Sketched Learning from Random Features Moments

Nicolas Keriven

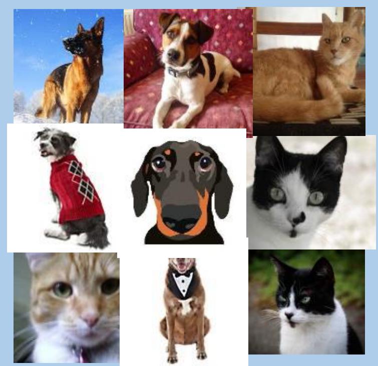
Ecole Normale Supérieure (Paris)
CFM-ENS chair in Data Science

(thesis with Rémi Gribonval at Inria Rennes)

Imaging in Paris, Apr. 5th 2018

Context: machine learning

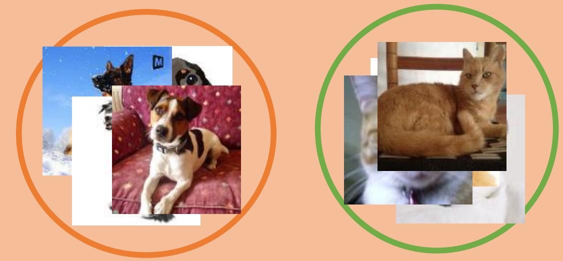
Database



Learning →

Task

- Clustering



- Classification

= cat

- etc…

Context: machine learning

Large **database**

Learning →

Task

- Clustering

- Classification

= cat

- etc...

Context: machine learning

Large database

Learning

Slow, costly

Task

- Clustering

- Classification

= cat

- etc…

Context: machine learning

***Large* database**

***Distributed* database**

Learning

Slow, costly

Task

- Clustering

- Classification

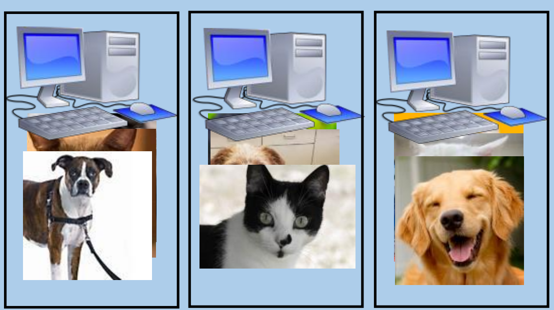
= cat

- etc...

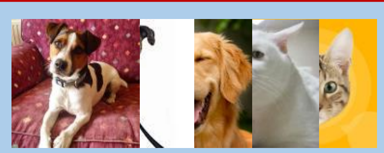
Context: machine learning

Large **database**

Distributed **database**



Data *Stream*

…  …

Learning

Slow, costly

Task

- Clustering 
- Classification  = cat
- etc…

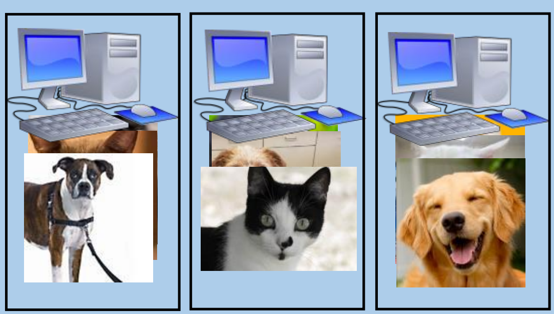
Context: machine learning

Large **database**

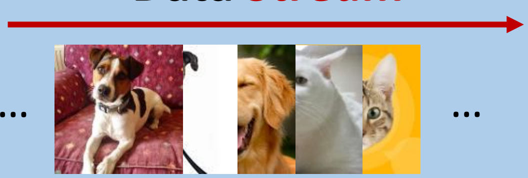
Learning

Slow, costly

Distributed **database**



Data ***Stream***

…  …

Task

- Clustering

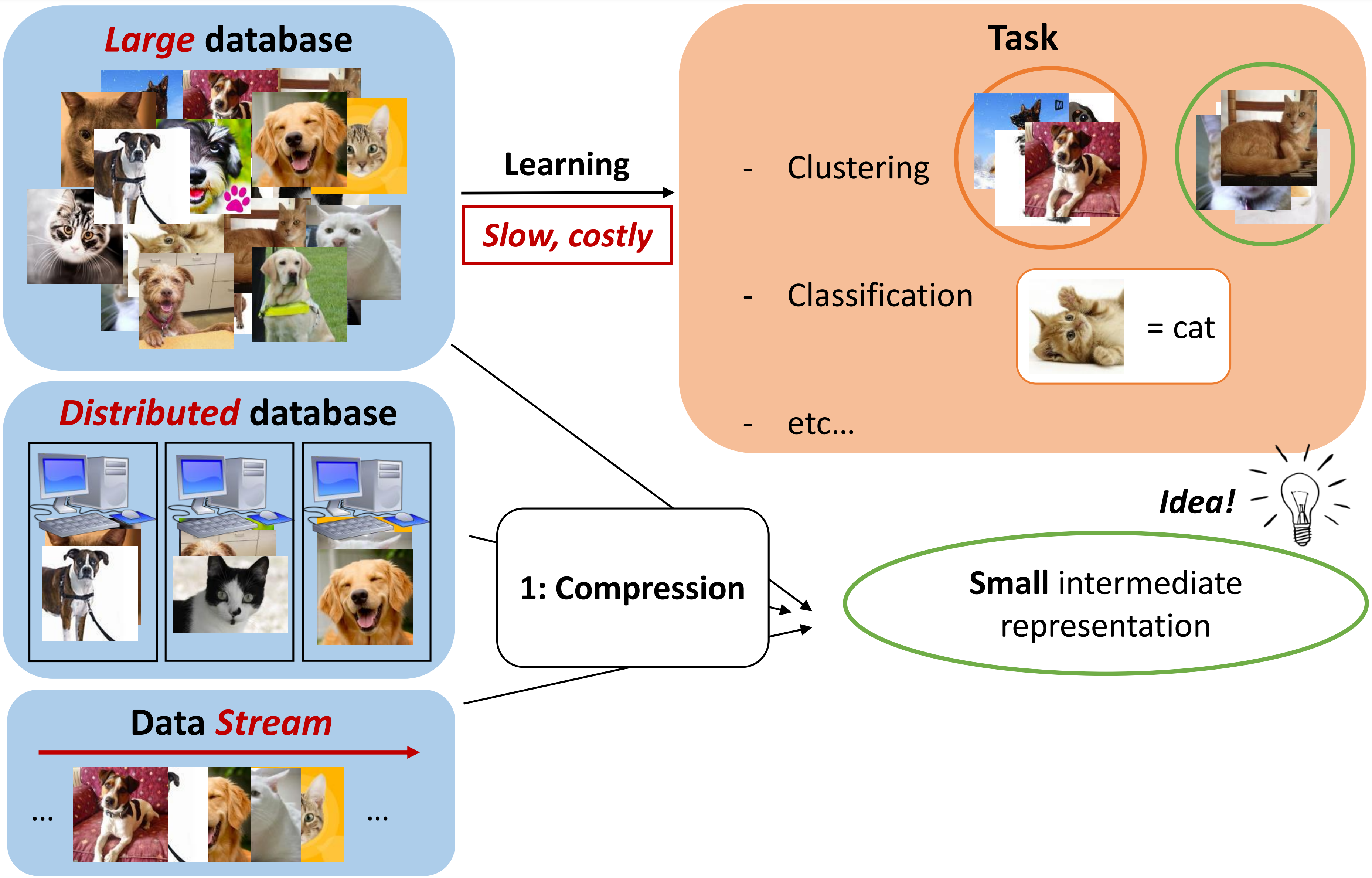
- Classification

- etc…

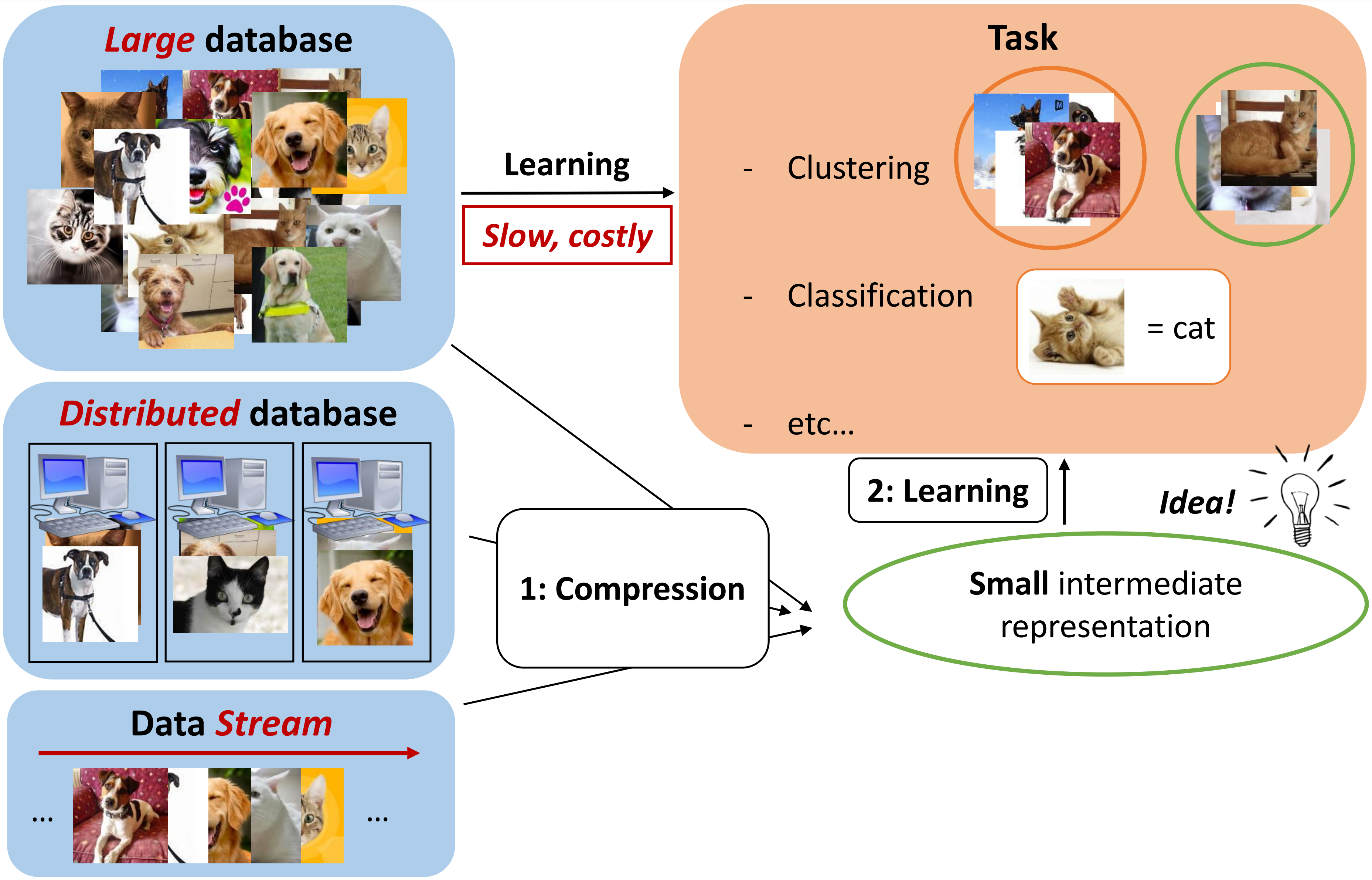
Idea!

