of its entries are nonzero.



Questions:

- is such an assumption realistic in practice? **YES**

- what is the gain of doing such an assumption?

What is the gain of assuming sparsity?

In the two papers

- E. Candès, J. Romberg, and T. Tao. "*Robust Uncertainty principles: Exact signal reconstruction from highly incomplete frequency information.*" IEEE Trans. Information Theory, 2006

- D. Donoho. "*Compressed sensing.*" IEEE Trans. Information Theory, 2006

a new method was released, allowing one to significantly reduce the number of measurements that are actually needed to recover a *sparse* signal., and giving raise to a entirely new fields of research, the '*Compressed sensing*' .

- One considers $x_0 \in \mathbb{R}^d$ (unknown) and one assumes it is s-sparse

- One makes m linear measurements of x_0. That is, one observes $Ax_0 \in \mathbb{R}^m$, where $A = \left(\begin{array}{c} a_1 \\ \hline \vdots \\ \hline a_m \end{array}\right) \in \mathbb{R}^{m \times d}$.

- Can we 'reconstruct' x_0 from Ax_0? If yes:

* what is the minimal value required for m?

* how to choose the matrix A?

Proposition. Assume that $m \geq 2s$ and that any $2s$ raws of A are linearly independent (such a matrix is easy to build). Then, any s-sparse vector $x_0 \in \mathbb{R}^d$ can be uniquely recovered from the knowledge of A and of $Ax_0 \in \mathbb{R}^m$.

Proof. If x_0 and x_0' are both s-sparse, then their difference $x_0 - x_0'$ is $2s$-sparse. If, furthermore, they satisfy $Ax_0' = Ax_0$, then $x_0' = x_0$ necessarily. ∎

If the conditions of the proposition are satisfied, to recover x_0 from A and Ax_0 we are thus left to solve a minimization problem:

$$(P_0): \qquad \min_x \|x\|_0 \quad \text{subject to } Ax = Ax_0,$$

where $\|x\|_0$ is the cardinality of the support of x.

Is the problem (already) over?

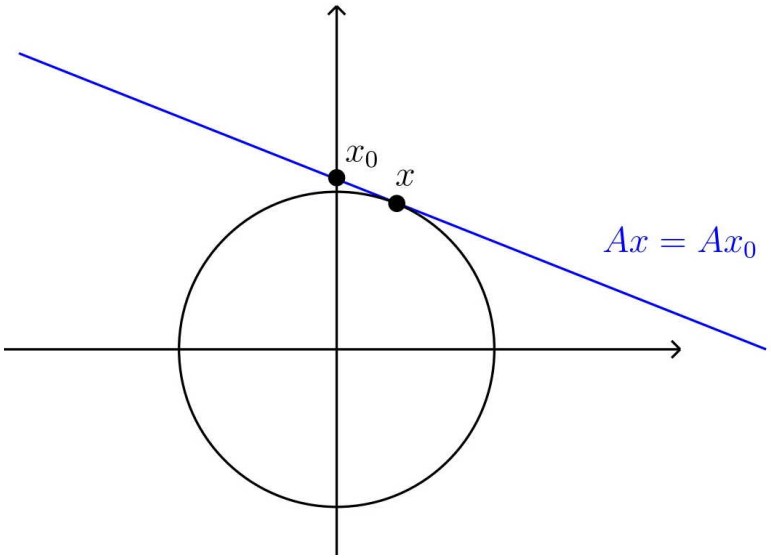
- In order to solve (P_0), we have to consider all the possible supports for x_0 and then to solve the corresponding systems.

- For instance, suppose that $d = 1000$ and $s = 10$. We have to solve $\binom{1000}{10} \geq 10^{20}$ linear systems of size 10×10. Each such system can be solved in 10^{-10} seconds. Then, the time required to solve (P_0) is around 10^{10} seconds, i.e., more than... 300 years!

- What is easy and quick, contrary to (P_0), is to solve

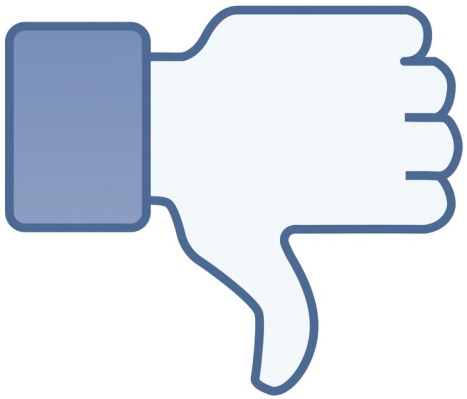
$$(P_2): \qquad \min_x \|x\|_2 \quad \text{subject to } Ax = Ax_0.$$

- Indeed (least square method): one can check that the solution of (P_2) is explicitely given by $x = {}^tA(A^tA)^{-1}Ax_0$.

