Sketched Learning from Random Features Moments

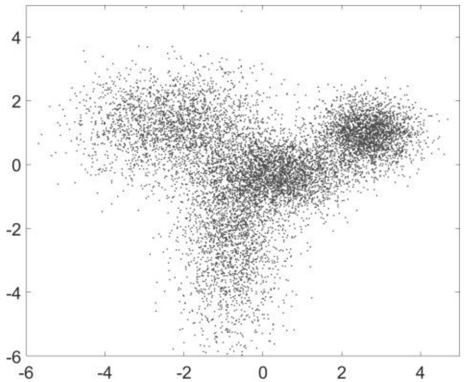
Nicolas Keriven

Ecole Normale Supérieure (Paris)
CFM-ENS chair in Data Science

(thesis with Rémi Gribonval at Inria Rennes)

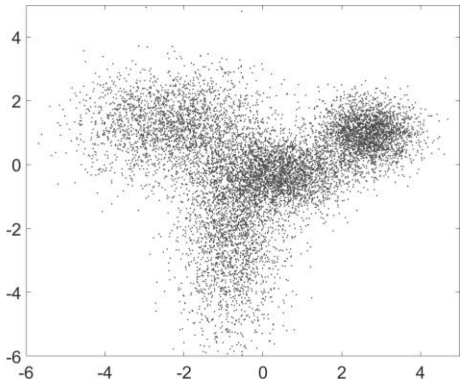
ISMP, July 6th 2018

Compressive learning



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

Compressive learning



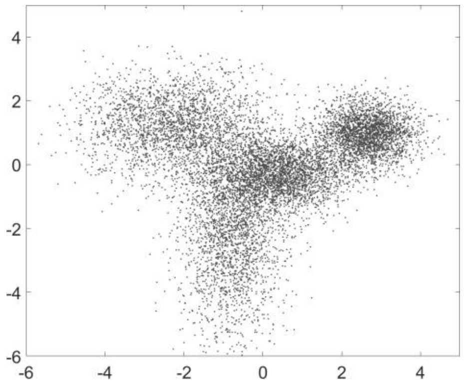
$x_1, \ldots, x_n \in \mathbb{R}^d$

Compression $\longrightarrow$ $\hat{\mathbf{z}}$ $\xrightarrow{\text{Learning}}$

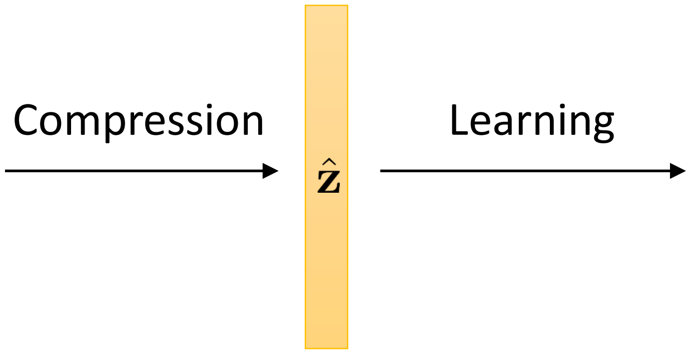
Linear sketch $z \in \mathbb{R}^m$

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

Compressive learning



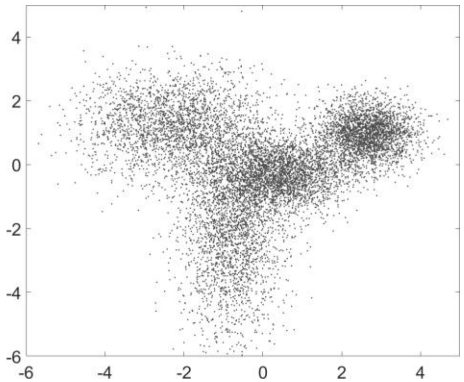
$x_1, \ldots, x_n \in \mathbb{R}^d$



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



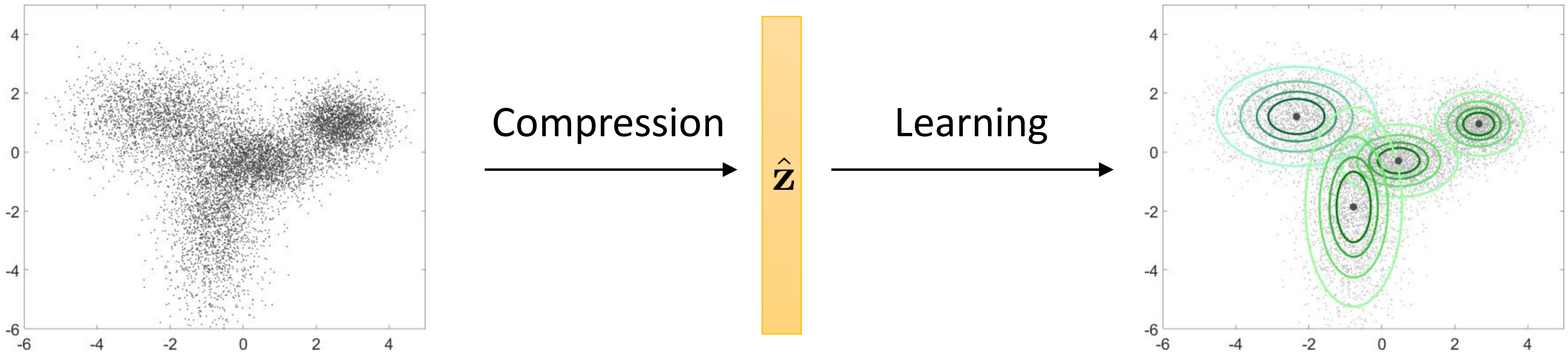
$$x_1, \ldots, x_n \in \mathbb{R}^d$$

Compression $\longrightarrow$ $\hat{\mathbf{z}}$ $\longrightarrow$ Learning

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



$x_1, \ldots, x_n \in \mathbb{R}^d$ ***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…
- **In this talk**: unsupervised learning

How-to: build a sketch

What is a sketch ?

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

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

How-to: build a sketch

What is a sketch ?

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

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

What is contained in a sketch ?

How-to: build a sketch

What is a sketch ?

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

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

What is contained in a sketch ?

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

How-to: build a sketch

What is a sketch ?

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

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

What is contained in a sketch ?

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

How-to: build a sketch

What is a sketch ?

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

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

What is contained in a sketch ?

- $\Phi(x) = x$: mean
- $\Phi(x) = x^k$: k^{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**

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

What is contained in a sketch ?

- $\Phi(x) = x$: mean
- $\Phi(x) = x^k$: k^{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

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

What is contained in a sketch ?

- $\Phi(x) = x$: mean
- $\Phi(x) = x^k$: k^{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**

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

What is contained in a sketch ?

- $\Phi(x) = x$: mean
- $\Phi(x) = x^k$: k^{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**

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

What is contained in a sketch ?

- $\Phi(x) = x$: mean
- $\Phi(x) = x^k$: k^{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: $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \pi^\star$

How-to: build a sketch

What is a sketch ?

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

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

What is contained in a sketch ?

- $\Phi(x) = x$: mean
- $\Phi(x) = x^k$: k^{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: $x_1, \dots, x_n \overset{i.i.d.}{\sim} \pi^\star$
- 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**

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

What is contained in a sketch ?

- $\Phi(x) = x$: mean
- $\Phi(x) = x^k$: k^{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: $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \pi^\star$
- Linear operator: $\mathcal{A}\pi = \mathbb{E}_{X\sim\pi}\Phi(X)$
- « Noisy » linear measurement:

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

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

How-to: build a sketch

What is a sketch ?

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

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

What is contained in a sketch ?