Small intermediate representation

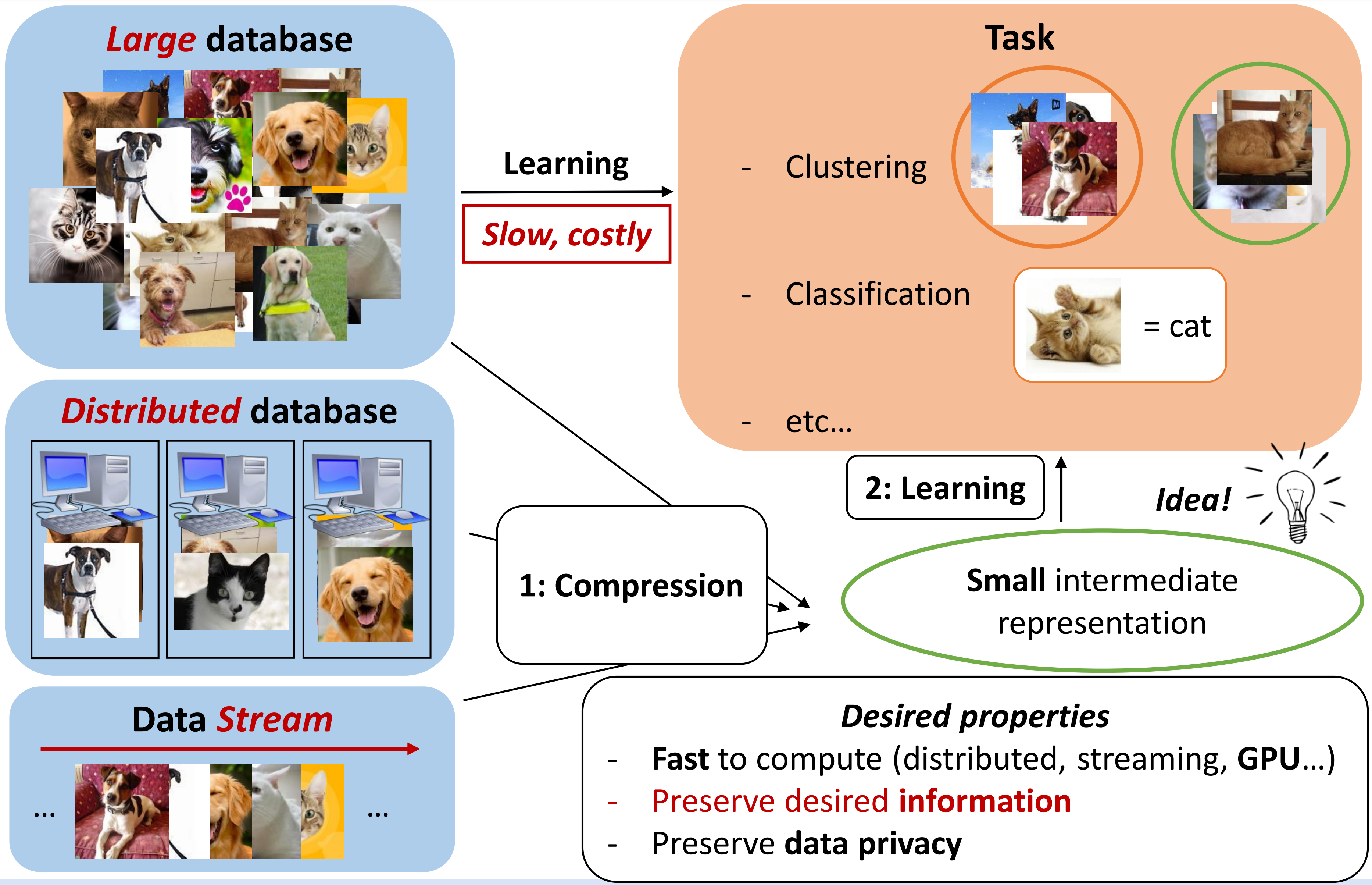
Context: machine learning



Context: machine learning

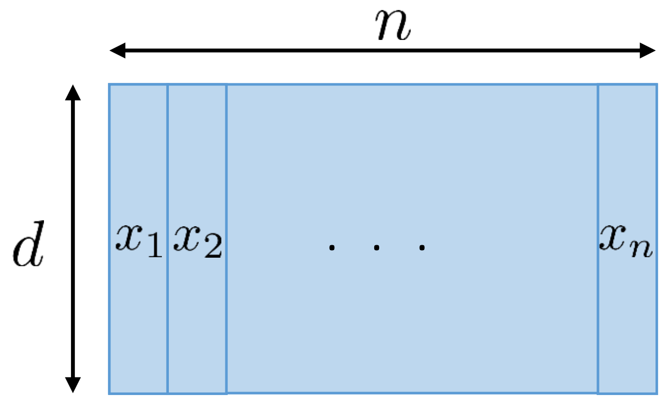


Context: machine learning



Three compression schemes

Feature extraction

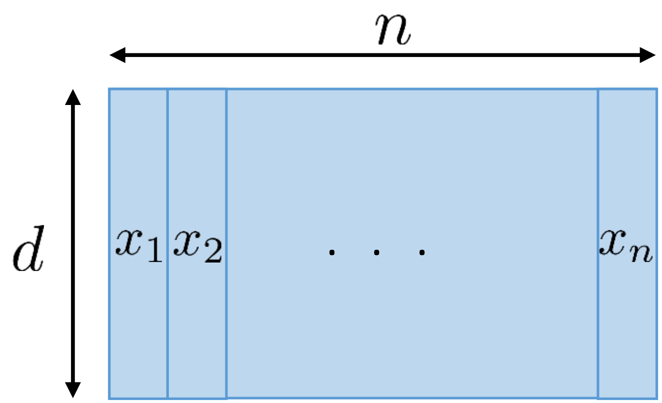


Data = Collection of vectors

Three compression schemes

Database

Feature extraction →



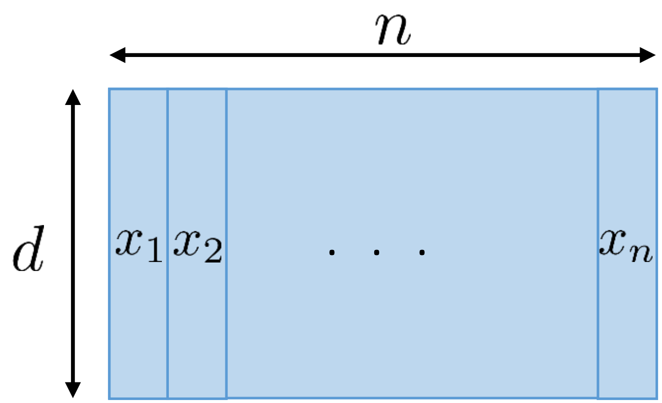
Data = Collection of vectors

Compression ?

Three compression schemes

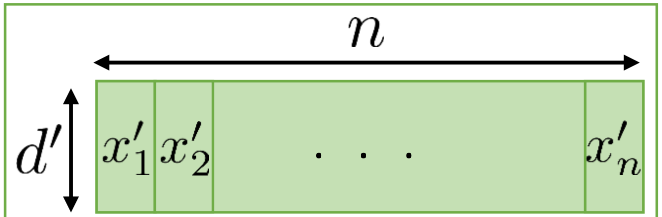
Database

Feature extraction



Data = Collection of vectors

Compression ?



Dimensionality reduction

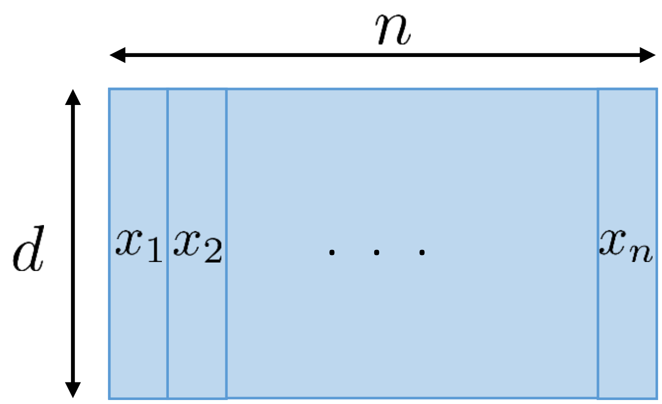
See eg *[Calderbank 2009, Boutsidis 2010]*

- Random Projection
- Feature selection

Three compression schemes

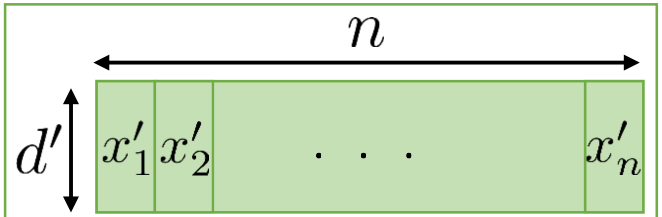
Database

Feature extraction



Data = Collection of vectors

Compression ?



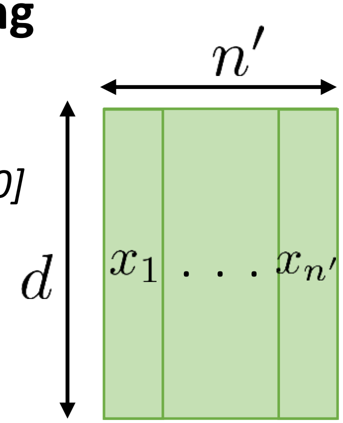
Dimensionality reduction

See eg *[Calderbank 2009, Boutsidis 2010]*

- Random Projection
- Feature selection

Subsampling coresets

See eg *[Feldman 2010]*

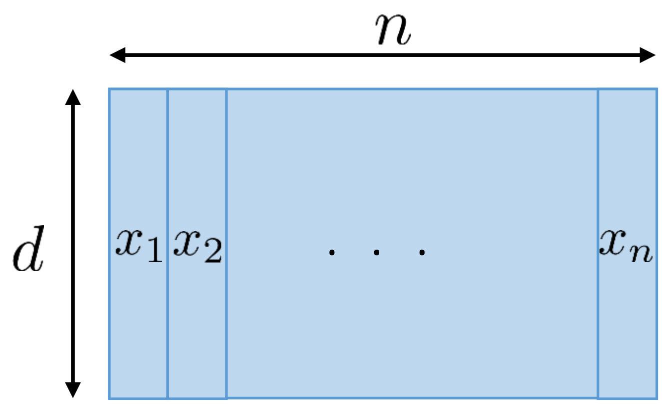


- Uniform sampling (naive)
- Adaptive sampling…

Three compression schemes

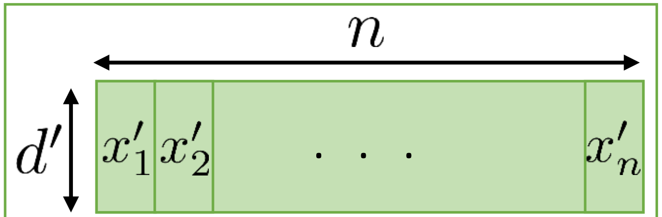
Database

Feature extraction →



Data = Collection of vectors

Compression ?



