

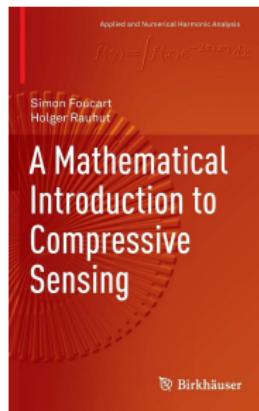
# **Compressive sensing, generalized sparsity**

Nicolas Keriven (w/ slides by Claude Petit and Aline Roumy)

December 6, 2023

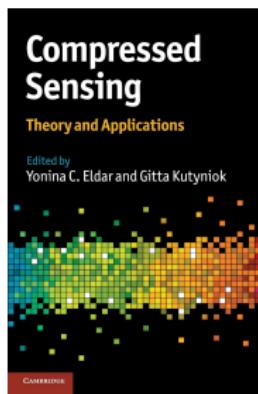
# Course material

S. Foucart, H. Rauhut, **A mathematical introduction to compressive sensing**, Birkhäuser, 2013.



# Course material

Compressed Sensing: Theory and Applications, Edited by Y.C. Eldar and G. Kutyniok, Cambridge University Press, 2012.



# Introduction

# Outline

## Introduction

- What is Compressive Sensing?

- Notations (Reminder)

- Problem formulation

## Recovery guarantees

- NSP

- Dual certificate

- Coherence

- RIP

## Recovering with random matrices?

## Concentration inequalities and proving the RIP

## Beyond Sparsity

- Total Variation

- Structured sparsity

- Matrix completion and Low-rank regularization

## Infinite-dimensional signals: superresolution and compressive learning

- Continuous sparsity: superresolution

- Generalized sparsity: sketching

# What is compressive sensing?

**Compressive sensing:** a way to acquire (or sense or sample) and compress data.

**Classical =**

sampling then compression

**Compressive sensing =**

sampling **AND** compression

# What is compressive sensing?

**Compressive sensing:** a way to acquire (or sense or sample) and compress data.

**Classical =**

sampling then compression

**Compressive sensing =**

sampling **AND** compression

**Several names exist:**

- compressed sensing
- compressed sampling
- compressive sampling
- **compressive sensing.** More accurate. Chosen in this course.  
The one of the reference book.

# What is compressive sensing?

**Compressive sensing:** a way to acquire (or sense or sample) and compress data.

**Classical =**

sampling then compression

**Compressive sensing =**

sampling **AND** compression

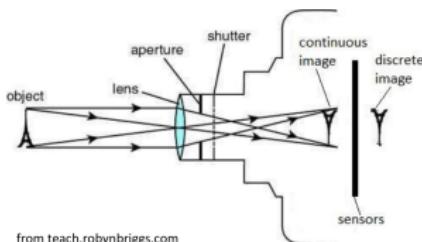
**Several names exist:**

- compressed sensing
- compressed sampling
- compressive sampling
- **compressive sensing.** More accurate. Chosen in this course.  
The one of the reference book.

**Is CS "old school"?** It can seem that way, since the advent of learning-based systems. But the principles and notions are fundamental, and underpin even the most advanced Deep Neural Net.

# Digital camera

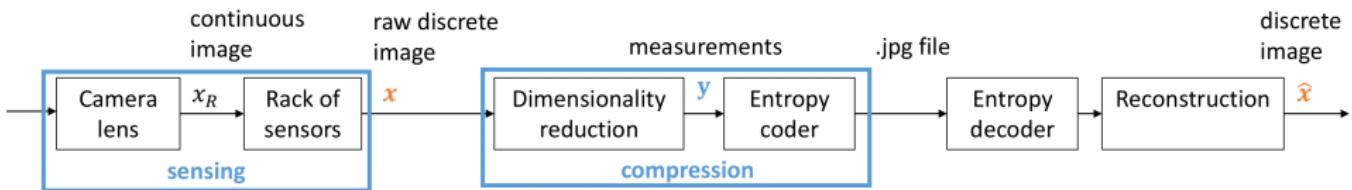
Digital camera: converts an image into **digital** data and **compress** it.



$$x_R: [0,1]^2 \rightarrow \mathbb{R}^3$$

$$\textcolor{brown}{x}: \{1, N_a\} \times \{1, N_b\} \rightarrow \{0, 255\}^3$$

$$\textcolor{blue}{y}: \{1, M_a\} \times \{1, M_b\} \rightarrow \{0, 255\}^3$$



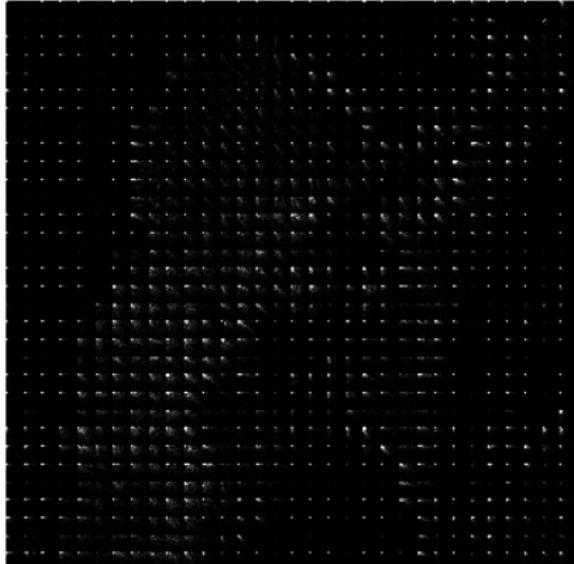
# Compression

How to compress a signal? Key concept: **few degrees of freedom in the transform domain → sparsity**

Left: image



Right: discrete cosine transform of image



# Classical Compression: dimensionality reduction with $s$ -term approximation

1. Transform the signal  $c = \Phi x$
2. Dimensionality reduction: keep the  $s$  coefficients  $c_s$  with largest absolute value (non-linear!)
3. Reconstruction:  $\hat{x} = \Phi^{-1} c_s$

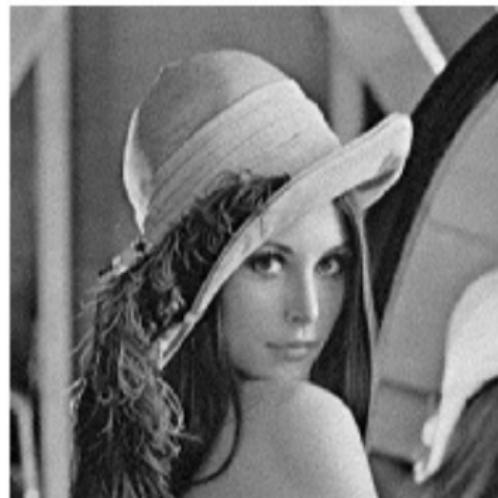
# Classical Compression: dimensionality reduction with $s$ -term approximation

1. Transform the signal  $c = \Phi x$
2. Dimensionality reduction: keep the  $s$  coefficients  $c_s$  with largest absolute value (non-linear!)
3. Reconstruction:  $\hat{x} = \Phi^{-1} c_s$

**Left:** 1% kept

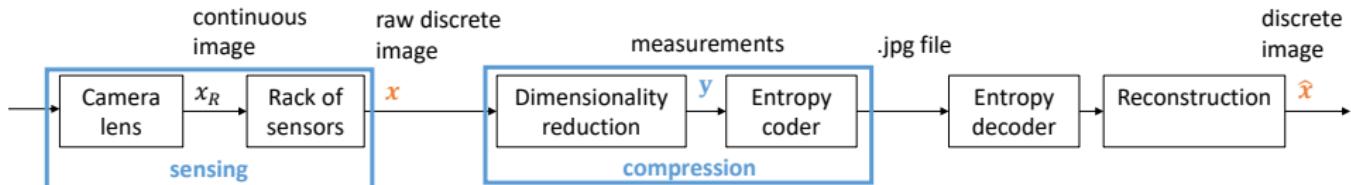


**Right:** 5% kept



# Classical vs compressive sensing

## Classical

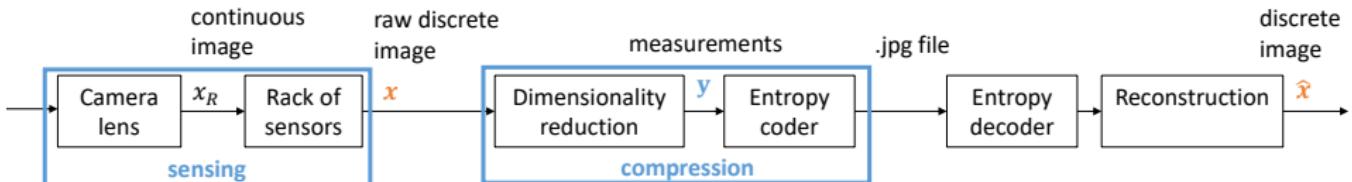


take **lots of samples** (Shannon-Nyquist),  
throw **most** of the coefficients away

$$\begin{aligned} \textcolor{orange}{x}: [1, N_a] \times [1, N_b] &\rightarrow \{0, 255\}^3 \\ \textcolor{blue}{y}: [1, M_a] \times [1, M_b] &\rightarrow \{0, 255\}^3 \\ (M_a M_b \ll N_a N_b) \end{aligned}$$

# Classical vs compressive sensing

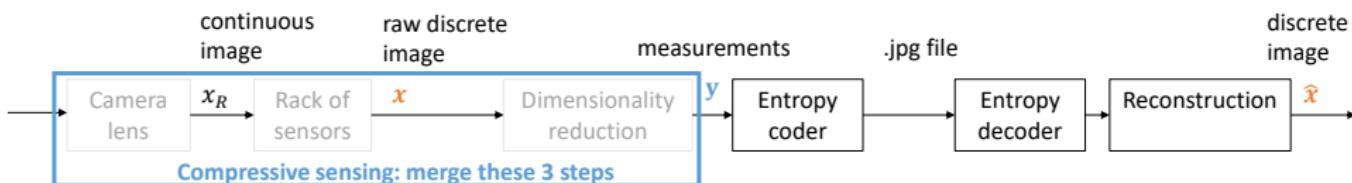
## Classical



take **lots of samples** (Shannon-Nyquist),  
throw **most** of the coefficients away

$$\begin{aligned} \mathbf{x} &: [1, N_a] \times [1, N_b] \rightarrow \{0, 255\}^3 \\ \mathbf{y} &: [1, M_a] \times [1, M_b] \rightarrow \{0, 255\}^3 \\ (M_a M_b) &\ll (N_a N_b) \end{aligned}$$

**Compressive sensing:** can we acquire less data in the first place?  
and still recover  $\hat{x}$ ?



# Can we sample signals at the “Information Rate”?

Yes, we can!



Wikipedia.

E. J. Candes and T. Tao, 2005  
“Decoding by linear programming”



Wikipedia.

D. L. Donoho, 2006  
“Compressed sensing”

# Can we sample signals at the “Information Rate”?

Yes, we can!



Wikipedia.

E. J. Candes and T. Tao, 2005  
“Decoding by linear programming”



Wikipedia.

D. L. Donoho, 2006  
“Compressed sensing”

→ We can sense/sample signals directly proportionally to their complexity/sparsity

## A bit of history...

1795 **Prony's method**: estimation of parameters of a small number of complex exponentials

## A bit of history...

1795 [Prony's method](#): estimation of parameters of a small number of complex exponentials

1900s [Carathéodory](#) show that *any* linear combination of  $k$  sinusoids are uniquely determined by *any*  $2k$  samples (independent of the frequency!)

## A bit of history...

- 1795 [Prony's method](#): estimation of parameters of a small number of complex exponentials
- 1900s [Carathéodory](#) show that *any* linear combination of  $k$  sinusoids are uniquely determined by *any*  $2k$  samples (independent of the frequency!)
- 1936 [Beurling](#): recovering a signal from a *partial* observation of its Fourier transform

## A bit of history...

1795 [Prony's method](#): estimation of parameters of a small number of complex exponentials

1900s [Carathéodory](#) show that *any* linear combination of  $k$  sinusoids are uniquely determined by *any*  $2k$  samples (independent of the frequency!)

1936 [Beurling](#): recovering a signal from a *partial* observation of its Fourier transform

2000s [Blu, Marzino, Vetterli](#): generalization to certain signals, recovering  $k$  signals from  $2k$  samples

## A bit of history...

1795 **Prony's method**: estimation of parameters of a small number of complex exponentials

1900s **Carathéodory** show that *any* linear combination of  $k$  sinusoids are uniquely determined by *any*  $2k$  samples (independent of the frequency!)

1936 **Beurling**: recovering a signal from a *partial* observation of its Fourier transform

2000s **Blu, Marzino, Vetterli**: generalization to certain signals, recovering  $k$  signals from  $2k$  samples

2005 - 2008 **Candès, Tao, Romberg, Donoho**: Compressive Sensing

# More difference between Classical sampling and CS

## Classical

- sample *then* compress
- infinite-length, continuous signals
- sampling pointwise *in time*
- recovery is “simple” interpolation

## Compressive Sensing

- directly gather *less* samples
- Finite-dimensional signals (mostly...?)
- inner-product of the *entire* signal with a *few* vectors
- recovery is *complex*, *non-linear*

# Compressive Sensing in a nutshell

To recover an  $s$ -sparse signal in dimension  $n$ , take  $m \gtrsim s \log(n/s)$  measurements with a *random, dense* sensing operator.

# Outline

## Introduction

What is Compressive Sensing?

Notations (Reminder)

Problem formulation

## Recovery guarantees

NSP

Dual certificate

Coherence

RIP

## Recovering with random matrices?

Concentration inequalities and proving the RIP

## Beyond Sparsity

Total Variation

Structured sparsity

Matrix completion and Low-rank regularization

## Infinite-dimensional signals: superresolution and compressive learning

Continuous sparsity: superresolution

Generalized sparsity: sketching

# Norms

## Definition ( $l_p$ -norm)

The  $l_p$ -norm of  $x \in \mathbb{R}^n$ ,  $p > 1$  is defined as

$$\|x\|_p = \begin{cases} \left( \sum_{i=1}^n |x_i|^p \right)^{1/p} & p \in [1, \infty) \\ \max_i |x_i| & p = \infty \end{cases}$$

If  $p < 1$ , definition still valid, but triangle inequality not satisfied  
⇒ **quasi-norm**.

## Definition (support and $l_0$ -norm)

The **support** of a vector  $x$  is the index set of its non-zero entries, i.e.

$$\text{supp}(x) = \{j \in [n] : x_j \neq 0\}, \text{ where } [n] = \{1, 2, \dots, n\}$$

The  **$l_0$ -norm** of  $x$  is defined as

$$\|x\|_0 = \text{card}(\text{supp}(x))$$

$\|x\|_0$  counts the **number of non-zero entries** of  $x$ .

$\|.\|_0$  is **not even a quasi-norm**: it is not homogeneous  $\|\lambda x\|_0 \neq |\lambda| \|x\|_0$

# Sparsity definition

## Definition ( $s$ -sparse)

A signal  $x \in \mathbb{R}^n$  is said to be  $s$ -sparse if it has at most  $s$  non-zero entries, i.e.  
 $\|x\|_0 \leq s$ .

## Definition ( $\Sigma_s$ )

We define  $\Sigma_s$  as the set containing all  $s$ -sparse signals, i.e.

$$\Sigma_s = \{x \in \mathbb{R}^n : \|x\|_0 \leq s\}.$$

**Note 1:** Sparsity is a highly nonlinear model ( $\Sigma_s$  is not a linear subspace: it is a union of subspaces)

**Note 2:** in many practical cases,  $x$  is not sparse itself, but it has a sparse representation in some basis  $\Phi$ . We still say that  $x$  is  $s$ -sparse, with the understanding that we can write  $x = \Phi u$ , and  $\|u\|_0 \leq s$ .

# Approximate sparsity

- A sparse signal can be **represented exactly** giving the **positions** and **values** of its  $s$  nonzero components

## Approximate sparsity

- A sparse signal can be **represented exactly** giving the **positions** and **values** of its  $s$  nonzero components
- Real-world signals are **rarely exactly sparse.**

# Approximate sparsity

- A sparse signal can be **represented exactly** giving the **positions** and **values** of its  $s$  nonzero components
- Real-world signals are **rarely exactly sparse**.
  - ▶ generalize the definition from “sparse” to “**compressible**” signals,

# Approximate sparsity

- A sparse signal can be **represented exactly** giving the **positions** and **values** of its  $s$  nonzero components
- Real-world signals are **rarely exactly sparse.**
  - ▶ generalize the definition from “**sparse**” to “**compressible**” signals,
  - ▶ describe the **representation error**, i.e. the error incurred representing just  $s$  components of the signal.



# Best $s$ -term approximation

The **best  $s$ -term approximation** picks the  $s$  components that minimize the representation error

## Definition (best $s$ -term approximation)

For  $p > 0$ , the  $l_p$ -error incurred by the **best  $s$ -term approximation** to a vector  $x \in \mathbb{R}^n$  is given by

$$\sigma_s(x)_p = \min_{\hat{x} \in \Sigma_s} \|x - \hat{x}\|_p$$

- If  $x \in \Sigma_s$ , then  $\sigma_s(x)_p = 0$  for any  $p$ .
- $\hat{x}$  is easy to compute: keep the  $s$  largest elements (in absolute value), set all others to zero.
- $\hat{x}$  does not depend on  $p$ , but  $\sigma_s(x)_p$  does!

# Sparsity support

Suppose  $x \in \mathbb{R}^n$ . Let  $S \subset [n]$  and  $S^c = [n] \setminus S$

- $S$ : **sparsity support** of  $x$ , i.e. the locations of the nonzero coefficients of  $x \rightarrow$  the **key** in recovering  $x$ !
- $S^c$ : set of locations of the 0 coefficients
- $S$  for compressible signal: set of locations of the  $s$  largest coefficients

# Sparsity support

Suppose  $x \in \mathbb{R}^n$ . Let  $S \subset [n]$  and  $S^c = [n] \setminus S$

- $S$ : **sparsity support** of  $x$ , i.e. the locations of the nonzero coefficients of  $x \rightarrow$  the **key** in recovering  $x$ !
- $S^c$ : set of locations of the 0 coefficients
- $S$  for compressible signal: set of locations of the  $s$  largest coefficients

## Notation

For  $x \in \mathbb{R}^n$ ,

$$x_S \in \begin{cases} \mathbb{R}^n & \text{by setting the entries of } x \text{ indexed by } S^c \text{ to 0} \\ \mathbb{R}^{|S|} & \text{by keeping only the entries indexed by } S \end{cases}$$

and for  $M \in \mathbb{R}^{m \times n}$ ,

$$M_S \in \begin{cases} \mathbb{R}^{m \times n} & \text{by setting the columns of } M \text{ indexed by } S^c \text{ to 0} \\ \mathbb{R}^{m \times |S|} & \text{by keeping only the columns indexed by } S \end{cases}$$

Clear from the context. For an  $s$ -sparse vector with support  $S$ ,

$$Mx = M_S x_S$$

# Outline

## Introduction

What is Compressive Sensing?

Notations (Reminder)

Problem formulation

Recovery guarantees

NSP

Dual certificate

Coherence

RIP

Recovering with random matrices?

Concentration inequalities and proving the RIP

Beyond Sparsity

Total Variation

Structured sparsity

Matrix completion and Low-rank regularization

Infinite-dimensional signals: superresolution and compressive learning

Continuous sparsity: superresolution

Generalized sparsity: sketching

# Sensing process model

## Modeling the dependency between signal and measurement

Let  $x \in R^n$  be a **s-sparse signal** to be recovered.

Let  $y \in R^m$ ,  $m < n$ , be **linear measurements** of the signal as

$$y = Mx$$

with  $M \in R^{m \times n}$ , being the **sensing matrix**.

$$\begin{matrix} y \\ \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{matrix} = \begin{matrix} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{matrix} \cdot \begin{matrix} x \\ \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{matrix}$$

The diagram illustrates the sensing process. On the left, a vertical vector  $y$  is shown as a stack of colored squares (purple, yellow, green, red). In the center, the equation  $y = Mx$  is displayed, where  $M$  is represented by a  $5 \times 10$  grid of colored squares. The grid has a repeating pattern of columns: purple, yellow, green, red, blue, magenta, green, red, blue, magenta. To the right of the multiplication sign is a vertical vector  $x$  consisting of 10 white squares, representing a sparse signal. The dot product of  $y$  and  $M$  results in the vector  $x$ .

# Reconstruction: problem formulation

## problem formulation

Given measurement  $y$ , sensing matrix  $M$  and the model  $y = Mx$ , Recover  $x$ ,  $s$ -sparse.

$$\begin{matrix} \text{y} \\ \text{M} \\ \text{x} \end{matrix} = \begin{matrix} \text{y} \\ \text{M} \\ \text{x} \end{matrix}$$

The diagram illustrates the linear system  $y = Mx$ . On the left, there is a vertical vector  $y$  composed of four colored squares: purple, yellow, green, and red. In the center is a 4x4 grid labeled  $M$ , containing smaller colored squares (purple, yellow, green, red) at various positions. To the right is a vertical vector  $x$  composed of four colored squares: red, white, purple, and orange. The equation  $y = Mx$  is written above the grid  $M$ .

# Reconstruction: problem formulation

## problem formulation

Given measurement  $y$ , sensing matrix  $M$  and the model  $y = Mx$ , Recover  $x$ ,  $s$ -sparse.

## Difficulties?

- Underdetermined system  $\Rightarrow$  infinitely many solutions.  
 $\rightarrow$  exploit the sparsity assumption of  $x$ .

# Minimum $l_0$ -norm solution

$$\hat{x} = \arg \min_{z \in \mathbb{R}^n} ||z||_0 \text{ subject to } Mz = y$$

- Reminder: recover all  $s$ -sparse  $x$  **iff all sets of  $2s$  columns of  $M$  are linearly independent** (so-called EsR condition)

# Minimum $l_0$ -norm solution

$$\hat{x} = \arg \min_{z \in \mathbb{R}^n} ||z||_0 \text{ subject to } Mz = y$$

- Reminder: recover all  $s$ -sparse  $x$  **iff all sets of  $2s$  columns of  $M$  are linearly independent** (so-called EsR condition)
- **Complexity?**
  - ▶ Problem is non-convex
  - ▶ Problem is **NP-hard!**
    - for a given  $s$ , try all possible  $\binom{n}{s}$  supports, estimate the  $s$  nonzero values of  $x$ , check if constraint is satisfied
    - **infeasible** for practical problem sizes

# Practical philosophies

$$\hat{x} = \arg \min_{z \in \mathbb{R}^n} \|z\|_0 \text{ subject to } Mz = y$$

Greedy  
algorithms

Focus on  $\|x\|_0$

MP, OMP, OLS...

Thresholding  
algorithms

Focus on  $y \sim Mx$

IHT...

Convex relaxation  
algorithms

Solve a nicer problem

BP, LASSO...

see course C. Elvira

## **Recovery guarantees**

# Outline

Introduction

- What is Compressive Sensing?

- Notations (Reminder)

- Problem formulation

## Recovery guarantees

- NSP

- Dual certificate

- Coherence

- RIP

Recovering with random matrices?

Concentration inequalities and proving the RIP

## Beyond Sparsity

- Total Variation

- Structured sparsity

- Matrix completion and Low-rank regularization

Infinite-dimensional signals: superresolution and compressive learning

- Continuous sparsity: superresolution

- Generalized sparsity: sketching

## Reminder: $\ell_0$ and EsR

Decoder for sparse vectors:

$$\min \|z\|_0 \text{ s.t. } y = Mz \quad (P_0)$$

## Reminder: $\ell_0$ and EsR

Decoder for sparse vectors:

$$\min \|z\|_0 \text{ s.t. } y = Mz \quad (P_0)$$

$(P_0)$  achieves exact recovery of all  $s$ -sparse vectors iff

*All subsets of  $2s$  columns of  $M$  are linearly independent* (EsR)

## Reminder: $\ell_0$ and EsR

Decoder for sparse vectors:

$$\min \|z\|_0 \text{ s.t. } y = Mz \quad (P_0)$$

$(P_0)$  achieves exact recovery of all  $s$ -sparse vectors iff

*All subsets of  $2s$  columns of  $M$  are linearly independent* (EsR)

But  $(P_0)$  is NP-hard, so one solution is to solve instead

$$\min \|z\|_p^p \text{ s.t. } y = Mz \quad (P_p)$$

→ You have seen  $p = 1$  (aka Basis Pursuit) as a convex relaxation of  $\ell_0$ , but the general case is also interesting. Other possibilities: greedy approaches, etc.

## Reminder: $\ell_0$ and EsR

Decoder for sparse vectors:

$$\min \|z\|_0 \text{ s.t. } y = Mz \quad (P_0)$$

$(P_0)$  achieves exact recovery of all  $s$ -sparse vectors iff

*All subsets of  $2s$  columns of  $M$  are linearly independent* (EsR)

But  $(P_0)$  is NP-hard, so one solution is to solve instead

$$\min \|z\|_p^p \text{ s.t. } y = Mz \quad (P_p)$$

→ You have seen  $p = 1$  (aka Basis Pursuit) as a convex relaxation of  $\ell_0$ , but the general case is also interesting. Other possibilities: greedy approaches, etc.

**Recovery guarantees?** Many... In this section: Null Space Property

## Reminder: solution set

Recall: The set of potential solutions is

$$\{x' \mid y = Mx'\} = \{x + v \mid v \in \ker(M)\}$$

## Reminder: solution set

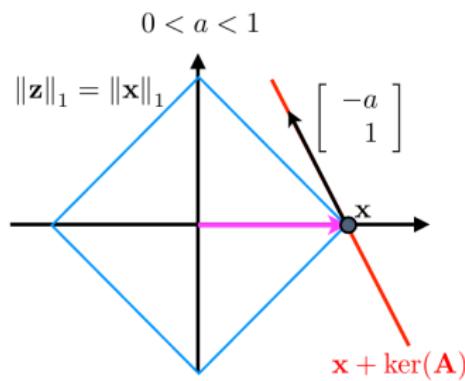
Recall: The set of potential solutions is

$$\{x' \mid y = Mx'\} = \{x + v \mid v \in \ker(M)\}$$

- The key lies in studying the *kernel*, or *null space* of  $M$
- Easy:  $x$  is the unique solution of  $(P_p)$  iff  $\|x\|_p^p < \|x + v\|_p^p \forall v \in \ker(M) \setminus \{0\}$

# Illustration

$p = 1$ : Success! and convex problem :)



$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathbb{R}^2$$

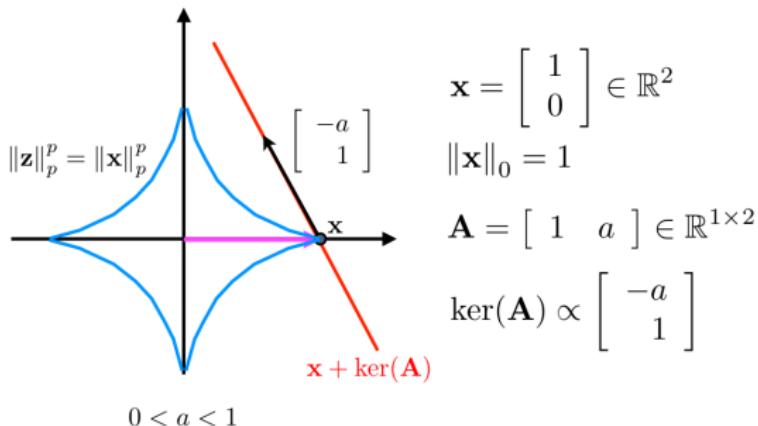
$$\|\mathbf{x}\|_0 = 1$$

$$\mathbf{A} = \begin{bmatrix} 1 & a \end{bmatrix} \in \mathbb{R}^{1 \times 2}$$

$$\text{ker}(\mathbf{A}) \propto \begin{bmatrix} -a \\ 1 \end{bmatrix}$$

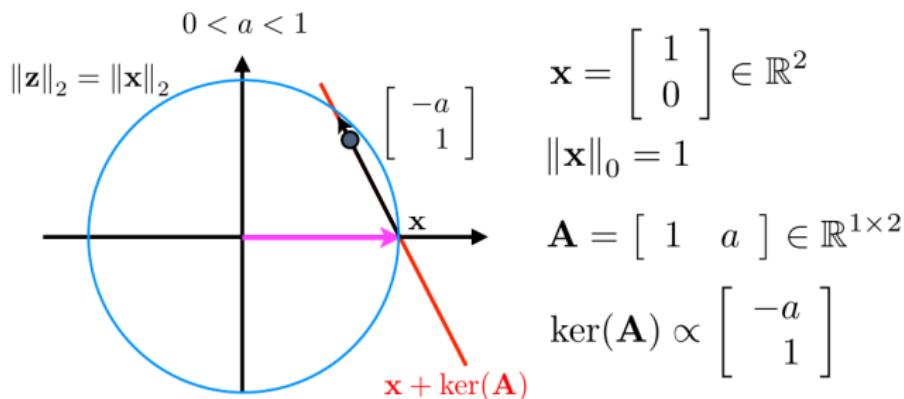
# Illustration

$0 < p < 1$ : Success! But non-convex...



# Illustration

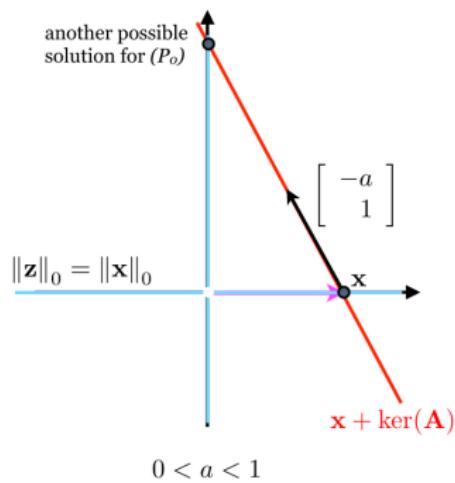
$p = 2$ : Fail... In fact,  $p > 1$  will “always” fail to recover sparse vectors



# Illustration

$p = 0$ : Fail!! ...for this example. “But I thought  $p = 0$  was the best?” →

**Uniform vs non-uniform recovery:** more to come.



# Null Space Property

## Null Space Property (F&R, Thm 4.4)

Take  $p \leq 1$ .

“Every  $s$ -sparse vectors  $x$  **supported on  $S$**  is the solution of  $(P_p)$  with  $y = Mx$ ”  
**iff**

$$\|v_S\|_p^p < \|v_{S^c}\|_p^p \quad \forall v \in \ker(M) \setminus \{0\} \quad (NSP_p(S))$$

# Null Space Property

## Null Space Property (F&R, Thm 4.4)

Take  $p \leq 1$ .