- $\Phi(x) = x$: mean
- $\Phi(x) = x^k$: k^{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: $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \pi^\star$
- Linear operator: $\mathcal{A}\pi = \mathbb{E}_{X \sim \pi}\Phi(X)$
- « Noisy » linear measurement:

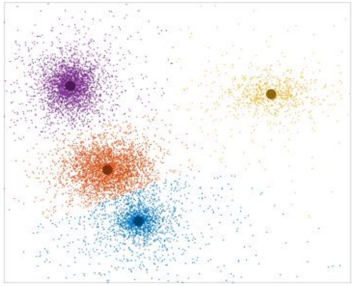
$$\boxed{\hat{\mathbf{z}} = \mathcal{A}\pi^\star + \hat{\mathbf{e}}}$$

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

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

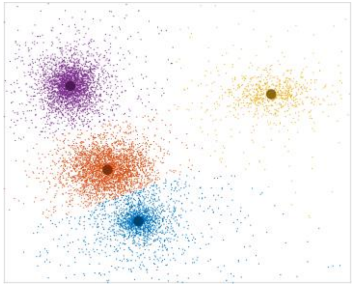
Example of applications *[Keriven 2016,2017]*

Retrieving mixture of Diracs from a sketch= k-means



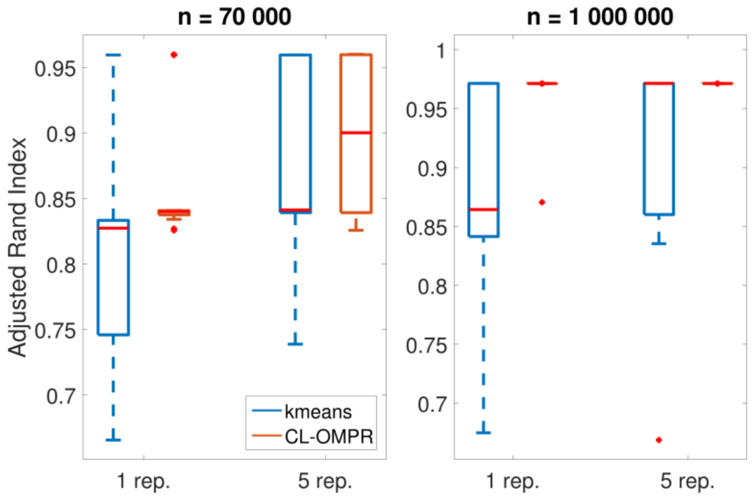
Example of applications *[Keriven 2016,2017]*

Retrieving mixture of Diracs from a sketch= k-means



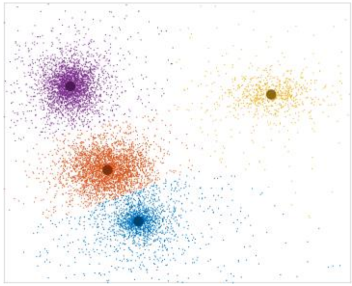
Application: Spectral clustering

for MNIST classification



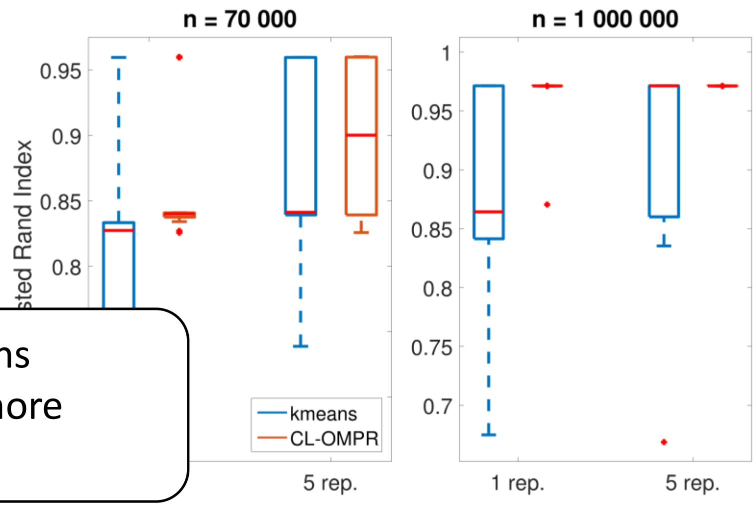
Example of applications *[Keriven 2016,2017]*

Retrieving mixture of Diracs from a sketch= k-means



Application: Spectral clustering

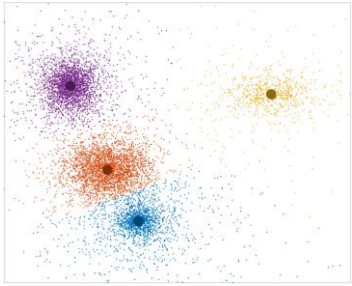
for MNIST classification



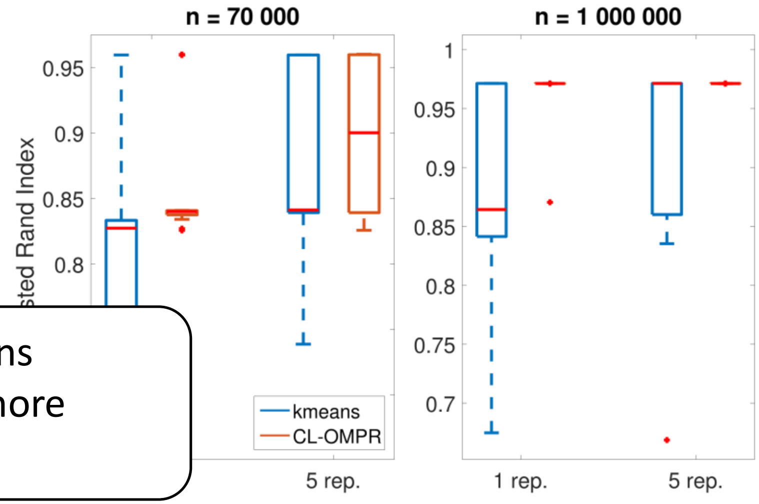
- Twice faster than k-means
- 4 orders of magnitude more memory efficient

Example of applications *[Keriven 2016,2017]*

Retrieving mixture of Diracs from a sketch= k-means

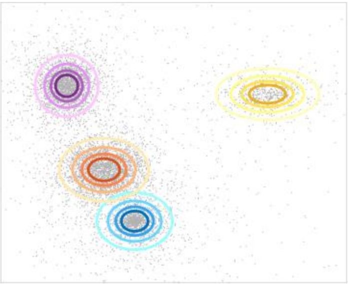


Application:
Spectral clustering
for MNIST classification



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

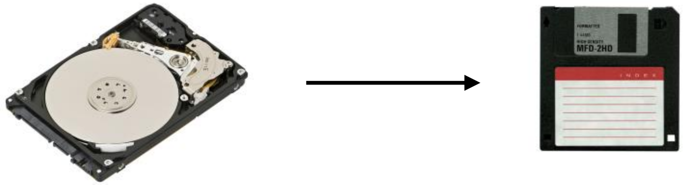
Retrieving GMMs from a sketch



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

Error:

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



In this talk

Q: Theoretical guarantees ?

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

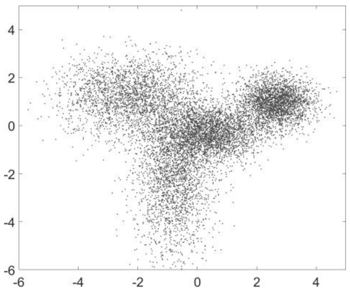
Outline

1. Information-preservation guarantees: a RIP analysis
   Joint work with **R. Gribonval, G. Blanchard, Y. Traonmilin**

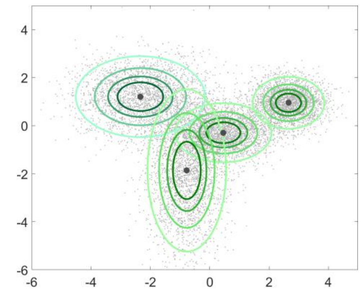
2. Total variation regularization: a dual certificate analysis