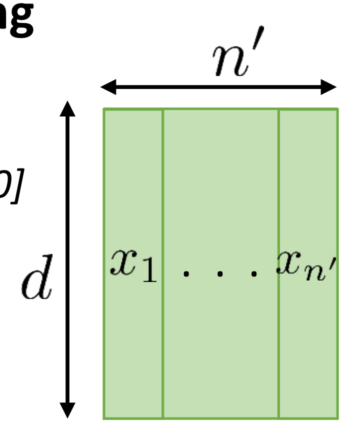
Dimensionality reduction

See eg *[Calderbank 2009, Boutsidis 2010]*

- Random Projection
- Feature selection

Subsampling coresets

See eg *[Feldman 2010]*

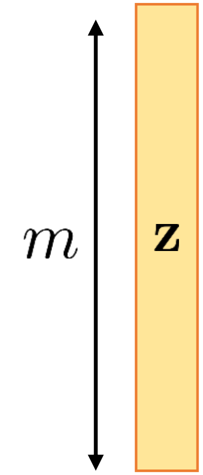


- Uniform sampling (naive)
- Adaptive sampling…

Linear sketch

See *[Thaper 2002]* *[Cormode 2011]*

Distributed, streaming



- Hash tables, histograms
- **Sketching for 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: $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \pi^\star$
- 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: $x_1, \dots, x_n \overset{i.i.d.}{\sim} \pi^\star$
- 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:

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

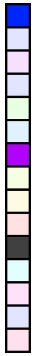
Noise $\hat{\mathbf{e}} = \hat{\mathbb{E}}\Phi(X) - \mathbb{E}_{\pi^\star}\Phi(X)$ ***small***

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

Compressive Sensing: sparsity ?

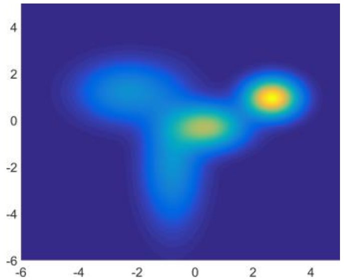
Compressive Sensing:

Classical compressive sensing



$\mathbf{x}$

Sketched learning in this talk



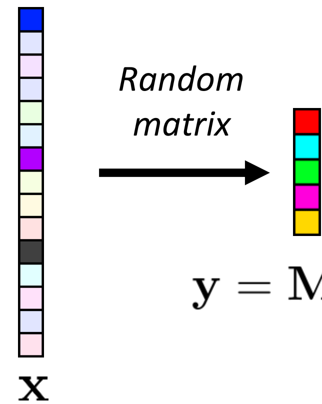
$\pi^\star$

Compressive Sensing: sparsity ?

Compressive Sensing:

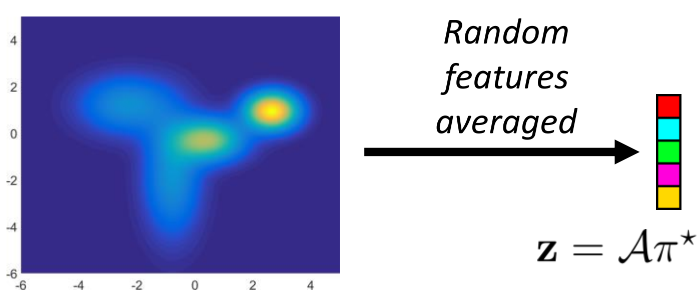
- Dimensionality reduction, random operator

Classical compressive sensing



$$\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{e}$$

Sketched learning in this talk



$$\pi^\star$$

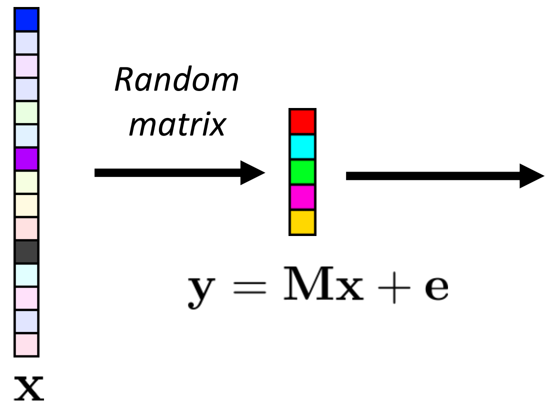
$$\mathbf{z} = \mathcal{A}\pi^\star + \mathbf{e}$$

Compressive Sensing: sparsity ?

Compressive Sensing:

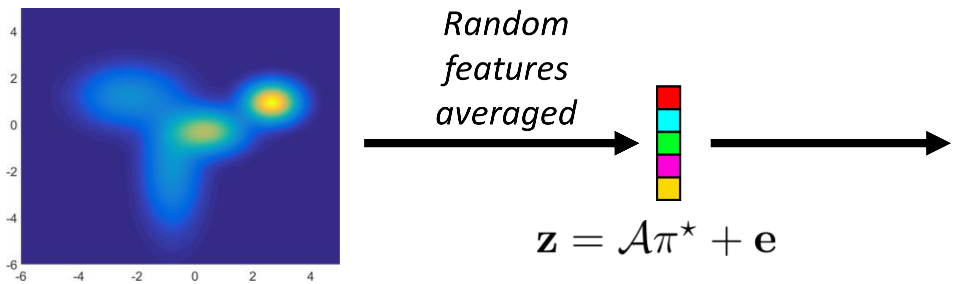
- Dimensionality reduction, random operator
- (Ill-posed) **inverse problem**: *density estimation*

Classical compressive sensing



$$\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{e}$$

Sketched learning in this talk



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

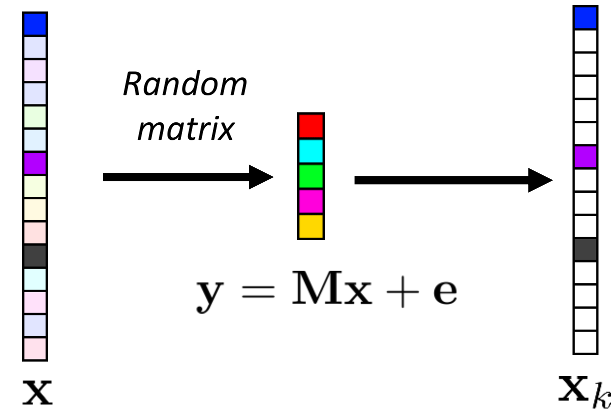
$\pi^\star$

Compressive Sensing: sparsity ?

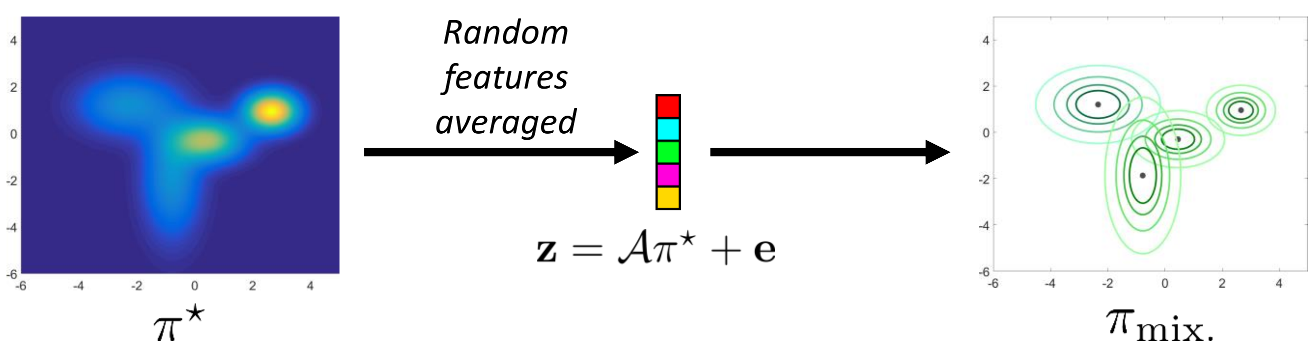
Compressive Sensing:

- Dimensionality reduction, random operator
- (Ill-posed) **inverse problem**: *density estimation*
- **Sparsity**: « simple » densities (mixture model)

Classical compressive sensing

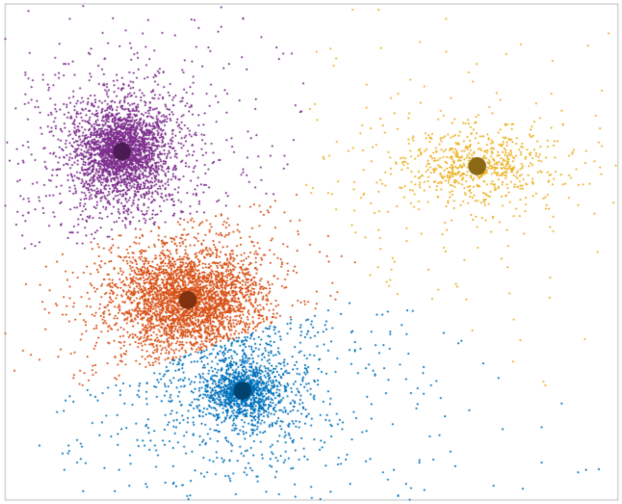


Sketched learning in this talk



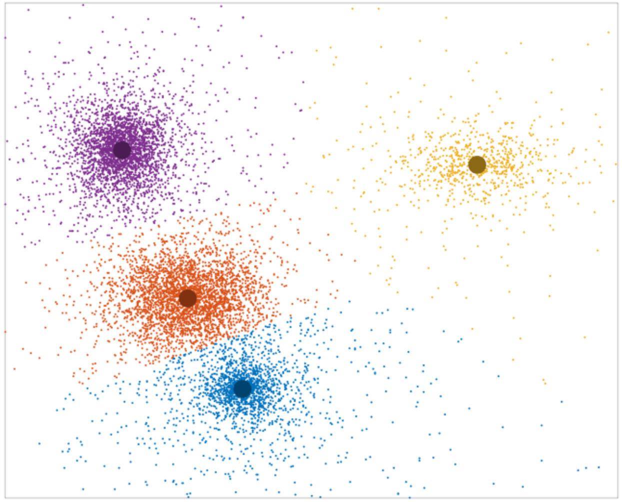
Result: Compressive k-means [Keriven et al 2017]