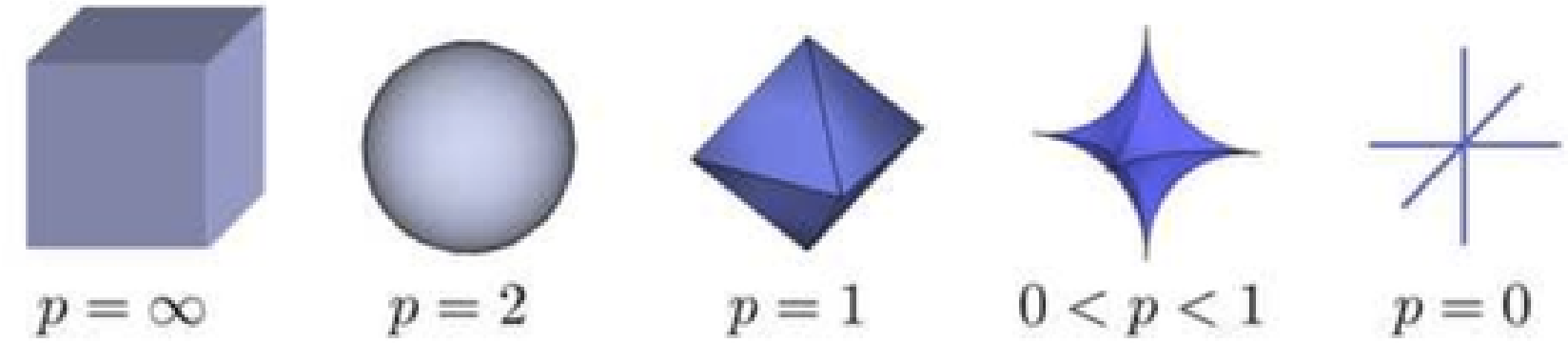


Unfortunately, especially in high dimension, the solution x of (P_2) is likely to be very far away from the expected solution x_0.

So, despite being easy to implement, this approach is of no help to solve our problem.



We have to find another idea!