3. Conclusion, outlooks

Recall: Linear inverse problem

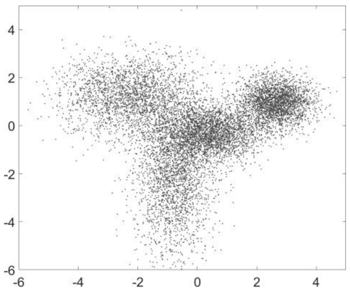


$\xrightarrow{\Phi}$ $\hat{\mathbf{z}}$ $\longrightarrow$

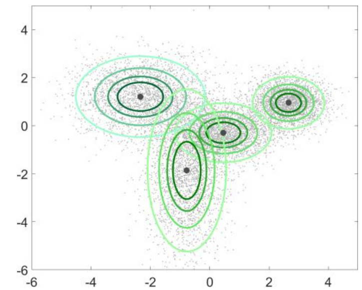


$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X)$$

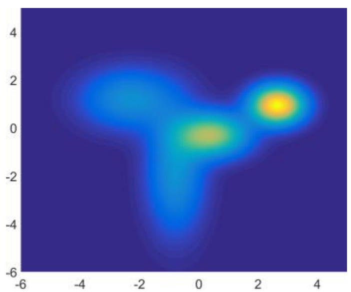
Recall: Linear inverse problem



$\xrightarrow{\Phi}$ $\hat{\mathbf{z}}$ $\longrightarrow$



$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X)$



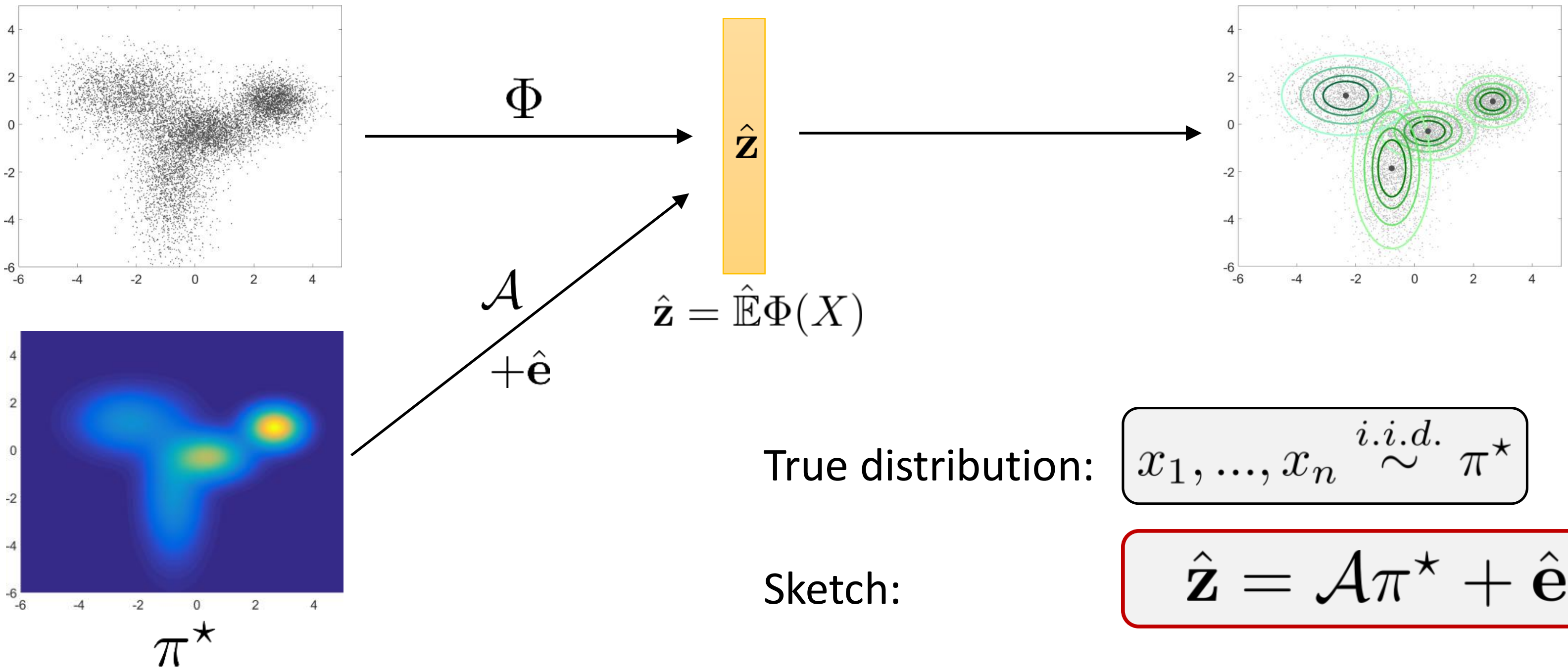
$\pi^\star$

$\mathcal{A}$ $+\hat{\mathbf{e}}$

True distribution: $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \pi^\star$

Sketch: $\boxed{\hat{\mathbf{z}} = \mathcal{A}\pi^\star + \hat{\mathbf{e}}}$

Recall: Linear inverse problem

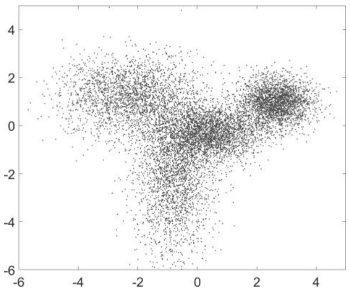


True distribution: $\boxed{x_1, \ldots, x_n \overset{i.i.d.}{\sim} \pi^\star}$

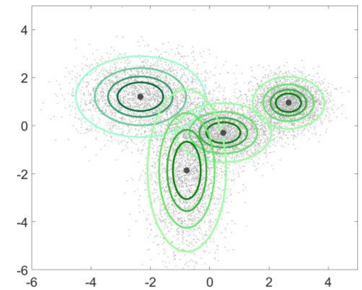
Sketch: $\boxed{\hat{\mathbf{z}} = \mathcal{A}\pi^\star + \hat{\mathbf{e}}}$

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

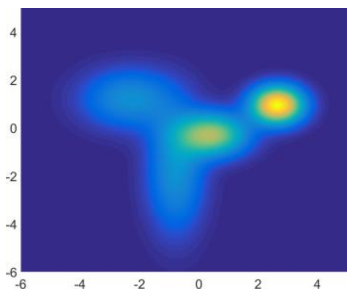
Recall: Linear inverse problem



Φ $\longrightarrow$ $\hat{\mathbf{z}}$ $\xrightarrow{\text{Best algorithm possible}}$



$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X)$



$\pi^\star$

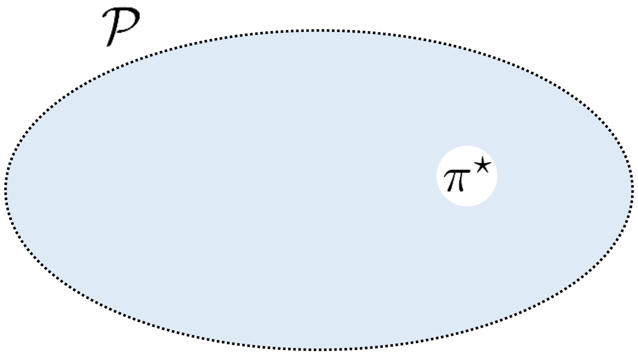
$\mathcal{A}$ $+\hat{\mathbf{e}}$

True distribution: $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \pi^\star$

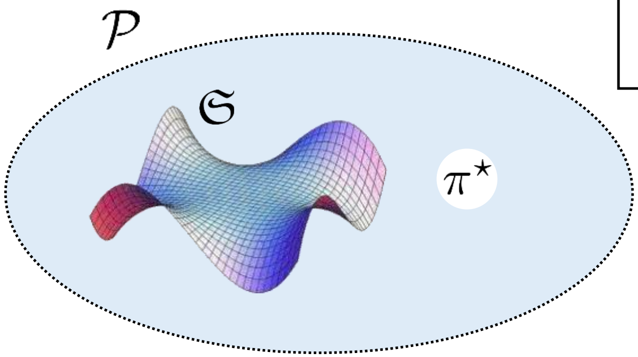
Sketch: $\hat{\mathbf{z}} = \mathcal{A}\pi^\star + \hat{\mathbf{e}}$