“Every  $s$ -sparse vectors  $x$  **supported on  $S$**  is the solution of  $(P_p)$  with  $y = Mx$ ”  
**iff**

$$\|v_S\|_p^p < \|v_{S^c}\|_p^p \quad \forall v \in \ker(M) \setminus \{0\} \quad (NSP_p(S))$$

Proof sketch: (*already seen*)

# Null Space Property

## Null Space Property (F&R, Thm 4.4)

Take  $p \leq 1$ .

“Every  $s$ -sparse vectors  $x$  supported on  $S$  is the solution of  $(P_p)$  with  $y = Mx$ ”  
**iff**

$$\|v_S\|_p^p < \|v_{S^c}\|_p^p \quad \forall v \in \ker(M) \setminus \{0\} \quad (\text{NSP}_p(S))$$

Proof sketch: (*already seen*)

⇐ Take  $x$  supported on  $S$  and  $z \neq x$  with  $Mz = Mx$ , and  $v = x - z \in \ker(M) \setminus \{0\}$ .  
Then, by the *quasi-triangle inequality*,

$$\begin{aligned} \|x\|_p^p &\leq \|x - z_S\|_p^p + \|z_S\|_p^p = \|v_S\|_p^p + \|z_S\|_p^p \\ &< \|v_{S^c}\|_p^p + \|z_S\|_p^p = \|z_{S^c}\|_p^p + \|z_S\|_p^p = \|z\|_p^p \end{aligned}$$

Hence  $\|x\|_p^p$  is minimal

# Null Space Property

## Null Space Property (F&R, Thm 4.4)

Take  $p \leq 1$ .

"Every  $s$ -sparse vectors  $x$  supported on  $S$  is the solution of  $(P_p)$  with  $y = Mx$ "  
**iff**

$$\|v_S\|_p^p < \|v_{S^c}\|_p^p \quad \forall v \in \ker(M) \setminus \{0\} \quad (\text{NSP}_p(S))$$

Proof sketch: (already seen)

$\Leftarrow$  Take  $x$  supported on  $S$  and  $z \neq x$  with  $Mz = Mx$ , and  $v = x - z \in \ker(M) \setminus \{0\}$ .  
Then, by the *quasi-triangle inequality*,

$$\begin{aligned} \|x\|_p^p &\leq \|x - z_S\|_p^p + \|z_S\|_p^p = \|v_S\|_p^p + \|z_S\|_p^p \\ &< \|v_{S^c}\|_p^p + \|z_S\|_p^p = \|z_{S^c}\|_p^p + \|z_S\|_p^p = \|z\|_p^p \end{aligned}$$

Hence  $\|x\|_p^p$  is minimal

$\Rightarrow$  Take  $v \in \ker(M) \setminus \{0\}$ . By hypothesis,  $x = v_S$  is the unique minimizer of  $(P_p)$  with  $y = Mv_S$ . Since we have also:  $v \neq 0 \Rightarrow v_S \neq -v_{S^c}$ , and  
 $Av = 0 \Rightarrow A(-v_{S^c}) = Av_S = y$ , then  $\|v_S\|_p^p < \|v_{S^c}\|_p^p$

# Null Space Property

Immediately, we have:

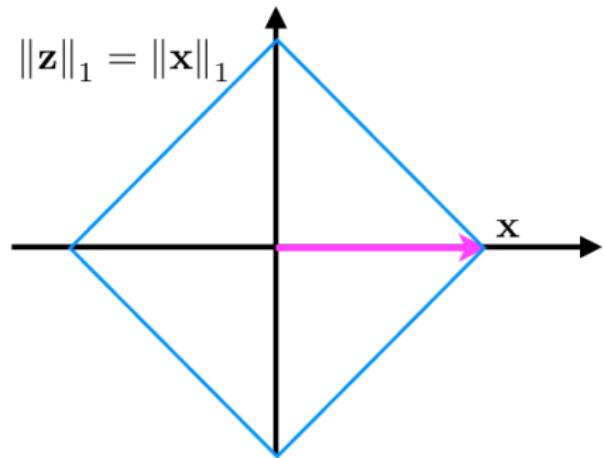
## Null Space Property (F&R, Thm 4.5)

Take  $p \leq 1$ . Every  $s$ -sparse vector  $x$  is the solution of  $(P_p)$  with  $y = Mx$  iff

$$\forall S \text{ with } \text{card}(S) = s, \text{NSP}_p(S) \text{ is verified} \quad (\text{NSP}_p(s))$$

- In particular  $\text{NSP}_0(s)$  is equivalent to EsR ! (Homework)

## NSP example



$$\|\mathbf{z}\|_1 = \|\mathbf{x}\|_1$$

$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathbb{R}^2$$

$$\|\mathbf{x}\|_0 = 1, S = \{1\}$$

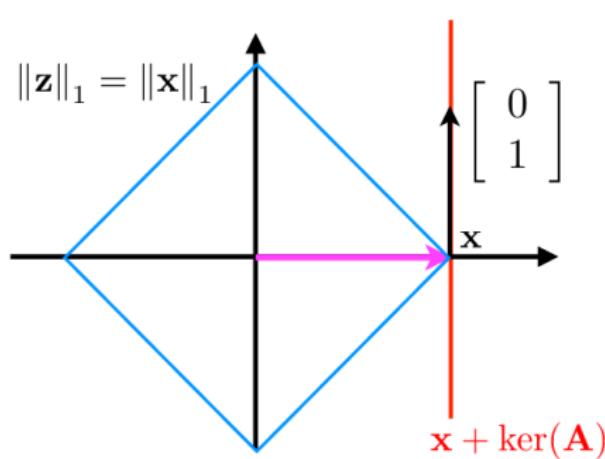
$$\mathbf{A} = \begin{bmatrix} 1 & a \end{bmatrix} \in \mathbb{R}^{1 \times 2}$$

$$\ker(\mathbf{A}) \propto \begin{bmatrix} -a \\ 1 \end{bmatrix}$$

$$(\text{NSP}_p(S)) : \|\mathbf{v}_S\|_1^1 < \|\mathbf{v}_{\bar{S}}\|_1^1 \equiv |a| < 1$$

## NSP example

$a = 0$  : NSP OK!



$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathbb{R}^2$$
$$\|\mathbf{x}\|_0 = 1, S = \{1\}$$

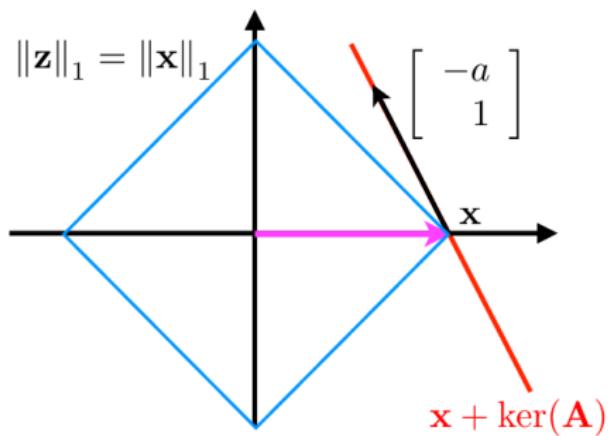
$$\mathbf{A} = \begin{bmatrix} 1 & a \end{bmatrix} \in \mathbb{R}^{1 \times 2}$$

$$\ker(\mathbf{A}) \propto \begin{bmatrix} -a \\ 1 \end{bmatrix}$$

$$(\text{NSP}_p(S)) : \|\mathbf{v}_S\|_1^1 < \|\mathbf{v}_{\bar{S}}\|_1^1 \equiv |a| < 1$$

## NSP example

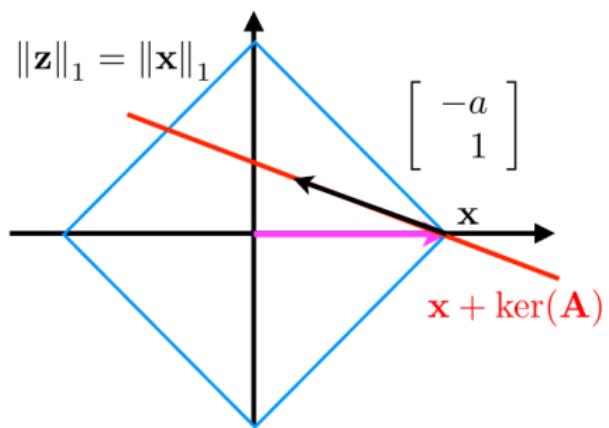
$0 < a < 1 : \text{NSP OK!}$



$$(\text{NSP}_p(S)) : \|\mathbf{v}_S\|_1^1 < \|\mathbf{v}_{\bar{S}}\|_1^1 \equiv |a| < 1$$

## NSP example

$a > 1$  : failure NSP!



$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathbb{R}^2$$
$$\|\mathbf{x}\|_0 = 1, S = \{1\}$$

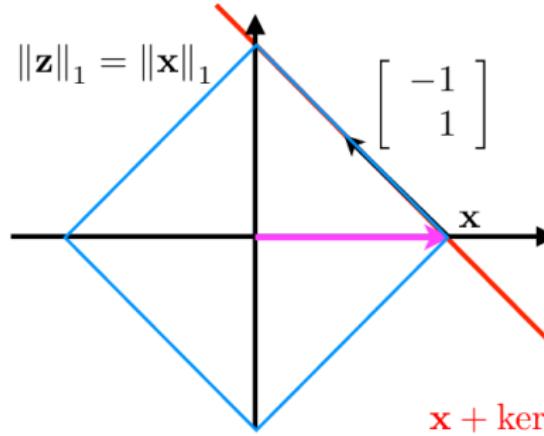
$$\mathbf{A} = \begin{bmatrix} 1 & a \end{bmatrix} \in \mathbb{R}^{1 \times 2}$$

$$\ker(\mathbf{A}) \propto \begin{bmatrix} -a \\ 1 \end{bmatrix}$$

$$(\text{NSP}_p(S)) : \|\mathbf{v}_S\|_1^1 < \|\mathbf{v}_{\bar{S}}\|_1^1 \equiv |a| < 1$$

# NSP example

$a = 1$  : failure NSP!



$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathbb{R}^2$$
$$\|\mathbf{x}\|_0 = 1, S = \{1\}$$

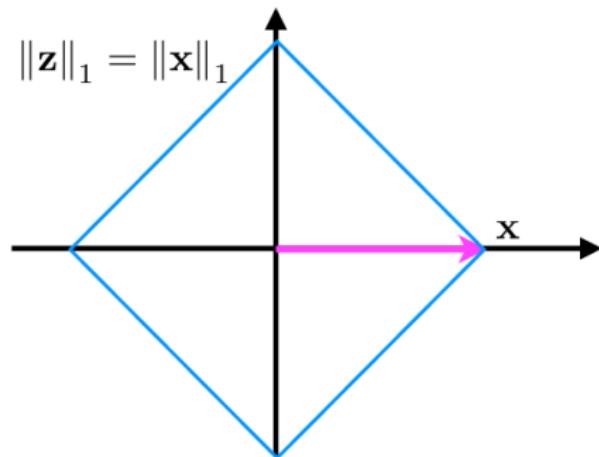
$$\mathbf{A} = \begin{bmatrix} 1 & a \end{bmatrix} \in \mathbb{R}^{1 \times 2}$$

$$\ker(\mathbf{A}) \propto \begin{bmatrix} -a \\ 1 \end{bmatrix}$$

$$(\text{NSP}_p(S)) : \|\mathbf{v}_S\|_1^1 < \|\mathbf{v}_{\bar{S}}\|_1^1 \equiv |a| < 1$$

Here  $x$  is a solution. Modified NSP:  $\|\mathbf{v}_S\|_p^p \leq \|\mathbf{v}_{S^c}\|_p^p$

## NSP example



$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathbb{R}^2$$

$$\|\mathbf{x}\|_0 = 1, S = \{1\}$$

$$\mathbf{A} = \begin{bmatrix} 1 & a \end{bmatrix} \in \mathbb{R}^{1 \times 2}$$

$$\ker(\mathbf{A}) \propto \begin{bmatrix} -a \\ 1 \end{bmatrix}$$

$$(\text{NSP}_p(S)) : \|\mathbf{v}_S\|_1^1 < \|\mathbf{v}_{\bar{S}}\|_1^1 \equiv |a| < 1$$

Can we have  $\text{NSP}_p(1)$ ?

# Instance vs Uniform recovery

$x$  is the unique solution... is equivalent to

...for a given  $x$   $\|x\|_p^p < \|x + v\|_p^p$  for all  $v \in \ker(M) \setminus \{0\}$

...for all  $x$  supported on  $S$   $NSP_p(S)$

...for all  $s$ -sparse  $x$   $NSP_p(x)$

- Rk: NSP is a “worst-case” condition. Failure means that *some*  $x$  cannot be recovered.

# Some properties of the NSP

Sparser is better!

- If  $S' \subset S$ , then  $NSP_p(S) \Rightarrow NSP_p(S')$
- If  $s' \leq s$ , then  $NSP_p(s) \Rightarrow NSP_p(s')$

Proof: easy.

# Some properties of the NSP

Sparser is better!

- If  $S' \subset S$ , then  $NSP_p(S) \Rightarrow NSP_p(S')$
- If  $s' \leq s$ , then  $NSP_p(s) \Rightarrow NSP_p(s')$

Proof: easy.

Preservation under shuffling (of  $y!$ ), rescaling, new observations

If  $NSP_p(S)$  (idem for  $NSP_p(s)$ ) is satisfied for  $M$ , it is also satisfied for

- $M' = GM$  with  $G \in \mathbb{R}^{m \times m}$  invertible (shuffling, rescaling)
  - ▶ Proof: the kernel does not change!
- $M' \in \mathbb{R}^{m' \times n}$  is obtained by adding rows to  $M$  (new observations)
  - ▶ Proof:  $\ker(M') \subset \ker(M)$

# Some properties of the NSP

Is  $\ell_0$  really the best?

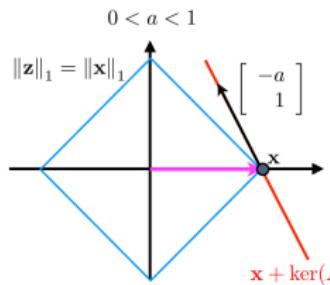
- Let  $q < p$ . Do we have  $NSP_p(S) \Rightarrow NSP_q(S)$ ?

# Some properties of the NSP

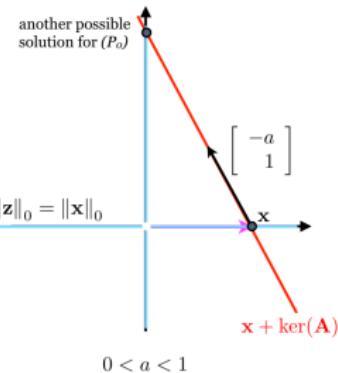
Is  $\ell_0$  really the best?

- Let  $q < p$ . Do we have  $NSP_p(S) \Rightarrow NSP_q(S)$ ?

► **NO!** Think about the previous example  $|a|^p < 1^p$ : we have  $|a| < 1$  but not  $|a|^0 < 1^0$



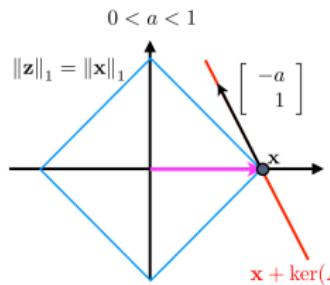
$$\begin{aligned}\mathbf{x} &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathbb{R}^2 \\ \|\mathbf{x}\|_0 &= 1 \\ \mathbf{A} &= \begin{bmatrix} 1 & a \end{bmatrix} \in \mathbb{R}^{1 \times 2} \\ \ker(\mathbf{A}) &\propto \begin{bmatrix} -a \\ 1 \end{bmatrix}\end{aligned}$$



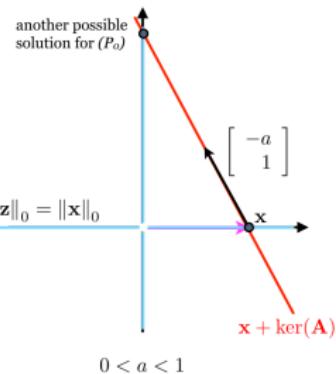
# Some properties of the NSP

Is  $\ell_0$  really the best?

- Let  $q < p$ . Do we have  $NSP_p(S) \Rightarrow NSP_q(S)$ ?
  - NO!** Think about the previous example  $|a|^p < 1^p$ : we have  $|a| < 1$  but not  $|a|^0 < 1^0$



$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathbb{R}^2$$
$$\|\mathbf{x}\|_0 = 1$$
$$\mathbf{A} = \begin{bmatrix} 1 & a \end{bmatrix} \in \mathbb{R}^{1 \times 2}$$
$$\ker(\mathbf{A}) \propto \begin{bmatrix} -a \\ 1 \end{bmatrix}$$

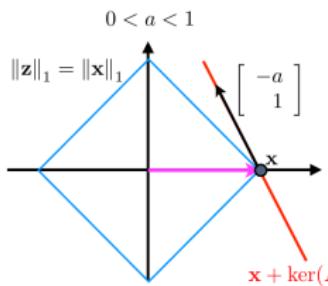


- Let  $q < p$ . Do we have  $NSP_p(s) \Rightarrow NSP_q(s)$ ?

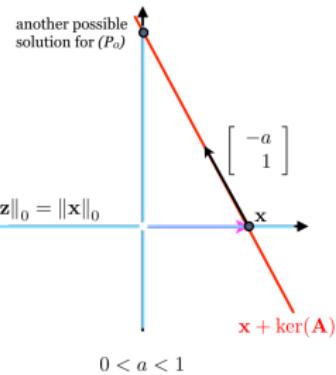
# Some properties of the NSP

Is  $\ell_0$  really the best?

- Let  $q < p$ . Do we have  $NSP_p(S) \Rightarrow NSP_q(S)$ ?
  - NO!** Think about the previous example  $|a|^p < 1^p$ : we have  $|a| < 1$  but not  $|a|^0 < 1^0$



$$\begin{aligned} \mathbf{x} &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathbb{R}^2 \\ \|\mathbf{x}\|_0 &= 1 \\ \mathbf{A} &= \begin{bmatrix} 1 & a \end{bmatrix} \in \mathbb{R}^{1 \times 2} \\ \ker(\mathbf{A}) &\propto \begin{bmatrix} -a \\ 1 \end{bmatrix} \end{aligned}$$



- Let  $q < p$ . Do we have  $NSP_p(s) \Rightarrow NSP_q(s)$ ?

**YES :)** [F&R Thm 4.10]

→  $p$  small is indeed better for **uniform** recovery, for instance recovery we cannot conclude.

# Stability

In practice, we have seen that  $x$  is not exactly sparse, but **compressible**: fast decaying  $\sigma_s(x)_1 = \min_{z \in \Sigma_s} \|x - z\|_1$ . Can we handle that? This is called **stability**.



# Stability (for $\ell_1$ )

## Stable NSP

$M$  satisfies the **stable NSP** with constant  $0 < \rho < 1$  relative to  $S$  if

$$\|v_S\|_1 \leq \rho \|v_{S^c}\|_1 \quad \forall v \in \ker(M) \quad (\rho\text{-NSP}_1(S))$$

- stronger than  $\text{NSP}_1(S)$  since  $\rho < 1$
- using a “ $\leq$ ” this time does not matter
- $\rho\text{-NSP}_1(s)$  when  $\rho\text{-NSP}_1(S)$  for all  $s$ -support  $S$
- NB: here we do  $\ell_1$ , there are extensions for  $\ell_p$

# Stability (for $\ell_1$ )

## Stable NSP

$M$  satisfies the **stable NSP** with constant  $0 < \rho < 1$  relative to  $S$  if

$$\|v_S\|_1 \leq \rho \|v_{S^c}\|_1 \quad \forall v \in \ker(M) \quad (\rho\text{-NSP}_1(S))$$

- stronger than  $\text{NSP}_1(S)$  since  $\rho < 1$
- using a “ $\leq$ ” this time does not matter
- $\rho\text{-NSP}_1(s)$  when  $\rho\text{-NSP}_1(S)$  for all  $s$ -support  $S$
- NB: here we do  $\ell_1$ , there are extensions for  $\ell_p$

## Thm [F&R 4.12]

Suppose that  $\rho\text{-NSP}_1(s)$  is satisfied. Then for **any**  $x$ , a solution  $z$  of  $(P_1)$  with  $y = Mx$  satisfies

$$\|x - z\|_1 \leq \frac{2(1+\rho)}{1-\rho} \sigma_s(x)_1$$

- $s$  appears in the bound and the NSP, not in any assumption on  $x$ !
- $(P_1)$  still uses the “exact” constraint  $y = Mz$

# Robustness

In practice, there is (always!) measurement **noise!** Bounded for some norm  $\|\cdot\|$ .

$$y = Mx + e \text{ with } \|e\| \leq \eta$$

Recovery guarantees in the presence of noise if called **robustness**.

# Robustness

In practice, there is (always!) measurement **noise!** Bounded for some norm  $\|\cdot\|$ .

$$y = Mx + e \text{ with } \|e\| \leq \eta$$

Recovery guarantees in the presence of noise if called **robustness**.

$(P_p)$  must be replaced. For instance, by **Constrained BP** (CBP)

$$\min_z \|z\|_p^p \text{ s.t. } \|Mz - y\| \leq \eta \quad (P_{p,\eta})$$

→ still convex for  $p = 1$

→ recall that this is *one* formulation, other (more or less “equivalent”) include BPD and Lasso. See course of T. Guyard

# Robustness (for $\ell_1$ )

## Robust NSP

A matrix  $M$  satisfies the **robust NSP** with constant  $0 < \rho < 1$  and  $\tau > 0$  relative to  $S$  and  $\|\cdot\|$  if

$$\|v_S\|_1 \leq \rho \|v_{S^c}\|_1 + \tau \|Mv\| \quad \forall v \in \mathbb{R}^n \quad ((\rho, \tau)\text{-NSP}_1(S))$$

- Now for all  $v$  and not only  $v$  in kernel!
- stronger than  $\rho\text{-NSP}_1(S)$  (take  $v$  in kernel)
- again,  $(\rho, \tau)\text{-NSP}_1(s)$  when  $(\rho, \tau)\text{-NSP}_1(S)$  for all  $s$ -support  $S$

# Robustness (for $\ell_1$ )

## Robust NSP

A matrix  $M$  satisfies the **robust NSP** with constant  $0 < \rho < 1$  and  $\tau > 0$  relative to  $S$  and  $\|\cdot\|$  if

$$\|v_S\|_1 \leq \rho \|v_{S^c}\|_1 + \tau \|Mv\| \quad \forall v \in \mathbb{R}^n \quad ((\rho, \tau)\text{-NSP}_1(S))$$

- Now for all  $v$  and not only  $v$  in kernel!
- stronger than  $\rho\text{-NSP}_1(S)$  (take  $v$  in kernel)
- again,  $(\rho, \tau)\text{-NSP}_1(s)$  when  $(\rho, \tau)\text{-NSP}_1(S)$  for all  $s$ -support  $S$

## Thm [F&R 4.19]

Suppose that  $(\rho, \tau)\text{-NSP}_1(s)$  is satisfied. Then for **any**  $x$ , a solution  $z$  of  $(P_{1,\eta})$  with  $y = Mx + e$  where  $\|e\| \leq \eta$  satisfies

$$\|x - z\|_1 \leq \frac{2(1 + \rho)}{1 - \rho} \sigma_s(x)_1 + \frac{4\tau}{1 - \rho} \eta$$

# Outline

Introduction

- What is Compressive Sensing?

- Notations (Reminder)

- Problem formulation

## Recovery guarantees

- NSP

- Dual certificate

- Coherence

- RIP

Recovering with random matrices?

Concentration inequalities and proving the RIP

## Beyond Sparsity

- Total Variation

- Structured sparsity

- Matrix completion and Low-rank regularization

Infinite-dimensional signals: superresolution and compressive learning

- Continuous sparsity: superresolution

- Generalized sparsity: sketching

# Dual certificate

**Duality** is at the core of the proofs in:

- the [original paper](#) “Decoding by linear programming” by Candès & Tao
- [modern approaches](#) (see “a RIPless theory...” by Plan & Candès), or even generalized sparsity (see later)

# Dual certificate

**Duality** is at the core of the proofs in:

- the [original paper](#) “Decoding by linear programming” by Candès & Tao
- [modern approaches](#) (see “a RIPless theory...” by Plan & Candès), or even generalized sparsity (see later)

**Dual certificates** are often a [useful proof intermediates](#). We'll just give a glance at them here.

## Dual certificate: intuition

Take BP denoising:

$$\min_z \frac{1}{2} \|Mz - y\|_2^2 + \lambda \|z\|_1$$

## Dual certificate: intuition

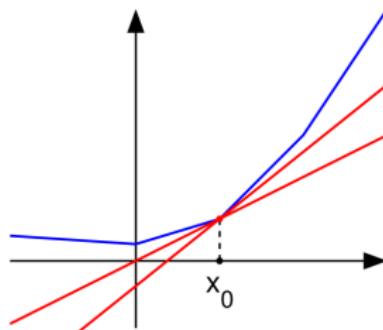
Take BP denoising:

$$\min_z \frac{1}{2} \|Mz - y\|_2^2 + \lambda \|z\|_1$$

By convexity and first-order conditions, every solution  $x$  satisfies

$$-\frac{1}{\lambda} M^\top (Mx - y) \in \partial \|x\|_1$$

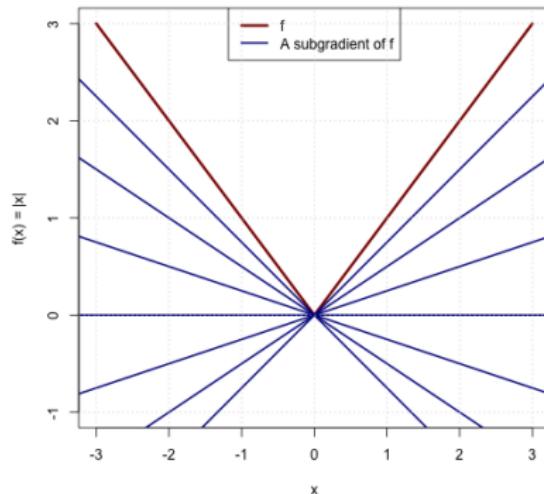
where  $\partial f(x)$  is the **subgradient** of the convex function  $f$  in  $x$  (ie such that  $\frac{f(z) - f(x)}{z - x} \geq \partial f(x)$  for all  $z$ ).



# Dual certificate: intuition

What is the subgradient of  $\ell_1$  ?

$$c \in \partial \|x\|_1 \Leftrightarrow \begin{cases} c_i = \text{sign}(x_i) & \text{if } x_i \neq 0 \\ |c_i| \leq 1 & \text{otherwise} \end{cases}$$



## Dual certificate and exact recovery

We have at least

$$\begin{aligned} -\frac{1}{\lambda}M^\top(Mx - y) &\in \partial\|x\|_1 \Rightarrow \text{Im}(M^\top) \cap \partial\|x\|_1 \neq \emptyset \\ \Rightarrow \exists c = M^\top u \text{ s.t. } &\begin{cases} c_i = \text{sign}(x_i) & \text{on } S \\ |c_i| \leq 1 & \text{on } S^c \end{cases} \end{aligned}$$

# Dual certificate and exact recovery

We have at least

$$\begin{aligned} -\frac{1}{\lambda} M^\top (Mx - y) &\in \partial \|x\|_1 \Rightarrow \text{Im}(M^\top) \cap \partial \|x\|_1 \neq \emptyset \\ &\Rightarrow \exists c = M^\top u \text{ s.t. } \begin{cases} c_i = \text{sign}(x_i) & \text{on } S \\ |c_i| \leq 1 & \text{on } S^c \end{cases} \end{aligned}$$

And in fact, using some duality machinery,

## Theorem (Candès & Tao 2006, Thm 1.4)

Take a support  $S$ , and  $x$  supported on  $S$ . If  $M_S$  has full column rank, and there exists  $c = M^\top u$  such that

$$\begin{cases} c_i = \text{sign}(x_i) & \text{on } S \\ |c_i| < 1 & \text{on } S^c \end{cases}$$

Then  $x$  is the unique solution of BP.

# Dual certificates

## Theorem (Candès & Tao 2006, Thm 1.4)

Take a support  $S$ , and  $x$  supported on  $S$ . If  $M_S$  has full column rank, and there exists  $c = M^\top u$  such that

$$\begin{cases} c_i = \text{sign}(x_i) & \text{on } S \\ |c_i| < 1 & \text{on } S^c \end{cases}$$

Then  $x$  is the unique solution of BP.

- such a  $c$  is called a **dual certificate**
- This is a *non-uniform* recovery condition: depends on the **sign vector** of  $x$  (ie, the existence of  $c$  guarantees recovery of all  $x$  with the same sign vector)
- The existence/construction of  $c$  is then proved by other means (RIP...)

# Stability Robustness

## Theorem (F& R Exo 4.17)

Take a support  $S$ , and  $x$  supported on  $S$ . Take  $M$  with normalized columns. If  $\|M_S^\top M_S - Id\|_2 \leq \alpha$ , and there exists  $c = M^\top u$  such that  $\|u\| \leq \gamma\sqrt{s}$  and

$$\begin{cases} c_i = \text{sign}(x_i) & \text{on } S \\ |c_i| \leq \beta < 1 & \text{on } S^c \end{cases}$$

Then, the solution of constrained BP with  $\|e\| \leq \eta$  satisfies

$$\|x - \hat{x}\|_2 \leq C\sigma_s(x)_1 + D\sqrt{s}\eta$$

# Outline

Introduction

- What is Compressive Sensing?

- Notations (Reminder)

- Problem formulation

## Recovery guarantees

- NSP

- Dual certificate

- Coherence

- RIP

Recovering with random matrices?

Concentration inequalities and proving the RIP

Beyond Sparsity

- Total Variation

- Structured sparsity

- Matrix completion and Low-rank regularization

Infinite-dimensional signals: superresolution and compressive learning

- Continuous sparsity: superresolution

- Generalized sparsity: sketching

# Why coherence?

- $\text{NSP}_p(s)$  is true if  $\text{NSP}_p(S)$  is true **for all**  $S$   
→ Checking them is combinatorial!
- Is there a more convenient criterion?  
→ Yes, coherence (which you have already seen in C. Elvira's course).