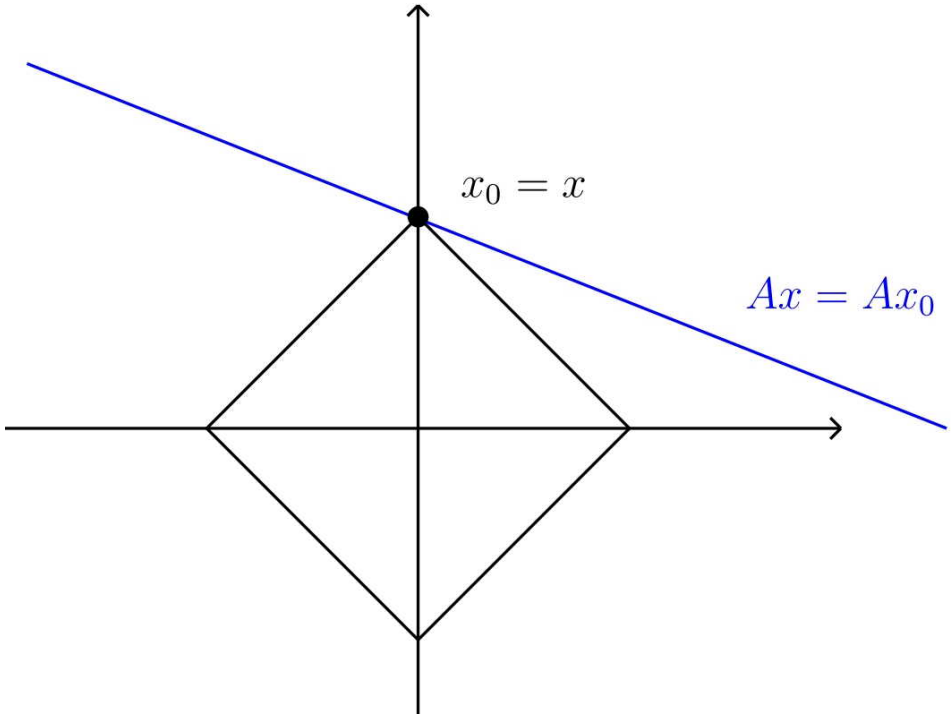


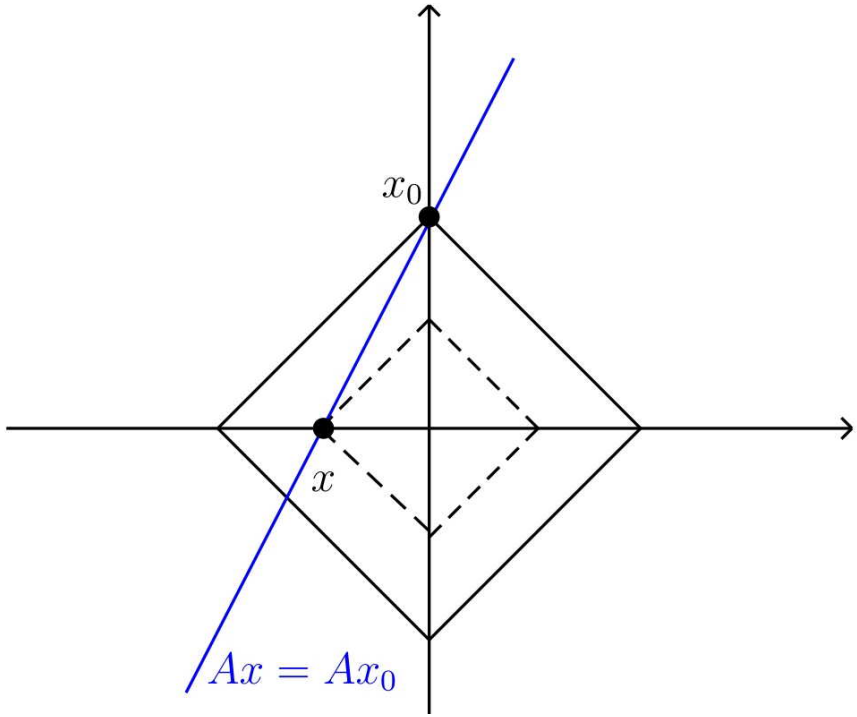
Use the ℓ_1 norm, that is, consider

$$(P_1): \qquad \min_x \|x\|_1 \quad \text{subject to } Ax = Ax_0.$$

In practice, one can solve (P_1) by using the simplex algorithm, which is quick and efficient!

Why does it work?





A famous and representative result in the theory of *compressed sensing* is the following theorem (or "how to solve a deterministic problem by introducing randomness")

Theorem (à la Candès, Romberg and Tao). Consider an integer $m \geq 2\beta s \log d + s$ where $\beta > 1$ is fixed. Assume that $A \in \mathcal{M}_{m\times d}(\mathbb{R})$ is *Gaussian*, more precisely that its entries are independent $N(0, 1/m)$ random variables. Finally, let x_0 be an s-sparse vector of $\mathbb{R}^d$. Then, with probability at least

$$1 - \frac{2}{d^{f(\beta,s)}},$$

one has that x_0 is the unique minimizer to the program

$$(P_1): \qquad \min_x \|x\|_1 \quad \text{subject to } Ax = Ax_0.$$

The fonction f is given by $f(\beta, s) = \left[\sqrt{\frac{\beta}{2s} + \beta} - \sqrt{\frac{\beta}{2s}}\right]^2$. It is increasing in s (for fixed β) and in β (for fixed s).

A classical experiment (following Donoho and Tanner)

- Fix a large d, say $d = 100$.