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

Information preservation guarantees

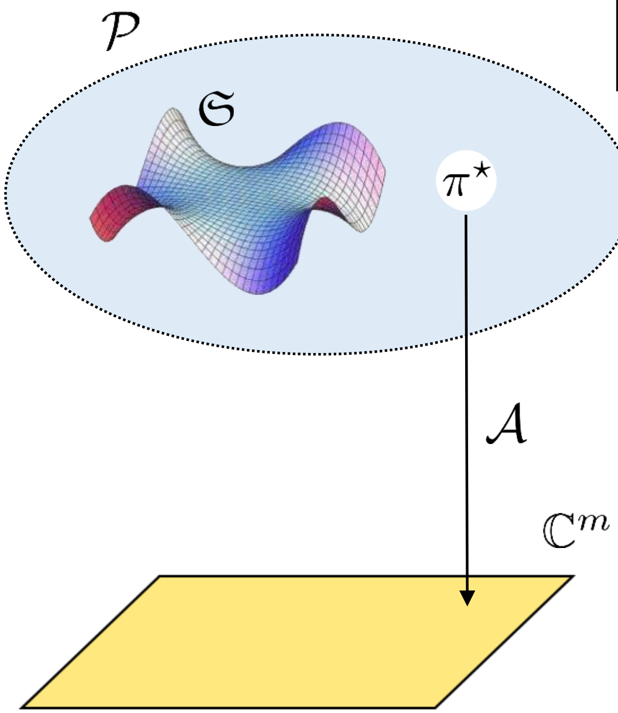


Information preservation guarantees



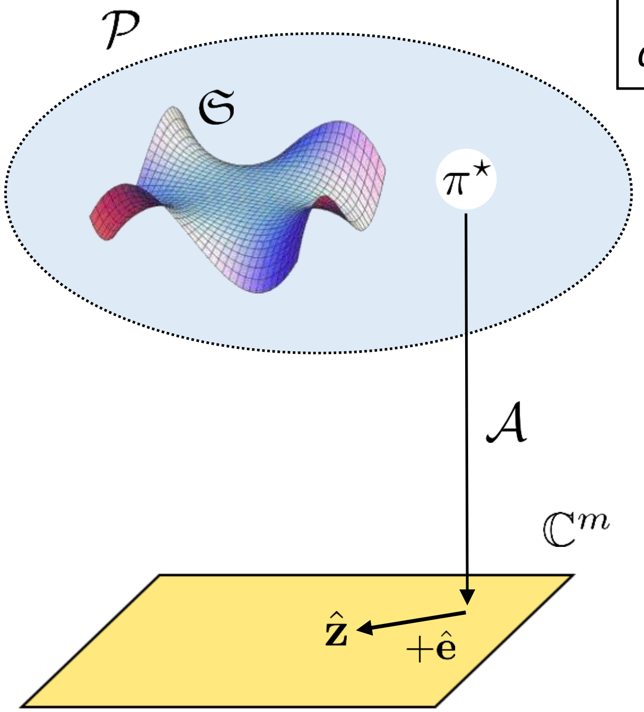
$\mathfrak{S}$: Model set of « simple » distributions (eg. GMMs)

Information preservation guarantees



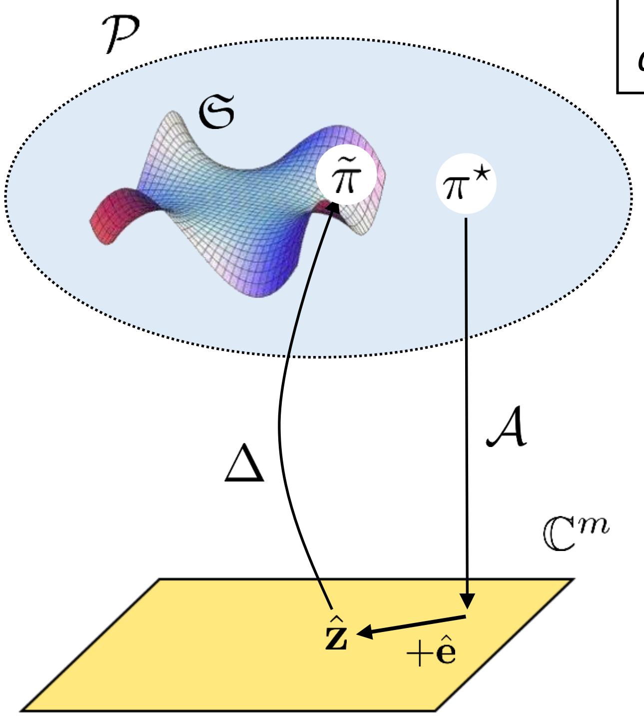
$\mathfrak{S}$ *: Model set of « simple » distributions (eg. GMMs)*

Information preservation guarantees



$\mathfrak{S}$: *Model set of « simple » distributions (eg. GMMs)*

Information preservation guarantees



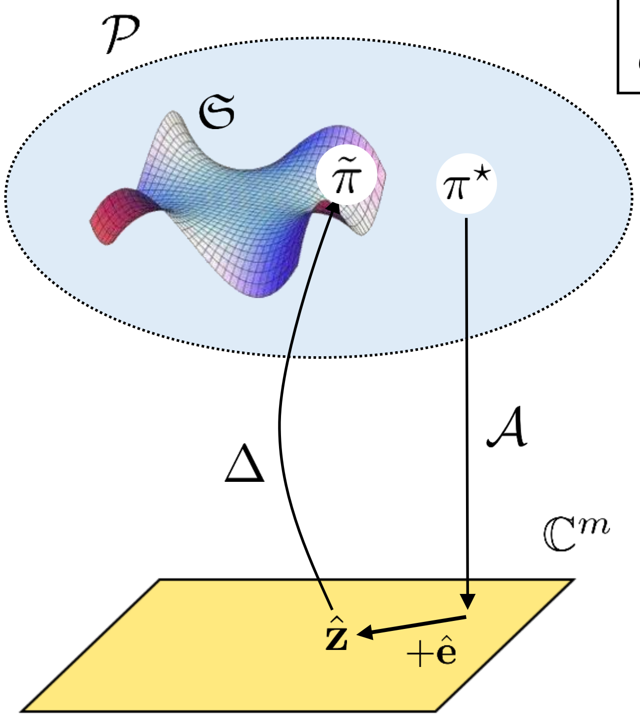
$\mathfrak{S}$: Model set of « simple » distributions (eg. GMMs)

Goal

Prove the existence of a *decoder* Δ robust to noise and stable to modeling error.

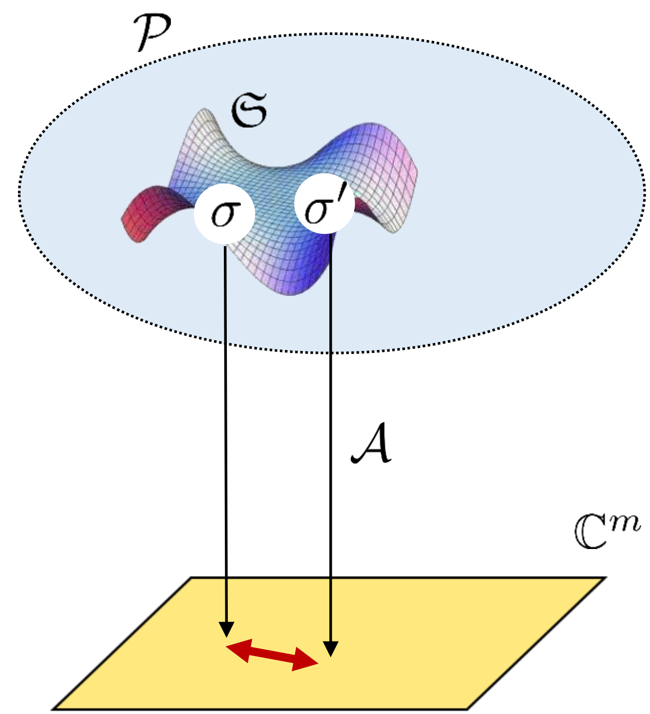
« Instance-optimal » decoder

Information preservation guarantees



$\mathfrak{S}$ *: Model set of « simple » distributions (eg. GMMs)*

$$\Longrightarrow$$



Goal

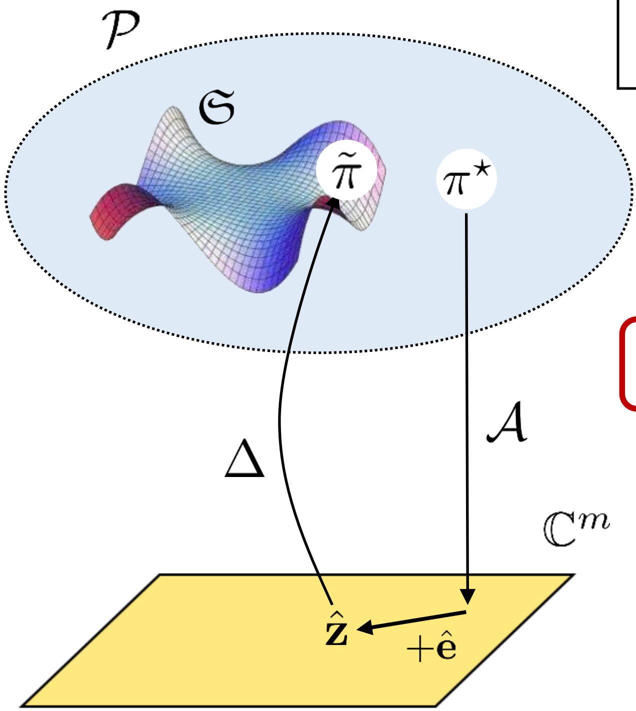
Prove the existence of a *decoder* Δ robust to noise and stable to modeling error.