Mixture of Diracs = k-means



Result: Compressive k-means [Keriven et al 2017]

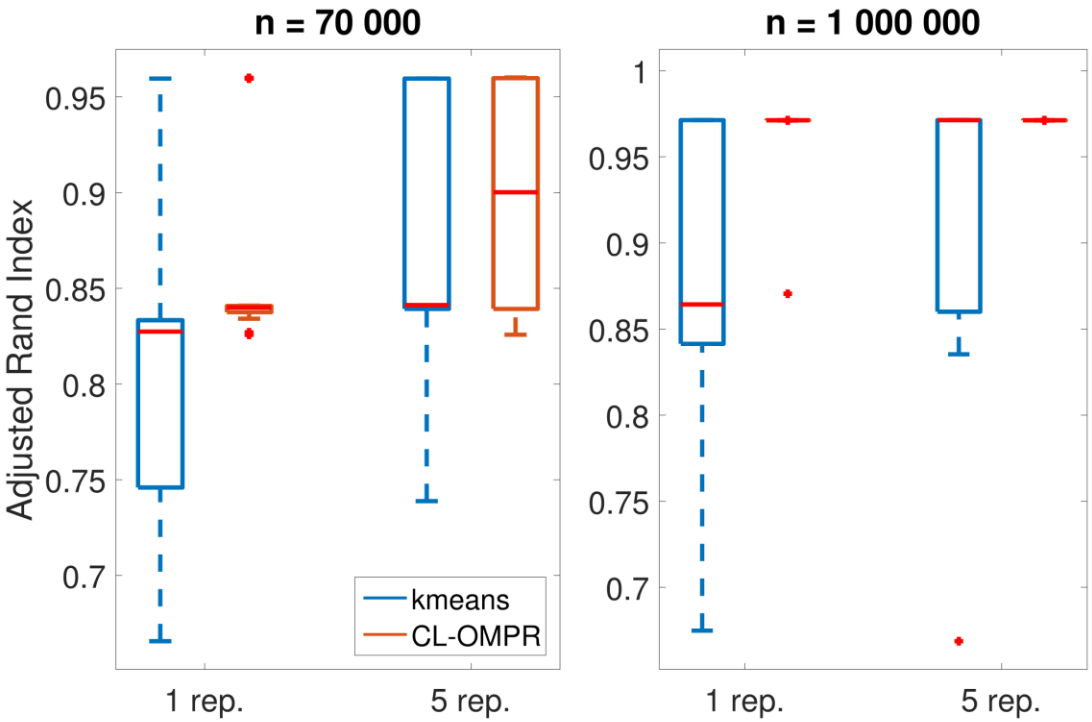
Mixture of Diracs = k-means



Application: Spectral clustering

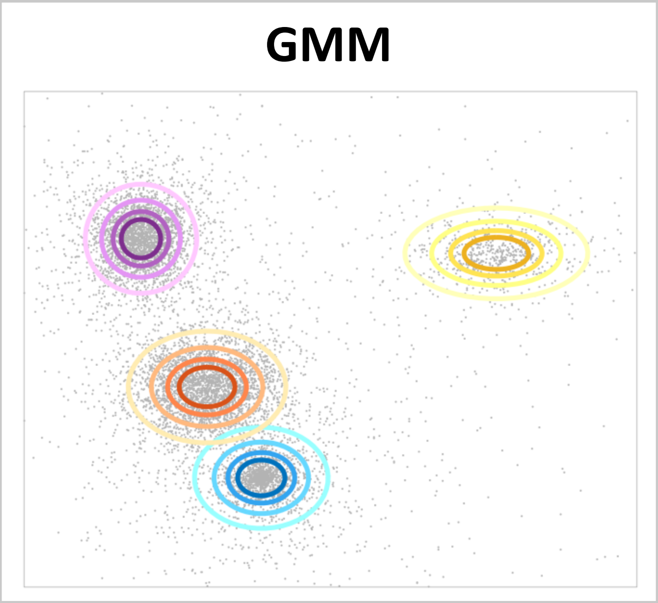
for MNIST classification [Uw 2001]

Classif. Perf.