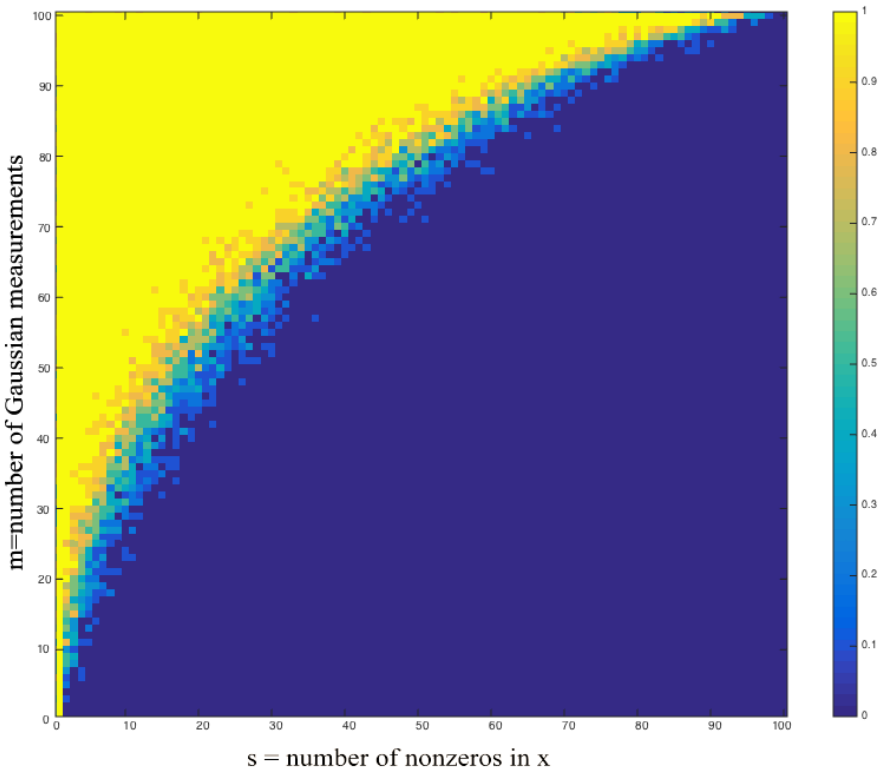
- Consider a pair $(s, m) \in \{1, \dots, d\}^2$ (the values for s and m will then vary).

- Pick a s-sparse vector $x_0 \in \mathbb{R}^d$ at random.

- Compute Ax_0 with $A \in \mathcal{M}_{m,d}(\mathbb{R})$ a random Gaussian matrix. Apply the simplex algorithm. If you (don't) get x_0, then consider it is a success (failure).

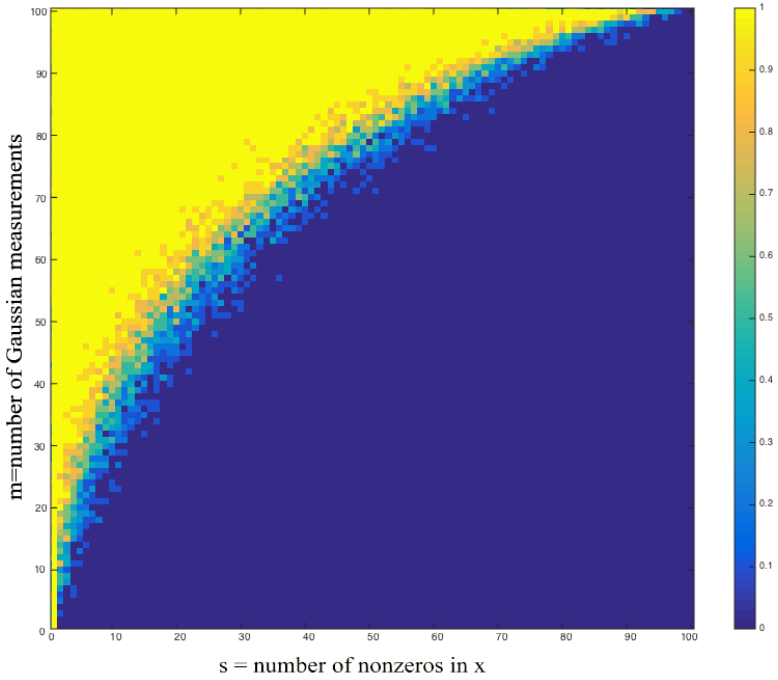
- For each possible value of s and m, repeat this experiment 10 times, and color the point of coordinates (s, m) with the rule:

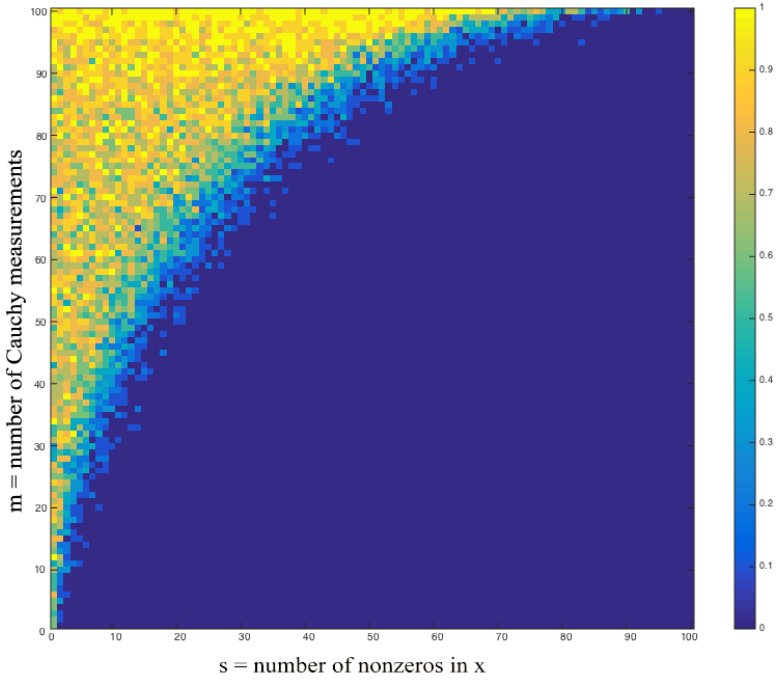
10 successes $\rightarrow$ ● ... 5 successes $\rightarrow$ ● ... no success $\rightarrow$ ●



One observes a strong **phase transition**. The equation for the boundary is very close to $m = 2s\log(d/s)+2s$ and, as such, agrees with Candès, Romberg and Tao's result.

An observation (Gaussian vs Cauchy measurements)





In order to understand the previous threshold phenomenon (in the Gaussian case only), we will analyze it in a more general framework.

It will require different mathematical tools, mainly coming from three distinct areas:

* Gaussian analysis

* geometry of convex cones

* concentration of measure

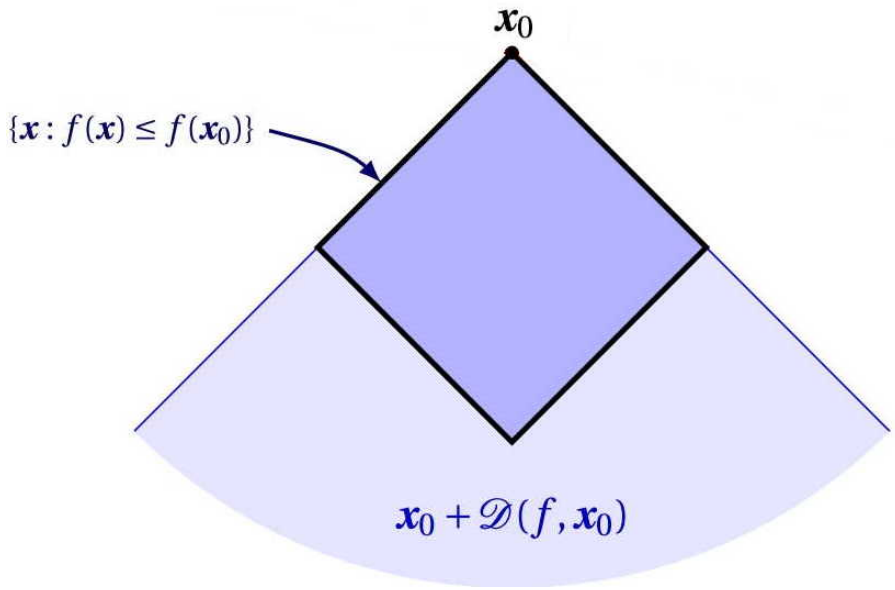
- D. Amelunxen, M. Lotz, M.B. McCoy, and J.A. Tropp. **"Living on the edge: phase transitions in convex programs with random data."** *Inform. Inference*, to appear.

- M.B. McCoy and J.A. Tropp. **"From Steiner formulas for cones to concentration of intrinsic volumes"**. *Discrete Comput. Geom.*, 2014.