# Why coherence?

- $\text{NSP}_p(s)$  is true if  $\text{NSP}_p(S)$  is true **for all**  $S$   
→ Checking them is combinatorial!
- Is there a more convenient criterion?  
→ Yes, coherence (which you have already seen in C. Elvira's course).

## Coherence

Take  $M$  **with  $\ell_2$  normalized columns**. The coherence of  $M$  is

$$\mu = \max_{i \neq j} |m_i^\top m_j| \tag{1}$$

- Intuitively, although  $M$  is “fat”, we want its columns to be *as less correlated as possible*, to be “distinguishable” within the mix  $y = \sum_{i \in S} m_i x_i$
- $\mu \leq 1$  by Cauchy-Schwartz

# Babel function

The coherence itself does not take into account sparsity size. We can look at the more precise notion:

## Babel function

$$\mu_1(s) = \max_i \max_{S \text{ of size } s, i \notin S} \left\{ \sum_{j \in S} |m_i^\top m_j| \right\} \quad (2)$$

- $\mu_1(1) = \mu$
- $\mu \leq \mu_1(s) \leq s\mu$

# Some properties of the coherence

(recall we take  $M$  with  $\ell_2$ -normalized columns)

- (F&R, Thm 5.7):

$$\mu \geq \sqrt{\frac{n-m}{m(n-1)}} \sim \frac{1}{\sqrt{m}}$$

when  $m \ll n$ . Equality holds iff:

- ▶  $M$  is **equiangular**:  $|m_i^\top m_j| = c$  constant  $\forall i \neq j$
- ▶  $M$  is a **tight frame**:  $MM^\top \propto Id_m$

# Some properties of the coherence

(recall we take  $M$  with  $\ell_2$ -normalized columns)

- (F&R, Thm 5.7):

$$\mu \geq \sqrt{\frac{n-m}{m(n-1)}} \sim \frac{1}{\sqrt{m}}$$

when  $m \ll n$ . Equality holds iff:

- ▶  $M$  is **equiangular**:  $|m_i^\top m_j| = c$  constant  $\forall i \neq j$
- ▶  $M$  is a **tight frame**:  $MM^\top \propto Id_m$

- (F&R, Thm 5.8):

$$\mu_1(s) \geq s \sqrt{\frac{n-m}{m(n-1)}}$$

whenever  $s \leq \sqrt{n-1}$ . Equality holds in the same conditions.

# Some properties of the coherence

(recall we take  $M$  with  $\ell_2$ -normalized columns)

- (F&R, Thm 5.7):

$$\mu \geq \sqrt{\frac{n-m}{m(n-1)}} \sim \frac{1}{\sqrt{m}}$$

when  $m \ll n$ . Equality holds iff:

- ▶  $M$  is **equiangular**:  $|m_i^\top m_j| = c$  constant  $\forall i \neq j$
- ▶  $M$  is a **tight frame**:  $MM^\top \propto Id_m$

- (F&R, Thm 5.8):

$$\mu_1(s) \geq s \sqrt{\frac{n-m}{m(n-1)}}$$

whenever  $s \leq \sqrt{n-1}$ . Equality holds in the same conditions.

- (F&R Prop 5.13): For each prime number  $m \geq 5$ , there is an explicit **complex** matrix with coherence  $\mu = \frac{1}{\sqrt{m}}$ .

# Guarantees for OMP and BP

F&R, Thm 5.14, 5.15

Let  $M$  with normalized columns. If

$$\mu_1(s) + \mu_1(s - 1) < 1$$

then every  $s$ -sparse vector  $x$  is recovered from  $y = Mx$

- after at most  $s$  iteration of OMP
- via Basis Pursuit ( $\ell_1$  minimization)

# Guarantees for OMP and BP

F&R, Thm 5.14, 5.15

Let  $M$  with normalized columns. If

$$\mu_1(s) + \mu_1(s - 1) < 1$$

then every  $s$ -sparse vector  $x$  is recovered from  $y = Mx$

- after at most  $s$  iteration of OMP
  - via Basis Pursuit ( $\ell_1$  minimization)
- 
- Since  $\mu_1(s) \leq s\mu$ , this condition is implied by the stronger but more classical condition

$$\mu \leq \frac{1}{2s - 1}$$

(which you have seen before)

- The proof uses NSP<sub>1</sub>( $s$ )...

# Guarantees for IHT

F&R, Thm 5.17

Let  $M$  with normalized columns. If

$$2\mu_1(s) + \mu_1(s - 1) < 1$$

then every  $s$ -sparse vector  $x$  is recovered from  $y = Mx$  after at most  $s$  iteration of IHT.

# Guarantees for IHT

F&R, Thm 5.17

Let  $M$  with normalized columns. If

$$2\mu_1(s) + \mu_1(s - 1) < 1$$

then every  $s$ -sparse vector  $x$  is recovered from  $y = Mx$  after at most  $s$  iteration of IHT.

- This condition is implied by

$$\mu \leq \frac{1}{3s - 1}$$

- stronger assumptions than OMP or BP !

# Quadratic bottleneck

Recall that we have seen that

$$\mu \geq \sqrt{\frac{n-m}{m(n-1)}} \sim \frac{1}{\sqrt{m}}$$

and that the recovery condition reads

$$\mu < \frac{1}{2s-1}$$

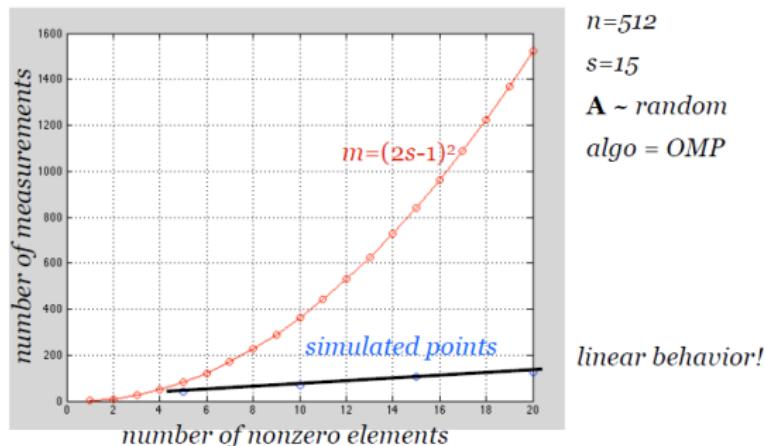
# Quadratic Bottleneck

Hence, the condition yields

$$m \gtrsim O(s^2)$$

i.e. a **quadratic rate**. This is fine, but we can do **much better**:  $m$  **linear** in  $s$  up to log factors.

→ For coherence-based methods, this condition is *tight*. This is the **quadratic bottleneck**. We need new tools!



# Outline

Introduction

- What is Compressive Sensing?

- Notations (Reminder)

- Problem formulation

## Recovery guarantees

- NSP

- Dual certificate

- Coherence

- RIP

Recovering with random matrices?

Concentration inequalities and proving the RIP

Beyond Sparsity

- Total Variation

- Structured sparsity

- Matrix completion and Low-rank regularization

Infinite-dimensional signals: superresolution and compressive learning

- Continuous sparsity: superresolution

- Generalized sparsity: sketching

# The restricted isometry property (RIP) : definition

## Definition (RIP)

Let  $\epsilon > 0$ ,  $s \in \mathbb{N}$ . A matrix  $M \in \mathbb{R}^{m,n}$  with  $m \leq n$  is  $(\epsilon, s)$ -RIP if

$$\forall x \in \Sigma_s, (1 - \epsilon) \|x\|_2^2 \leq \|Mx\|_2^2 \leq (1 + \epsilon) \|x\|_2^2 \quad (3)$$

# The restricted isometry property (RIP) : definition

## Definition (RIP)

Let  $\epsilon > 0$ ,  $s \in \mathbb{N}$ . A matrix  $M \in \mathbb{R}^{m,n}$  with  $m \leq n$  is  $(\epsilon, s)$ -RIP if

$$\forall x \in \Sigma_s, (1 - \epsilon) \|x\|_2^2 \leq \|Mx\|_2^2 \leq (1 + \epsilon) \|x\|_2^2 \quad (3)$$

- $M$  “preserves” the Euclidean norm **of  $s$ -sparse vectors**  $\rightarrow$  it is “almost” an **isometry** (hence the name)
- $M \in \mathbb{R}^{m,n}$  with  $m \ll n$  is  $(\epsilon, s)$ -RIP if

$$\forall x \in \Sigma_s \setminus \{0\}, \left| \frac{\|Mx\|_2^2 - \|x\|_2^2}{\|x\|_2^2} \right| \leq \epsilon \quad (4)$$

# RIP constant

- the “RIP constant”  $\epsilon_s$  of order  $s$  is the smallest constant  $\epsilon$  such that  $(\epsilon, s)$ -RIP holds:  $\dots \epsilon_s \leq \epsilon_{s+1} \dots$

# RIP constant

- the “RIP constant”  $\epsilon_s$  of order  $s$  is the smallest constant  $\epsilon$  such that  $(\epsilon, s)$ -RIP holds:  $\dots \epsilon_s \leq \epsilon_{s+1} \dots$

## F&R, thm 6.8

One has

$$m \geq c \frac{s}{\epsilon_s^2}$$

provided  $n \geq Cm$  and  $\epsilon_s \leq \epsilon^*$ , where  $c, C, \epsilon^*$  are constant that only depend on each other.

→ the RIP constant, sparsity, and  $m$  are linked.

## Recovery with the RIP: $\ell_0$ and BP

Note that, like EsR, we do not act on  $s$  columns of  $M$ , but at least  $2s$ , since we are looking at *differences* of  $s$ -sparse vectors.

- Exo: which  $(\cdot, \cdot)$ -RIP implies EsR, aka  $\ell_0$  recovery?

# Recovery with the RIP: $\ell_0$ and BP

Note that, like EsR, we do not act on  $s$  columns of  $M$ , but at least  $2s$ , since we are looking at *differences* of  $s$ -sparse vectors.

- Exo: which  $(\cdot, \cdot)$ -RIP implies EsR, aka  $\ell_0$  recovery?

## Robustness and stability, F&R thm 6.13

If

$$\epsilon_{2s} < \frac{4}{\sqrt{41}} \approx 0.62$$

then  $M$  satisfies the  $\ell_2$ -robust NSP<sub>1</sub>( $s$ ) with constant  $\rho, \tau$  that only depends on  $\epsilon_{2s}$ .

As a consequence, we have all the NSP results of the previous section.

# Recovery with the RIP: IHT

## Theorem (Optimality of IHT for RIP matrices )

If  $M$  is  $(\epsilon, 3s)$ -RIP, then

$$\|x^{l+1} - x\| \leq 2\epsilon \|x^l - x\| \leq \dots \leq (2\epsilon)^{l+1} \|x^0 - x\| \quad (5)$$

where  $x^l$  are the iterates of IHT.

In particular, if  $\epsilon_{3s} < \frac{1}{2}$ , the iterates  $x^l$  converge to  $x$

- Here it's not "in  $s$  iterations"!
- Stability and Robustness: see F&R Thm 6.21. Condition:  $\epsilon_{6s} < 1/\sqrt{3}$ .

# Recovery with the RIP: OMP

Here it becomes complicated...

## F&R thm 6.25

If

$$\epsilon_{13s} < 1/6$$

then the sequence  $x^l$  produced by OMP from  $y = Ax + e$  satisfies, for any  $S$ ,

$$\|y - Ax^{12s}\|_2 \lesssim \|Ax_{S^c} + e\|_2$$

Not even recovery of  $x$ !

But if

$$\epsilon_{26s} < 1/6$$

then

$$\|x - x^{24s}\|_1 \lesssim \sigma_s(x)_1 + \sqrt{s}\|e\|_2$$

In practice, generally much better than that! No need to do  $24s$  iterations...

**Recovering with random matrices?**

# Intuition

What is a good sensing matrix? How much measurements to satisfy the RIP?

Recall **coherence** (for  $\ell_2$ -normalized matrices), which must be as small as possible

$$\mu = \max_{i \neq j} |m_i^\top m_j|$$

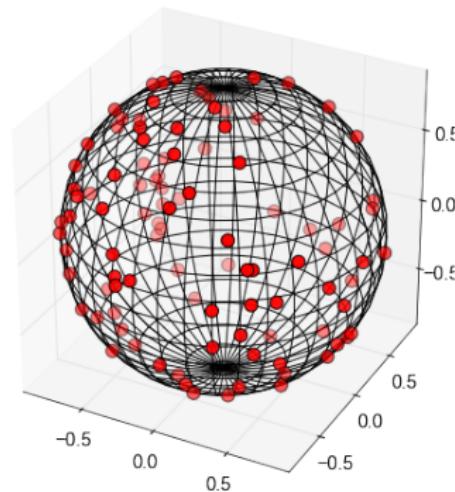
# Intuition

What is a good sensing matrix? How much measurements to satisfy the RIP?

Recall **coherence** (for  $\ell_2$ -normalized matrices), which must be as small as possible

$$\mu = \max_{i \neq j} |m_i^\top m_j|$$

- we need  $n$  vectors that are “well-spread” on the  $m$ -dimensional sphere.
- Deterministic constructions are possible, but complicated... We will rely on **random matrices**, and show the desired property **with high probability**



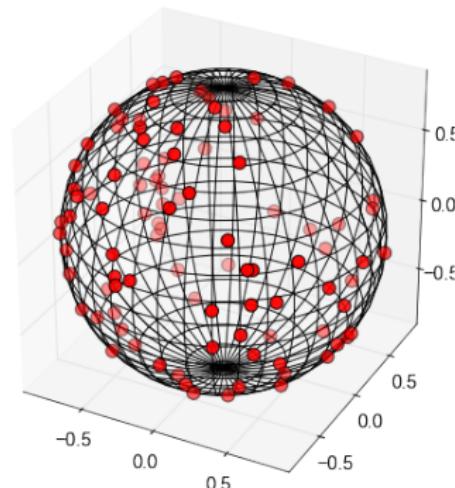
# Intuition

What is a good sensing matrix? How much measurements to satisfy the RIP?

Recall **coherence** (for  $\ell_2$ -normalized matrices), which must be as small as possible

$$\mu = \max_{i \neq j} |m_i^\top m_j|$$

- we need  $n$  vectors that are “well-spread” on the  $m$ -dimensional sphere.
- Deterministic constructions are possible, but complicated... We will rely on **random matrices**, and show the desired property **with high probability**



- recall: coherence = quadratic bottleneck. We will study RIP instead.
- you have seen polynomial/exponential systems, that verify EsR. But not enough for RIP!

# Gaussian matrices

One of the most used/convenient construction is **Gaussian matrices**. All entries are independent:

$$m_{ij} \sim \mathcal{N}(0, 1/m)$$

# Gaussian matrices

One of the most used/convenient construction is **Gaussian matrices**. All entries are independent:

$$m_{ij} \sim \mathcal{N}(0, 1/m)$$

Variance is  $1/m$  to have proper scaling

$$\mathbb{E}\|Mx\|_2^2 = \|x\|_2^2$$

Basis for proving the RIP: “show that  $\|Mx\|_2^2$  is close to its expectation with high probability (for all sparse vectors)” see after

# RIP for Gaussian matrices

## F&R Thm 9.27

Let  $M$  be a Gaussian matrix (with variance  $1/m$ ). For  $\epsilon, \delta > 0$ , assume

$$m \geq 2\epsilon^{-2}(s \log(en/s) + \log(2/\delta))$$

Then, with probability at least  $1 - \delta$ ,

$$\delta_s \leq 2 \left( 1 + \frac{1}{\sqrt{2 \log(en/s)}} \right) \epsilon + \left( 1 + \frac{1}{\sqrt{2 \log(en/s)}} \right)^2 \epsilon^2$$

# RIP for Gaussian matrices

## F&R Thm 9.27

Let  $M$  be a Gaussian matrix (with variance  $1/m$ ). For  $\epsilon, \delta > 0$ , assume

$$m \geq 2\epsilon^{-2}(s \log(en/s) + \log(2/\delta))$$

Then, with probability at least  $1 - \delta$ ,

$$\delta_s \leq 2 \left( 1 + \frac{1}{\sqrt{2 \log(en/s)}} \right) \epsilon + \left( 1 + \frac{1}{\sqrt{2 \log(en/s)}} \right)^2 \epsilon^2$$

- Examine all the rates in  $m$ : which are fast, which are slow?

# RIP for Gaussian matrices

## F&R Thm 9.27

Let  $M$  be a Gaussian matrix (with variance  $1/m$ ). For  $\epsilon, \delta > 0$ , assume

$$m \geq 2\epsilon^{-2}(s \log(en/s) + \log(2/\delta))$$

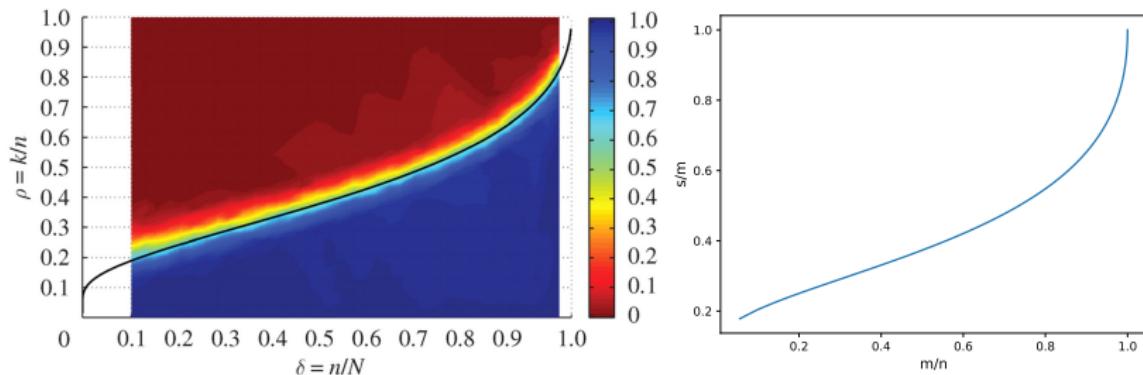
Then, with probability at least  $1 - \delta$ ,

$$\delta_s \leq 2 \left( 1 + \frac{1}{\sqrt{2 \log(en/s)}} \right) \epsilon + \left( 1 + \frac{1}{\sqrt{2 \log(en/s)}} \right)^2 \epsilon^2$$

- Examine all the rates in  $m$ : which are fast, which are slow?
- With more or less the same rate, one can also prove the robust NSP (Thm 9.29)

# Phase transition

Almost linear rate  $m \sim s \log(en/s)$ . **This is the well-known rate of Compressed Sensing.**



Exo: produce the theoretical curve above (right)

# Rademacher and subgaussian matrices

Gaussian rv may not be easily found in physical measurement systems!  
More “natural”: Rademacher variables

$$m_{ij} \sim \text{Unif}(\{1/\sqrt{m}, -1/\sqrt{m}\})$$

→ much, **much** more convenient on a computer! (just addition subtraction)

# Rademacher and subgaussian matrices

Gaussian rv may not be easily found in physical measurement systems!  
More “natural”: Rademacher variables

$$m_{ij} \sim \text{Unif}(\{-1/\sqrt{m}, 1/\sqrt{m}\})$$

→ much, **much** more convenient on a computer! (just addition subtraction)  
More generally, *centered* “subgaussian” random variables w/ parameters  
 $\kappa, \beta$  (includes Gaussian and bounded rv like rademacher)

$$\mathbb{P}(|m_{ij}| \geq t) \leq \beta e^{-\kappa t^2}$$

# RIP for subgaussian matrices

## F&R Thm 9.2

Let  $M$  be a subgaussian matrix of centered rv with variance  $1/m$ . We have  $\epsilon_s \leq \epsilon$  with probability at least  $1 - \delta$ , provided

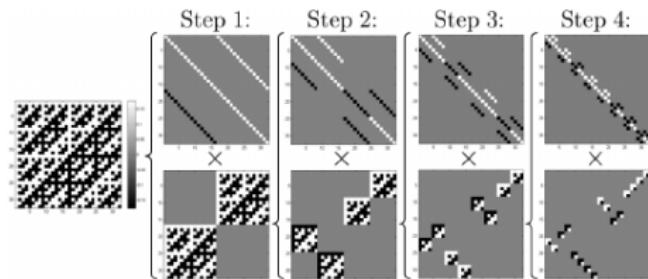
$$m \gtrsim \epsilon^{-2} (s \log(en/s) + \log(2/\delta))$$

- Basically the same result than Gaussian random matrices

# Beyond iid Matrices

Random matrices with iid entries are convenient theoretically, but **not really in practice**. **Active research field!**

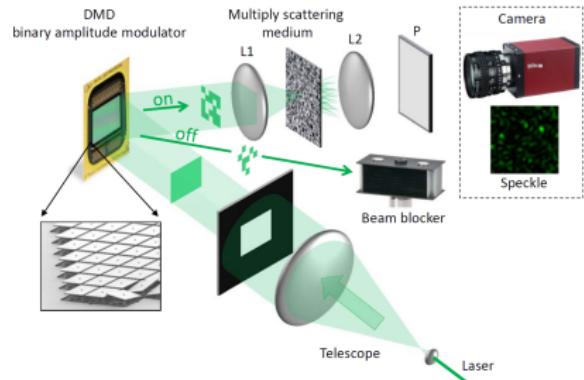
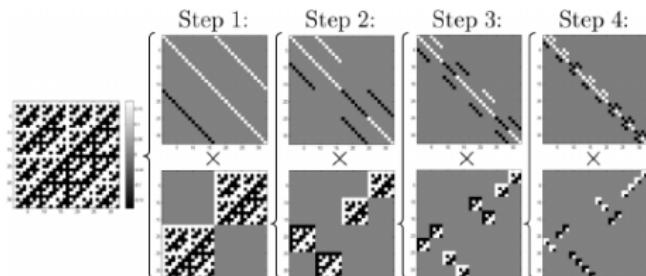
- Dense, unstructured matrices (costly!). Could we have “FFT”-like instead? Maybe, but **structure is bad for the RIP! (trade-off)**



# Beyond iid Matrices

Random matrices with iid entries are convenient theoretically, but **not really in practice**. **Active research field!**

- Dense, unstructured matrices (costly!). Could we have “FFT”-like instead? Maybe, but **structure is bad for the RIP! (trade-off)**
- Multiplication by large matrices is costly . Can we have “physical” random transforms? Maybe optics...



## Concentration inequalities and proving the RIP

# Concentration inequalities

We have seen that, with iid centered rv with variance  $1/m$ , we have  
 $\mathbb{E}\|Mx\|_2^2 = \|x\|_2^2$ .

# Concentration inequalities

We have seen that, with iid centered rv with variance  $1/m$ , we have  
 $\mathbb{E}\|Mx\|_2^2 = \|x\|_2^2$ .

$$\begin{aligned}\mathbb{E}\|Mx\|_2^2 &= \mathbb{E} \sum_{i=1}^m \left( \sum_{j=1}^n m_{ij} x_j \right)^2 = \mathbb{E} \sum_{i=1}^m \sum_{j,k=1}^n m_{ij} m_{ik} x_j x_k \\ &= \sum_{i=1}^m \sum_{j,k=1}^n \mathbb{E}(m_{ij} m_{ik}) x_j x_k = \sum_{i=1}^m \sum_{j=1}^n \mathbb{E}(m_{ij}^2) x_j^2 \\ &= \sum_{j=1}^n \sum_{i=1}^m \frac{1}{m} x_j^2 = \sum_j x_j^2 = \|x\|_2^2\end{aligned}$$

# Concentration inequalities

We have seen that, with iid centered rv with variance  $1/m$ , we have  
 $\mathbb{E}\|Mx\|_2^2 = \|x\|_2^2$ .

$$\begin{aligned}\mathbb{E}\|Mx\|_2^2 &= \mathbb{E} \sum_{i=1}^m \left( \sum_{j=1}^n m_{ij} x_j \right)^2 = \mathbb{E} \sum_{i=1}^m \sum_{j,k=1}^n m_{ij} m_{ik} x_j x_k \\ &= \sum_{i=1}^m \sum_{j,k=1}^n \mathbb{E}(m_{ij} m_{ik}) x_j x_k = \sum_{i=1}^m \sum_{j=1}^n \mathbb{E}(m_{ij}^2) x_j^2 \\ &= \sum_{j=1}^n \sum_{i=1}^m \frac{1}{m} x_j^2 = \sum_j x_j^2 = \|x\|_2^2\end{aligned}$$

The main proof technique for the RIP is to **bound the deviation of  $\|Mx\|_2^2$  from its expectation** with high probability. This is done with **concentration inequalities**.

# Basic concentration inequalities

**Markov's inequality:** Given a non-negative random variable  $X$  with finite mean

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}, \quad \forall t > 0. \quad \text{Decay in } \mathcal{O}\left(\frac{1}{t}\right)$$

# Basic concentration inequalities

**Markov's inequality:** Given a non-negative random variable  $X$  with finite mean

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}, \quad \forall t > 0. \quad \text{Decay in } \mathcal{O}\left(\frac{1}{t}\right)$$

**Chebyshev's inequality:** Given a random variable  $X$  with finite variance

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\text{var}(X)}{t^2}, \quad \forall t > 0. \quad \text{Decay in } \mathcal{O}\left(\frac{1}{t^2}\right)$$

# Basic concentration inequalities

**Markov's inequality:** Given a non-negative random variable  $X$  with finite mean

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}, \quad \forall t > 0. \quad \text{Decay in } \mathcal{O}\left(\frac{1}{t}\right)$$

**Chebyshev's inequality:** Given a random variable  $X$  with finite variance

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\text{var}(X)}{t^2}, \quad \forall t > 0. \quad \text{Decay in } \mathcal{O}\left(\frac{1}{t^2}\right)$$

**Chernoff bound:** Given a random variable  $X$  with mean  $\mu$

$$\mathbb{P}(X - \mu \geq t) \leq \frac{\mathbb{E}[e^{\lambda|X-\mu|}]}{e^{\lambda t}}, \quad \forall t, \lambda > 0. \quad \text{Decay in } \mathcal{O}(e^{-\lambda t})$$

as soon as  $\mathbb{E}[e^{\lambda|X-\mu|}]$  is finite.

# Basic concentration inequalities

**Markov's inequality:** Given a non-negative random variable  $X$  with finite mean

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}, \quad \forall t > 0. \quad \text{Decay in } \mathcal{O}\left(\frac{1}{t}\right)$$

**Chebyshev's inequality:** Given a random variable  $X$  with finite variance

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\text{var}(X)}{t^2}, \quad \forall t > 0. \quad \text{Decay in } \mathcal{O}\left(\frac{1}{t^2}\right)$$

**Chernoff bound:** Given a random variable  $X$  with mean  $\mu$

$$\mathbb{P}(X - \mu \geq t) \leq \frac{\mathbb{E}[e^{\lambda|X-\mu|}]}{e^{\lambda t}}, \quad \forall t, \lambda > 0. \quad \text{Decay in } \mathcal{O}(e^{-\lambda t})$$

as soon as  $\mathbb{E}[e^{\lambda|X-\mu|}]$  is finite.

Also see Hoeffding's inequality, Bernstein inequality, McDiarmid inequality... [book Boucheron Lugosi Massart]

# Concentration inequality

## Theorem (Concentration of Gaussian Matrices)

Let  $x \in \mathbb{R}^n$ . For a Gaussian matrix  $M$ ,

$$\forall 0 \leq \epsilon \leq 3, \quad \mathbb{P}_M \left( \left| \frac{\|Mx\|_2^2}{\|x\|_2^2} - 1 \right| > \epsilon \right) \leq 2e^{-\frac{m\epsilon^2}{6}} \quad (6)$$

Equivalently: with probability  $1 - \delta$ , we have

$$(1 - \epsilon)\|x\|_2^2 \leq \|Mx\|_2^2 \leq (1 + \epsilon)\|x\|_2^2$$

provided  $m \gtrsim \epsilon^{-2} \log(1/\delta)$ .

# Concentration inequality

## Theorem (Concentration of Gaussian Matrices)

Let  $x \in \mathbb{R}^n$ . For a Gaussian matrix  $M$ ,

$$\forall 0 \leq \epsilon \leq 3, \quad \mathbb{P}_M \left( \left| \frac{\|Mx\|_2^2}{\|x\|_2^2} - 1 \right| > \epsilon \right) \leq 2e^{-\frac{m\epsilon^2}{6}} \quad (6)$$

Equivalently: with probability  $1 - \delta$ , we have

$$(1 - \epsilon)\|x\|_2^2 \leq \|Mx\|_2^2 \leq (1 + \epsilon)\|x\|_2^2$$

provided  $m \gtrsim \epsilon^{-2} \log(1/\delta)$ .

We have the RIP **for one vector  $x$** . We need this **for all  $s$ -sparse vectors simultaneously!** This will be done using **union bounds** and **covering numbers of finite-dimensional compact sets (!)**

# Union bound for finite set

## Union Bound

For two events  $A, B$ , it is obvious that

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

# Union bound for finite set

## Union Bound

For two events  $A, B$ , it is obvious that

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

*Exo: how to apply this to obtain concentration for **several** vectors?*

# Union bound for finite set

## Union Bound

For two events  $A, B$ , it is obvious that

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

*Exo: how to apply this to obtain concentration for several vectors?*

We obtain the so-called Johnson-Lindenstrauss lemma.

## Lemma (Johnson-Lindenstrauss)

For a Gaussian matrix  $M$ , let  $0 \leq \epsilon \leq 3, \delta > 0$ .

Let  $\mathcal{Q}$  be a **finite set** of vectors  $\subset \mathbb{R}^n$ . If  $m \geq \frac{6}{\epsilon^2} \log \frac{2|\mathcal{Q}|}{\delta}$ , then

$$\mathbb{P}_M \left( \sup_{x \in \mathcal{Q}} \left| \frac{\|Mx\|_2^2}{\|x\|_2^2} - 1 \right| \leq \epsilon \right) \geq 1 - \delta$$

# Covering numbers

## Covering numbers

A compact set in dimension  $d$  can be covered with  $O(r^{-d})$  balls of radius  $r$ .