« Instance-optimal » decoder

Lower Restricted Isometry Property

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

Information preservation guarantees



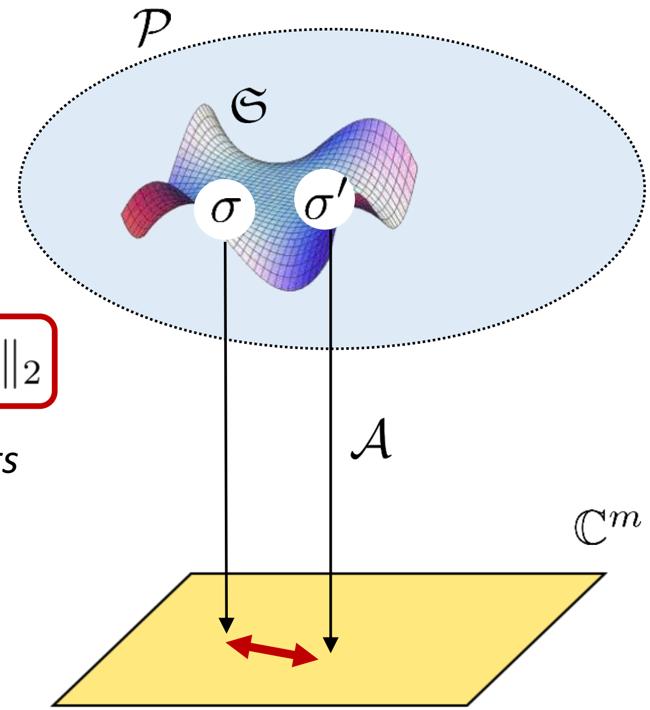
$\mathfrak{S}$ *: Model set of « simple » distributions (eg. GMMs)*

$$\Longrightarrow$$

$$\Longleftarrow$$

$$\boxed{\Delta(\hat{\mathbf{z}}) \in \arg\min_{\sigma\in\mathfrak{S}} \|\hat{\mathbf{z}} - \mathcal{A}\sigma\|_2}$$

Generalized Method of Moments



Goal

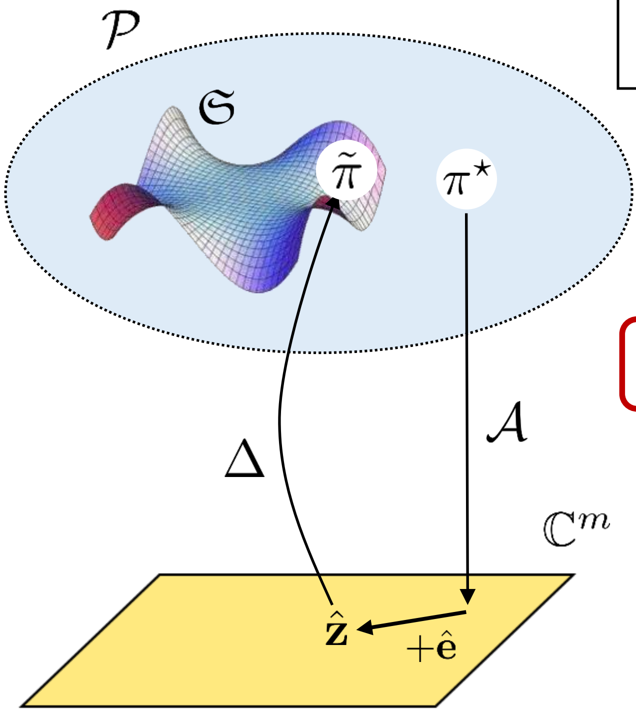
Prove the existence of a *decoder* Δ robust to noise and stable to modeling error.

« Instance-optimal » decoder

Lower Restricted Isometry Property

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

Information preservation guarantees



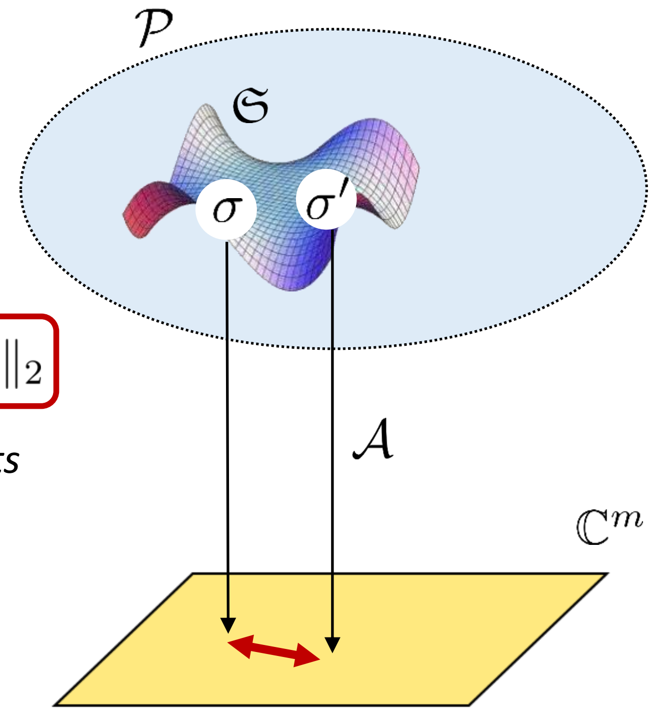
$\mathfrak{S}$: Model set of « simple » distributions (eg. GMMs)

$\Longrightarrow$

$\Longleftarrow$

$$\boxed{\Delta(\hat{\mathbf{z}}) \in \arg\min_{\sigma \in \mathfrak{S}} \|\hat{\mathbf{z}} - \mathcal{A}\sigma\|_2}$$

Generalized Method of Moments



Goal

Prove the existence of a *decoder* Δ robust to noise and stable to modeling error.

Lower Restricted Isometry Property

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

« Instance-optimal » decoder

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

Proving the LRIP

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

Proving the LRIP

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

1 **Pointwise LRIP**

Construction of $\mathcal{A}$:

- Kernel mean *[Gretton 2006, Borgwardt 2006]*
- Random features *[Rahimi 2007]*

$\forall \sigma, \sigma', \text{ w.h.p. on } \mathcal{A}, \text{ LRIP.}$

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

1 Pointwise LRIP

Construction of $\mathcal{A}$:

- Kernel mean *[Gretton 2006, Borgwardt 2006]*
- Random features *[Rahimi 2007]*

$\forall \sigma, \sigma', \text{ w.h.p. on } \mathcal{A}, \text{ LRIP.}$

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

1 Pointwise LRIP

Construction of $\mathcal{A}$:

- Kernel mean *[Gretton 2006, Borgwardt 2006]*
- Random features *[Rahimi 2007]*

$$\forall \sigma, \sigma', \text{ w.h.p. on } \mathcal{A}, \text{ LRIP.}$$

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 $\mathfrak{S}$

$$\text{w.h.p. on } \mathcal{A}, \forall \sigma, \sigma', \text{ LRIP.}$$

Main result [Keriven 2016]

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 $\boxed{m \geq C \quad \times \quad \log(\text{cov. num.})}$,

Main result [Keriven 2016]

Main hypothesis

The *normalized secant set* $\mathcal{S}(\mathfrak{S})$ has finite covering numbers.

Result

For $\boxed{m \geq \underbrace{C}_{\text{Pointwise concentration}} \times \log(\underbrace{\text{cov. num.}}_{\text{Dimensionality of the model}})}$,

Main result [Keriven 2016]

Main hypothesis

The *normalized secant set* $\mathcal{S}(\mathfrak{S})$ has finite covering numbers.

Result

For $\boxed{m \geq C \times \log(\text{cov. num.})}$,

Pointwise concentration

Dimensionality of the model

W.h.p.

$$\|\pi^\star - \Delta(\hat{\mathbf{z}})\| \lesssim d(\pi^\star, \mathfrak{S}) + \mathcal{O}(1/\sqrt{n})$$

Main result [Keriven 2016]

Main hypothesis

The *normalized secant set* $\mathcal{S}(\mathfrak{S})$ has finite covering numbers.