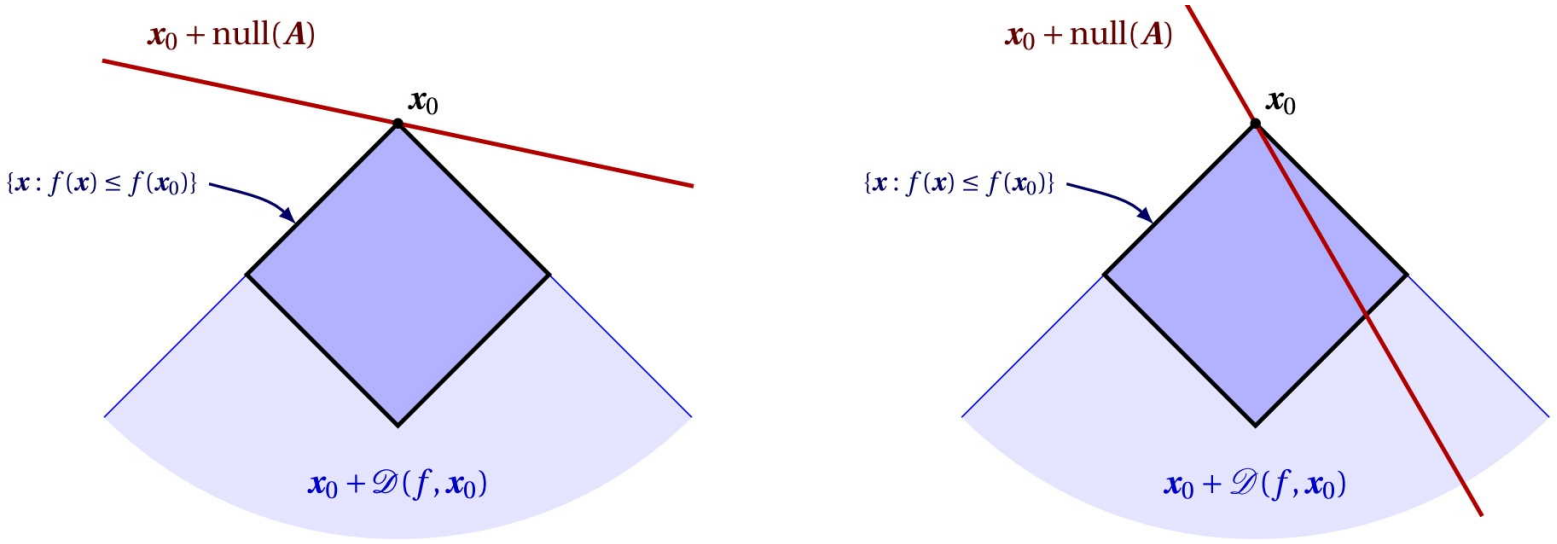
Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function and let us consider the minimization problem:

$$(P): \qquad \min_x f(x) \quad \text{subject to } Ax = Ax_0.$$

Definition. The descent cone of f at x_0 is $\mathcal{D}(f, x_0) = \{y \in \mathbb{R}^d : \exists \tau > 0 \text{ s.t. } f(x_0 + \tau y) \leq f(x_0)\}$.



Fact 1. One has that x_0 is the unique solution to (P) if and only if $\mathcal{D}(f, x_0) \cap \mathsf{null}(A) = \{0\}$. $(\mathsf{null}A = \ker A)$



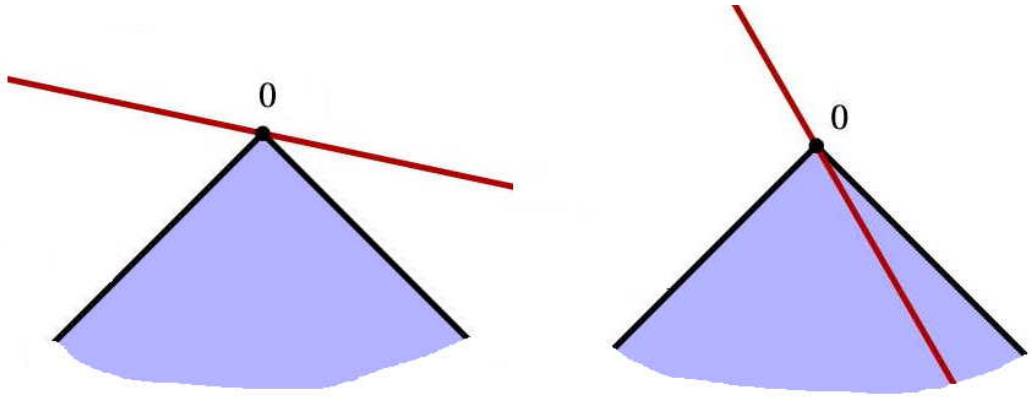
Fact 2. Since A is Gaussian, its law is invariant by any orthogonal transformation. As a result, $\mathsf{null}(A)$ is distributed as QL_{d-m}, where Q is chosen at random in $O(d)$ and $L_{d-m} \subset \mathbb{R}^d$ stands for any (fixed) subspace of dimension $d - m$.

After translation by $-x_0$, our problem can be interpreted as a question coming from *stochastic geometry*.