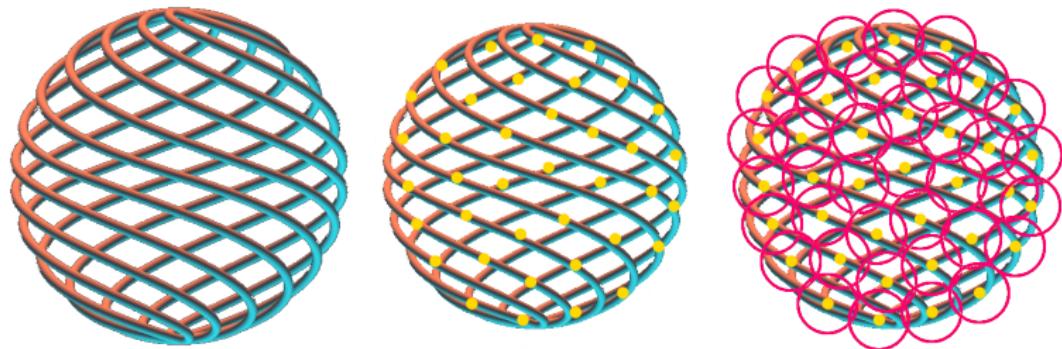
# Covering numbers

## Covering numbers

A compact set in dimension  $d$  can be covered with  $O(r^{-d})$  balls of radius  $r$ .

Finishing the proof of the RIP (exo):

1. Covering the set  $\{x/\|x\|_2, x \in \Sigma_s\}$  with balls of appropriate size



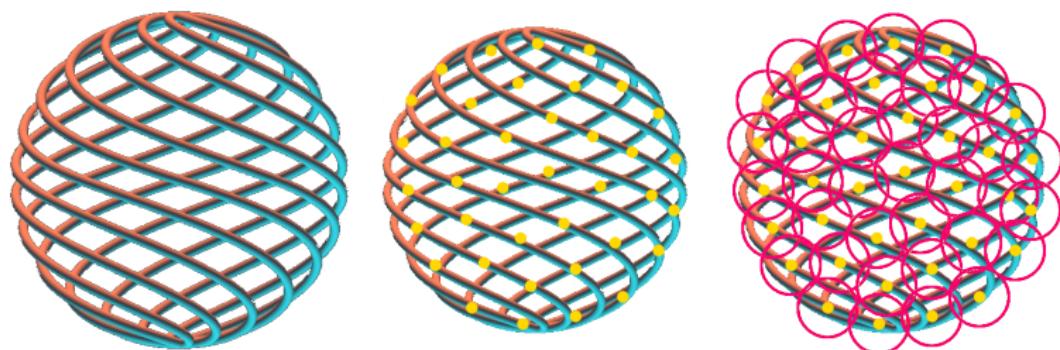
# Covering numbers

## Covering numbers

A compact set in dimension  $d$  can be covered with  $O(r^{-d})$  balls of radius  $r$ .

Finishing the proof of the RIP (exo):

1. Covering the set  $\{x/\|x\|_2, x \in \Sigma_s\}$  with balls of appropriate size
2. Applying the JL lemma for the center of each ball



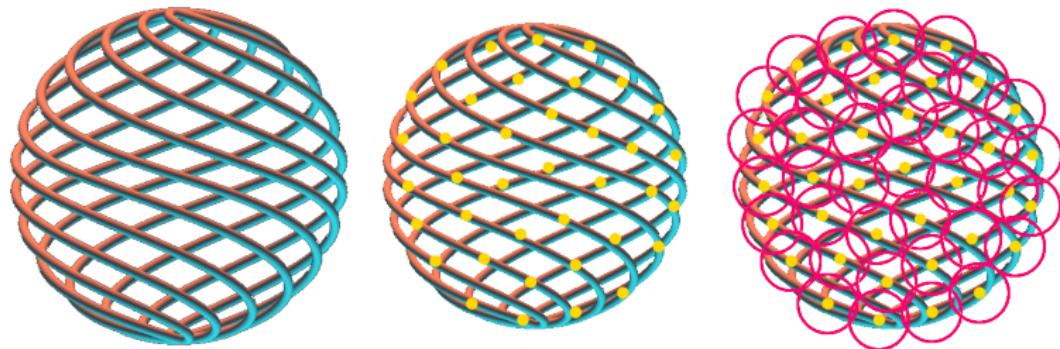
# Covering numbers

## Covering numbers

A compact set in dimension  $d$  can be covered with  $O(r^{-d})$  balls of radius  $r$ .

Finishing the proof of the RIP (exo):

1. Covering the set  $\{x/\|x\|_2, x \in \Sigma_s\}$  with balls of appropriate size
2. Applying the JL lemma for the center of each ball
3. Conclude with triangular inequality.



# **Beyond Sparsity**

# Outline

Introduction

- What is Compressive Sensing?

- Notations (Reminder)

- Problem formulation

Recovery guarantees

- NSP

- Dual certificate

- Coherence

- RIP

Recovering with random matrices?

Concentration inequalities and proving the RIP

**Beyond Sparsity**

- Total Variation

- Structured sparsity

- Matrix completion and Low-rank regularization

Infinite-dimensional signals: superresolution and compressive learning

- Continuous sparsity: superresolution

- Generalized sparsity: sketching

# Total variation model

- Reminder: images are sparse in [wavelet/DCT basis](#).



# Total variation model

- Reminder: images are sparse in [wavelet/DCT basis](#).



- Other model: images are formed by **constant-by-part values**.  
→ the **difference of neighboring pixels** is a sparse signal

This is the “Rudin-Osher-Fatemi” (ROF) model [1992]. It uses the **total variation**:

$$\|x\|_{TV} = \|\nabla x\|_1 = \sum_{i,j \text{ “neighbours”}} |x_i - x_j|$$

→ Several possibilities for “neighbours” (4-neighbours, 8-neighbours, graph...) → [discrete “gradient”](#)

# Total variation application

- Denoising:  $\min_x \|y - x\|^2 + \lambda \|\nabla x\|_1$



→ directly proximal operator of TV (... no closed-form)

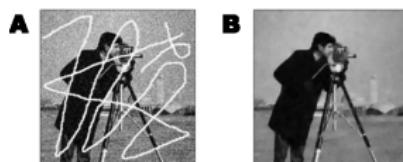
# Total variation application

- Denoising:  $\min_x \|y - x\|^2 + \lambda \|\nabla x\|_1$



→ directly proximal operator of TV (... no closed-form)

- Inpainting:  $\min_x \|y - Mx\|^2 + \lambda \|\nabla x\|_1$



→ Even more complicated (Prox-GD asks for the computation of proximal operator at each operation)

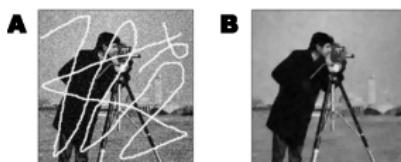
# Total variation application

- Denoising:  $\min_x \|y - x\|^2 + \lambda \|\nabla x\|_1$



→ directly proximal operator of TV (... no closed-form)

- Inpainting:  $\min_x \|y - Mx\|^2 + \lambda \|\nabla x\|_1$



→ Even more complicated (Prox-GD asks for the computation of proximal operator at each operation)

- There exists powerful convex optimization algorithms (Chambolle-Pock)

# Outline

Introduction

- What is Compressive Sensing?

- Notations (Reminder)

- Problem formulation

Recovery guarantees

- NSP

- Dual certificate

- Coherence

- RIP

Recovering with random matrices?

Concentration inequalities and proving the RIP

**Beyond Sparsity**

- Total Variation

- Structured sparsity

- Matrix completion and Low-rank regularization

Infinite-dimensional signals: superresolution and compressive learning

- Continuous sparsity: superresolution

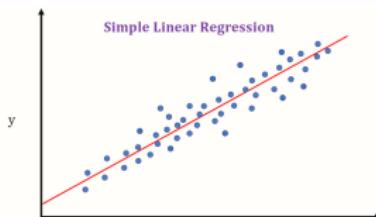
- Generalized sparsity: sketching

# Linear Regression and Model selection

The term “LASSO” comes from **Statistics**. There, and in Machine Learning, sparsity is used for **model selection**. Just different notations and interpretation!

# Linear Regression and Model selection

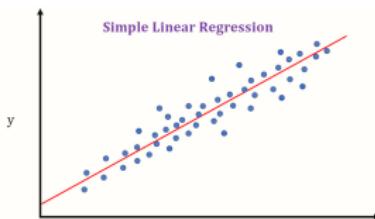
The term “LASSO” comes from **Statistics**. There, and in Machine Learning, sparsity is used for **model selection**. Just different notations and interpretation!



- seek to **linearly regress** the variable  $y$  from **high-dimensional** covariates  $x^1, \dots, x^d$ , but we hypothesize that **not all covariates have influence** on  $y$ .

# Linear Regression and Model selection

The term “LASSO” comes from **Statistics**. There, and in Machine Learning, sparsity is used for **model selection**. Just different notations and interpretation!



- seek to **linearly regress** the variable  $y$  from **high-dimensional** covariates  $x^1, \dots, x^d$ , but we hypothesize that **not all covariates have influence** on  $y$ .
- We have examples  $(y_i, x_i)_{i=1}^n$  with  $x_i = [x_i^1, \dots, x_i^d]$ , but potentially  $d > n$
- We solve:

$$\min_{\beta \in \mathbb{R}^d} \|Y - X^\top \beta\|_2^2 + \lambda \|\beta\|_1$$

where  $x_i$  are the columns of  $X$ .  $\beta$  is the regression vector.

- We **never** have exact recovery here. Different type of guarantees...

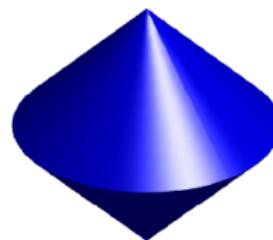
# Group Sparsity

Sometimes we know in advance that **group of covariates** have influence **together** or not.

- Group coordinates of  $\beta$  together:  $\beta_{g_1}, \dots, \beta_{g_p}$  with group  $\beta_{g_l} = [\beta_1^l, \dots, \beta_{i_l}^l]$
- Group Lasso: new regularization term

$$\|\beta\|_{gL} = \sum_{l=1}^p \|\beta_{g_l}\|_2$$

→  $\ell_1$ -norm of  $\ell_2$ -norms of groups. Group of size 1: regular sparsity.

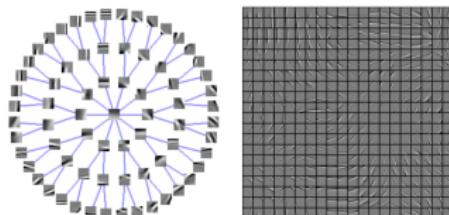


$$\Omega(\beta) = \|\beta_{\{1,2\}}\|_2 + |\beta_3|.$$

- Need different optimization algorithms...

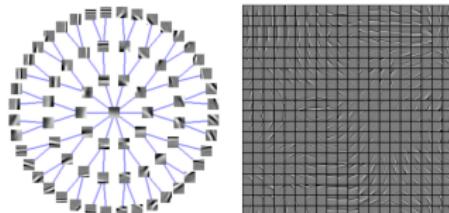
# Beyond group sparsity: structured sparsity

- Structured dictionary for image patches

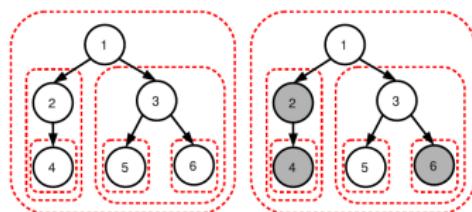


# Beyond group sparsity: structured sparsity

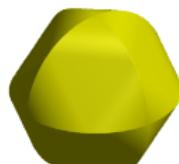
- Structured dictionary for image patches



- Hierarchical norm, overlapping groups, union of groups...



$$\Omega(\beta) = \|\beta\|_2 + |\beta_2| + |\beta_3|.$$



$$\psi(\beta) \text{ with } \mathcal{G} = \{\{1,2\}, \{2,3\}, \{1,3\}\}.$$

# Outline

Introduction

- What is Compressive Sensing?

- Notations (Reminder)

- Problem formulation

Recovery guarantees

- NSP

- Dual certificate

- Coherence

- RIP

Recovering with random matrices?

Concentration inequalities and proving the RIP

**Beyond Sparsity**

- Total Variation

- Structured sparsity

- Matrix completion and Low-rank regularization**

Infinite-dimensional signals: superresolution and compressive learning

- Continuous sparsity: superresolution

- Generalized sparsity: sketching

# Matrix completion

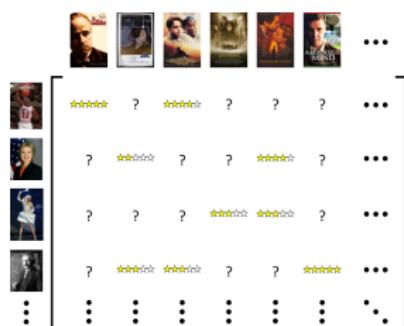
- **Recommender system:** given users as row and items as columns, fill with preferences. Known preference → missing values. Also called “collaborative filtering”



The “Netflix prize”: a huge competition (100 480 507 ratings, 480 189 users, 17 770 movies) in 2008, with a 1M dollars prize. Greatly boosted the interest in collaborative filtering...

# Matrix completion

- **Recommender system:** given users as row and items as columns, fill with preferences. Known preference → missing values. Also called “collaborative filtering”

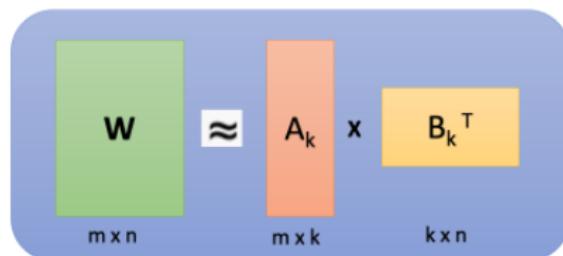


The “Netflix prize”: a huge competition (100 480 507 ratings, 480 189 users, 17 770 movies) in 2008, with a 1M dollars prize. Greatly boosted the interest in collaborative filtering...

- An instance of **matrix completion** and **missing data inputation**, two important fields

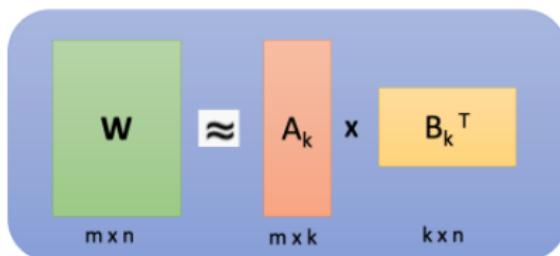
# Low-rank

- For such problems, the **low-rank assumption** can seem natural.



# Low-rank

- For such problems, the **low-rank assumption** can seem natural.



→ Each **item**  $i$  is represented by a “feature vector”  $a_i \in \mathbb{R}^k$ , each **user**  $j$  by a “preference vector”  $b_j \in \mathbb{R}^k$ , and the score is the scalar product between the two:

$$w_{ij} = a_i^\top b_j$$

# Nuclear norm

- As with regular sparsity, choosing in advance the rank  $k$  can be problematic: we'd rather prefer a regularization term! (like  $\ell_0$  or  $\ell_1$ )

$$\min_W \left\| \underbrace{\hat{W}}_{\text{Obs.}} - \underbrace{\mathcal{M}}_{\text{Mask}}(W) \right\|_F^2 + \lambda \|W\|_?$$

# Nuclear norm

- As with regular sparsity, choosing in advance the rank  $k$  can be problematic: we'd rather prefer a regularization term! (like  $\ell_0$  or  $\ell_1$ )

$$\min_W \left\| \underbrace{\hat{W}}_{\text{Obs.}} - \underbrace{\mathcal{M}}_{\text{Mask}}(W) \right\|_F^2 + \lambda \|W\|_?$$

- For low-rank matrices, it is the pseudo-norm:  
 $\|W\|_0 = \#\text{nnz}\{\sigma_i(W) \mid 1 \leq i \leq \min(n, m)\}$ , where  $\sigma_i(W) \geq 0$  are the singular values of  $W$ .

# Nuclear norm

- As with regular sparsity, choosing in advance the rank  $k$  can be problematic: we'd rather prefer a regularization term! (like  $\ell_0$  or  $\ell_1$ )

$$\min_W \left\| \underbrace{\hat{W}}_{\text{Obs.}} - \underbrace{\mathcal{M}}_{\text{Mask}}(W) \right\|_F^2 + \lambda \|W\|_?$$

- For low-rank matrices, it is the pseudo-norm:  
 $\|W\|_0 = \#\text{nnz}\{\sigma_i(W) \mid 1 \leq i \leq \min(n, m)\}$ , where  $\sigma_i(W) \geq 0$  are the singular values of  $W$ .
- Its convex relaxation is the nuclear norm:

$$\|W\|_* = \sum_{i=1}^{\min(n, m)} \sigma_i(W)$$

# Nuclear norm

- As with regular sparsity, choosing in advance the rank  $k$  can be problematic: we'd rather prefer a regularization term! (like  $\ell_0$  or  $\ell_1$ )

$$\min_W \left\| \underbrace{\hat{W}}_{\text{Obs.}} - \underbrace{\mathcal{M}}_{\text{Mask}}(W) \right\|_F^2 + \lambda \|W\|_?$$

- For low-rank matrices, it is the pseudo-norm:  
 $\|W\|_0 = \#\text{nnz}\{\sigma_i(W) \mid 1 \leq i \leq \min(n, m)\}$ , where  $\sigma_i(W) \geq 0$  are the singular values of  $W$ .
- Its convex relaxation is the nuclear norm:

$$\|W\|_* = \sum_{i=1}^{\min(n, m)} \sigma_i(W)$$

- Nuclear norm minimization (and variants) is still an active research field. Main problem: computational complexity. The SVD is super costly! Many (many) alternative approach...

# Infinite-dimensional signals: super-resolution and compressive learning

# Continuous sparsity

For now, we have seen:

- **Sensing continuous signals:** Shannon-Nyquist theorem. Can recover **every bandlimited signal**, no notion of “sparsity”

# Continuous sparsity

For now, we have seen:

- **Sensing continuous signals:** Shannon-Nyquist theorem. Can recover **every bandlimited signal**, no notion of “sparsity”
- **Compressive sensing:** measure **discrete** sparse signal directly proportionally to their sparsity.

# Continuous sparsity

For now, we have seen:

- **Sensing continuous signals:** Shannon-Nyquist theorem. Can recover **every bandlimited signal**, no notion of “sparsity”
- **Compressive sensing:** measure **discrete** sparse signal directly proportionally to their sparsity.

Can we do sparsity **for continuous signals?**

# Continuous sparsity

For now, we have seen:

- **Sensing continuous signals:** Shannon-Nyquist theorem. Can recover **every bandlimited signal**, no notion of “sparsity”
- **Compressive sensing:** measure **discrete** sparse signal directly proportionally to their sparsity.

Can we do sparsity **for continuous signals?**

- What does it mean for a continuous signal to be “sparse”?
- Can we have continuous equivalent of  $\ell_0, \ell_1 \dots$ ?

# Outline

Introduction

- What is Compressive Sensing?

- Notations (Reminder)

- Problem formulation

Recovery guarantees

- NSP

- Dual certificate

- Coherence

- RIP

Recovering with random matrices?

Concentration inequalities and proving the RIP

Beyond Sparsity

- Total Variation

- Structured sparsity

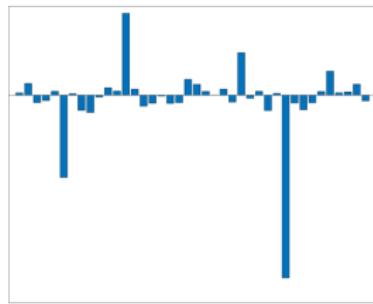
- Matrix completion and Low-rank regularization

**Infinite-dimensional signals: superresolution and compressive learning**

- Continuous sparsity: superresolution

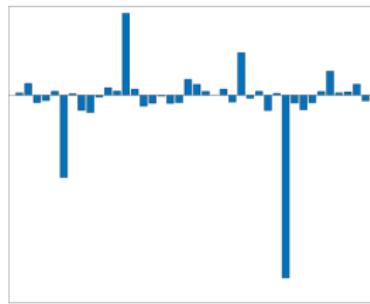
- Generalized sparsity: sketching

# Discrete compressive sensing



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

# Discrete compressive sensing

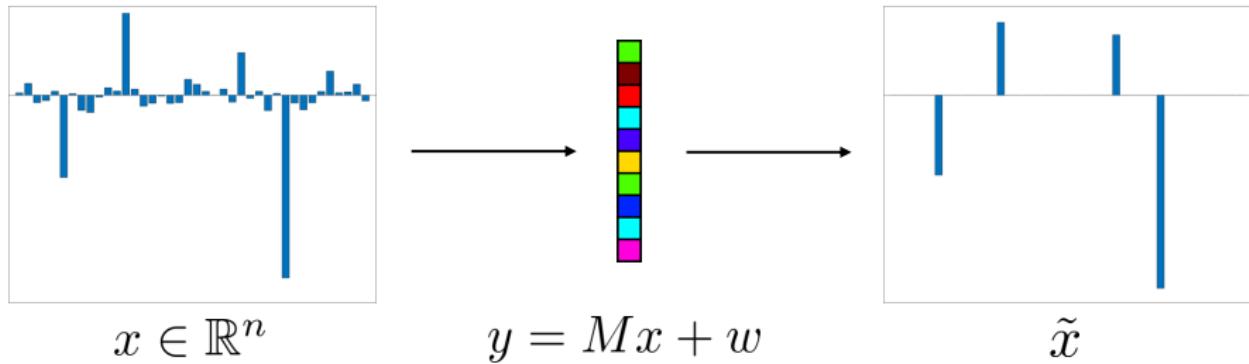


$$x \in \mathbb{R}^n$$

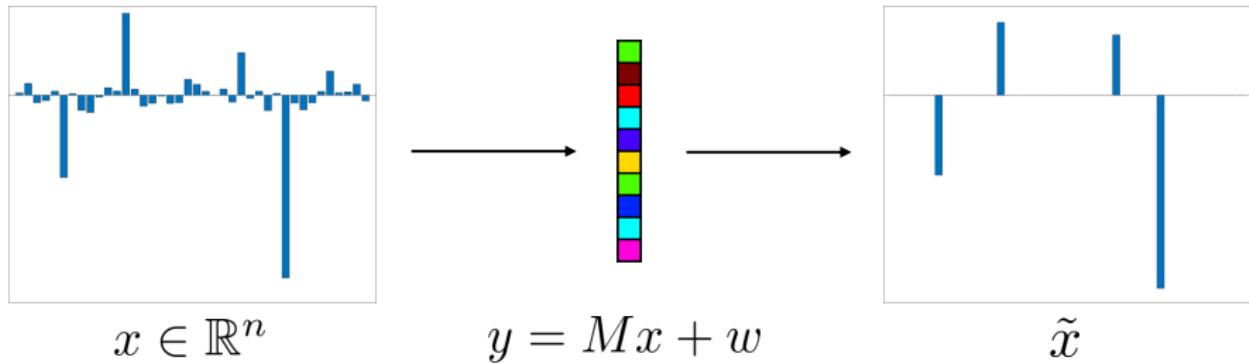


$$y = Mx + w$$

# Discrete compressive sensing

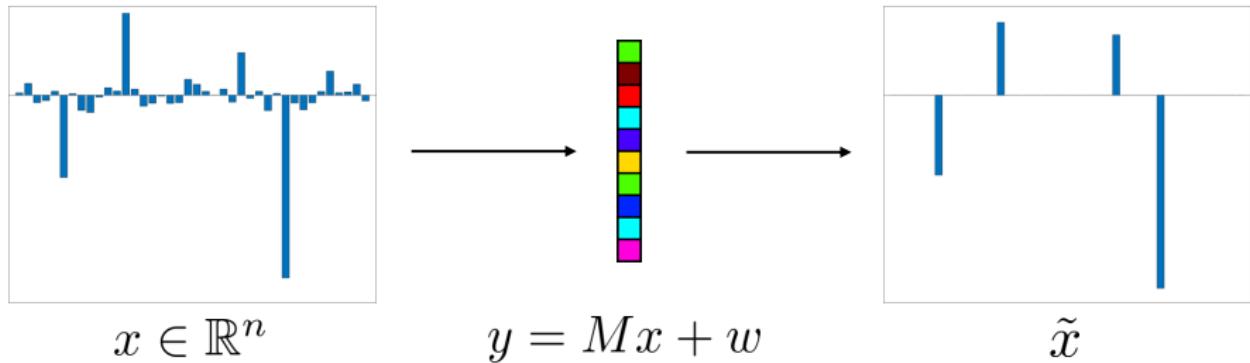


# Discrete compressive sensing



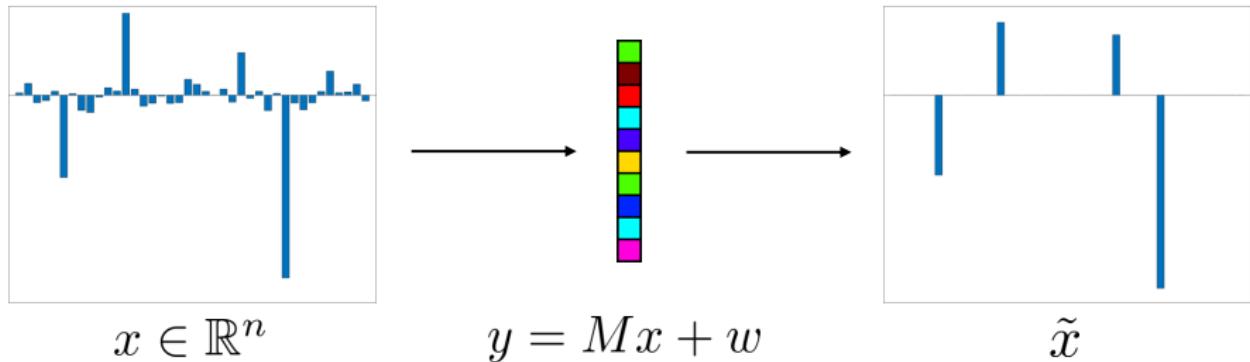
- **Signal:** vector

# Discrete compressive sensing



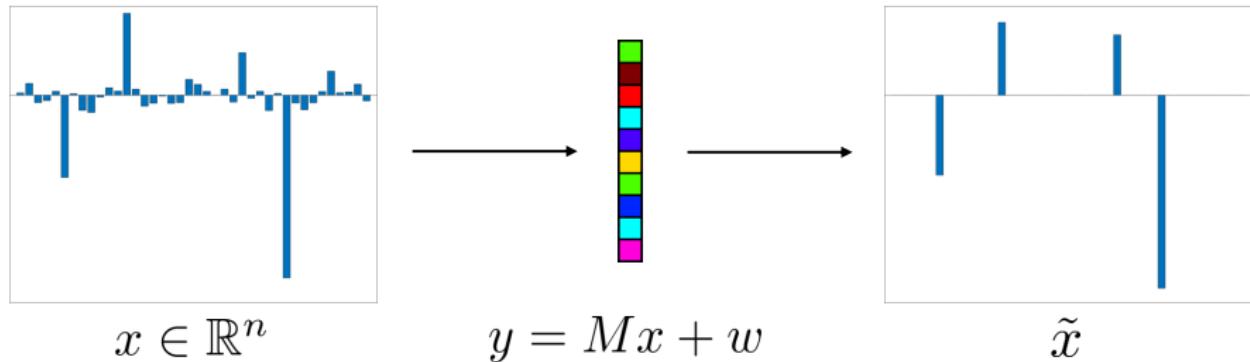
- **Signal:** vector
- **Sparsity:** few non-zeros coefficients

# Discrete compressive sensing



- **Signal:** vector
- **Sparsity:** few non-zeros coefficients
- **Dimensionality reduction** (often random matrix)

# Discrete compressive sensing

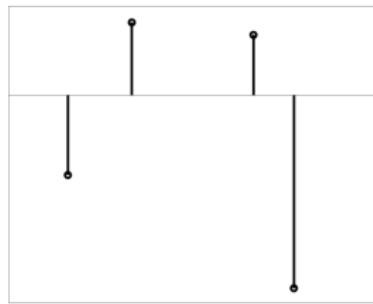


- **Signal:** vector
- **Sparsity:** few non-zeros coefficients
- **Dimensionality reduction** (often random matrix)
- **Recovery:** convex relaxation

$$\min_{\|x\|_0 \leq s} \|Mx - y\| \longrightarrow \boxed{\min_x \frac{1}{2} \|Mx - y\|_2^2 + \lambda \|x\|_1}$$

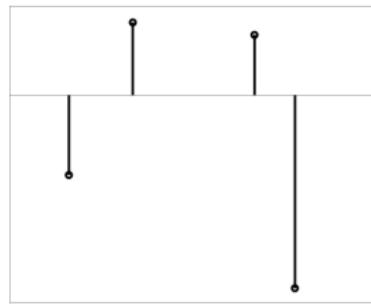
LASSO

# Continuous sparsity ?



$$\mu \in \mathcal{M}(\mathcal{X})$$

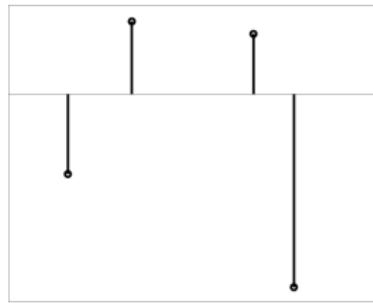
# Continuous sparsity ?