Result

For $\boxed{m \geq C \times \log(\text{cov. num.})}$,

Pointwise concentration

Dimensionality of the model

W.h.p.

Modeling error

Empirical noise

$$\|\pi^\star - \Delta(\hat{\mathbf{z}})\| \lesssim d(\pi^\star, \mathfrak{S}) + \mathcal{O}(1/\sqrt{n})$$

Main hypothesis

The *normalized secant set* $\mathcal{S}(\mathfrak{S})$ has finite covering numbers.

Result

Does not depend on n *!*

For $\boxed{m \geq C \times \log(\text{cov. num.})}$,

Pointwise concentration

Dimensionality of the model

W.h.p.

Modeling error

Empirical noise

$$\|\pi^\star - \Delta(\hat{\mathbf{z}})\| \lesssim d(\pi^\star, \mathfrak{S}) + \mathcal{O}(1/\sqrt{n})$$

Does not depend on m *!*

Main result [Keriven 2016]

Main hypothesis

The *normalized secant set* $\mathcal{S}(\mathfrak{S})$ has finite covering numbers.

Result

Does not depend on n *!*

For $\boxed{m \geq C \times \log(\text{cov. num.})}$,

Pointwise concentration

Dimensionality of the model

W.h.p.

Modeling error

Empirical noise

$$\|\pi^\star - \Delta(\hat{\mathbf{z}})\| \lesssim d(\pi^\star, \mathfrak{S}) + \mathcal{O}(1/\sqrt{n})$$

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

- ε - separated centroids
- M- bounded domain for centroids

Application

k-means with mixtures of Diracs

Hypotheses

- ε - separated centroids
- M- bounded domain for centroids

(no assumption on the ***data****)*

Application

k-means with mixtures of Diracs

(*no assumption on the* ***data***)

Hypotheses

- ε - separated centroids
- M- bounded domain for centroids

Sketch

- *Adjusted* Random Fourier features *(for technical reasons)*

Application

k-means with mixtures of Diracs

Hypotheses

- ε - separated centroids
- M- bounded domain for centroids

(no assumption on the ***data****)*

Sketch

- *Adjusted* Random Fourier features *(for technical reasons)*

Result

- W.r.t. **k-means usual cost (SSE)**

Application

k-means with mixtures of Diracs

*(no assumption on the **data**)*

Hypotheses

- ε - 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

$$m \geq \mathcal{O}\left(k^2 d \cdot \texttt{polylog}(k,d) \log(M/\varepsilon)\right)$$

Application

k-means with mixtures of Diracs

(*no assumption on the* ***data***)

Hypotheses

- ε - 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

$$m \geq \mathcal{O}\left(k^2 d \cdot \texttt{polylog}(k,d) \log(M/\varepsilon)\right)$$

GMM with known covariance

Application

k-means with mixtures of Diracs

(*no assumption on the **data***)

Hypotheses

- ε - 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

$$m \geq \mathcal{O}\left(k^2 d \cdot \texttt{polylog}(k,d) \log(M/\varepsilon)\right)$$

GMM with known covariance

Hypotheses

- Sufficiently separated means
- Bounded domain for means

Application

k-means with mixtures of Diracs

(no assumption on the ***data***)

Hypotheses

- ε - 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

$$m \geq \mathcal{O}\left(k^2 d \cdot \texttt{polylog}(k,d) \log(M/\varepsilon)\right)$$

GMM with known covariance

Hypotheses

- Sufficiently separated means
- Bounded domain for means

Sketch

- Fourier features

Application

k-means with mixtures of Diracs

*(no assumption on the **data**)*

Hypotheses

- ε - 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

$$m \geq \mathcal{O}\left(k^2 d \cdot \mathtt{polylog}(k,d) \log(M/\varepsilon)\right)$$

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

- ε - separated centroids
- M- bounded domain for centroids

*(no assumption on the **data**)*

Sketch

- *Adjusted* Random Fourier features *(for technical reasons)*

Result

- W.r.t. **k-means usual cost (SSE)**

Sketch size

$$m \geq \mathcal{O}\left(k^2 d \cdot \texttt{polylog}(k,d) \log(M/\varepsilon)\right)$$

GMM with known covariance

Hypotheses

- Sufficiently separated means
- Bounded domain for means

Sketch

- Fourier features

Result

- With respect to **log-likelihood**

Sketch size

$$m \geq \mathcal{O}(k^2 d \cdot \texttt{polylog}(k,d))$$

Application

k-means with mixtures of Diracs

Hypotheses

(*no assumption on the* ***data***)

- ε - 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

$$m \geq \mathcal{O}\left(k^2 d \cdot \texttt{polylog}(k,d) \log(M/\varepsilon)\right)$$

GMM with known covariance

Hypotheses

- Sufficiently separated means
- Bounded domain for means

Sketch

- Fourier features

Result

- With respect to **log-likelihood**

Sketch size

$$m \geq \mathcal{O}(k^2 d \cdot \texttt{polylog}(k,d))$$

Compared to Generalized Method of moments, ***different*** *guarantees*

Outline

1. Information-preservation guarantees:
a RIP analysis

2. Total variation regularization:
a dual certificate analysis
Joint work with **C. Poon, G. Peyré**

3. Conclusion, outlooks

Total Variation regularization

Previously: RIP analysis

Minimization: moment matching

$$\boxed{\min_{\theta,w} \| \sum w_i \mathcal{A}\pi_{\theta_i} - \hat{\mathbf{z}}\|_2}$$

Total Variation regularization

Previously: RIP analysis

Minimization: moment matching

$$\boxed{\min_{\theta,w} \| \sum w_i \mathcal{A}\pi_{\theta_i} - \hat{\mathbf{z}} \|_2}$$

- Must know k
- **Non-convex !**

Total Variation regularization

Previously: RIP analysis

Minimization: moment matching

$$\boxed{\min_{\theta,w} \|\sum w_i \mathcal{A}\pi_{\theta_i} - \hat{\mathbf{z}}\|_2}$$

- Must know k
- **Non-convex !**

Convex relaxation (« *super resolution* »): Beurling-LASSO (BLASSO) *[DeCastro 2015]*

$$\boxed{\min_\mu \frac{1}{2}\|\int (\mathcal{A}\pi_\theta) d\mu(\theta) - \hat{\mathbf{z}}\|_2^2 + \lambda\|\mu\|_{\mathrm{TV}}}$$

- μ : Radon measure
- $\|\cdot\|_{\mathrm{TV}}$: Total variation (« L1 norm »)

Total Variation regularization

Previously: RIP analysis

Minimization: moment matching

$$\boxed{\min_{\theta,w} \| \sum w_i \mathcal{A}\pi_{\theta_i} - \hat{\mathbf{z}}\|_2}$$

- Must know k
- **Non-convex !**

Convex relaxation (« *super resolution* »): Beurling-LASSO (BLASSO) *[DeCastro 2015]*

$$\boxed{\min_\mu \frac{1}{2}\|\int (\mathcal{A}\pi_\theta) d\mu(\theta) - \hat{\mathbf{z}}\|_2^2 + \lambda\|\mu\|_{\mathrm{TV}}}$$

- μ : Radon measure
- $\|\cdot\|_{\mathrm{TV}}$: Total variation (« L1 norm »)

Questions:

- Is the measure μ sparse ? $\mu = \sum \tilde{w}_i \delta_{\tilde{\theta}_i}$
- Does it have the right number of components ?
- Does it recover the true w_i, θ_i ?