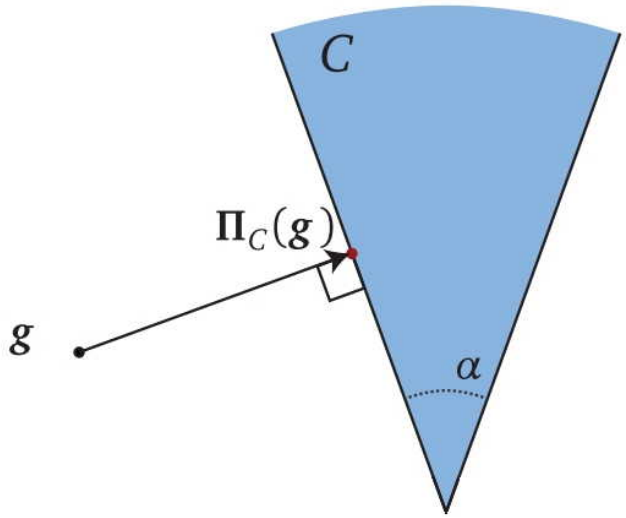
Problem: A closed convex cone C and a subspace $L_{d-m} \subset \mathbb{R}^d$ of dimension $d-m$ being given, compute the probability that

$$C \cap QL_{d-m} \neq \{0\},$$

where Q is chosen at random in $O(d)$.



Towards the Crofton formula (polyhedral case)

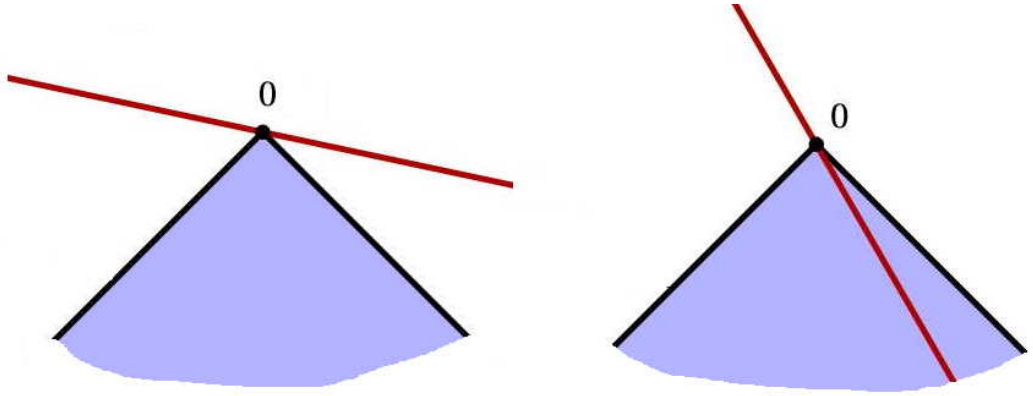


$$v_0(C) = \tfrac{1}{2} - \tfrac{\alpha}{2\pi},\ v_1(C) = \tfrac{1}{2},\ v_2(C) = \tfrac{\alpha}{2\pi}$$

Let $\Pi_C : \mathbb{R}^d \to C$ be the projection onto the polyhedral cone C. Define the *intrinsic volumes* $\{v_k(C)\}_{k=0,\ldots,d}$ of C as

$$v_k(C) = P\{\Pi_C(\mathrm{g}) \text{ belongs to } k\text{-dimensional face of } C\},$$

where g denotes a standard Gaussian vector of $\mathbb{R}^d$.



One has

$$P(\text{blue cone} \cap Q(\text{red line}) \neq \{0\}) = 2\,\frac{\text{angle}}{2\Pi} = 2v_2(C)$$

Crofton's formula. Provided C is not a subspace, one has

$$\begin{aligned} P(C \cap QL_{d-m} \neq \{0\}) &= 2 \sum_{\substack{j=m+1 \\ j-m-1 \text{ even}}}^{d} v_j(C) \\ &= 2v_{m+1}(C) + 2v_{m+3}(C) + \dots . \end{aligned}$$

Intrinsic volumes of a closed convex cone are positive and sum to 1. One can therefore consider a random variable V_C defined as

$$P(V_C = k) = v_k(C), \quad k = 0, 1, \ldots, d.$$

And by playing a little bit with the Crofton's formula and the definition of V_C, one obtains the 'interlacing property':

$$P(V_C \leq m-1) \leq P(C \cap QL_{d-m} = \{0\}) \leq P(V_C \leq m).$$

Thus: $\boxed{P(x_0 \text{ is the unique solution of (P)}) \approx P(V_C \leq m)}$.