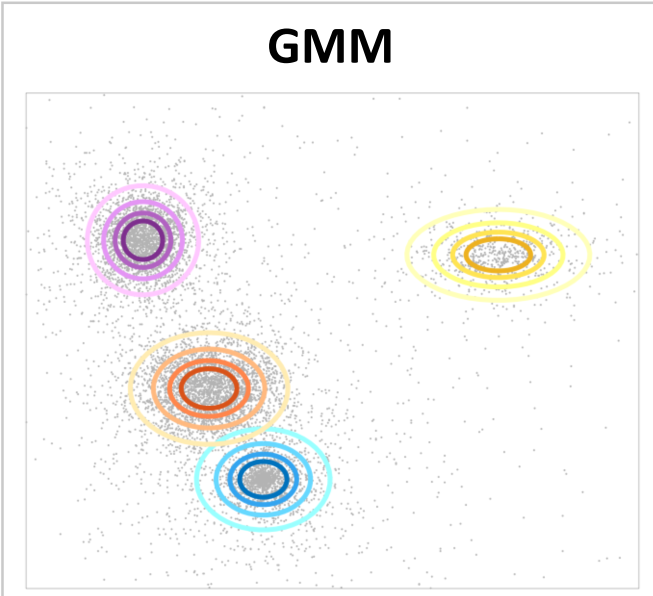


- Twice faster than k-means
- 4 orders of magnitude more memory efficient

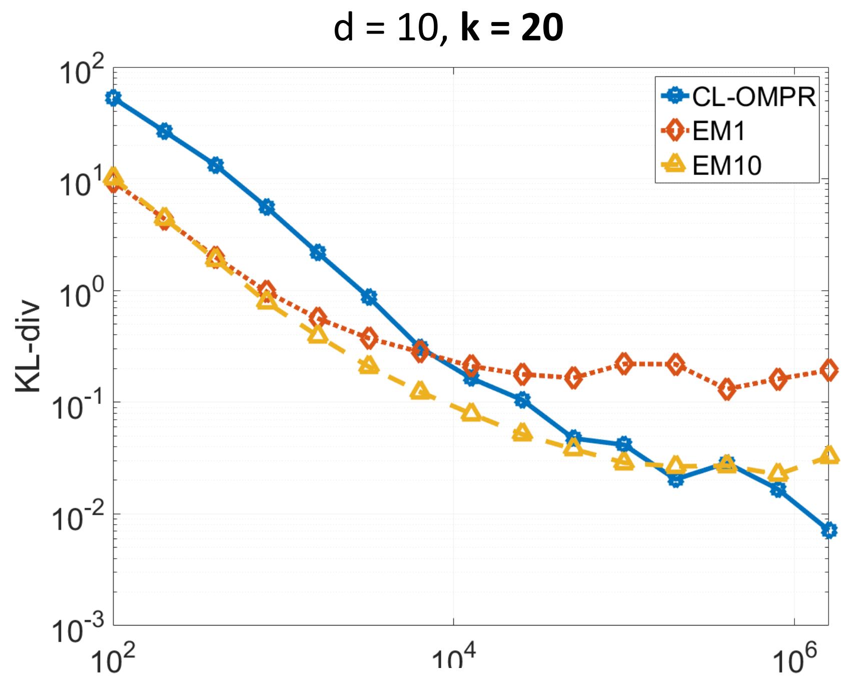
Gaussian mixture models



Gaussian mixture models

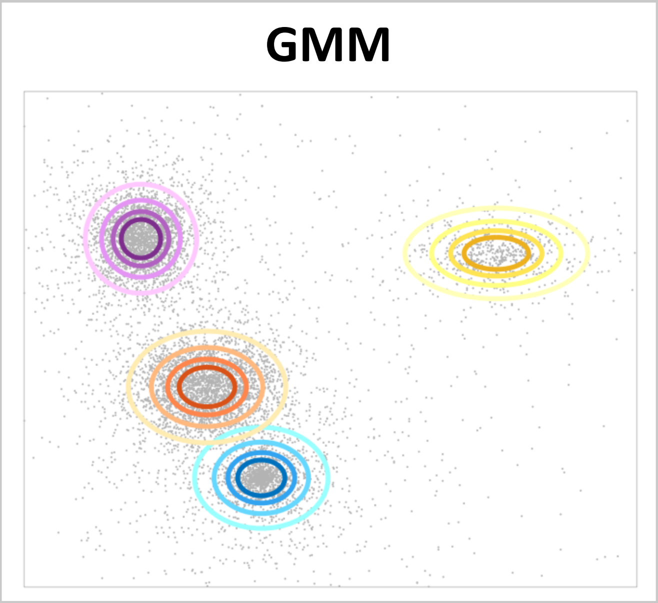


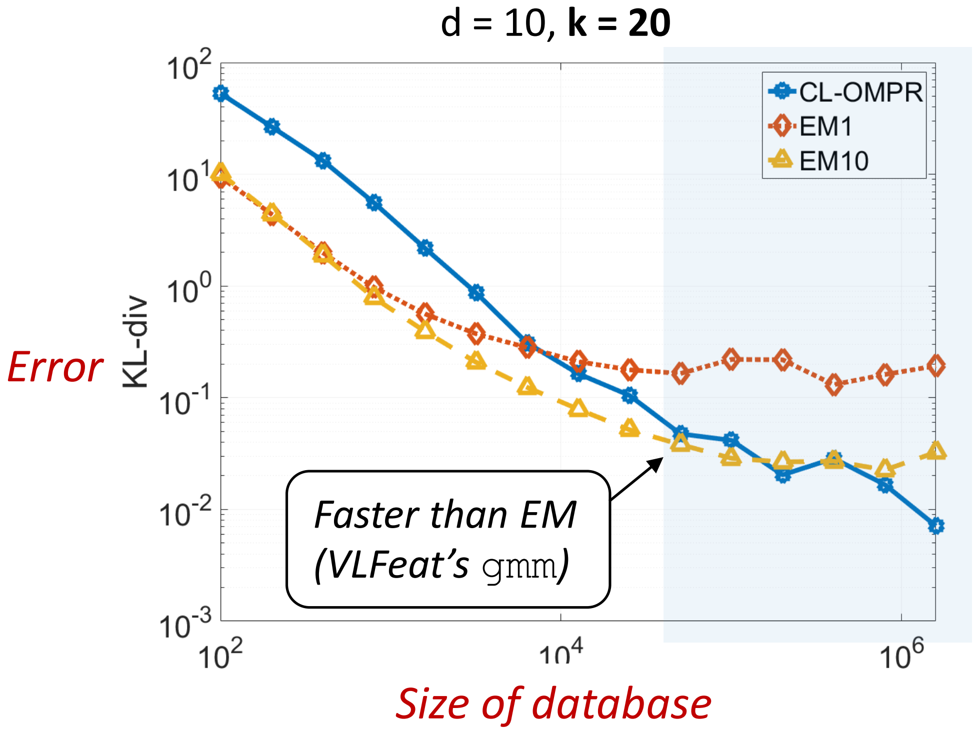
Error



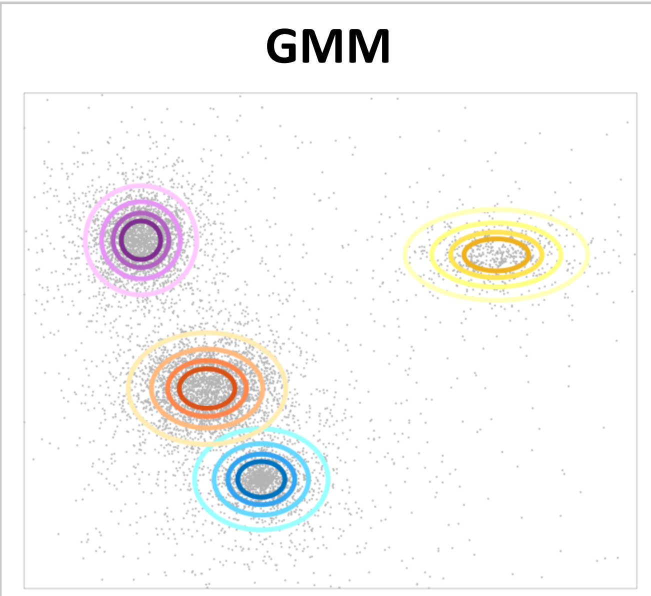
Size of database

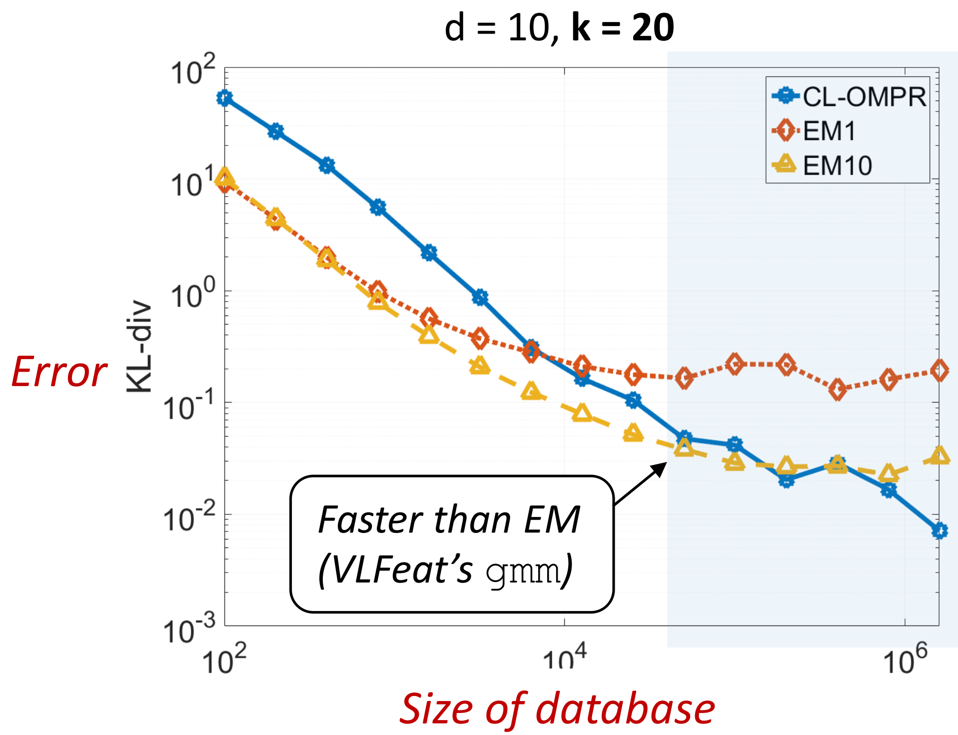
Gaussian mixture models





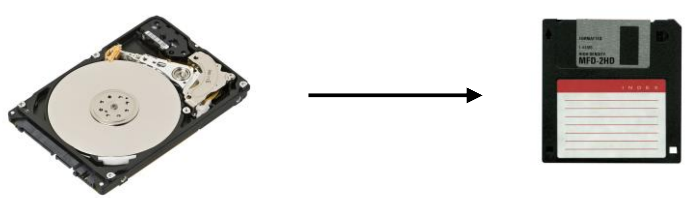
Gaussian mixture models





Application: ***speaker verification*** *[Reynolds 2000]* (d=12, k=64)

- EM on 300 000 vectors : **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

② Total variation regularization: a dual certificate analysis

③ Conclusion, outlooks

Outline

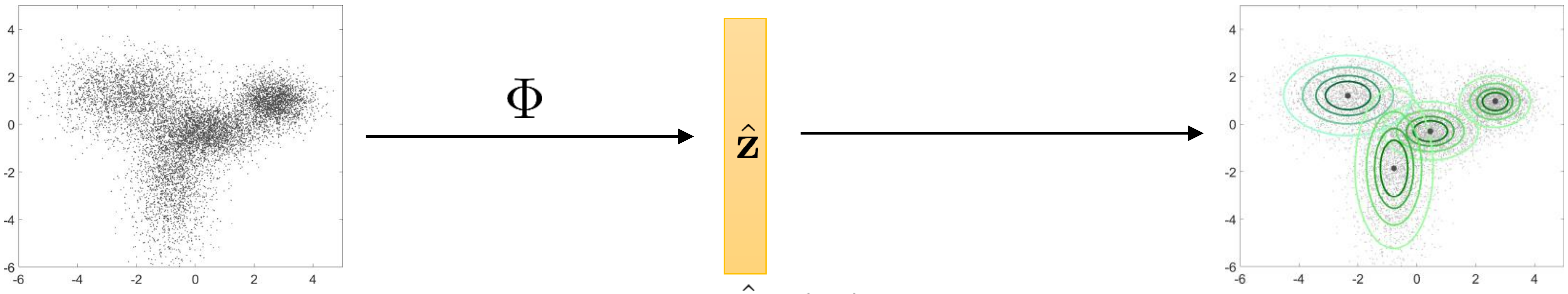
① Information-preservation guarantees: a RIP analysis

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

② Total variation regularization: a dual certificate analysis

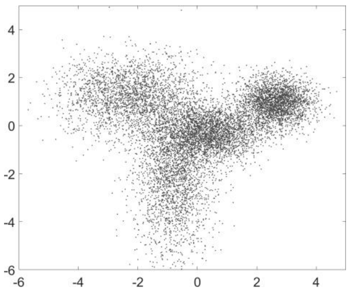
③ Conclusion, outlooks

Recall: Linear inverse problem



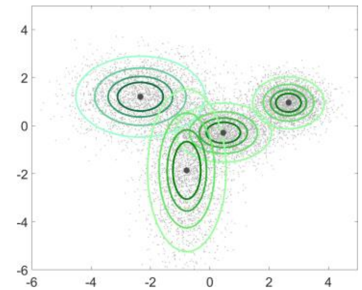
$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X)$$

Recall: Linear inverse problem



Φ

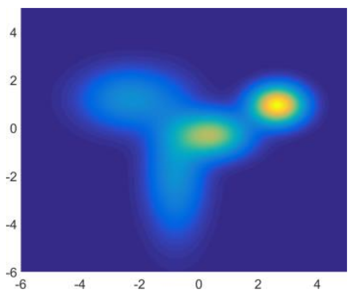
$\hat{\mathbf{z}}$



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

$\mathcal{A}$

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

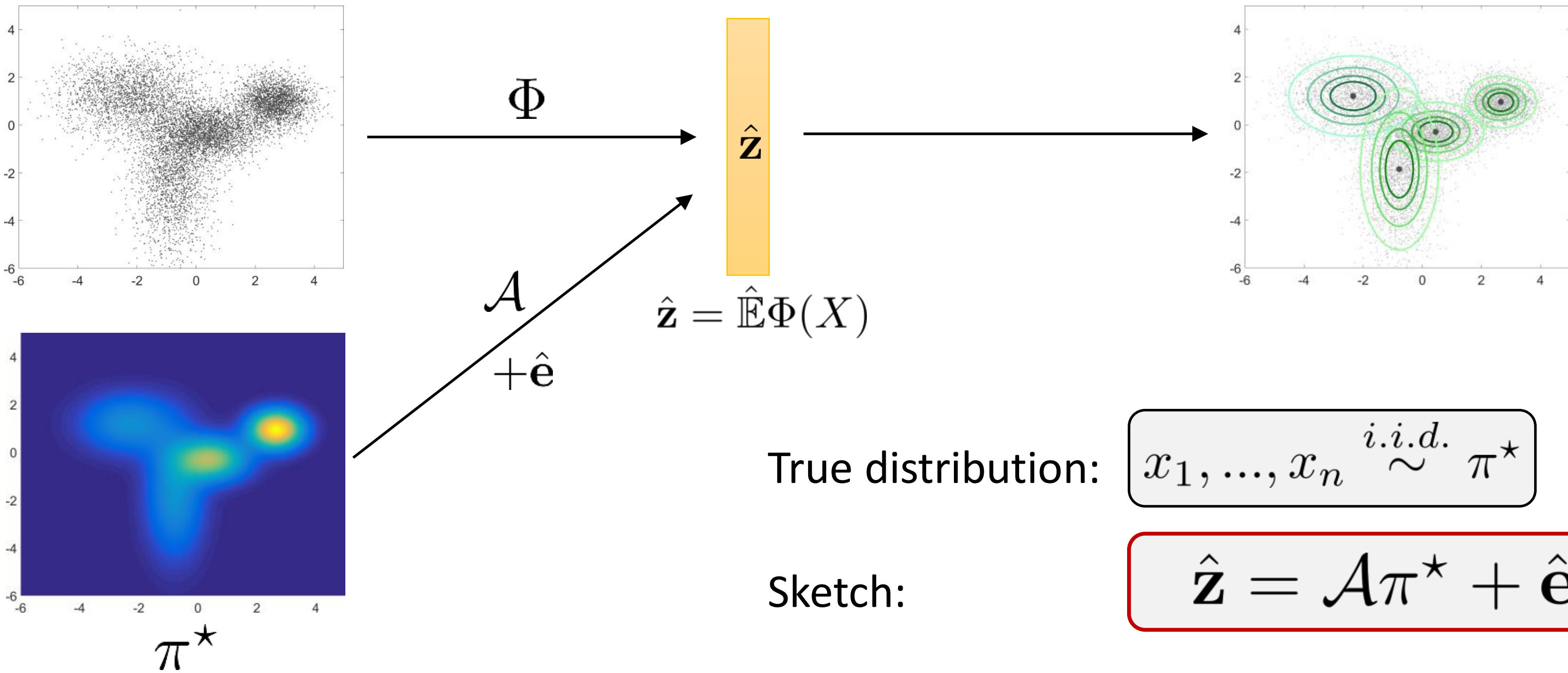


$\pi^\star$

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

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

Recall: Linear inverse problem

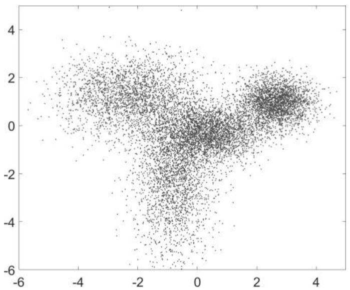


True distribution: $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \pi^\star$