Dual certificates

Dual certificate analysis:
(= Lagrange multiplier)

Dual certificates

Dual certificate analysis:
(= Lagrange multiplier)

Function $\eta(\theta) = \langle \mathbf{h}, \mathcal{A}\pi_\theta \rangle_{\mathbb{C}^m}$

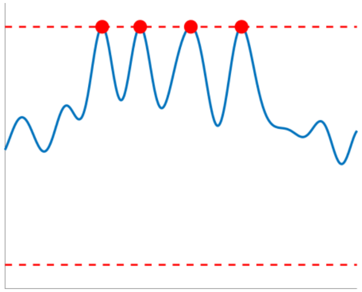
Dual certificates

Dual certificate analysis:
(= Lagrange multiplier)

Function $\eta(\theta) = \langle \mathbf{h}, \mathcal{A}\pi_\theta \rangle_{\mathbb{C}^m}$

Such that:

- $\eta(\theta_i) = 1$
- $|\eta(\theta)| < 1$ otherwise
- $\nabla^2 \eta(\theta_i) \prec 0$



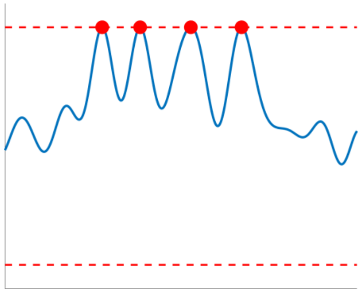
Dual certificates

Dual certificate analysis:
(= Lagrange multiplier)

Function $\eta(\theta) = \langle \mathbf{h}, \mathcal{A}\pi_\theta \rangle_{\mathbb{C}^m}$

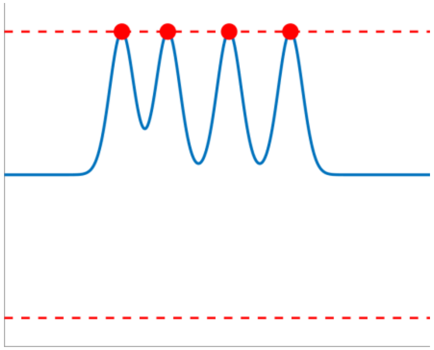
Such that:

- $\eta(\theta_i) = 1$
- $|\eta(\theta)| < 1$ otherwise
- $\nabla^2 \eta(\theta_i) \prec 0$



Step 1: study full kernel *[Candes 2013]*

Assume θ_i *sufficiently separated*



$m = \infty$

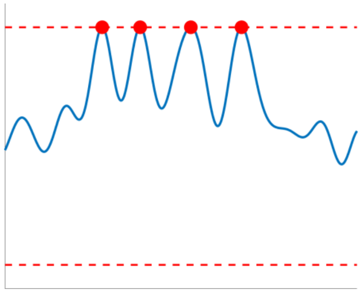
Dual certificates

Dual certificate analysis:
(= Lagrange multiplier)

Function $\eta(\theta) = \langle \mathbf{h}, \mathcal{A}\pi_\theta \rangle_{\mathbb{C}^m}$

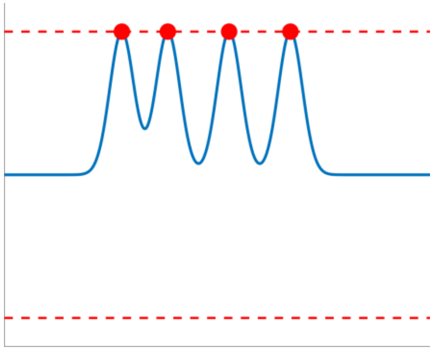
Such that:

- $\eta(\theta_i) = 1$
- $|\eta(\theta)| < 1$ otherwise
- $\nabla^2 \eta(\theta_i) \prec 0$



Step 1: study full kernel *[Candes 2013]*

Assume θ_i *sufficiently separated*



$m = \infty$

Step 2: bounding the deviations

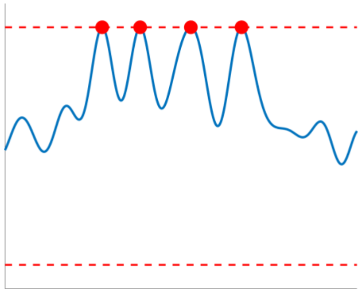
Dual certificates

Dual certificate analysis:
(= Lagrange multiplier)

Function $\eta(\theta) = \langle \mathbf{h}, \mathcal{A}\pi_\theta \rangle_{\mathbb{C}^m}$

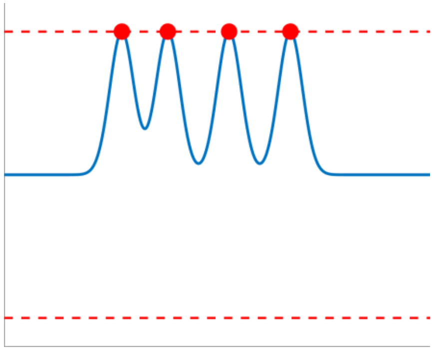
Such that:

- $\eta(\theta_i) = 1$
- $|\eta(\theta)| < 1$ otherwise
- $\nabla^2 \eta(\theta_i) \prec 0$



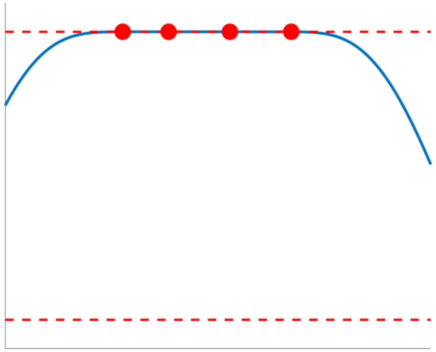
Step 1: study full kernel *[Candes 2013]*

Assume θ_i sufficiently separated



$m = \infty$

Step 2: bounding the deviations



m=10

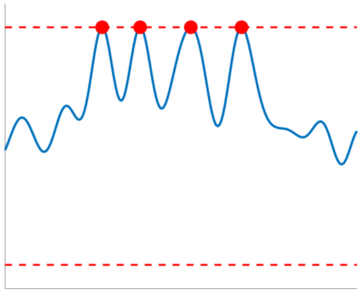
Dual certificates

Dual certificate analysis:
(= Lagrange multiplier)

Function $\eta(\theta) = \langle \mathbf{h}, A\pi_\theta \rangle_{\mathbb{C}^m}$

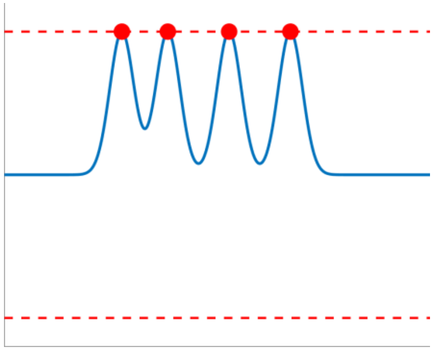
Such that:

- $\eta(\theta_i) = 1$
- $|\eta(\theta)| < 1$ otherwise
- $\nabla^2 \eta(\theta_i) \prec 0$



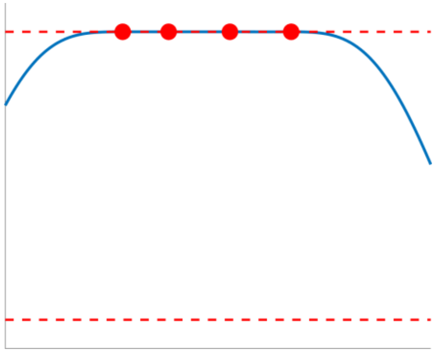
Step 1: study full kernel *[Candes 2013]*

Assume θ_i sufficiently separated

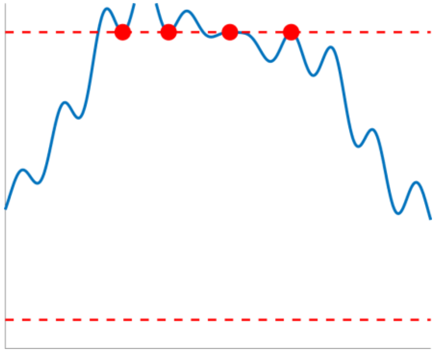


$m = \infty$

Step 2: bounding the deviations



m=10



m=20

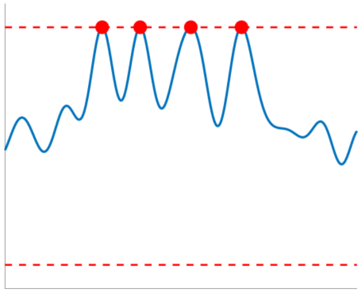
Dual certificates

Dual certificate analysis:
(= Lagrange multiplier)

Function $\eta(\theta) = \langle \mathbf{h}, A\pi_\theta \rangle_{\mathbb{C}^m}$

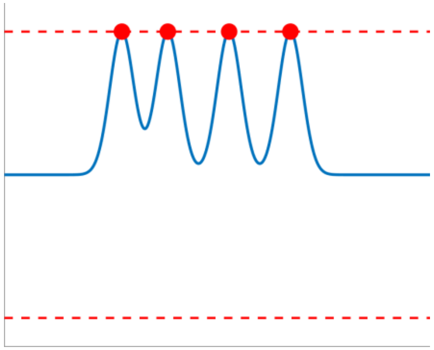
Such that:

- $\eta(\theta_i) = 1$
- $|\eta(\theta)| < 1$ otherwise
- $\nabla^2 \eta(\theta_i) \prec 0$



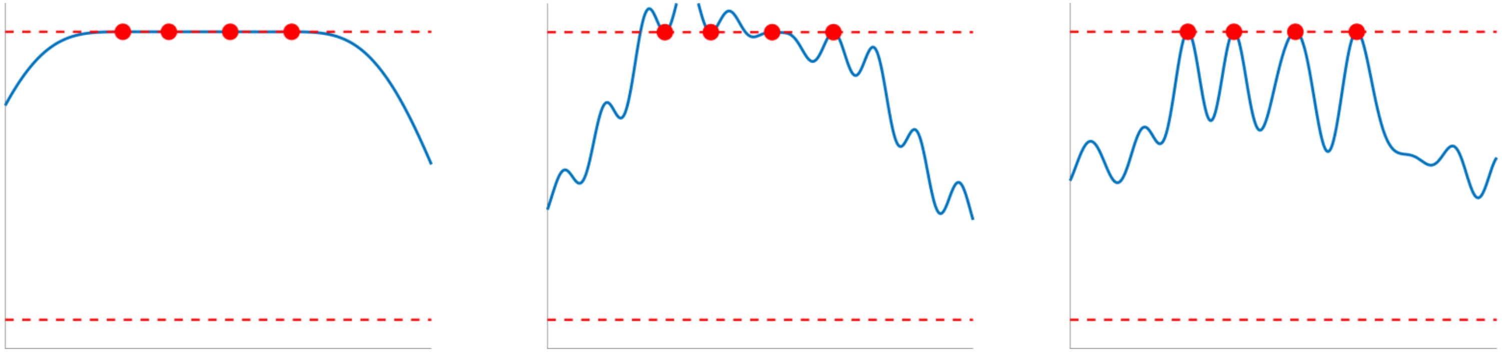
Step 1: study full kernel *[Candes 2013]*

Assume θ_i sufficiently separated



$m = \infty$

Step 2: bounding the deviations



m=10 *m=20* *m=50*

Results for separated GMM

1: Ideal scaling in sparsity

Results for separated GMM

1: Ideal scaling in sparsity

$$m \geq \mathcal{O}(kd^4 \cdot \texttt{polylog}(k,d))$$

Results for separated GMM

1: Ideal scaling in sparsity

$$m \geq \mathcal{O}(\underset{\substack{\uparrow \\ \textit{In progress...}}}{k} d^4 \cdot \texttt{polylog}(k, d))$$

Results for separated GMM

1: Ideal scaling in sparsity

$$m \geq \mathcal{O}(kd^4 \cdot \texttt{polylog}(k,d))$$

In progress…

- $\tilde{\mu}$ ***not necessarily right number of components***, but:
- Mass of $\tilde{\mu}$ concentrated around true θ_i
- (weak) robustness to modelling error
- *Proof*: infinite-dimensional golfing scheme (new)

Results for separated GMM

1: Ideal scaling in sparsity

$$m \geq \mathcal{O}(kd^4 \cdot \texttt{polylog}(k,d))$$

In progress…

- $\tilde{\mu}$ ***not necessarily right number of components***, but:
- Mass of $\tilde{\mu}$ concentrated around true θ_i
- (weak) robustness to modelling error
- *Proof*: infinite-dimensional golfing scheme (new)

Assumption: data are *actually* drawn from a GMM…

2: *Minimal norm* certificate

[Duval, Peyré 2015]

$$m \geq \mathcal{O}(k^2 d^3 \cdot \texttt{polylog}(k,d))$$

In progress…

Results for separated GMM

1: Ideal scaling in sparsity

$$m \geq \mathcal{O}(kd^4 \cdot \texttt{polylog}(k,d))$$

In progress…

- $\tilde{\mu}$ ***not necessarily right number of components***, but:
- Mass of $\tilde{\mu}$ concentrated around true θ_i
- (weak) robustness to modelling error
- *Proof*: infinite-dimensional golfing scheme (new)

Assumption: data are *actually* drawn from a GMM…

2: *Minimal norm* certificate

[Duval, Peyré 2015]

$$m \geq \mathcal{O}(k^2 d^3 \cdot \texttt{polylog}(k,d))$$

In progress…

- when *n* high enough: $\tilde{\mu}$ **sparse, with right number of components**
- $\tilde{\theta}_i \xrightarrow[n\to\infty]{} \theta_i$
- Proof: adaptation of *[Tang, Recht 2013]*

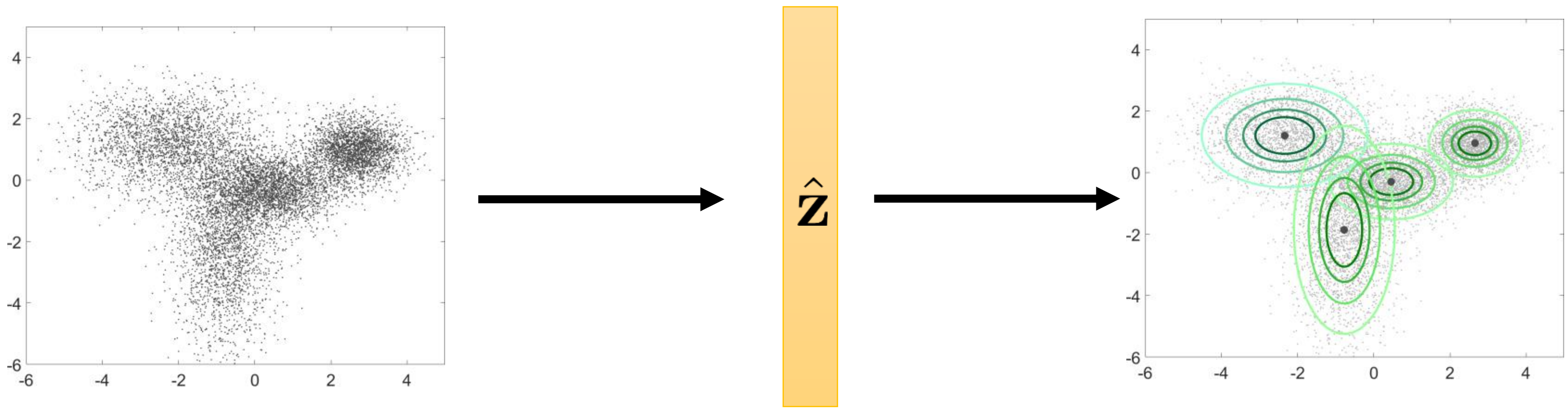
Outline

1. Information-preservation guarantees:
   a RIP analysis

2. Total variation regularization:
   a dual certificate analysis

3. Conclusion, outlooks

Sketch learning



- Sketching :
  - Streaming, distributed learning
  - Original view on data compression and generalized moments
  - Combines random features and kernel mean with infinite dimensional Compressive sensing

Summary, outlooks

- **RIP analysis**
  - Information preservation guarantees
  - Fine control on noise, modeling error (instance optimal decoder) and recovery metrics
  - Necessary and sufficient conditions

Summary, outlooks

- **RIP analysis**
  - Information preservation guarantees
  - Fine control on noise, modeling error (instance optimal decoder) and recovery metrics
  - Necessary and sufficient conditions

- **Dual certificate analysis**
  - Convex minimization
  - In some cases, automatically guess the right number of components

Summary, outlooks

- **RIP analysis**
  - Information preservation guarantees
  - Fine control on noise, modeling error (instance optimal decoder) and recovery metrics
  - Necessary and sufficient conditions

- **Dual certificate analysis**
  - Convex minimization
  - In some cases, automatically guess the right number of components

- **Outlooks**
  - Algorithms for TV minimization
  - Other features Φ (not necessarily random…)
  - Other « sketched » learning tasks
  - Multilayer sketches ?

Thank you !

- Gribonval, Blanchard, Keriven, Traonmilin. **Compressive Statistical Learning with Random Feature Moments**. 2017. <arXiv:1706.07180>
- Keriven. **Sketching for Large-Scale Learning of Mixture Models**. *PhD Thesis.* <tel-01620815>
- Poon, Keriven, Peyré. **A Dual Certificates Analysis of Compressive Off-the-Grid Recovery.** 2018. <arXiv:1802.08464>
- **Code, applications**: nkeriven.github.io