$$y = \Phi\mu + w$$

$$\mu \in \mathcal{M}(\mathcal{X})$$

# Continuous sparsity ?



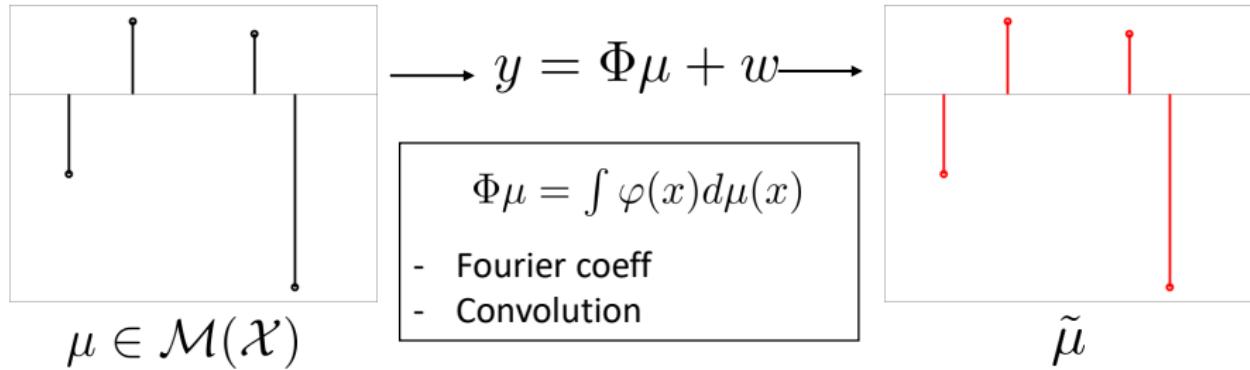
$$\mu \in \mathcal{M}(\mathcal{X})$$

$$\longrightarrow y = \Phi\mu + w$$

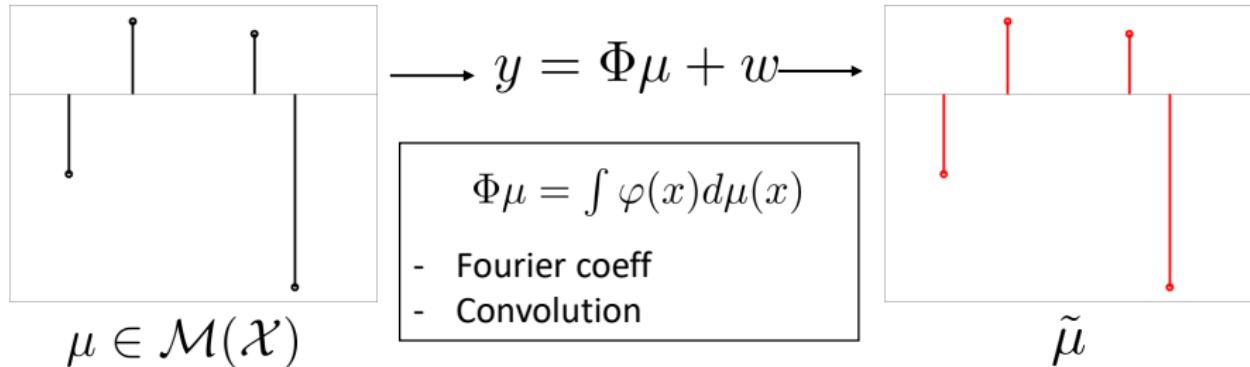
$$\Phi\mu = \int \varphi(x)d\mu(x)$$

- Fourier coeff
- Convolution

# Continuous sparsity ?

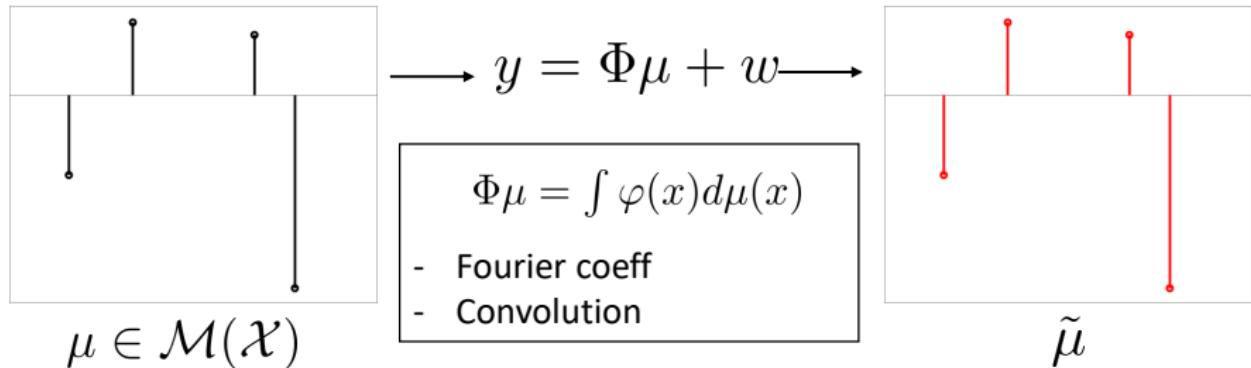


# Continuous sparsity ?



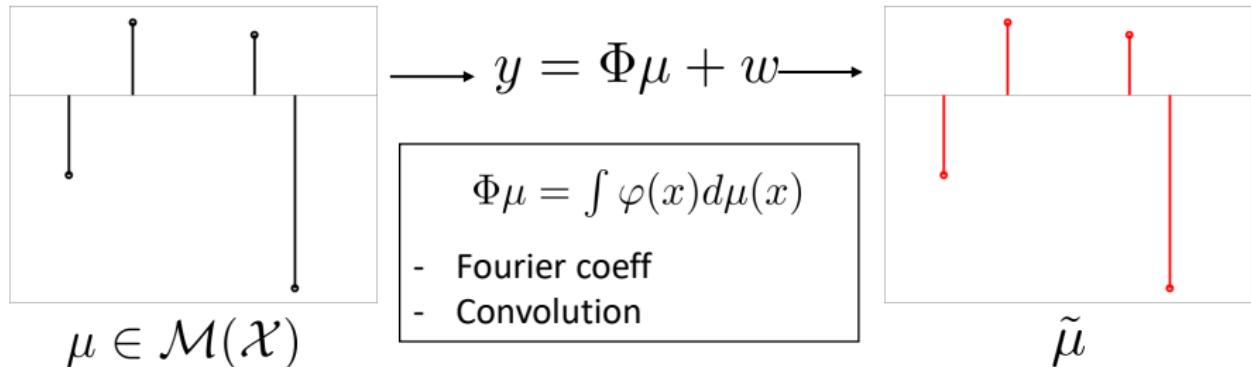
- **Signal:** Radon measure

# Continuous sparsity ?



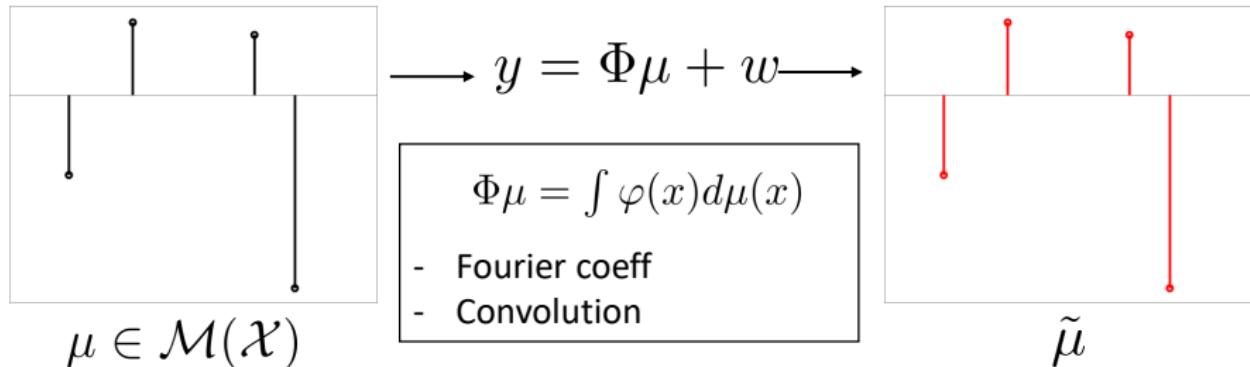
- **Signal:** Radon measure
- **Sparsity:**  $\mu_0 = \sum_i a_i \delta_{x_i}$

# Continuous sparsity ?



- **Signal:** Radon measure
- **Sparsity:**  $\mu_0 = \sum_i a_i \delta_{x_i}$
- **Dimensionality reduction** (e.g. first Fourier coefficients)

# Continuous sparsity ?

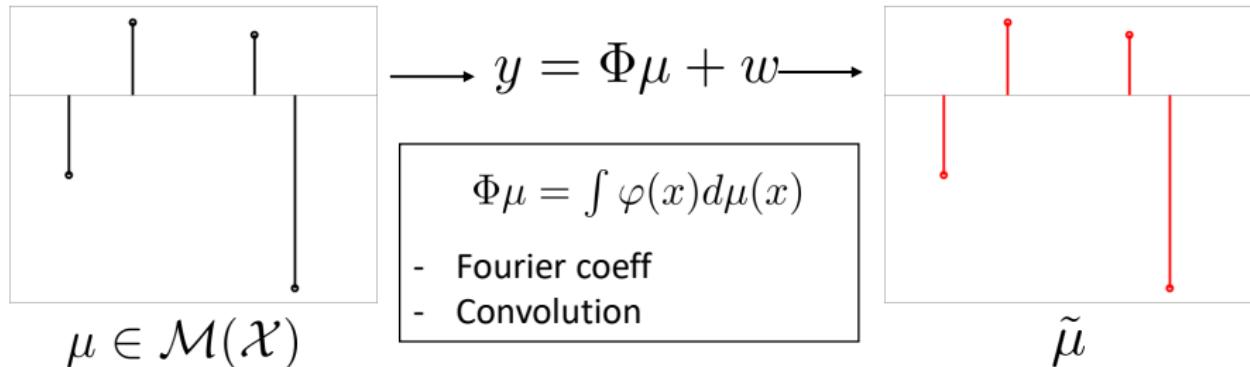


- **Signal:** Radon measure
- **Sparsity:**  $\mu_0 = \sum_i a_i \delta_{x_i}$
- **Dimensionality reduction** (e.g. first Fourier coefficients)
- **Recovery:** convex relaxation?

$$\min_{a,x} \left\| \Phi\left(\sum_i a_i \delta_{x_i}\right) - y \right\|$$

See Keriven 2017, Gribonval 2017

# Continuous sparsity ?



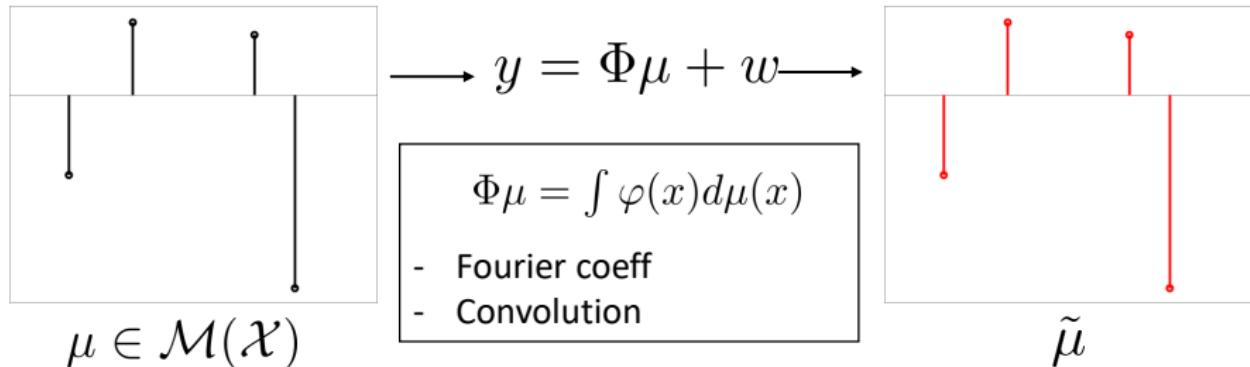
- **Signal:** Radon measure
- **Sparsity:**  $\mu_0 = \sum_i a_i \delta_{x_i}$
- **Dimensionality reduction** (e.g. first Fourier coefficients)
- **Recovery:** convex relaxation?

$\min_{a,x} \|\Phi(\sum_i a_i \delta_{x_i}) - y\|$  →  
See Keriven 2017, Gribonval 2017

**BLASSO [De Castro, Gamboa 2012]**

$$\min_{\mu} \frac{1}{2} \|\Phi\mu - y\|^2 + \lambda |\mu|(\mathcal{X})$$

# Continuous sparsity ?



- **Signal:** Radon measure
- **Sparsity:**  $\mu_0 = \sum_i a_i \delta_{x_i}$
- **Dimensionality reduction** (e.g. first Fourier coefficients)
- **Recovery:** convex relaxation?

$$\min_{a,x} \|\Phi(\sum_i a_i \delta_{x_i}) - y\|$$

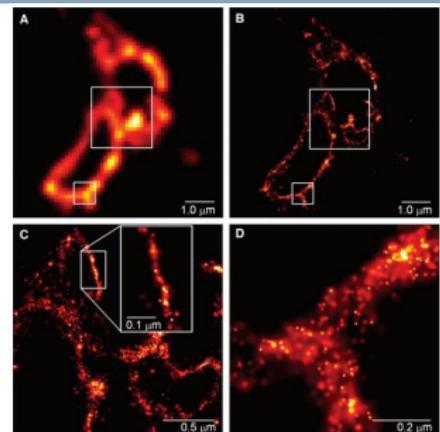
See Keriven 2017, Gribonval 2017

**BLASSO [De Castro, Gamboa 2012]**

$$\min_{\mu} \frac{1}{2} \|\Phi\mu - y\|^2 + \lambda |\mu|(\mathcal{X})$$

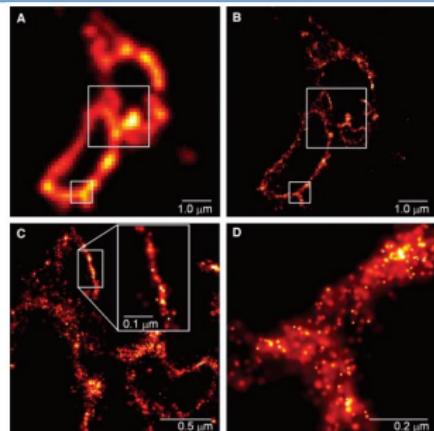
Other approaches: « Prony-like » ESPRIT, MUSIC... (but only 1d noiseless Fourier)

# Example of applications



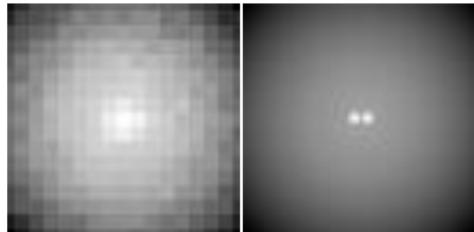
**Fluorescence microscopy (3D)**  
**PALM, STORM... [Betzig 2006]**

# Example of applications

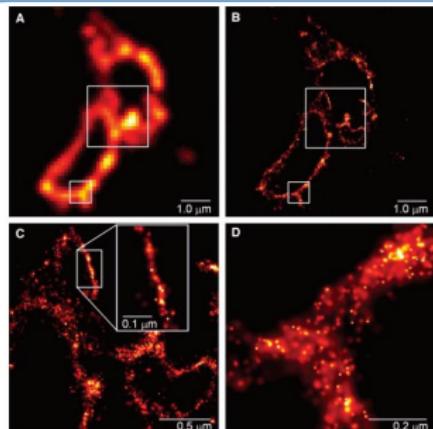


**Fluorescence microscopy (3D)**  
PALM, STORM... [Betzig 2006]

**Astronomy (2D)**  
[Puschmann 2017]



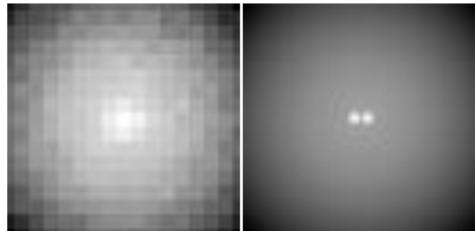
# Example of applications



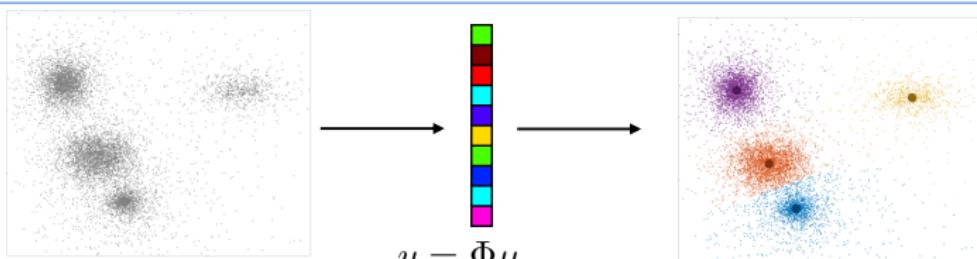
Fluorescence microscopy (3D)  
PALM, STORM... [Betzig 2006]

## Astronomy (2D)

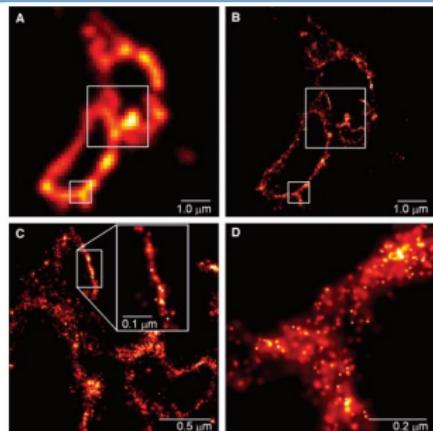
[Puschmann 2017]



Compressive  
mixture model  
learning  
(many D)  
[Keriven 2017]



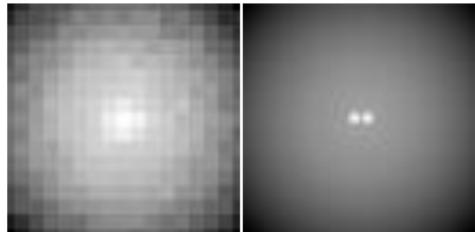
# Example of applications



Fluorescence microscopy (3D)  
PALM, STORM... [Betzig 2006]

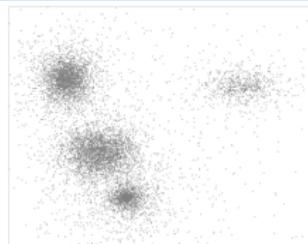
## Astronomy (2D)

[Puschmann 2017]

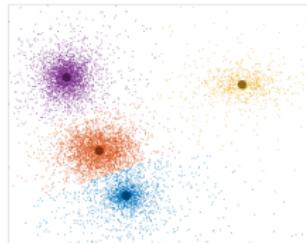


- Neuro-imaging with EEG (3D) [Gramfort 2013]
- 1-layer neural network (many D) [Bach 2017]
- Radar
- Geophysics
- ...

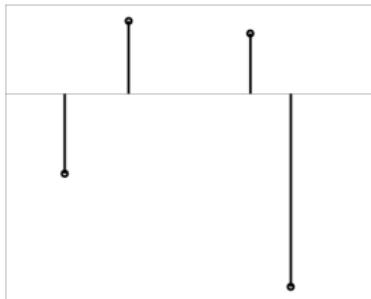
Compressive  
mixture model  
learning  
(many D)  
[Keriven 2017]



$$y = \Phi \mu_{\text{emp.}}$$

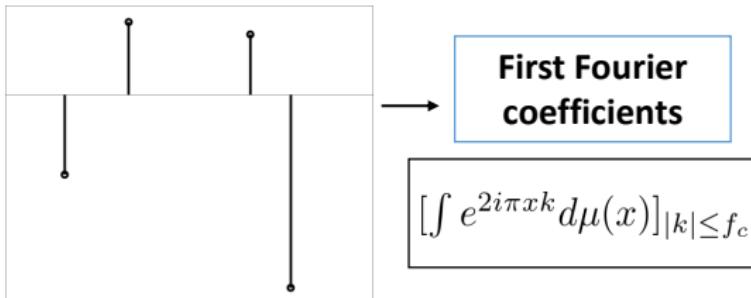


# A seminal result [Candès, Fernandez-Granda 2012]



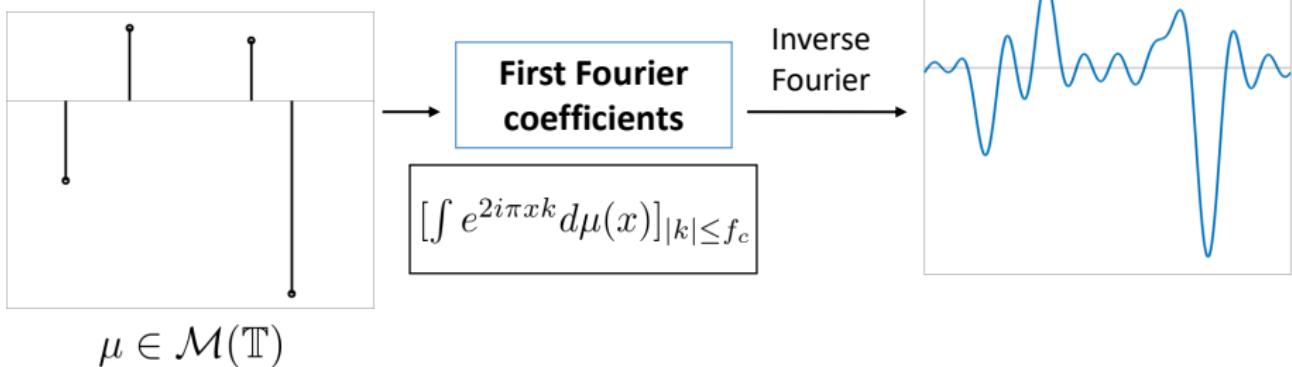
$$\mu \in \mathcal{M}(\mathbb{T})$$

# A seminal result [Candès, Fernandez-Granda 2012]

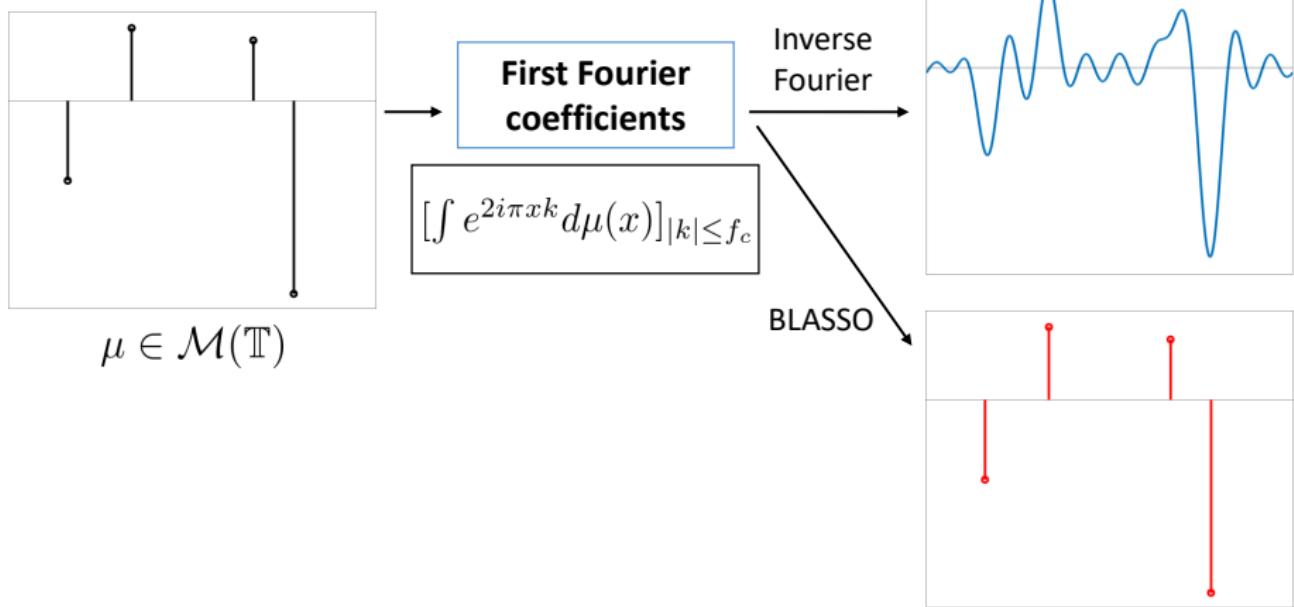


$$\mu \in \mathcal{M}(\mathbb{T})$$

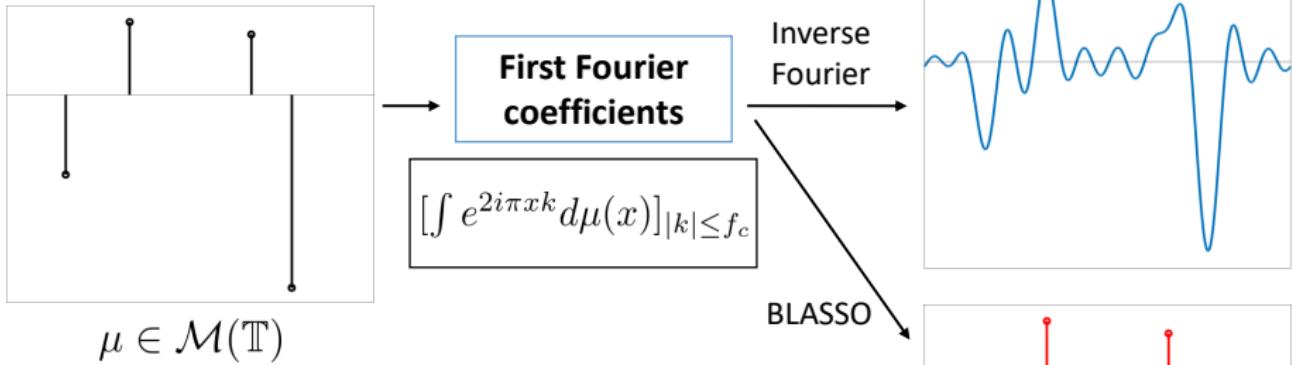
# A seminal result [Candès, Fernandez-Granda 2012]



# A seminal result [Candès, Fernandez-Granda 2012]

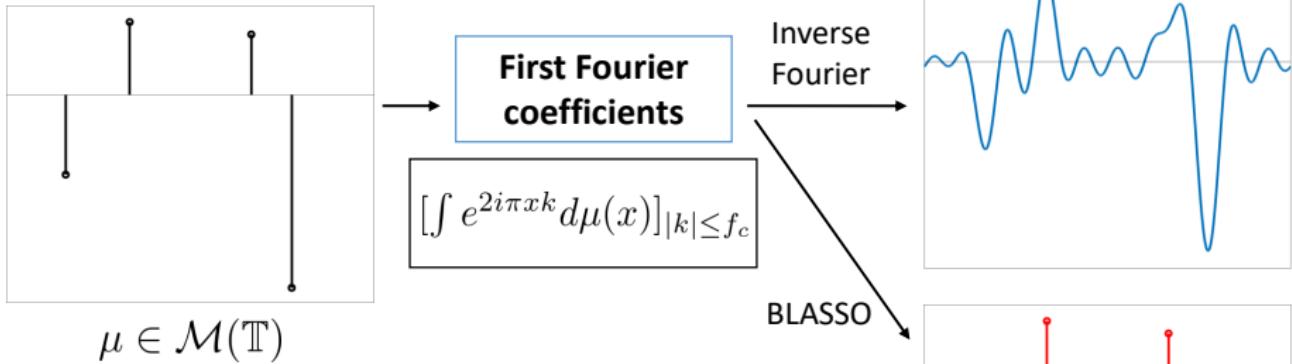


# A seminal result [Candès, Fernandez-Granda 2012]



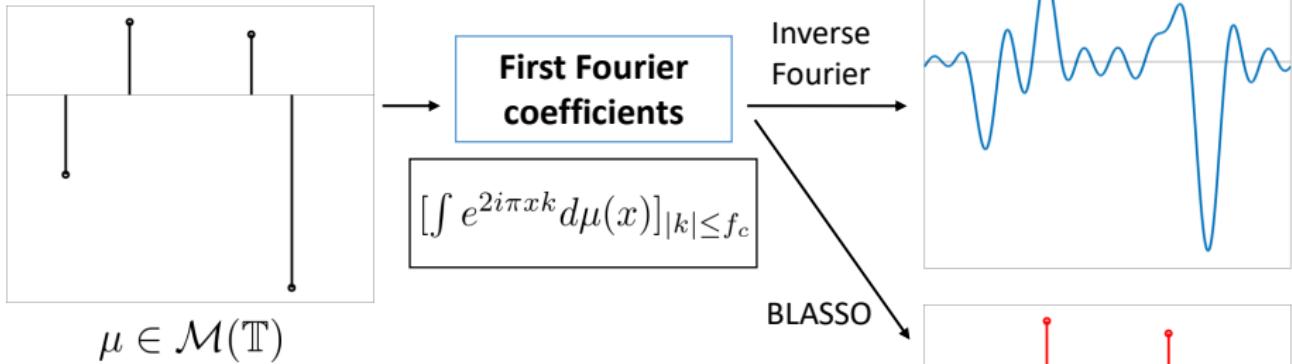
- **Regular Fourier on the Torus**
- Minimal separation  $\Delta \geq \mathcal{O}(1/f_c)$

# A seminal result [Candès, Fernandez-Granda 2012]



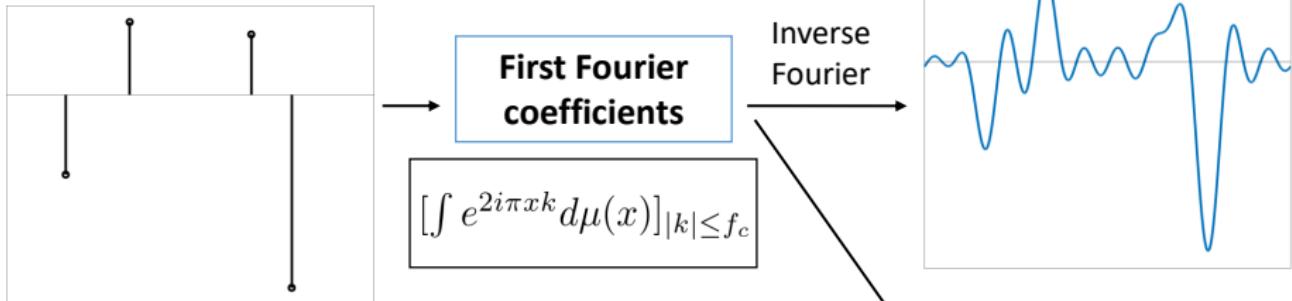
- **Regular Fourier on the Torus**
- Minimal separation  $\Delta \geq \mathcal{O}(1/f_c)$
- **Reconstruction:** formulated as **SDP** (other: Frank-Wolfe, greedy, Prony-like...)

# A seminal result [Candès, Fernandez-Granda 2012]

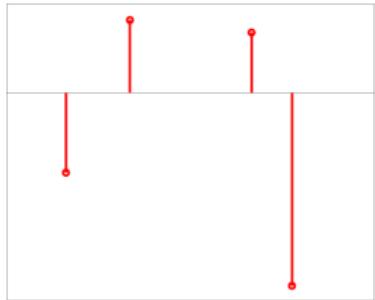


- **Regular Fourier on the Torus**
- Minimal separation  $\Delta \geq \mathcal{O}(1/f_c)$
- **Reconstruction:** formulated as **SDP** (other: Frank-Wolfe, greedy, Prony-like...)
- **Noise:** weak convergence (mass of  $\tilde{\mu}$  **concentrated** around true Diracs)

# A seminal result [Candès, Fernandez-Granda 2012]



- **Regular Fourier on the Torus**
- Minimal separation  $\Delta \geq \mathcal{O}(1/f_c)$
- **Reconstruction:** formulated as **SDP** (other: Frank-Wolfe, greedy, Prony-like...)
- **Noise:** weak convergence (mass of  $\tilde{\mu}$  **concentrated** around true Diracs)



Many extensions  
since...