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

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

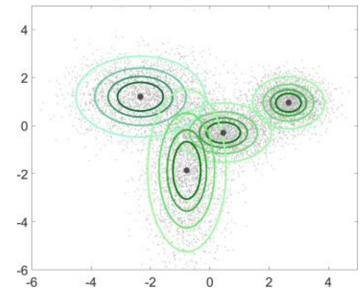
Recall: Linear inverse problem



Φ

$\hat{\mathbf{z}}$

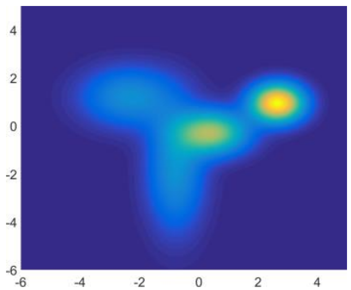
Best algorithm possible



$\mathcal{A}$

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

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



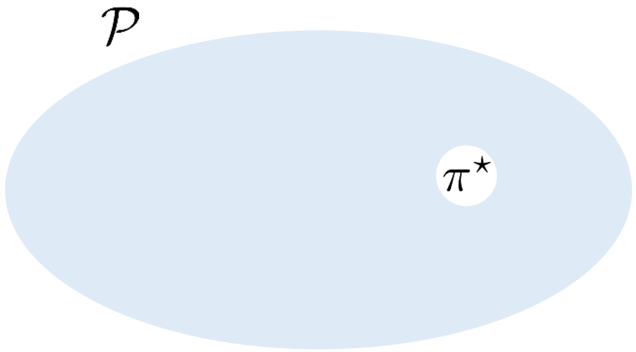
$\pi^\star$

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

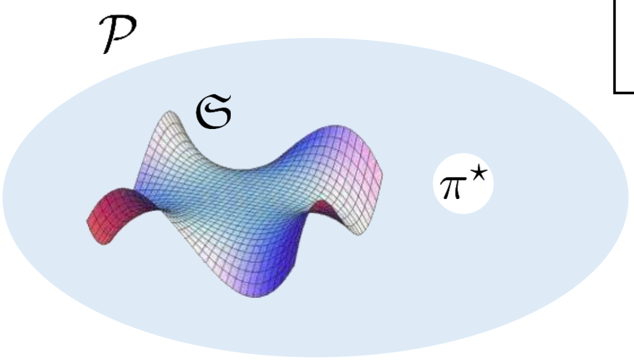
Sketch: $\hat{\mathbf{z}} = \mathcal{A}\pi^\star + \hat{\mathbf{e}}$

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

Information preservation guarantees



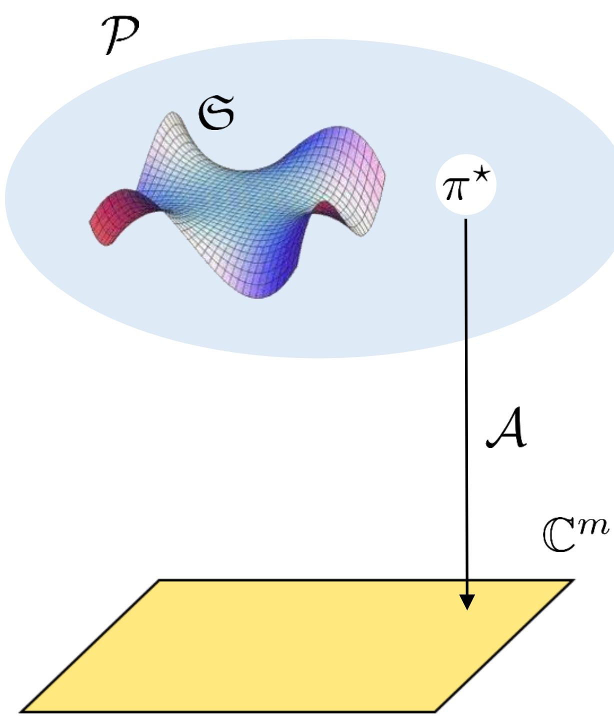
Information preservation guarantees



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

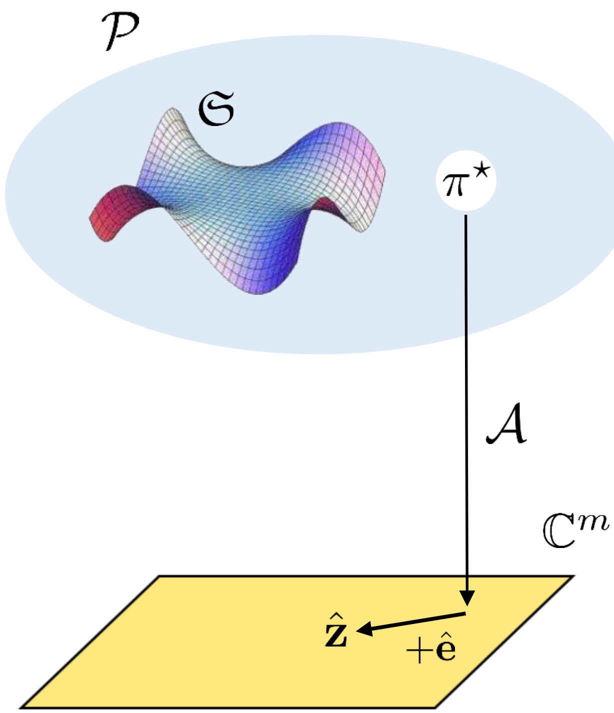
Information preservation guarantees

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

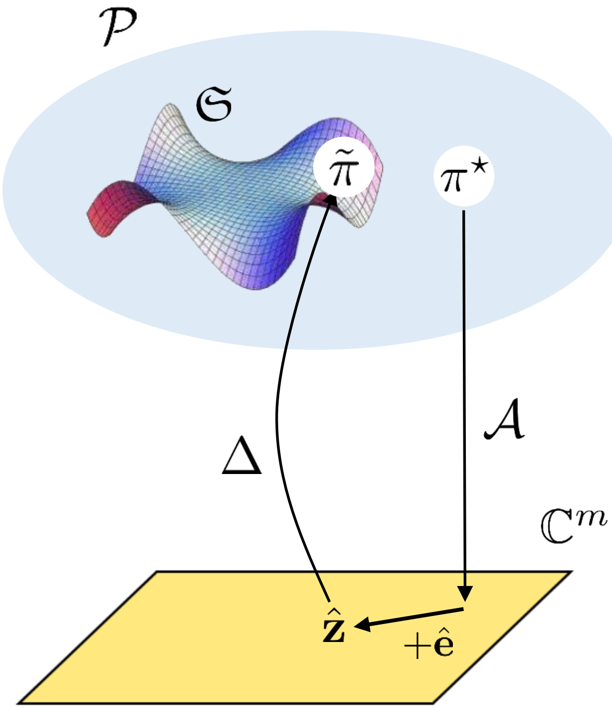


Information preservation guarantees

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



Information preservation guarantees



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

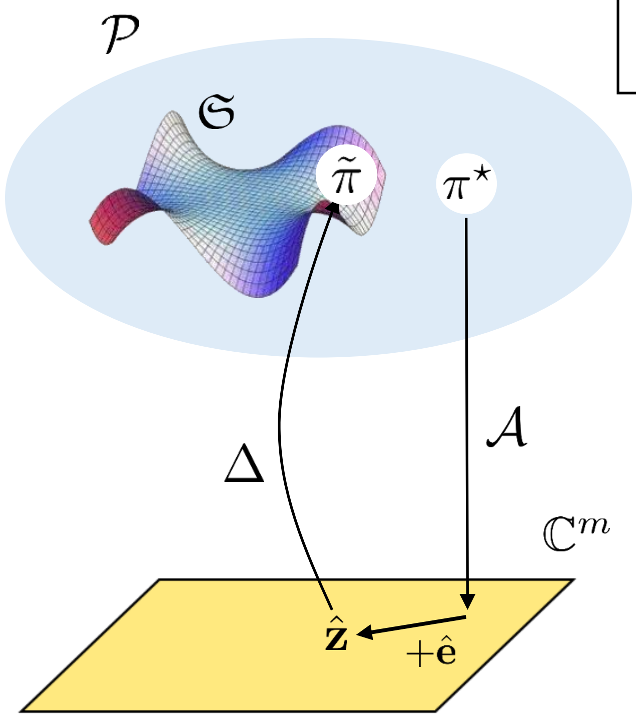
Goal

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

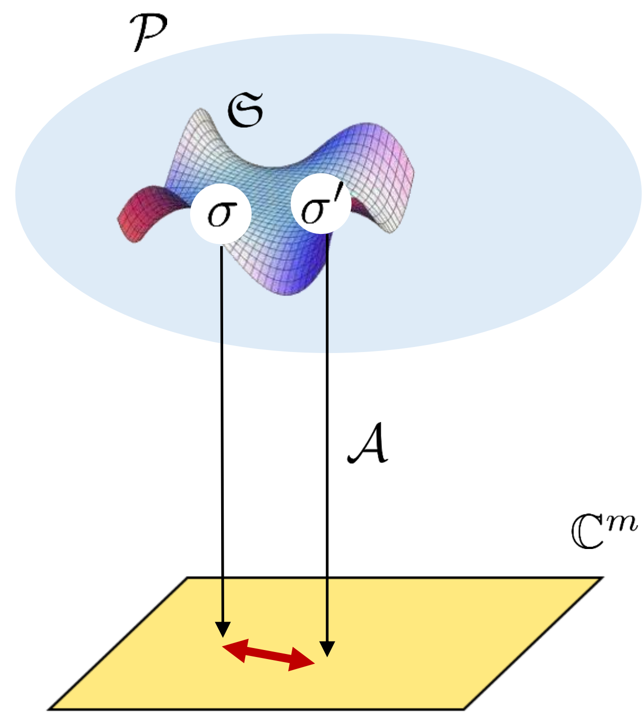
« Instance-optimal » decoder

Information preservation guarantees

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



$\Longrightarrow$



Goal

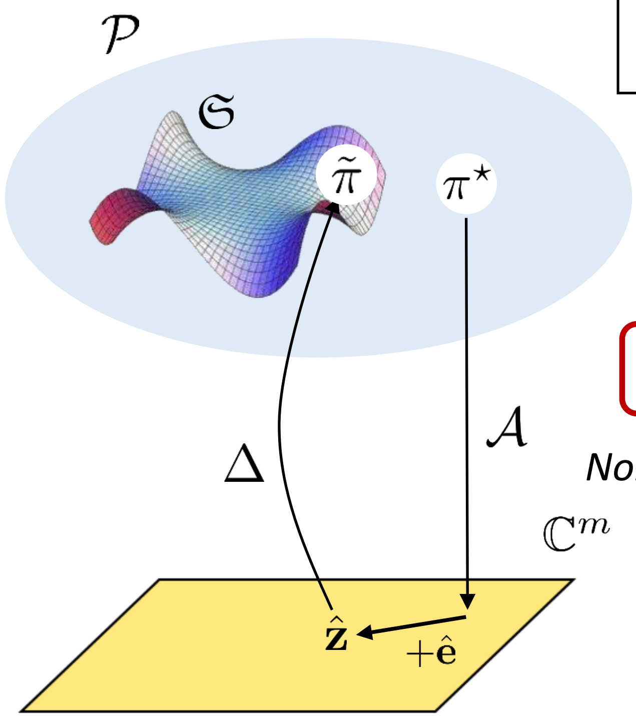
Prove the existence of a *decoder* Δ robust to noise and stable to modeling error.