We are now left to study $P(V_C \leq m)$. To do so, we shall rely on a last and final ingredient, the *Master Steiner Formula* of McCoy and Tropp (2013).

This formula is particularly useful in our context, as it provides a clear bridge between the abstract random variable V_C and the concrete random variable $\|\Pi_C(\mathbf{g})\|^2$, where Π_C is the projection onto the cone C and $\mathbf{g} \sim N(0, I_d)$.

It is our gateway towards classical results of concentration of measure (Talagrand, Ledoux, ...). By pushing this idea further, one can mathematically prove the phase transition (see Amelunxen *et al.* and McCoy and Tropp).

In this talk and from now on, we won't interested in bounds but, instead, in *exact* asymptotic for $P(V_C \leq m)$ ($\to$ CLT)

Master Steiner Formula (McCoy, Tropp) Let $\Pi_C : \mathbb{R}^d \to C$ be the projection onto the closed convex cone C. Let $\mathbf{g} \sim N(0, I_d)$. Let $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ be a measurable function. One has

$$E[\varphi(\|\Pi_C(\mathbf{g})\|^2)] = \sum_{k=0}^{d} E[\varphi(X_k)]P(V_C = k),$$

where X_k is distributed according to the χ^2 law with k degrees of freedom.

Otherwise stated, $\boxed{\|\Pi_C(\mathbf{g})\|^2 \overset{\text{(law)}}{=} \sum_{i=1}^{V_C} \eta_i^2}$, where $\eta_1, \eta_2, \ldots \sim N(0,1)$ are independent and also independent from V_C.

Corollary/Definition: Statistical dimension δ_C of a closed convex cone C is defined as $E[\|\Pi_C(\mathbf{g})\|^2] = E[V_C]$.

Another corollary of the Master Steiner Formula is that

$$E[e^{\eta V_C}] = E[e^{\xi \|\Pi_C(\mathbf{g})\|^2}], \quad \text{with } \xi = \tfrac{1}{2}(1 - e^{-2\eta}).$$

An interesting consequence is the following. If C_d is a sequence of closed convex cone of $\mathbb{R}^d$ such that $E(V_{C_d}) = \delta_{C_d} \to \infty$ and $\liminf \mathsf{Var}(V_{C_d})/\delta_{C_d} > 0$ as $d \to \infty$, then

$$\frac{V_{C_d} - \delta_{C_d}}{\sqrt{\mathsf{Var}(V_{C_d})}} \to N(0,1) \quad \text{iff} \quad \frac{\|\Pi_{C_d}(\mathbf{g})\|^2 - \delta_{C_d}}{\sqrt{\mathsf{Var}(\|\Pi_{C_d}(\mathbf{g})\|^2)}} \to N(0,1).$$

Theorem (Goldstein, Nourdin, Peccati). Let $x_0 \in \mathbb{R}^d$, let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function and let $C = \{y \in \mathbb{R}^d : \exists \tau > 0 \text{ such that } f(x_0 + \tau y) \leq f(x_0)\}$ be the descent cone of f at x_0.

Consider the minimization problem

$$(P): \qquad \min_x f(x) \quad \text{subject to } Ax = Ax_0,$$

where $A \in \mathcal{M}_{m\times d}(\mathbb{R})$ is *Gaussian* (all its entries are independent $N(0,1)$ random variables) and where $\boxed{m = \lfloor \delta_C + t\sqrt{\mathsf{Var}(V_C)} \rfloor, \ \ t \in \mathbb{R}}$.

Suppose that $E(V_C) = \delta_C \to \infty$ and that $\liminf \mathsf{Var}(V_C)/\delta_C > 0$ as $d \to \infty$.

Then, as $d \to \infty$,

$$\boxed{P(x_0 \text{ is the unique solution of (P)}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{u^2}{2}} du + O(\frac{1}{\sqrt{\log \delta_C}}).}$$

Reading the phase transition.

For instance, selecting $m \geq \delta_C + 1.6\sqrt{\mathsf{Var}(V_C)}$, one has

$$P(x_0 \text{ is the unique solution of (P)}) \geq 0.95.$$

In contrast, for $m \leq \delta_C - 1.6\sqrt{\mathsf{Var}(V_C)}$,

$$P(x_0 \text{ is the unique solution of (P)}) \leq 0.05$$

It follows that the phase transition happens on an interval of length $3.2\sqrt{\mathsf{Var}(V_C)}$.

(*Remark*: A crude bound is $\mathsf{Var}(V_C) \leq 2\delta_C$.)