# Previous work, contribution

## Relevant previous works:

- [Tang, Recht 2013]:  $m \geq s \log(s) \log(f_c)$  **random** Fourier coefficients are sufficient

# Previous work, contribution

## Relevant previous works:

- [Tang, Recht 2013]:  $m \geq s \log(s) \log(f_c)$  **random** Fourier coefficients are sufficient
  - **Random signs assumption**
  - 1D discrete Fourier

# Previous work, contribution

## Relevant previous works:

- [Tang, Recht 2013]:  $m \geq s \log(s) \log(f_c)$  **random** Fourier coefficients are sufficient
  - **Random signs assumption**
  - 1D discrete Fourier
- [Bendory et al. 2016]: extension to other measurement operators

# Previous work, contribution

## Relevant previous works:

- [Tang, Recht 2013]:  $m \geq s \log(s) \log(f_c)$  **random** Fourier coefficients are sufficient
  - **Random signs assumption**
  - 1D discrete Fourier
- [Bendory et al. 2016]: extension to other measurement operators
  - Minimal separation does not take into account **geometry** of the meas. operator

# Previous work, contribution

## Relevant previous works:

- [Tang, Recht 2013]:  $m \geq s \log(s) \log(f_c)$  **random** Fourier coefficients are sufficient
  - **Random signs assumption**
  - 1D discrete Fourier
- [Bendory et al. 2016]: extension to other measurement operators
  - Minimal separation does not take into account **geometry** of the meas. operator
- [Duval, Peyré 2015]: In some cases, **support stability** in the **small noise regime**

# Previous work, contribution

## Relevant previous works:

- [Tang, Recht 2013]:  $m \geq s \log(s) \log(f_c)$  **random** Fourier coefficients are sufficient
  - **Random signs assumption**
  - 1D discrete Fourier
- [Bendory et al. 2016]: extension to other measurement operators
  - Minimal separation does not take into account **geometry** of the meas. operator
- [Duval, Peyré 2015]: In some cases, **support stability** in the small noise regime
  - **Noise level** under which support stability is achieved?

# Previous work, contribution

## Relevant previous works:

- [Tang, Recht 2013]:  $m \geq s \log(s) \log(f_c)$  **random** Fourier coefficients are sufficient
  - **Random signs assumption**
  - 1D discrete Fourier
- [Bendory et al. 2016]: extension to other measurement operators
  - Minimal separation does not take into account **geometry** of the meas. operator
- [Duval, Peyré 2015]: In some cases, **support stability** in the small noise regime
  - **Noise level** under which support stability is achieved?

## Contributions:

- Generalize to many multi-d measurement operators, express the minimal separation as a **geometry-aware Fisher metric**

# Previous work, contribution

## Relevant previous works:

- [Tang, Recht 2013]:  $m \geq s \log(s) \log(f_c)$  **random** Fourier coefficients are sufficient
  - **Random signs assumption**
  - 1D discrete Fourier
- [Bendory et al. 2016]: extension to other measurement operators
  - Minimal separation does not take into account **geometry** of the meas. operator
- [Duval, Peyré 2015]: In some cases, **support stability** in the small noise regime
  - **Noise level** under which support stability is achieved?

## Contributions:

- Generalize to many multi-d measurement operators, express the minimal separation as a **geometry-aware Fisher metric**
- 1: **Remove the random sign assumption** (*weak convergence*)

# Previous work, contribution

## Relevant previous works:

- [Tang, Recht 2013]:  $m \geq s \log(s) \log(f_c)$  **random Fourier coefficients** are sufficient
  - **Random signs assumption**
  - 1D discrete Fourier
- [Bendory et al. 2016]: extension to other measurement operators
  - Minimal separation does not take into account **geometry** of the meas. operator
- [Duval, Peyré 2015]: In some cases, **support stability** in the small noise regime
  - **Noise level** under which support stability is achieved?

## Contributions:

- Generalize to many multi-d measurement operators, express the minimal separation as a **geometry-aware Fisher metric**
- 1: **Remove the random sign assumption** (*weak convergence*)
- 2: Prove **support stability** when  $\|w\| \leq s^{-1}$  (*with random signs*)

# Outline

- 1 Background on dual certificates
- 2 Minimal separation and Fisher metric
- 3 Main results, applications
- 4 Conclusion, outlooks

# Dual certificates

**Random linear operator:**

$$\omega_1, \dots, \omega_m \stackrel{iid}{\sim} \Lambda$$

$$\Phi\mu = \frac{1}{\sqrt{m}} \left[ \int \varphi_{\omega_k}(x) d\mu(x) \right]_{k=1}^m$$

# Dual certificates

**Random linear operator:**

$$\omega_1, \dots, \omega_m \stackrel{iid}{\sim} \Lambda$$

$$\Phi\mu = \frac{1}{\sqrt{m}} \left[ \int \varphi_{\omega_k}(x) d\mu(x) \right]_{k=1}^m$$

**Noisy measurement:**

$$y = \Phi\mu_0 + w$$

# Dual certificates

**Random linear operator:**

$$\omega_1, \dots, \omega_m \stackrel{iid}{\sim} \Lambda$$

$$\Phi\mu = \frac{1}{\sqrt{m}} \left[ \int \varphi_{\omega_k}(x) d\mu(x) \right]_{k=1}^m$$

**Noisy measurement:**  $y = \Phi\mu_0 + w$

**The BLASSO problem:**

$$\min_{\mu} \frac{1}{2} \|\Phi\mu - y\|^2 + \lambda |\mu|(\mathcal{X})$$

# Dual certificates

**Random linear operator:**

$$\omega_1, \dots, \omega_m \stackrel{iid}{\sim} \Lambda$$

$$\Phi\mu = \frac{1}{\sqrt{m}} \left[ \int \varphi_{\omega_k}(x) d\mu(x) \right]_{k=1}^m$$

**Noisy measurement:**  $y = \Phi\mu_0 + w$

**The BLASSO problem:**

$$\min_{\mu} \frac{1}{2} \|\Phi\mu - y\|^2 + \lambda |\mu|(\mathcal{X})$$

**First-order conditions**

$\mu_0$  solution of  
BLASSO

$$\Leftrightarrow \frac{1}{\lambda} \Phi^*(\Phi\mu_0 - y) \in \partial|\mu_0|(\mathcal{X})$$

# Dual certificates

**Random linear operator:**

$$\omega_1, \dots, \omega_m \stackrel{iid}{\sim} \Lambda$$

$$\Phi\mu = \frac{1}{\sqrt{m}} \left[ \int \varphi_{\omega_k}(x) d\mu(x) \right]_{k=1}^m$$

**Noisy measurement:**  $y = \Phi\mu_0 + w$

**The BLASSO problem:**

$$\min_{\mu} \frac{1}{2} \|\Phi\mu - y\|^2 + \lambda |\mu|(\mathcal{X})$$

**First-order conditions**

$\mu_0$  solution of  
BLASSO

$$\Leftrightarrow \frac{1}{\lambda} \Phi^*(\Phi\mu_0 - y) \in \partial|\mu_0|(\mathcal{X})$$

**Dual certificate (noiseless case)**

$\mu_0$  solution of  
 $\min_{\Phi\mu=y} |\mu|(\mathcal{X})$

$$\Leftrightarrow \text{Im}(\Phi^*) \cap \partial|\mu_0|(\mathcal{X}) \neq \emptyset$$

# What does it look like ?

**What is a dual certificate ?**

$$\eta \in \text{Im}(\Phi^*) \cap \partial|\mu_0|(\mathcal{X})$$

# What does it look like ?

**What is a dual certificate ?**

$$\eta \in \text{Im}(\Phi^*) \cap \partial|\mu_0|(\mathcal{X})$$

$$\eta(x) = \sum_{k=1}^m h_k \varphi_{\omega_k}(x)$$

# What does it look like ?

What is a dual certificate ?

$$\eta \in \text{Im}(\Phi^*) \cap \partial|\mu_0|(\mathcal{X})$$

$$\eta(x) = \sum_{k=1}^m h_k \varphi_{\omega_k}(x)$$

**Case**  $\mu_0 = \sum_i a_i \pi_{x_i}$  :

# What does it look like ?

What is a dual certificate ?

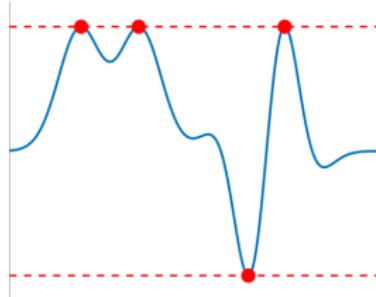
$$\eta \in \text{Im}(\Phi^*) \cap \partial|\mu_0|(\mathcal{X})$$

$$\eta(x) = \sum_{k=1}^m h_k \varphi_{\omega_k}(x)$$

**Case**  $\mu_0 = \sum_i a_i \pi_{x_i}$  :

$$\eta(x_i) = \text{sign}(a_i)$$

$$\|\eta\|_\infty \leq 1$$



# What does it look like ?

What is a dual certificate ?

$$\eta \in \text{Im}(\Phi^*) \cap \partial|\mu_0|(\mathcal{X})$$

$$\eta(x) = \sum_{k=1}^m h_k \varphi_{\omega_k}(x)$$

Case  $\mu_0 = \sum_i a_i \pi_{x_i}$  :

$$\eta(x_i) = \text{sign}(a_i)$$

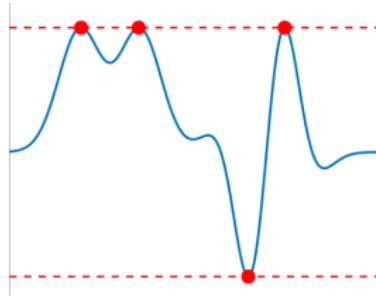
$$\|\eta\|_\infty \leq 1$$

**Non-degenerate** dual certif.

$$|\eta(x)| < 1$$

$$\text{sign}(a_i) \nabla^2 \eta(x_i) \prec 0$$

Ensures **uniqueness** and **robustness**...



# What does it look like ?

What is a dual certificate ?

$$\eta \in \text{Im}(\Phi^*) \cap \partial|\mu_0|(\mathcal{X})$$

$$\eta(x) = \sum_{k=1}^m h_k \varphi_{\omega_k}(x)$$

Intuitively:

Larger  $m \rightarrow$  easier interpolation

Case  $\mu_0 = \sum_i a_i \pi_{x_i}$  :

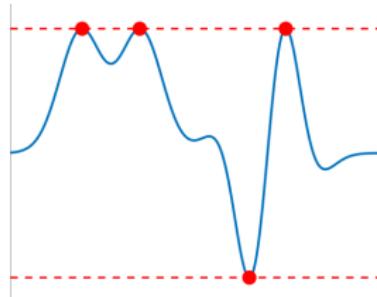
$$\eta(x_i) = \text{sign}(a_i)$$

$$\|\eta\|_\infty \leq 1$$

**Non-degenerate** dual certif.

$$|\eta(x)| < 1$$

$$\text{sign}(a_i) \nabla^2 \eta(x_i) \prec 0$$



Ensures **uniqueness** and **robustness**...

# Proof strategy

**Step 1:** Study the limit case  $m \rightarrow \infty$  to derive an appropriate notion of minimal separation

# Proof strategy

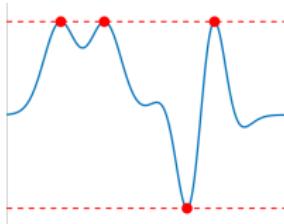
**Step 1:** Study the limit case  $m \rightarrow \infty$  to derive an appropriate notion of minimal separation

**Step 2:** bound the deviation for finite number of measurements

# Proof strategy

**Step 1:** Study the limit case  $m \rightarrow \infty$  to derive an appropriate notion of minimal separation

**Step 2:** bound the deviation for finite number of measurements

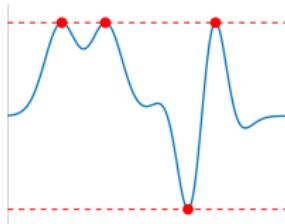


$$m \rightarrow \infty$$

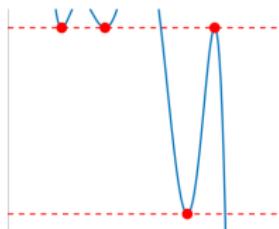
# Proof strategy

**Step 1:** Study the limit case  $m \rightarrow \infty$  to derive an appropriate notion of minimal separation

**Step 2:** bound the deviation for finite number of measurements



$m \rightarrow \infty$

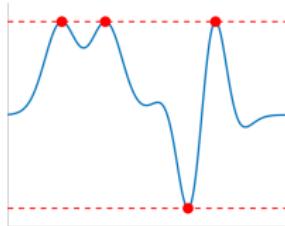


$m=10$

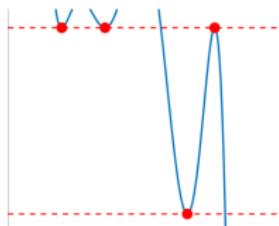
# Proof strategy

**Step 1:** Study the limit case  $m \rightarrow \infty$  to derive an appropriate notion of minimal separation

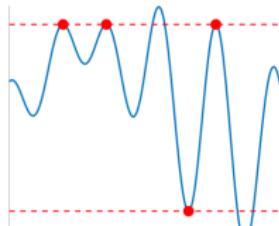
**Step 2:** bound the deviation for finite number of measurements



$m \rightarrow \infty$



$m=10$

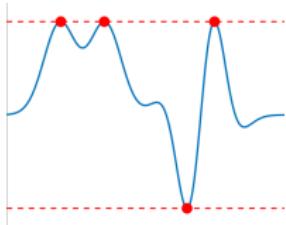


$m=50$

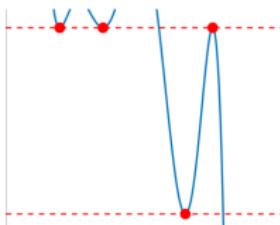
# Proof strategy

**Step 1:** Study the limit case  $m \rightarrow \infty$  to derive an appropriate notion of minimal separation

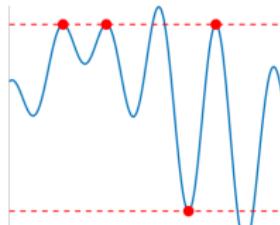
**Step 2:** bound the deviation for finite number of measurements



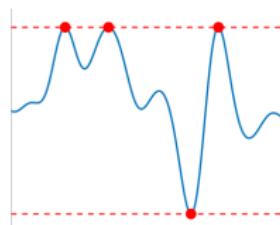
$m \rightarrow \infty$



$m=10$



$m=50$

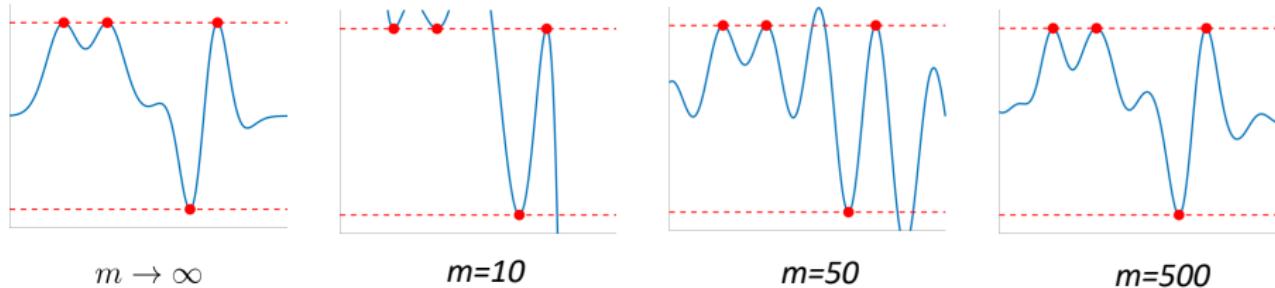


$m=500$

# Proof strategy

**Step 1:** Study the limit case  $m \rightarrow \infty$  to derive an appropriate notion of minimal separation

**Step 2:** bound the deviation for finite number of measurements



**Step 3:** recovery

- Adaptation of [Azaïs 2015] for weak convergence
- Quantitative Implicit Function Theorem [Denoyelle 2015] for support stability

# Outline

- 1 Background on dual certificates
- 2 Minimal separation and Fisher metric
- 3 Main results, applications
- 4 Conclusion, outlooks

# Limit covariance kernels

## How to construct a certificate ?

Study **limit covariance kernel** when  $m \rightarrow \infty$ :

$$\kappa(x, x') = \mathbb{E}_\omega \varphi_\omega(x) \varphi_\omega(x')$$

# Limit covariance kernels

## How to construct a certificate ?

Study **limit covariance kernel** when  $m \rightarrow \infty$ :

$$\kappa(x, x') = \mathbb{E}_\omega \varphi_\omega(x) \varphi_\omega(x')$$

Sub-sampled version:

$$\kappa(x, x') = \frac{1}{m} \sum_{k=1}^m \varphi_{\omega_k}(x) \varphi_{\omega_k}(x')$$

# Limit covariance kernels

## How to construct a certificate ?

Study **limit covariance kernel** when  $m \rightarrow \infty$ :

$$\kappa(x, x') = \mathbb{E}_\omega \varphi_\omega(x) \varphi_\omega(x')$$

Sub-sampled version:

$$\kappa(x, x') = \frac{1}{m} \sum_{k=1}^m \varphi_{\omega_k}(x) \varphi_{\omega_k}(x')$$

## Strategy under minimal separation

# Limit covariance kernels

## How to construct a certificate ?

Study **limit covariance kernel** when  $m \rightarrow \infty$ :

$$\kappa(x, x') = \mathbb{E}_\omega \varphi_\omega(x) \varphi_\omega(x')$$

Sub-sampled version:

$$\kappa(x, x') = \frac{1}{m} \sum_{k=1}^m \varphi_{\omega_k}(x) \varphi_{\omega_k}(x')$$

## Strategy under minimal separation

$$\eta \in \text{Span}\{\kappa(x_i, .), \partial\kappa(x_i, .)\} \subset \text{Im}(\Phi^*)$$

# Limit covariance kernels

## How to construct a certificate ?

Study **limit covariance kernel** when  $m \rightarrow \infty$ :

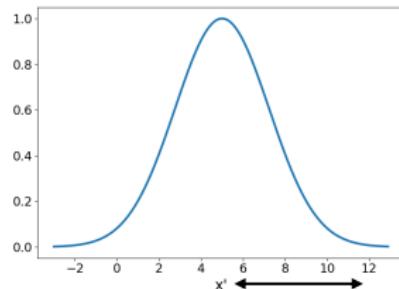
$$\kappa(x, x') = \mathbb{E}_\omega \varphi_\omega(x) \varphi_\omega(x')$$

Sub-sampled version:

$$\kappa(x, x') = \frac{1}{m} \sum_{k=1}^m \varphi_{\omega_k}(x) \varphi_{\omega_k}(x')$$

## Strategy under minimal separation

$$\eta \in \text{Span}\{\kappa(x_i, .), \partial \kappa(x_i, .)\} \subset \text{Im}(\Phi^*)$$



# Limit covariance kernels

## How to construct a certificate ?

Study **limit covariance kernel** when  $m \rightarrow \infty$ :

$$\kappa(x, x') = \mathbb{E}_\omega \varphi_\omega(x) \varphi_\omega(x')$$

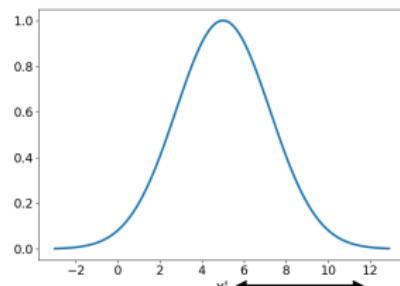
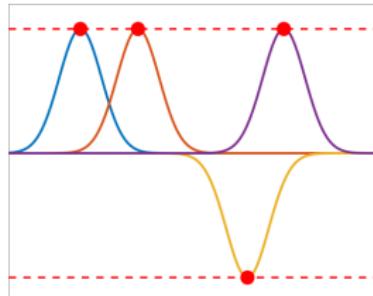
Sub-sampled version:

$$\kappa(x, x') = \frac{1}{m} \sum_{k=1}^m \varphi_{\omega_k}(x) \varphi_{\omega_k}(x')$$

## Strategy under minimal separation

$$\eta \in \text{Span}\{\kappa(x_i, .), \partial \kappa(x_i, .)\} \subset \text{Im}(\Phi^*)$$

1: kernel at each saturation point



Min. Separation  $\Delta$

# Limit covariance kernels

## How to construct a certificate ?

Study **limit covariance kernel** when  $m \rightarrow \infty$ :

$$\kappa(x, x') = \mathbb{E}_\omega \varphi_\omega(x) \varphi_\omega(x')$$

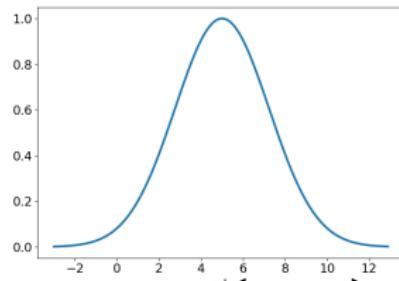
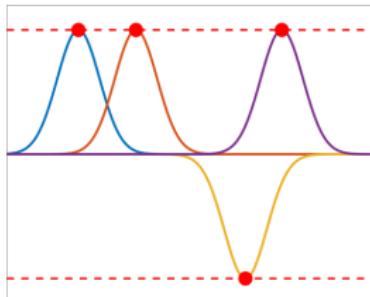
Sub-sampled version:

$$\kappa(x, x') = \frac{1}{m} \sum_{k=1}^m \varphi_{\omega_k}(x) \varphi_{\omega_k}(x')$$

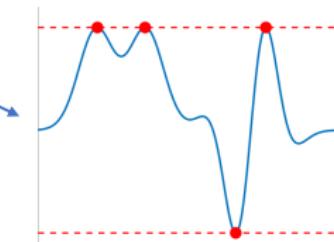
## Strategy under minimal separation

$$\eta \in \text{Span}\{\kappa(x_i, .), \partial \kappa(x_i, .)\} \subset \text{Im}(\Phi^*)$$

### 1: kernel at each saturation point



Min. Separation  $\Delta$



### 2: Small adjustments (minimal separation)

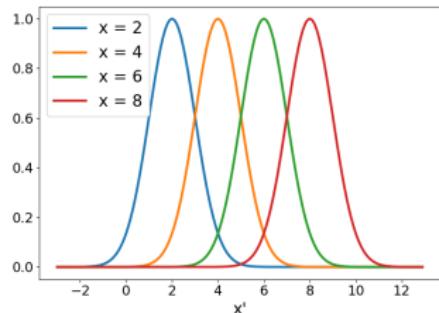
# Minimal separation and Fisher metric

Which metric for separation ?

# Minimal separation and Fisher metric

Which metric for separation ?

Classical case: translation-invariant kernel

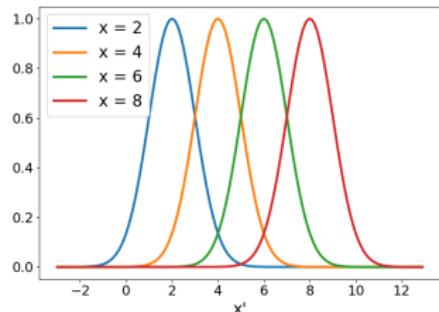


$$\kappa(x, x') = \kappa(x - x')$$

# Minimal separation and Fisher metric

Which metric for separation ?

Classical case: translation-invariant kernel



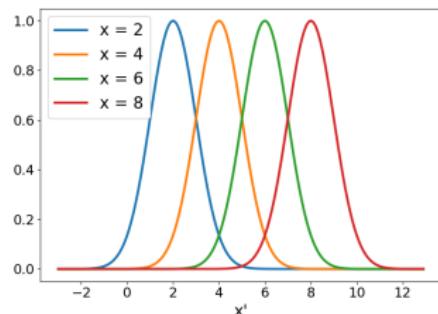
$$\kappa(x, x') = \kappa(x - x')$$

$\|x - x'\|$  natural

# Minimal separation and Fisher metric

Which metric for separation ?

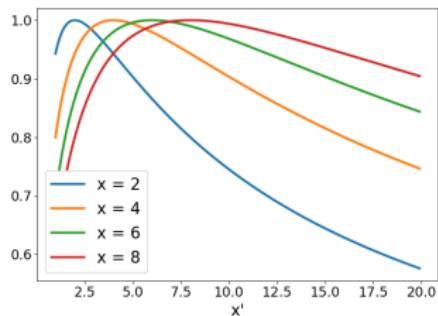
Classical case: translation-invariant kernel



$$\kappa(x, x') = \kappa(x - x')$$

$\|x - x'\|$  natural

Non translation-inv. ?

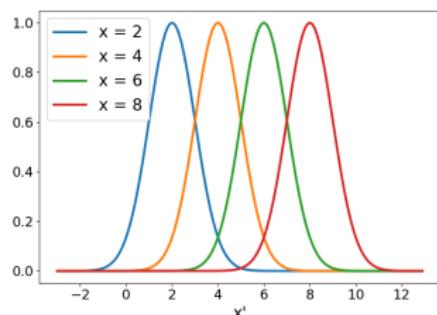


Kernel for microscopy

# Minimal separation and Fisher metric

Which metric for separation ?

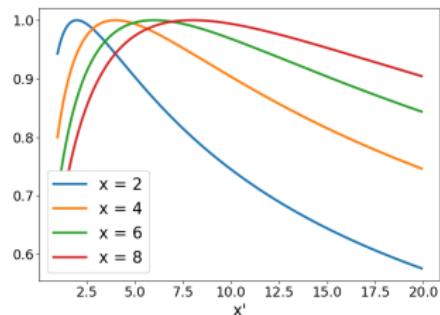
Classical case: translation-invariant kernel



$$\kappa(x, x') = \kappa(x - x')$$

$\|x - x'\|$  natural

Non translation-inv. ?



Kernel for microscopy

Riemannian metric associated to a kernel [Amari 99]:

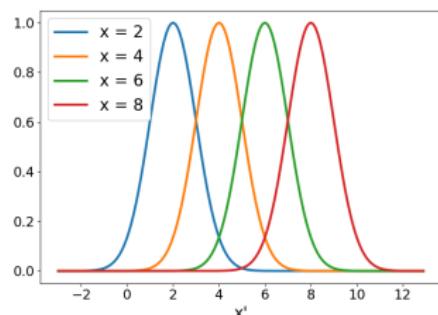
$$H_x = \nabla_1 \nabla_2 \kappa(x, x) : \text{metric tensor}$$

$$d_H(x, x') : \text{geodesic distance}$$

# Minimal separation and Fisher metric

Which metric for separation ?

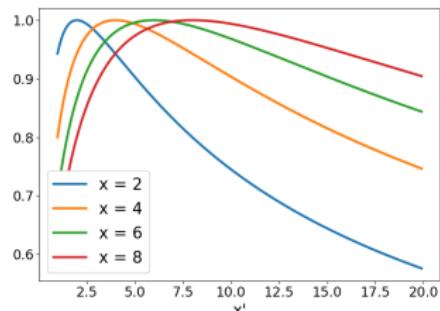
Classical case: translation-invariant kernel



$$\kappa(x, x') = \kappa(x - x')$$

$\|x - x'\|$  natural

Non translation-inv. ?