« Instance-optimal » decoder

Lower Restricted Isometry Property

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

Information preservation guarantees



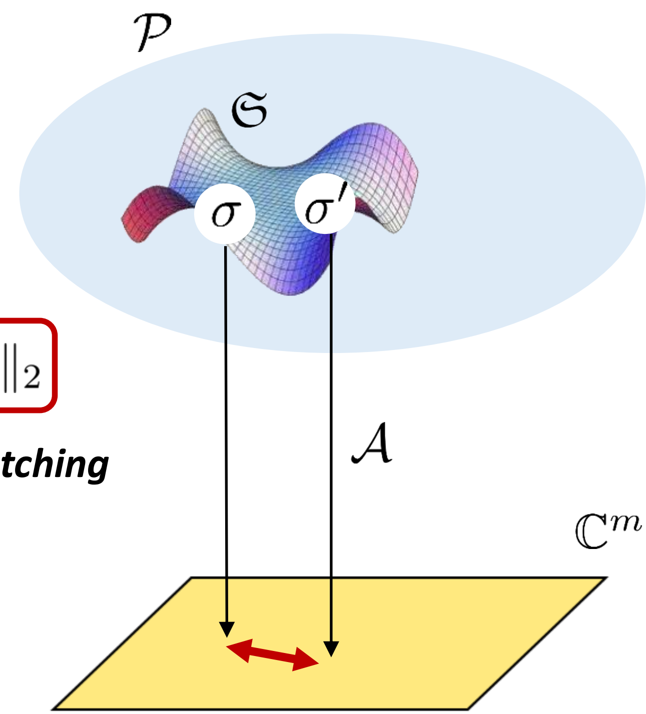
$\mathfrak{S}$ *: Model set of « simple » distributions (eg. GMMs)*

$\Longrightarrow$

$\Longleftarrow$

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

*Non-convex **generalized moment matching***



Goal

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

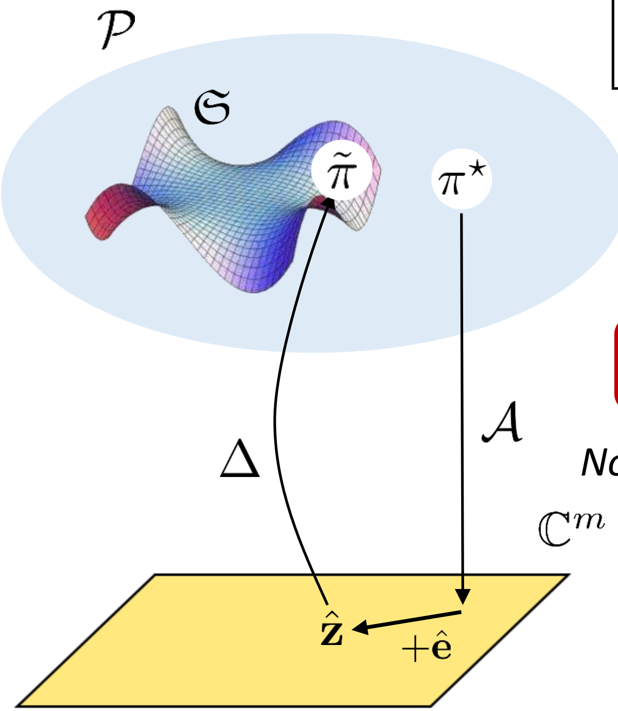
« Instance-optimal » decoder

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)*

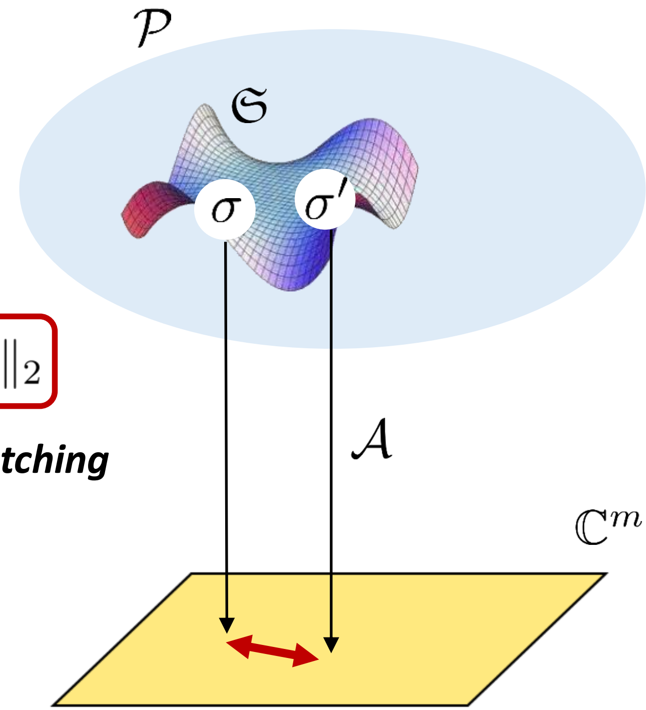


$\Longrightarrow$

$\Longleftarrow$

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

Non-convex ***generalized moment matching***



Goal

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

« Instance-optimal » decoder

Lower Restricted Isometry Property

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

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

Appropriate metric

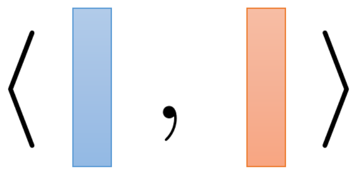
Goal: LRIP $\text{w.h.p. on } \mathcal{A}, \forall \sigma, \sigma' \in \mathfrak{S}, \; \|\sigma - \sigma'\| \lesssim \|\mathcal{A}\sigma - \mathcal{A}\sigma'\|_2.$

Appropriate metric

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

Reproducing kernel:

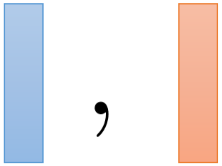
$\kappa(x, x')$



Appropriate metric

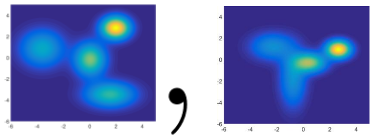
Goal: LRIP w.h.p. on $\mathcal{A}$, $\forall \sigma, \sigma' \in \mathfrak{S}$, $\|\sigma - \sigma'\| \lesssim \|\mathcal{A}\sigma - \mathcal{A}\sigma'\|_2$.

Reproducing kernel:

$\kappa(x, x')$ $\langle$  $\rangle$

Kernel mean

$$\kappa(\pi, \pi') = \mathbb{E}\kappa(X, X')$$

$\langle$  $\rangle$

Appropriate metric

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

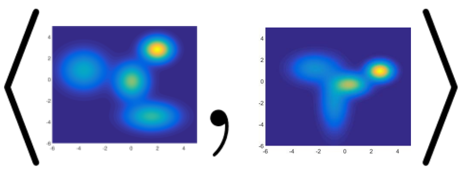
Reproducing kernel:

$\kappa(x, x')$ $\quad \langle$ ▮ , ▮ $\rangle$

Φ : random features *[Rahimi2007]* to approximate κ

Kernel mean

$$\kappa(\pi, \pi') = \mathbb{E}\kappa(X, X')$$

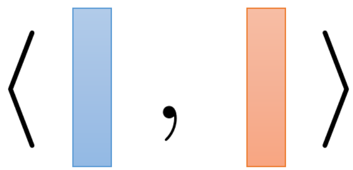


Appropriate metric

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

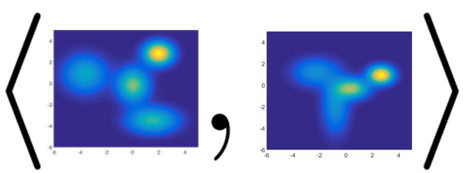
Reproducing kernel:

$\kappa(x, x')$



Kernel mean

$$\kappa(\pi, \pi') = \mathbb{E}\kappa(X, X')$$



Φ : random features *[Rahimi2007]* to approximate κ

$$\mathcal{A}\pi = \mathbb{E}_\pi \Phi(X)$$

Basis for LRIP

$$\|\pi - \pi'\|_\kappa^2 \approx \|\mathcal{A}\pi - \mathcal{A}\pi'\|_2^2$$

Proof strategy (1)

Reformulation of the LRIP

Goal: LRIP $\|\sigma - \sigma'\|_\kappa \lesssim \|\mathcal{A}(\sigma - \sigma')\|_2$

Proof strategy (1)

Reformulation of the LRIP

Goal: LRIP $\|\sigma - \sigma'\|_\kappa \lesssim \|\mathcal{A}(\sigma - \sigma')\|_2$

$$\Leftrightarrow \boxed{1 \lesssim \|\mathcal{A}(\frac{\sigma-\sigma'}{\|\sigma-\sigma'\|_\kappa})\|_2}$$

Proof strategy (1)

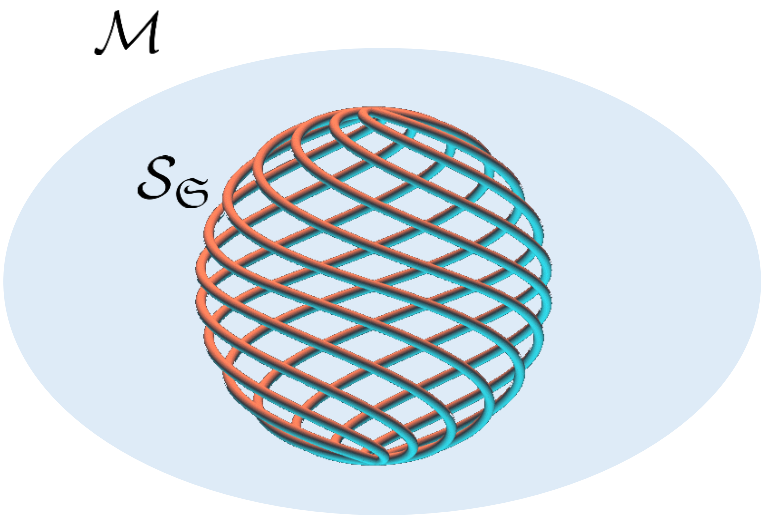
Reformulation of the LRIP

Goal: LRIP $\|\sigma - \sigma'\|_\kappa \lesssim \|\mathcal{A}(\sigma - \sigma')\|_2$

$$\Leftrightarrow 1 \lesssim \|\mathcal{A}(\tfrac{\sigma-\sigma'}{\|\sigma-\sigma'\|_\kappa})\|_2$$