Kernel for microscopy

Riemannian metric associated to a kernel [Amari 99]:

$$H_x = \nabla_1 \nabla_2 \kappa(x, x) : \text{metric tensor}$$

$$d_H(x, x') : \text{geodesic distance}$$

Thm: under some hypothesis, for  $d_H(x_i, x_j) \geq \Delta$ , there exists non-degenerate  $\eta$

# Examples

Kernel

Features

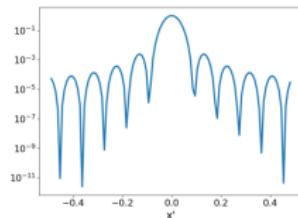
Fisher metric and minimal separation

# Examples

## Kernel

Discrete Fourier on Torus:

*Féjer kernel*



## Features

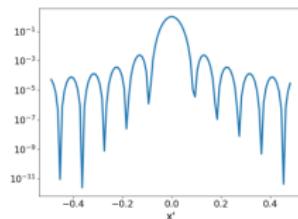
Fisher metric and minimal separation

# Examples

## Kernel

Discrete Fourier on Torus:

*Féjer kernel*



## Features

$$\varphi_\omega(x) = e^{2\pi i \omega^\top x}$$

$$\Lambda \propto \prod_{j=1}^d g_j(\omega_j)$$

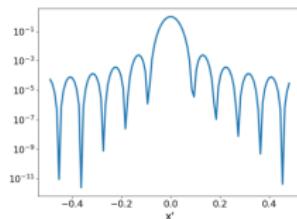
## Fisher metric and minimal separation

# Examples

## Kernel

Discrete Fourier on Torus:

*Féjer kernel*



## Features

$$\varphi_\omega(x) = e^{2\pi i \omega^\top x}$$

$$\Lambda \propto \prod_{j=1}^d g_j(\omega_j)$$

## Fisher metric and minimal separation

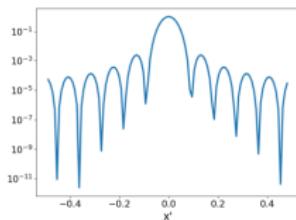
$$d_H(x, x') \propto \|x - x'\|_2$$

$$\Delta = \sqrt{d\sqrt{s}}/f_c$$

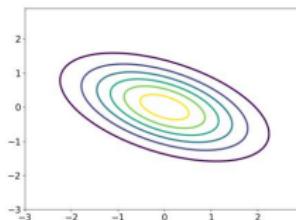
# Examples

## Kernel

Discrete Fourier on Torus:  
*Féjer kernel*



Continuous Gaussian Fourier:  
*Gaussian kernel*



## Features

$$\varphi_\omega(x) = e^{2\pi i \omega^\top x}$$

$$\Lambda \propto \prod_{j=1}^d g_j(\omega_j)$$

## Fisher metric and minimal separation

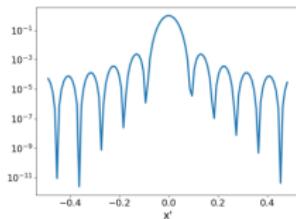
$$d_H(x, x') \propto \|x - x'\|_2$$

$$\Delta = \sqrt{d\sqrt{s}}/f_c$$

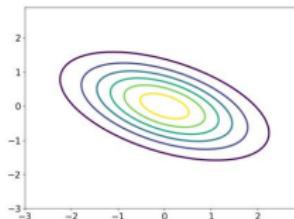
# Examples

## Kernel

Discrete Fourier on Torus:  
*Féjer kernel*



Continuous Gaussian Fourier:  
*Gaussian kernel*



## Features

$$\varphi_\omega(x) = e^{2\pi i \omega^\top x} \quad \varphi_\omega(x) = e^{i \omega^\top x}$$

$$\Lambda \propto \prod_{j=1}^d g_j(\omega_j) \quad \Lambda = \mathcal{N}(0, \Sigma^{-1})$$

## Fisher metric and minimal separation

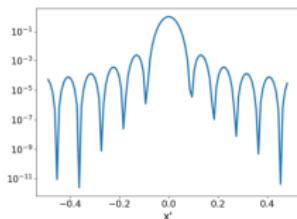
$$d_H(x, x') \propto \|x - x'\|_2$$

$$\Delta = \sqrt{d\sqrt{s}}/f_c$$

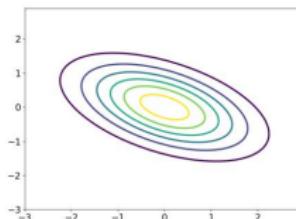
# Examples

## Kernel

Discrete Fourier on Torus:  
*Féjer kernel*



Continuous Gaussian Fourier:  
*Gaussian kernel*



## Features

$$\varphi_\omega(x) = e^{2\pi i \omega^\top x} \quad \varphi_\omega(x) = e^{i \omega^\top x}$$

$$\Lambda \propto \prod_{j=1}^d g_j(\omega_j) \quad \Lambda = \mathcal{N}(0, \Sigma^{-1})$$

## Fisher metric and minimal separation

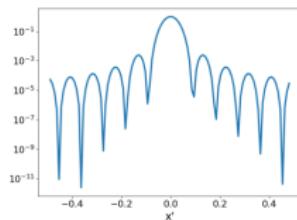
$$d_H(x, x') \propto \|x - x'\|_2 \quad d_H(x, x') = \|x - x'\|_{\Sigma^{-1}}$$

$$\Delta = \sqrt{d\sqrt{s}}/f_c \quad \Delta = \sqrt{\log(s)}$$

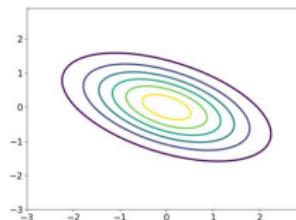
# Examples

## Kernel

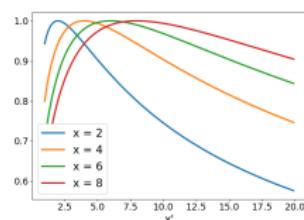
Discrete Fourier on Torus:  
*Féjer kernel*



Continuous Gaussian Fourier:  
*Gaussian kernel*



Microscopy (Laplace transform):



## Features

$$\varphi_\omega(x) = e^{2\pi i \omega^\top x} \quad \varphi_\omega(x) = e^{i \omega^\top x}$$

$$\Lambda \propto \prod_{j=1}^d g_j(\omega_j) \quad \Lambda = \mathcal{N}(0, \Sigma^{-1})$$

## Fisher metric and minimal separation

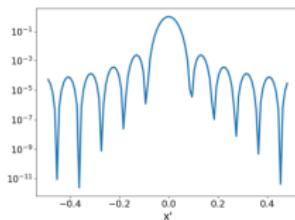
$$d_H(x, x') \propto \|x - x'\|_2 \quad d_H(x, x') = \|x - x'\|_{\Sigma^{-1}}$$

$$\Delta = \sqrt{d\sqrt{s}}/f_c \quad \Delta = \sqrt{\log(s)}$$

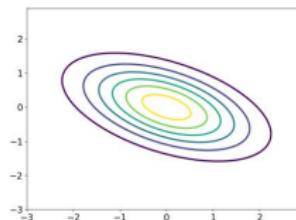
# Examples

## Kernel

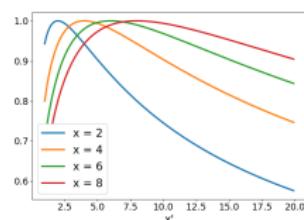
Discrete Fourier on Torus:  
*Féjer kernel*



Continuous Gaussian Fourier:  
*Gaussian kernel*



Microscopy (Laplace transform):



## Features

$$\varphi_\omega(x) = e^{2\pi i \omega^\top x}$$

$$\varphi_\omega(x) = e^{i \omega^\top x}$$

$$\varphi_\omega(x) = \prod_{j=1}^d \frac{x_i + \alpha_i}{\alpha_i} e^{-\omega^\top x}$$

$$\Lambda \propto \prod_{j=1}^d g_j(\omega_j)$$

$$\Lambda = \mathcal{N}(0, \Sigma^{-1})$$

$$\Lambda \propto e^{-\alpha^\top \omega}$$

## Fisher metric and minimal separation

$$d_H(x, x') \propto \|x - x'\|_2 \quad d_H(x, x') = \|x - x'\|_{\Sigma^{-1}}$$

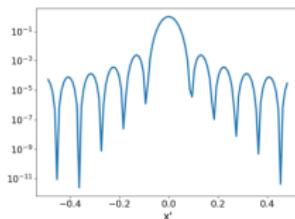
$$\Delta = \sqrt{d\sqrt{s}}/f_c$$

$$\Delta = \sqrt{\log(s)}$$

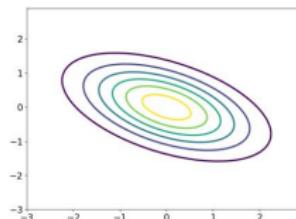
# Examples

## Kernel

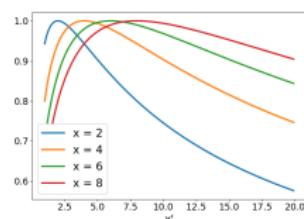
Discrete Fourier on Torus:  
*Féjer kernel*



Continuous Gaussian Fourier:  
*Gaussian kernel*



Microscopy (Laplace transform):



## Features

$$\varphi_\omega(x) = e^{2\pi i \omega^\top x}$$

$$\varphi_\omega(x) = e^{i \omega^\top x}$$

$$\varphi_\omega(x) = \prod_{j=1}^d \frac{x_i + \alpha_i}{\alpha_i} e^{-\omega^\top x}$$

$$\Lambda \propto \prod_{j=1}^d g_j(\omega_j)$$

$$\Lambda = \mathcal{N}(0, \Sigma^{-1})$$

$$\Lambda \propto e^{-\alpha^\top \omega}$$

## Fisher metric and minimal separation

$$d_H(x, x') \propto \|x - x'\|_2$$

$$d_H(x, x') = \|x - x'\|_{\Sigma^{-1}}$$

$$d_H(x, x') = \sqrt{\sum_j \left| \log \left( \frac{x_i + \alpha_i}{x'_i + \alpha_i} \right) \right|^2}$$

$$\Delta = \sqrt{d\sqrt{s}}/f_c$$

$$\Delta = \sqrt{\log(s)}$$

$$\Delta = d + \log(ds)$$

# Outline

- 1 Background on dual certificates
- 2 Minimal separation and Fisher metric
- 3 Main results, applications
- 4 Conclusion, outlooks

# Main results

**Thm:** Eliminating the random signs

# Main results

**Thm:** Eliminating the random signs

$$m \geq \mathcal{O}(\textcolor{red}{s} d^r \cdot \text{polylog}(s, d))$$

$\nwarrow$  Depend on kernel

# Main results

**Thm:** Eliminating the random signs

$$m \geq \mathcal{O}(\textcolor{red}{s} d^r \cdot \text{polylog}(s, d))$$

$\nwarrow$  Depend on kernel

- The recovered measure **concentrate** around true Diracs

# Main results

**Thm:** Eliminating the random signs

$$m \geq \mathcal{O}(\textcolor{red}{s} d^r \cdot \text{polylog}(s, d))$$

$\nwarrow$  Depend on kernel

- The recovered measure **concentrate** around true Diracs

- Proof: **golfing scheme**

[Gross 2009, Candès Plan 2011]



$$m = m_1 + m_2 + m_3 + \dots$$

# Main results

**Thm:** Eliminating the random signs

$$m \geq \mathcal{O}(\textcolor{red}{s} d^r \cdot \text{polylog}(s, d))$$

$\nwarrow$  Depend on kernel

- The recovered measure **concentrate** around true Diracs

- Proof: **golfing scheme**

[Gross 2009, Candès Plan 2011]



$$m = m_1 + m_2 + m_3 + \dots$$

**Thm:** Support stability

# Main results

**Thm:** Eliminating the random signs

$$m \geq \mathcal{O}(\textcolor{red}{s} d^r \cdot \text{polylog}(s, d))$$

$\nwarrow$  Depend on kernel

- The recovered measure **concentrate** around true Diracs

- Proof: **golfing scheme**

[Gross 2009, Candès Plan 2011]



$$m = m_1 + m_2 + m_3 + \dots$$

**Thm:** Support stability

$$m \geq \mathcal{O}(\textcolor{red}{s} d^r \cdot \text{polylog}(s, d))$$

*With random signs*

$$m \geq \mathcal{O}(\textcolor{red}{s}^{\frac{3}{2}} d^r \cdot \text{polylog}(s, d))$$

*Without random signs*

# Main results

**Thm:** Eliminating the random signs

$$m \geq \mathcal{O}(\textcolor{red}{s} d^r \cdot \text{polylog}(s, d))$$

Depend on kernel

- The recovered measure **concentrate** around true Diracs

- Proof: **golfing scheme**

[Gross 2009, Candès Plan 2011]



$$m = m_1 + m_2 + m_3 + \dots$$

**Thm:** Support stability

$$m \geq \mathcal{O}(\textcolor{red}{s} d^r \cdot \text{polylog}(s, d))$$

*With random signs*

$$m \geq \mathcal{O}(\textcolor{red}{s}^{\frac{3}{2}} d^r \cdot \text{polylog}(s, d))$$

*Without random signs*

- Quantified small noise** : if  $\lambda, \|w\| \leq \frac{\min_i |a_i|}{\textcolor{red}{s} d^q}$ , then:

# Main results

**Thm:** Eliminating the random signs

$$m \geq \mathcal{O}(\textcolor{red}{s} d^r \cdot \text{polylog}(s, d))$$

↑ Depend on kernel

- The recovered measure **concentrate** around true Diracs

- Proof: **golfing scheme**

[Gross 2009, Candès Plan 2011]



$$m = m_1 + m_2 + m_3 + \dots$$

**Thm:** Support stability

$$m \geq \mathcal{O}(\textcolor{red}{s} d^r \cdot \text{polylog}(s, d))$$

With random signs

$$m \geq \mathcal{O}(\textcolor{red}{s}^{\frac{3}{2}} d^r \cdot \text{polylog}(s, d))$$

Without random signs

- Quantified small noise** : if  $\lambda, \|w\| \leq \frac{\min_i |a_i|}{\textcolor{red}{s} d^q}$ , then:
- The recovered measure is formed of exactly  $\mathcal{S}$  Diracs

$$\sqrt{\sum_i |\tilde{a}_i - a_i|^2 + d_H(\tilde{x}_i, x_i)^2} \lesssim \frac{\sqrt{s}}{\min_i |a_i|} (\|w\| + \lambda)$$

# Applications

- Féjer kernel (discrete Fourier):  $m \geq sd^3, \|w\| \leq 1/(sd^3)$

# Applications

- Féjer kernel (discrete Fourier):  $m \geq sd^3, \|w\| \leq 1/(sd^3)$
- Microscopy with Laplace transform:  $m \geq sd^7, \|w\| \leq 1/(sd^5)$

# Outline

Introduction

- What is Compressive Sensing?

- Notations (Reminder)

- Problem formulation

Recovery guarantees

- NSP

- Dual certificate

- Coherence

- RIP

Recovering with random matrices?

Concentration inequalities and proving the RIP

Beyond Sparsity

- Total Variation

- Structured sparsity

- Matrix completion and Low-rank regularization

**Infinite-dimensional signals: superresolution and compressive learning**

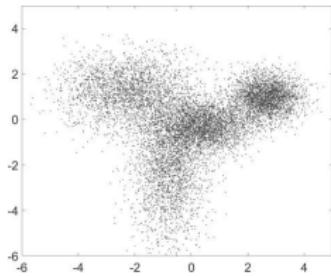
- Continuous sparsity: superresolution

- Generalized sparsity: sketching

# Generalized RIP?

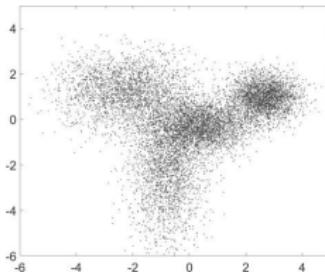
- We have seen dual certificates in continuous space
- Can we have guarantees for more general “low-dimensional” models?  
In infinite-dimension?
- Here: example of sketched learning

# Compressive learning

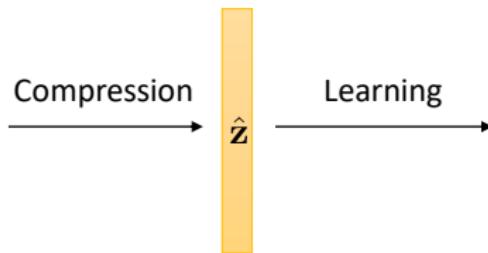


$x_1, \dots, x_n \in \mathbb{R}^d$

# Compressive learning



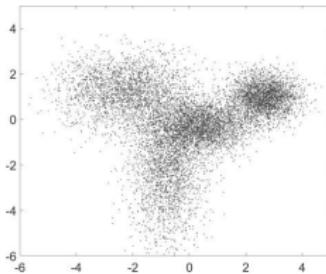
$x_1, \dots, x_n \in \mathbb{R}^d$



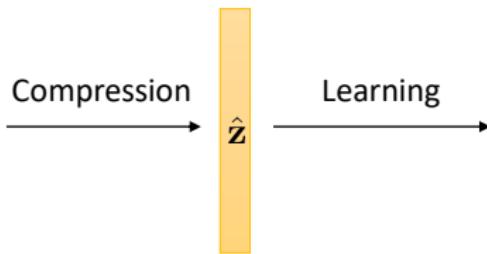
**Linear sketch**  $\mathbf{z} \in \mathbb{R}^m$

- **Sketched learning:** First **compress** data in a **linear sketch** [Cormode 2011], then learn

# Compressive learning



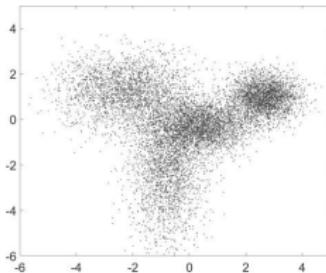
$x_1, \dots, x_n \in \mathbb{R}^d$



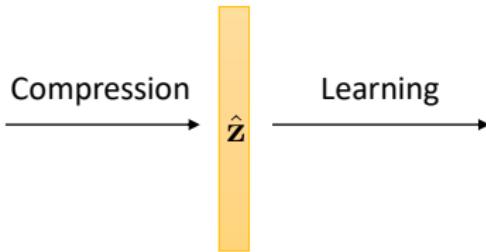
**Linear sketch**  $z \in \mathbb{R}^m$

- **Sketched learning:** First **compress** data in a **linear sketch** [Cormode 2011], then learn
  - Hash tables, count sketches, histograms...

# Compressive learning



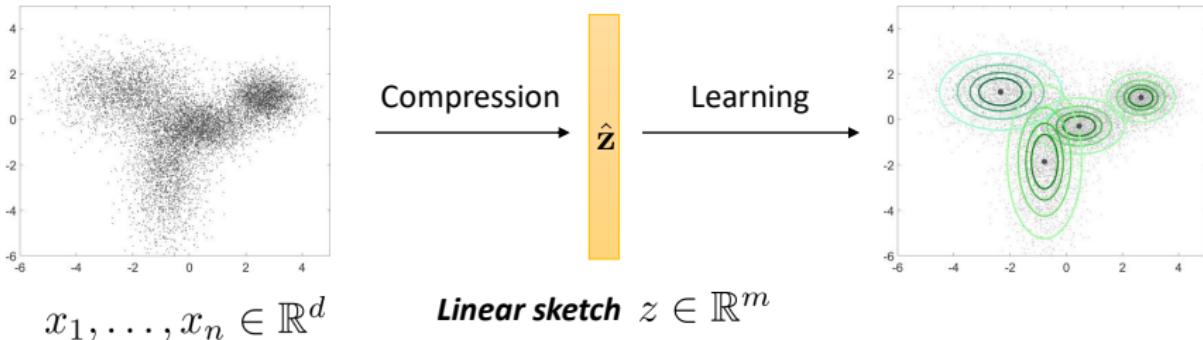
$x_1, \dots, x_n \in \mathbb{R}^d$



**Linear sketch**  $z \in \mathbb{R}^m$

- **Sketched learning:** First **compress** data in a **linear sketch** [Cormode 2011], then learn
  - Hash tables, count sketches, histograms...
- **Advantages:** **one-pass**, streaming, **distributed** compression, **data privacy**...

# Compressive learning



- **Sketched learning:** First **compress** data in a **linear sketch** [Cormode 2011], then learn
  - Hash tables, count sketches, histograms...
- **Advantages:** **one-pass**, streaming, **distributed** compression, **data privacy**...
- **In this talk:** unsupervised learning

# How-to: build a sketch

## What is a sketch ?

Any *linear* sketch = **empirical moments**

$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X) = \frac{1}{n} \sum_i \Phi(x_i)$$

# How-to: build a sketch

## What is a sketch ?

Any *linear* sketch = empirical moments

$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X) = \frac{1}{n} \sum_i \Phi(x_i)$$

## What is contained in a sketch ?

# How-to: build a sketch

## What is a sketch ?

Any *linear* sketch = empirical moments

$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X) = \frac{1}{n} \sum_i \Phi(x_i)$$

## What is contained in a sketch ?

- $\Phi(x) = x$  : mean

# How-to: build a sketch

## What is a sketch ?

Any *linear* sketch = empirical moments

$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X) = \frac{1}{n} \sum_i \Phi(x_i)$$

## What is contained in a sketch ?

- $\Phi(x) = x$  : mean
- $\Phi(x) = x^k$  :  $k^{\text{th}}$  moment

# How-to: build a sketch

## What is a sketch ?

Any *linear* sketch = **empirical moments**

$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X) = \frac{1}{n} \sum_i \Phi(x_i)$$

## What is contained in a sketch ?

- $\Phi(x) = x$  : mean
- $\Phi(x) = x^k$  :  $k^{\text{th}}$  moment
- $\Phi(x) = [1_{x \in B_i}]_{i=1}^m$  : histogram

# How-to: build a sketch

## What is a sketch ?

Any *linear* sketch = empirical moments

$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X) = \frac{1}{n} \sum_i \Phi(x_i)$$

## What is contained in a sketch ?

- $\Phi(x) = x$  : mean
- $\Phi(x) = x^k$  :  $k^{\text{th}}$  moment
- $\Phi(x) = [1_{x \in B_i}]_{i=1}^m$  : histogram
- Proposed: **kernel random features**  
*[Rahimi 2007]*  
(random proj. + non-linearity)

# How-to: build a sketch

## What is a sketch ?

Any *linear* sketch = empirical moments

$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X) = \frac{1}{n} \sum_i \Phi(x_i)$$

## What is contained in a sketch ?

- $\Phi(x) = x$  : mean
- $\Phi(x) = x^k$  :  $k^{\text{th}}$  moment
- $\Phi(x) = [1_{x \in B_i}]_{i=1}^m$  : histogram
- Proposed: **kernel random features**  
*[Rahimi 2007]*  
(random proj. + non-linearity)

## Questions:

- What information is preserved by the sketching ?
- How to retrieve this information ?
- What is a sufficient number of features ?

# How-to: build a sketch

## What is a sketch ?

Any **linear** sketch = **empirical moments**

$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X) = \frac{1}{n} \sum_i \Phi(x_i)$$

## What is contained in a sketch ?

- $\Phi(x) = x$  : mean
- $\Phi(x) = x^k$  :  $k^{\text{th}}$  moment
- $\Phi(x) = [1_{x \in B_i}]_{i=1}^m$  : histogram
- Proposed: **kernel random features**  
*[Rahimi 2007]*  
(random proj. + non-linearity)

## Questions:

- What information is preserved by the sketching ?
- How to retrieve this information ?
- What is a sufficient number of features ?

Intuition: sketching as a **linear embedding**

# How-to: build a sketch

## What is a sketch ?

Any **linear** sketch = **empirical moments**

$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X) = \frac{1}{n} \sum_i \Phi(x_i)$$

## What is contained in a sketch ?

- $\Phi(x) = x$  : mean
- $\Phi(x) = x^k$  :  $k^{\text{th}}$  moment
- $\Phi(x) = [1_{x \in B_i}]_{i=1}^m$  : histogram
- Proposed: **kernel random features**  
*[Rahimi 2007]*  
(random proj. + non-linearity)

## Questions:

- What information is preserved by the sketching ?
- How to retrieve this information ?
- What is a sufficient number of features ?

## Intuition: sketching as a **linear embedding**

- Assumption:  $x_1, \dots, x_n \stackrel{i.i.d.}{\sim} \pi^*$

# How-to: build a sketch

## What is a sketch ?

Any **linear** sketch = **empirical moments**

$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X) = \frac{1}{n} \sum_i \Phi(x_i)$$

## What is contained in a sketch ?

- $\Phi(x) = x$  : mean
- $\Phi(x) = x^k$  :  $k^{\text{th}}$  moment
- $\Phi(x) = [1_{x \in B_i}]_{i=1}^m$  : histogram
- Proposed: **kernel random features**  
*[Rahimi 2007]*  
(random proj. + non-linearity)

## Questions:

- What information is preserved by the sketching ?
- How to retrieve this information ?
- What is a sufficient number of features ?

## Intuition: sketching as a **linear embedding**

- Assumption:  $x_1, \dots, x_n \stackrel{i.i.d.}{\sim} \pi^\star$
- Linear operator:  $\mathcal{A}\pi = \mathbb{E}_{X \sim \pi} \Phi(X)$

# How-to: build a sketch

## What is a sketch ?

Any **linear** sketch = **empirical moments**

$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X) = \frac{1}{n} \sum_i \Phi(x_i)$$

## What is contained in a sketch ?

- $\Phi(x) = x$  : mean
- $\Phi(x) = x^k$  :  $k^{\text{th}}$  moment
- $\Phi(x) = [1_{x \in B_i}]_{i=1}^m$  : histogram
- Proposed: **kernel random features**  
*[Rahimi 2007]*  
(random proj. + non-linearity)

## Questions:

- What information is preserved by the sketching ?
- How to retrieve this information ?
- What is a sufficient number of features ?

## Intuition: sketching as a **linear embedding**

- Assumption:  $x_1, \dots, x_n \stackrel{i.i.d.}{\sim} \pi^*$
- Linear operator:  $\mathcal{A}\pi = \mathbb{E}_{X \sim \pi} \Phi(X)$
- « Noisy » linear measurement:

$$\hat{\mathbf{z}} = \mathcal{A}\pi^* + \hat{\mathbf{e}}$$

Noise  $\hat{\mathbf{e}} = \hat{\mathbb{E}}\Phi(X) - \mathbb{E}_{\pi^*} \Phi(X)$  **small**

# How-to: build a sketch

## What is a sketch ?

Any **linear** sketch = **empirical moments**

$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X) = \frac{1}{n} \sum_i \Phi(x_i)$$

## What is contained in a sketch ?

- $\Phi(x) = x$  : mean
- $\Phi(x) = x^k$  :  $k^{\text{th}}$  moment
- $\Phi(x) = [1_{x \in B_i}]_{i=1}^m$  : histogram
- Proposed: **kernel random features**  
*[Rahimi 2007]*  
(random proj. + non-linearity)

## Questions:

- What information is preserved by the sketching ?
- How to retrieve this information ?
- What is a sufficient number of features ?

## Intuition: sketching as a **linear embedding**

- Assumption:  $x_1, \dots, x_n \stackrel{i.i.d.}{\sim} \pi^*$
- Linear operator:  $\mathcal{A}\pi = \mathbb{E}_{X \sim \pi} \Phi(X)$
- « Noisy » linear measurement:

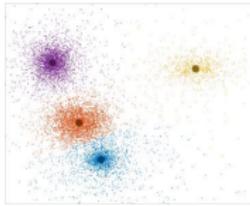
$$\hat{\mathbf{z}} = \mathcal{A}\pi^* + \hat{\mathbf{e}}$$

Noise  $\hat{\mathbf{e}} = \hat{\mathbb{E}}\Phi(X) - \mathbb{E}_{\pi^*} \Phi(X)$  **small**

Dimensionality-reducing, random, linear embedding: Compressive Sensing?

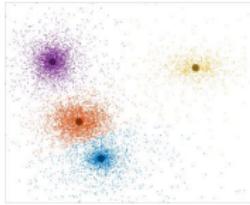
# Example of applications [Keriven 2016,2017]

**Retrieving mixture of Diracs  
from a sketch= k-means**

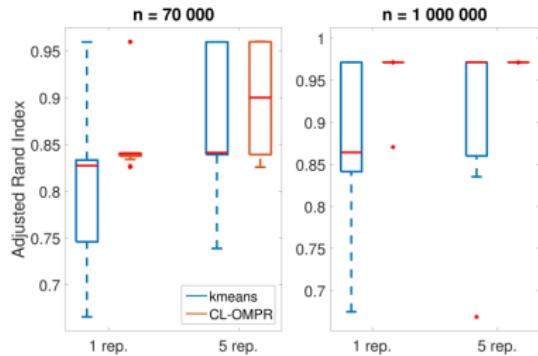


# Example of applications [Keriven 2016,2017]

Retrieving mixture of Diracs  
from a sketch= k-means

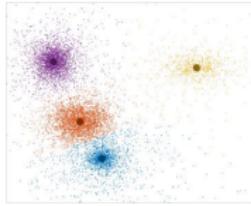


**Application:**  
*Spectral clustering*  
for MNIST  
classification



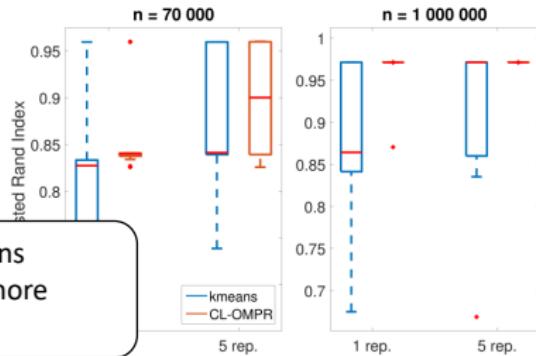
# Example of applications [Keriven 2016,2017]

Retrieving mixture of Diracs from a sketch= k-means



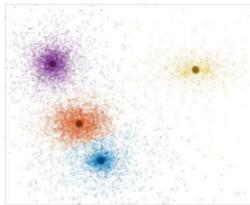
## Application: Spectral clustering for MNIST classification

- Twice faster than k-means
- 4 orders of magnitude more memory efficient



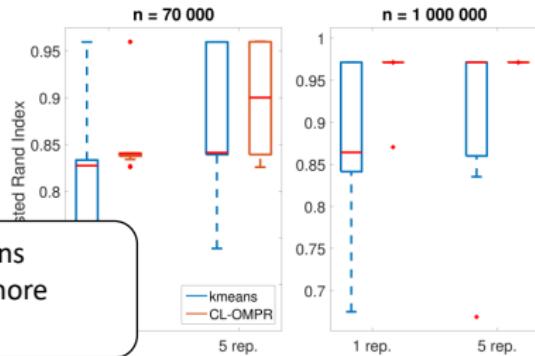
# Example of applications [Keriven 2016,2017]

Retrieving mixture of Diracs from a sketch= k-means

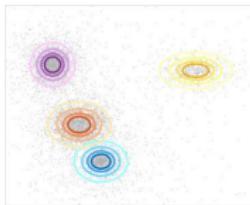


**Application:**  
*Spectral clustering*  
for MNIST classification

- Twice faster than k-means
- 4 orders of magnitude more memory efficient



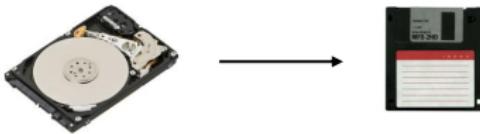
Retrieving GMMs from a sketch



**Application:** *speaker verification* [Reynolds 2000]

Error:

- EM on 300 000 samples : **29.53**
- **20kB** sketch computed on **50GB** database: **28.96**



## Q: Theoretical guarantees ?

- Inspired by Compressive Sensing:
  - 1: with the Restricted Isometry Property (RIP)
  - 2: with dual certificates

# Outline

1

## Information-preservation guarantees: a RIP analysis

Joint work with R. Gribonval, G. Blanchard, Y. Traonmilin

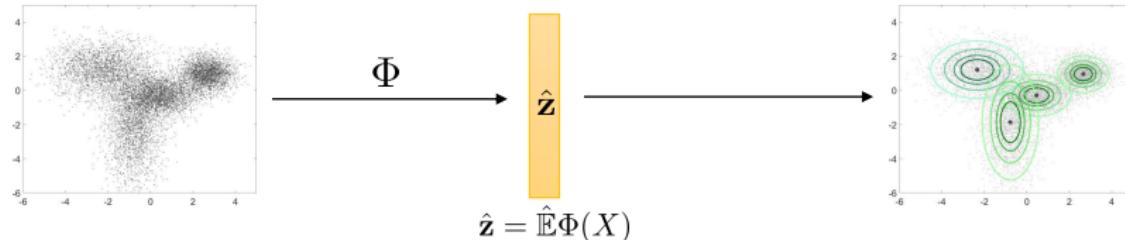
2

## Total variation regularization: a dual certificate analysis

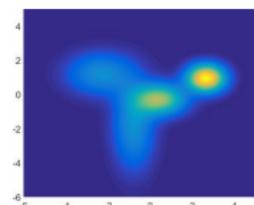
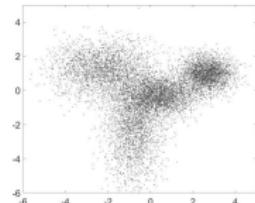
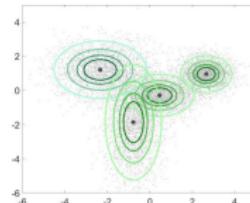
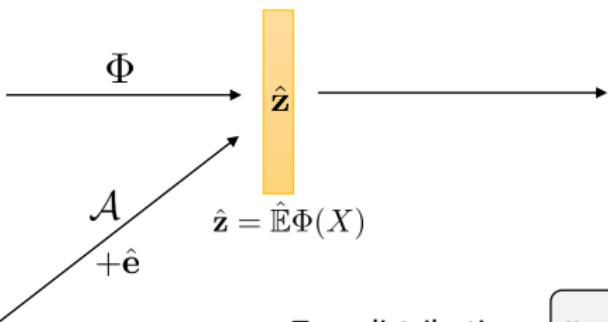
3

## Conclusion, outlooks

# Recall: Linear inverse problem



# Recall: Linear inverse problem

 $\pi^*$ 

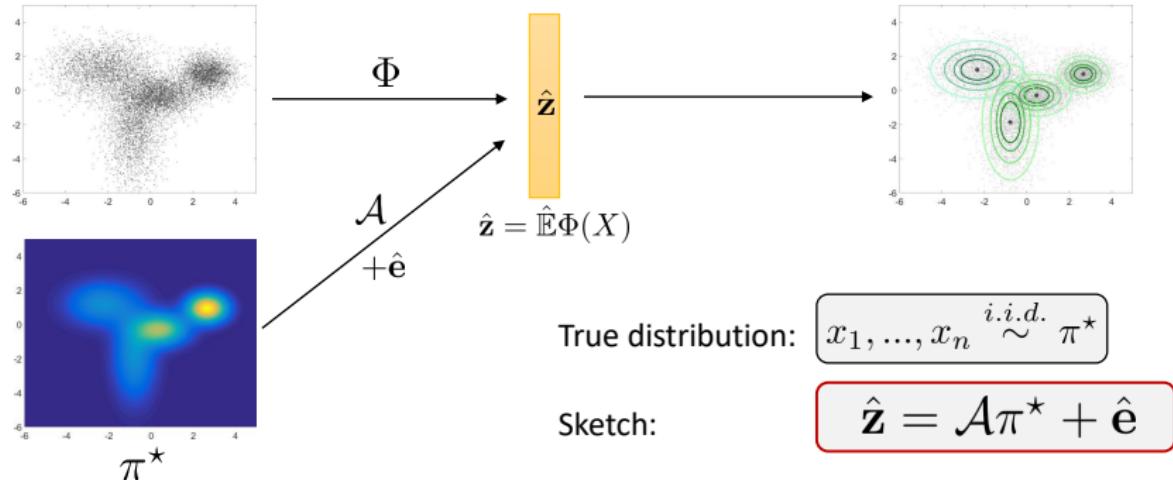
True distribution:

$$x_1, \dots, x_n \stackrel{i.i.d.}{\sim} \pi^*$$

Sketch:

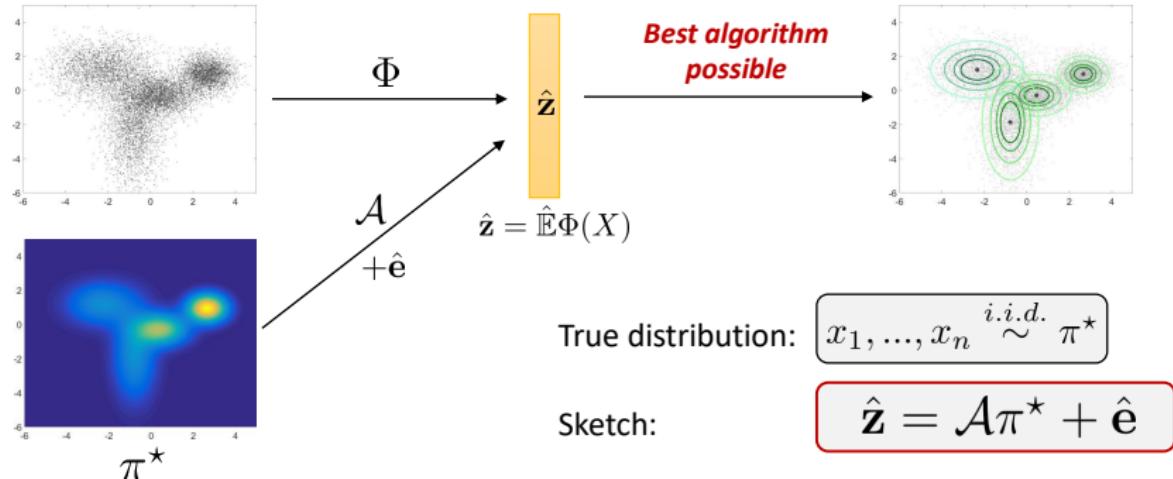
$$\hat{\mathbf{z}} = \mathcal{A}\pi^* + \hat{\mathbf{e}}$$

# Recall: Linear inverse problem



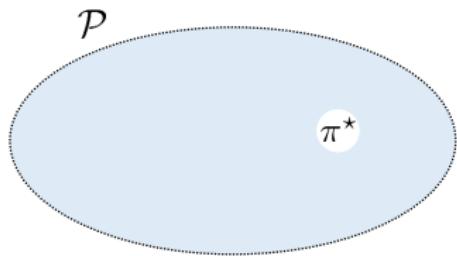
- Estimation problem = **linear inverse problem** on measures
- **Extremely ill-posed !**

# Recall: Linear inverse problem

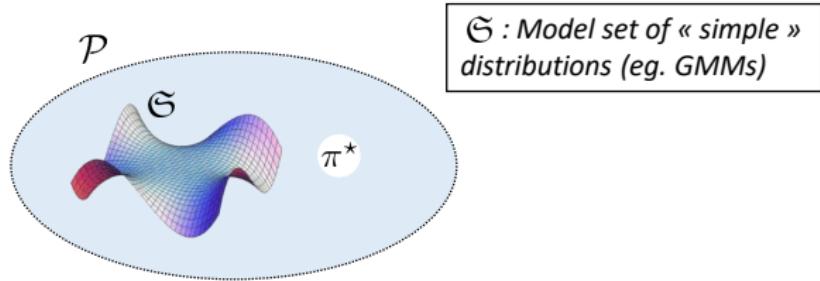


- Estimation problem = **linear inverse problem** on measures
- **Extremely ill-posed !**
- **Feasibility?** (*information-preservation*)

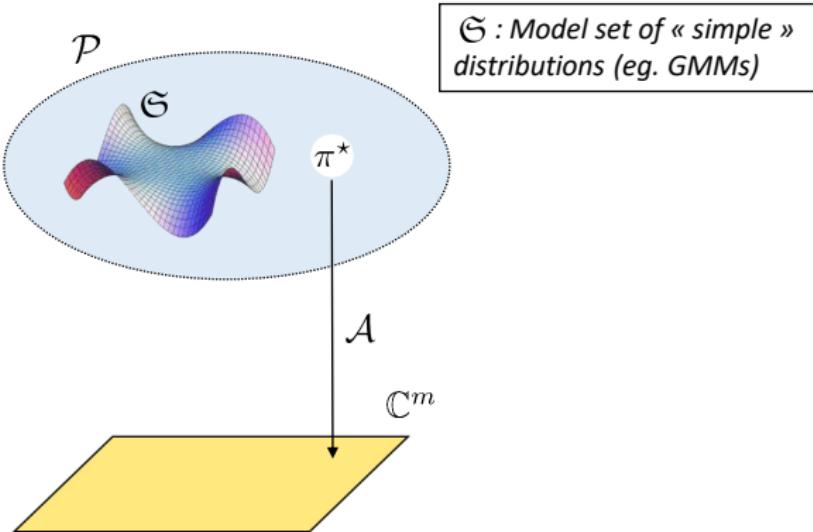
# Information preservation guarantees



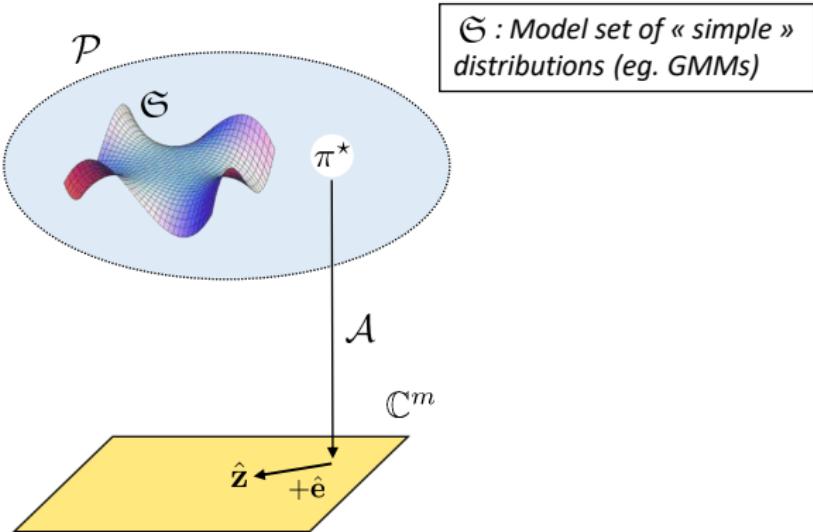
# Information preservation guarantees



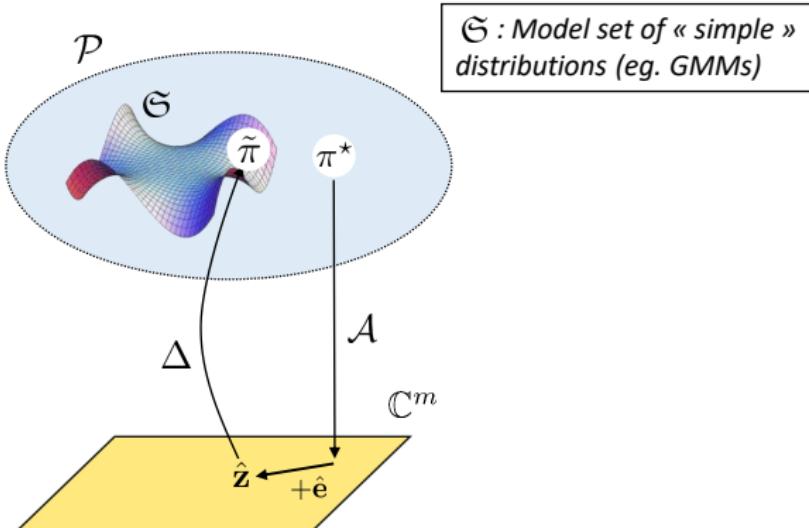
# Information preservation guarantees



# Information preservation guarantees



# Information preservation guarantees

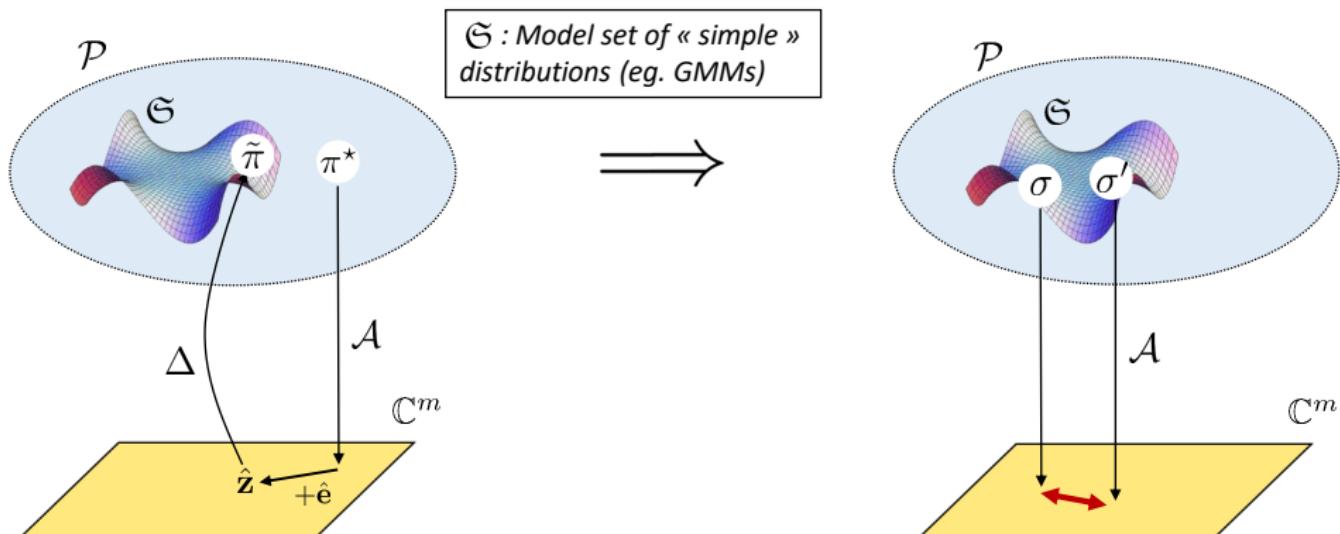


## Goal

Prove the existence of a *decoder*  $\Delta$  robust to **noise** and stable to **modeling error**.

« Instance-optimal » decoder

# Information preservation guarantees



## Goal

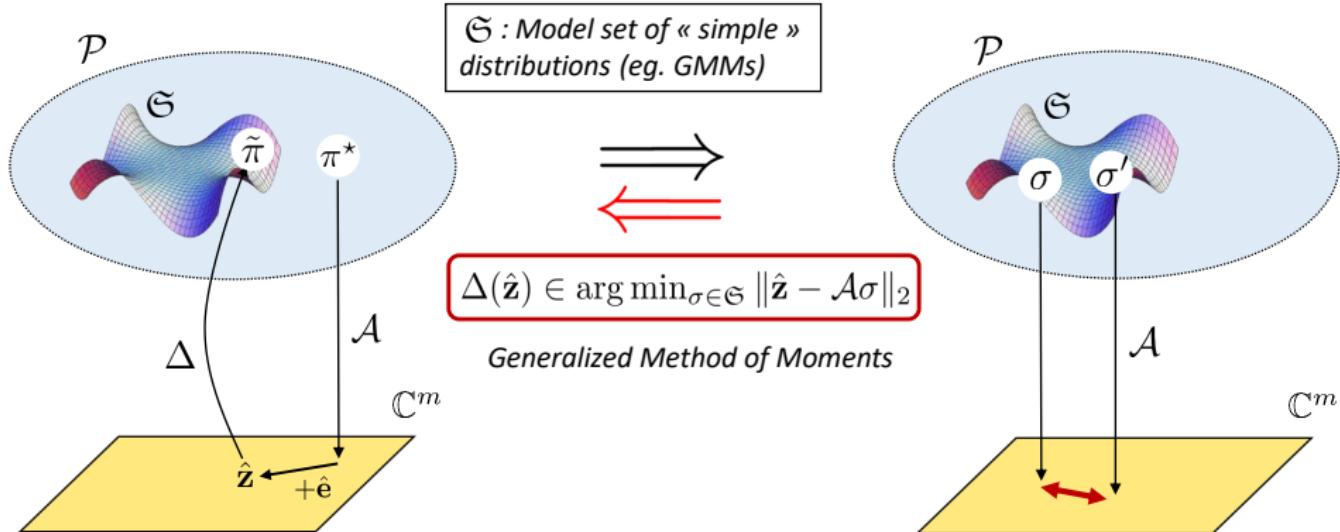
Prove the existence of a *decoder*  $\Delta$  robust to **noise** and stable to **modeling error**.

« Instance-optimal » decoder

## Lower Restricted Isometry Property

$$\|\sigma - \sigma'\| \lesssim \|A\sigma - A\sigma'\|_2$$

# Information preservation guarantees



## Goal

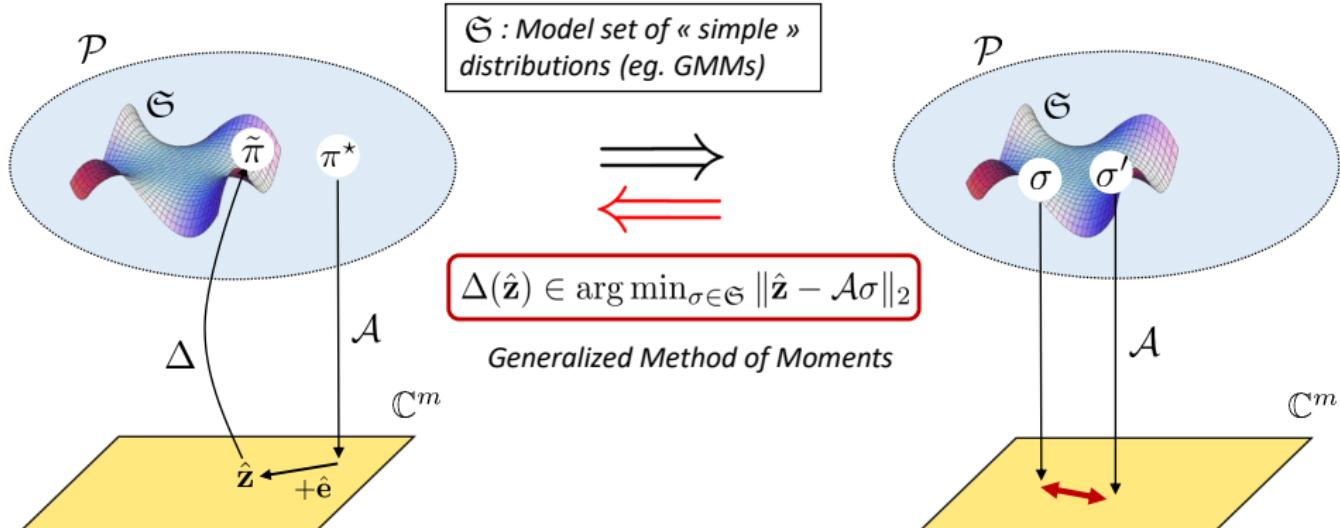
Prove the existence of a *decoder*  $\Delta$  robust to **noise** and stable to **modeling error**.

« Instance-optimal » decoder

## Lower Restricted Isometry Property

$$\|\sigma - \sigma'\| \lesssim \|A\sigma - A\sigma'\|_2$$

# Information preservation guarantees



## Goal

Prove the existence of a *decoder*  $\Delta$  robust to **noise** and stable to **modeling error**.

« Instance-optimal » decoder

## Lower Restricted Isometry Property

$$\|\sigma - \sigma'\| \lesssim \|\mathcal{A}\sigma - \mathcal{A}\sigma'\|_2$$

New goal: find/construct models  $\mathcal{S}$  and operators  $\mathcal{A}$  that satisfy the LRIP (w.h.p.)

# Proving the LRIP

**Goal: LRIP** w.h.p. on  $\mathcal{A}$ ,  $\forall \sigma, \sigma' \in \mathfrak{S}$ ,  $\|\sigma - \sigma'\| \lesssim \|\mathcal{A}\sigma - \mathcal{A}\sigma'\|_2$ .

# Proving the LRIP

**Goal: LRIP** w.h.p. on  $\mathcal{A}$ ,  $\forall \sigma, \sigma' \in \mathfrak{S}$ ,  $\|\sigma - \sigma'\| \lesssim \|\mathcal{A}\sigma - \mathcal{A}\sigma'\|_2$ .

## 1 Pointwise LRIP

### Construction of $\mathcal{A}$ :

- Kernel mean [Gretton 2006, Borgwardt 2006]
- Random features [Rahimi 2007]

$\forall \sigma, \sigma'$ , w.h.p. on  $\mathcal{A}$ , LRIP.

# Proving the LRIP

**Goal: LRIP** w.h.p. on  $\mathcal{A}$ ,  $\forall \sigma, \sigma' \in \mathfrak{S}$ ,  $\|\sigma - \sigma'\| \lesssim \|\mathcal{A}\sigma - \mathcal{A}\sigma'\|_2$ .

## 1 Pointwise LRIP

**Construction of  $\mathcal{A}$  :**

- Kernel mean [Gretton 2006, Borgwardt 2006]
- Random features [Rahimi 2007]

$\forall \sigma, \sigma'$ , w.h.p. on  $\mathcal{A}$ , LRIP.

## 2 Extension to LRIP

**Covering numbers** (compacity) of the normalized secant set  $\mathcal{S}(\mathfrak{S})$

# Proving the LRIP

**Goal: LRIP** w.h.p. on  $\mathcal{A}$ ,  $\forall \sigma, \sigma' \in \mathfrak{S}$ ,  $\|\sigma - \sigma'\| \lesssim \|\mathcal{A}\sigma - \mathcal{A}\sigma'\|_2$ .

## 1 Pointwise LRIP

**Construction of  $\mathcal{A}$  :**

- Kernel mean [Gretton 2006, Borgwardt 2006]
- Random features [Rahimi 2007]

$\forall \sigma, \sigma'$ , w.h.p. on  $\mathcal{A}$ , LRIP.

## 2 Extension to LRIP

**Covering numbers** (compacity) of the normalized secant set  $\mathcal{S}(\mathfrak{S})$

*Subset of a unit ball (infinite dimension)  
that only depends on  $\mathfrak{S}$*

# Proving the LRIP

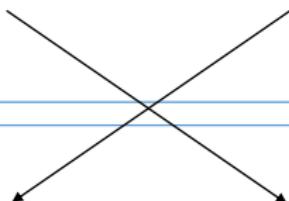
**Goal: LRIP** w.h.p. on  $\mathcal{A}$ ,  $\forall \sigma, \sigma' \in \mathfrak{S}$ ,  $\|\sigma - \sigma'\| \lesssim \|\mathcal{A}\sigma - \mathcal{A}\sigma'\|_2$ .

## 1 Pointwise LRIP

**Construction of  $\mathcal{A}$  :**

- Kernel mean [Gretton 2006, Borgwardt 2006]
- Random features [Rahimi 2007]

$\forall \sigma, \sigma'$ , w.h.p. on  $\mathcal{A}$ , LRIP.



w.h.p. on  $\mathcal{A}$ ,  $\forall \sigma, \sigma'$ , LRIP.

## 2 Extension to LRIP

**Covering numbers** (compactness) of the normalized secant set  $\mathcal{S}(\mathfrak{S})$

*Subset of a unit ball (infinite dimension)  
that only depends on  $\mathfrak{S}$*

## Main hypothesis

The *normalized secant set*  $\mathcal{S}(\mathfrak{S})$  has finite covering numbers.

# Main result [Keriven 2016]

## Main hypothesis

The *normalized secant set*  $\mathcal{S}(\mathfrak{S})$  has finite covering numbers.

## Result

For

$$m \geq C \times \log(\text{cov. num.}) ,$$

# Main result [Keriven 2016]

## Main hypothesis

The *normalized secant set*  $\mathcal{S}(\mathfrak{S})$  has finite covering numbers.

## Result

For  $m \geq C \times \log(\text{cov. num.})$ ,

Pointwise concentration

Dimensionality of the model

# Main result [Keriven 2016]

## Main hypothesis

The *normalized secant set*  $\mathcal{S}(\mathfrak{S})$  has finite covering numbers.

## Result

$$\text{For } m \geq C \times \log(\text{cov. num.}),$$

Pointwise concentration

Dimensionality of the model

W.h.p.

$$\|\pi^* - \Delta(\hat{\mathbf{z}})\| \lesssim d(\pi^*, \mathfrak{S}) + \mathcal{O}(1/\sqrt{n})$$

# Main result [Keriven 2016]

## Main hypothesis

The *normalized secant set*  $\mathcal{S}(\mathfrak{S})$  has finite covering numbers.

## Result

$$\text{For } m \geq C \times \log(\text{cov. num.}),$$

Pointwise concentration

Dimensionality of the model

W.h.p.

*Modeling error*

*Empirical noise*

$$\|\pi^* - \Delta(\hat{\mathbf{z}})\| \lesssim d(\pi^*, \mathfrak{S}) + \mathcal{O}(1/\sqrt{n})$$

# Main result [Keriven 2016]

## Main hypothesis

The *normalized secant set*  $\mathcal{S}(\mathfrak{S})$  has finite covering numbers.

### Result

$$\text{For } m \geq C \times \log(\text{cov. num.}),$$

Pointwise concentration

Dimensionality of the model

W.h.p.

Modeling error

Empirical noise

$$\|\pi^* - \Delta(\hat{\mathbf{z}})\| \lesssim d(\pi^*, \mathfrak{S}) + \mathcal{O}(1/\sqrt{n})$$

Does not depend on  $m$  !

# Main result [Keriven 2016]

## Main hypothesis

The *normalized secant set*  $\mathcal{S}(\mathfrak{S})$  has finite covering numbers.

### Result

For

$$m \geq C \times \log(\text{cov. num.}),$$

Pointwise concentration

Dimensionality of the model

W.h.p.

Modeling error

Empirical noise

$$\|\pi^* - \Delta(\hat{\mathbf{z}})\| \lesssim d(\pi^*, \mathfrak{S}) + \mathcal{O}(1/\sqrt{n})$$

Does not depend on  $m$  !

- **Classic Compressive Sensing:** finite dimension: **Known**
- **Here:** infinite dimension: **Technical**

# Application

## k-means with mixtures of Diracs

# Application

## k-means with mixtures of Diracs

### Hypotheses

- $\varepsilon$  - separated centroids
- $M$  - bounded domain for centroids

# Application

## k-means with mixtures of Diracs

### Hypotheses

*(no assumption  
on the data)*

- $\varepsilon$  - separated centroids
- $M$  - bounded domain for centroids

# Application

## k-means with mixtures of Diracs

### Hypotheses

*(no assumption  
on the data)*

- $\varepsilon$  - separated centroids
- $M$ - bounded domain for centroids

### Sketch

- Adjusted Random Fourier features (*for technical reasons*)

# Application

## k-means with mixtures of Diracs

### Hypotheses

*(no assumption  
on the data)*

- $\varepsilon$  - separated centroids
- $M$  - bounded domain for centroids

### Sketch

- Adjusted Random Fourier features (*for technical reasons*)

### Result

- W.r.t. k-means usual cost (SSE)

# Application

## k-means with mixtures of Diracs

### Hypotheses

*(no assumption  
on the data)*

- $\varepsilon$  - separated centroids
- $M$  - bounded domain for centroids

### Sketch

- Adjusted Random Fourier features (*for technical reasons*)

### Result

- W.r.t. k-means usual cost (SSE)

### Sketch size

$$m \geq \mathcal{O} (k^2 d \cdot \text{polylog}(k, d) \log(M/\varepsilon))$$

# Application

## k-means with mixtures of Diracs

### Hypotheses

*(no assumption  
on the data)*

- $\varepsilon$  - separated centroids
- $M$  - bounded domain for centroids

### Sketch

- Adjusted Random Fourier features (*for technical reasons*)

### Result

- W.r.t. k-means usual cost (SSE)

### Sketch size

$$m \geq \mathcal{O} (k^2 d \cdot \text{polylog}(k, d) \log(M/\varepsilon))$$

## GMM with known covariance

# Application

## k-means with mixtures of Diracs

### Hypotheses

(no assumption  
on the data)

- $\varepsilon$  - separated centroids
- $M$  - bounded domain for centroids

### Sketch

- Adjusted Random Fourier features (for technical reasons)

### Result

- W.r.t. k-means usual cost (SSE)

### Sketch size

$$m \geq \mathcal{O} (k^2 d \cdot \text{polylog}(k, d) \log(M/\varepsilon))$$

## GMM with known covariance

### Hypotheses

- Sufficiently separated means
- Bounded domain for means

# Application

## k-means with mixtures of Diracs

### Hypotheses

(no assumption  
on the data)

- $\varepsilon$  - separated centroids
- $M$  - bounded domain for centroids

### Sketch

- Adjusted Random Fourier features (for technical reasons)

### Result

- W.r.t. k-means usual cost (SSE)

### Sketch size

$$m \geq \mathcal{O} (k^2 d \cdot \text{polylog}(k, d) \log(M/\varepsilon))$$

## GMM with known covariance

### Hypotheses

- Sufficiently separated means
- Bounded domain for means

### Sketch

- Fourier features

# Application

## k-means with mixtures of Diracs

### Hypotheses

(no assumption  
on the data)

- $\varepsilon$  - separated centroids
- $M$ - bounded domain for centroids

### Sketch

- Adjusted Random Fourier features (for technical reasons)

### Result

- W.r.t. k-means usual cost (SSE)

### Sketch size

$$m \geq \mathcal{O} (k^2 d \cdot \text{polylog}(k, d) \log(M/\varepsilon))$$

## GMM with known covariance

### Hypotheses

- Sufficiently separated means
- Bounded domain for means

### Sketch

- Fourier features

### Result

- With respect to log-likelihood

# Application

## k-means with mixtures of Diracs

### Hypotheses

(no assumption  
on the data)

- $\varepsilon$  - separated centroids
- $M$  - bounded domain for centroids

### Sketch

- Adjusted Random Fourier features (for technical reasons)

### Result

- W.r.t. k-means usual cost (SSE)

### Sketch size

$$m \geq \mathcal{O}(\textcolor{red}{k^2 d} \cdot \text{polylog}(k, d) \log(M/\varepsilon))$$

## GMM with known covariance

### Hypotheses

- Sufficiently separated means
- Bounded domain for means

### Sketch

- Fourier features

### Result

- With respect to log-likelihood

### Sketch size

$$m \geq \mathcal{O}(\textcolor{red}{k^2 d} \cdot \text{polylog}(k, d))$$

# Application

## k-means with mixtures of Diracs

### Hypotheses

(no assumption  
on the data)

- $\varepsilon$  - separated centroids
- $M$  - bounded domain for centroids

### Sketch

- Adjusted Random Fourier features (for technical reasons)

### Result

- W.r.t. k-means usual cost (SSE)

### Sketch size

$$m \geq \mathcal{O}(\textcolor{red}{k^2 d} \cdot \text{polylog}(k, d) \log(M/\varepsilon))$$

## GMM with known covariance

### Hypotheses

- Sufficiently separated means
- Bounded domain for means

### Sketch

- Fourier features

### Result

- With respect to log-likelihood

### Sketch size

$$m \geq \mathcal{O}(\textcolor{red}{k^2 d} \cdot \text{polylog}(k, d))$$

Compared to Generalized Method of moments, **different** guarantees

# What about algorithms??

- See practical session next time!