Definition: Normalized Secant set

$$\mathcal{S}_{\mathfrak{S}} = \left\{ \tfrac{\sigma-\sigma'}{\|\sigma-\sigma'\|_\kappa};\ \sigma, \sigma' \in \mathfrak{S} \right\}$$



Proof strategy (1)

Reformulation of the LRIP

Goal: LRIP $\|\sigma - \sigma'\|_\kappa \lesssim \|\mathcal{A}(\sigma - \sigma')\|_2$

$$\Leftrightarrow \boxed{1 \lesssim \left\|\mathcal{A}\left(\frac{\sigma - \sigma'}{\|\sigma-\sigma'\|_\kappa}\right)\right\|_2}$$

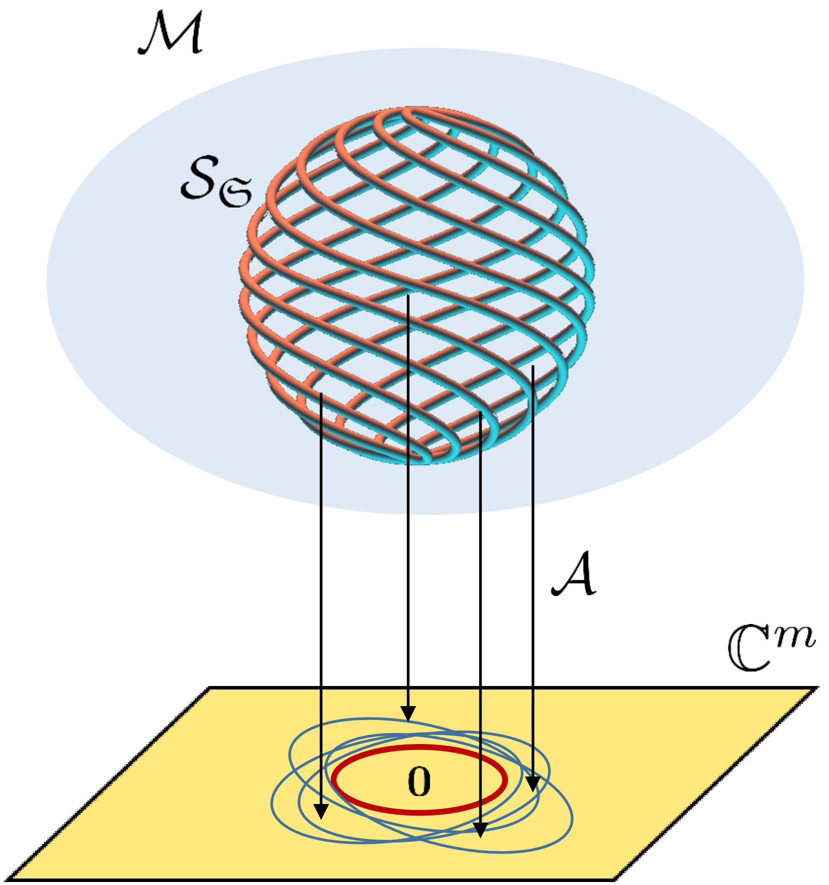
Definition: Normalized Secant set

$$\mathcal{S}_{\mathfrak{S}} = \left\{\frac{\sigma-\sigma'}{\|\sigma-\sigma'\|_\kappa};\ \sigma, \sigma' \in \mathfrak{S}\right\}$$

New goal

With high probability on $\mathcal{A}$:

for all $s \in \mathcal{S}_{\mathfrak{S}}$, $1 \lesssim \|\mathcal{A}s\|_2$.



Proof strategy (2)

Goal: LRIP $\text{w.h.p. on } \mathcal{A}, \forall s \in \mathcal{S}_{\mathfrak{S}}, 1 \lesssim \|\mathcal{A}s\|_2.$

Proof strategy (2)

Goal: LRIP $\text{w.h.p. on } \mathcal{A}, \forall s \in \mathcal{S}_{\mathfrak{S}}, 1 \lesssim \|\mathcal{A}s\|_2.$

① **Pointwise LRIP:**
Concentration inequality

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

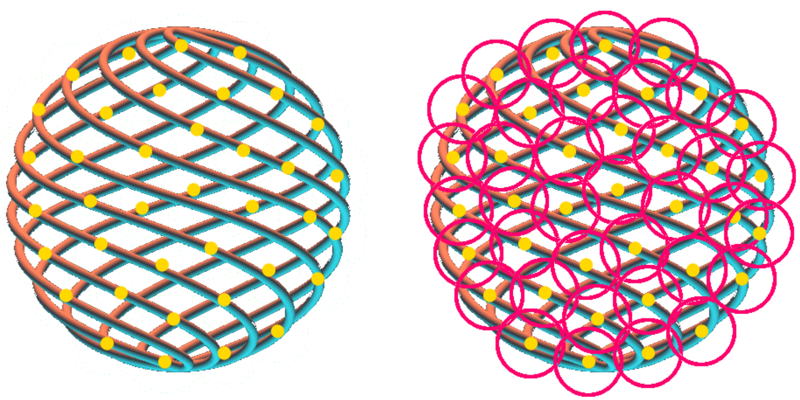
Proof strategy (2)

Goal: LRIP $\text{w.h.p. on } \mathcal{A},\ \forall s \in \mathcal{S}_{\mathfrak{G}},\ 1 \lesssim \|\mathcal{A}s\|_2.$

① **Pointwise LRIP:**
Concentration inequality

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

② **Extension to LRIP:**
covering numbers



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

Main result

Main hypothesis

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

Main result

Main hypothesis

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

Result

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

Main result

Main hypothesis

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

Result

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

Quality of pointwise LRIP

Dimensionality of the model

Main result

Main hypothesis

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

Result

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

Quality of pointwise LRIP

Dimensionality of the model

W.h.p.

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

Main result

Main hypothesis

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

Result

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

Quality of pointwise LRIP

Dimensionality of the model

W.h.p.

Modeling error

Empirical noise

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

Main result

Main hypothesis

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

Result

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

Quality of pointwise LRIP

Dimensionality of the model

W.h.p.

Modeling error

Empirical noise

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

- **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

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

Hypotheses

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

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

*(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)**

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

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

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

(*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

Result

- With respect to **log-likelihood**

Sketch size

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

Summary

With the RIP analysis:

Summary

With the RIP analysis:

- **Moment matching**: best decoder possible (instance optimal)
  - Information-preservation guarantees

Summary

With the RIP analysis:

- **Moment matching**: best decoder possible (instance optimal)
  - Information-preservation guarantees

- Fine control on modeling error, noise, and metrics
  - **Can incorporate k-means cost or log-likelihood**

Summary

With the RIP analysis:

- **Moment matching**: best decoder possible (instance optimal)
  - Information-preservation guarantees
- Fine control on modeling error, noise, and metrics
  - **Can incorporate k-means cost or log-likelihood**

Compressive Sensing:

Summary

With the RIP analysis:

- **Moment matching**: best decoder possible (instance optimal)
  - Information-preservation guarantees
- Fine control on modeling error, noise, and metrics
  - **Can incorporate k-means cost or log-likelihood**

Compressive Sensing:

- Random, dimensionality-reducing operator ✓

Summary

With the RIP analysis:

- **Moment matching**: best decoder possible (instance optimal)
  - Information-preservation guarantees
- Fine control on modeling error, noise, and metrics
  - **Can incorporate k-means cost or log-likelihood**

Compressive Sensing:

- Random, dimensionality-reducing operator ✓
- Sparsity ✓

Summary

With the RIP analysis:

- **Moment matching**: best decoder possible (instance optimal)
  - Information-preservation guarantees
- Fine control on modeling error, noise, and metrics
  - **Can incorporate k-means cost or log-likelihood**

Compressive Sensing:

- Random, dimensionality-reducing operator ✓
- Sparsity ✓
- The information is preserved ✓

Summary

With the RIP analysis:

- **Moment matching**: best decoder possible (instance optimal)
  - Information-preservation guarantees
- Fine control on modeling error, noise, and metrics
  - **Can incorporate k-means cost or log-likelihood**

Compressive Sensing:

- Random, dimensionality-reducing operator ✓
- Sparsity ✓
- The information is preserved ✓
- **Convex relaxation?** ✗

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

$$\min_\theta \|\mathcal{A}(\sum w_i \pi_{\theta_i}) - \hat{\mathbf{z}}\|_2$$

Total Variation regularization

Previously: RIP analysis

Minimization: moment matching

$$\min_\theta \|\mathcal{A}(\sum w_i \pi_{\theta_i}) - \hat{\mathbf{z}}\|_2$$

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

Total Variation regularization

Previously: RIP analysis

Minimization: moment matching

$$\min_\theta \|\mathcal{A}(\sum w_i \pi_{\theta_i}) - \hat{\mathbf{z}}\|_2$$

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

Convex relaxation (« *super resolution* »)

$$\min_\mu \frac{1}{2}\|\Psi\mu - \hat{\mathbf{z}}\|_2 + \lambda\|\mu\|_{\mathrm{TV}}$$

- μ : Radon measure
- $\Psi\mu = \int (\mathcal{A}\pi_\theta) d\mu(\theta)$
- $\|\cdot\|_{\mathrm{TV}}$: Total variation (« L1 norm »)

Total Variation regularization

Previously: RIP analysis

Minimization: moment matching

$$\min_\theta \|\mathcal{A}(\sum w_i \pi_{\theta_i}) - \hat{\mathbf{z}}\|_2$$

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

Convex relaxation (« *super resolution* »)

$$\min_\mu \frac{1}{2}\|\Psi\mu - \hat{\mathbf{z}}\|_2 + \lambda\|\mu\|_{\mathrm{TV}}$$

- μ : Radon measure
- $\Psi\mu = \int (\mathcal{A}\pi_\theta) d\mu(\theta)$
- $\|\cdot\|_{\mathrm{TV}}$: Total variation (« L1 norm »)

Convex:

- can be handled by eg Frank-Wolfe algorithm *[Boyd 2015]*, or in some cases as a SDP

Total Variation regularization

Previously: RIP analysis

Minimization: moment matching

$$\min_\theta \|\mathcal{A}(\sum w_i \pi_{\theta_i}) - \hat{\mathbf{z}}\|_2$$

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

Convex relaxation (« *super resolution* »)

$$\min_\mu \frac{1}{2}\|\Psi\mu - \hat{\mathbf{z}}\|_2 + \lambda\|\mu\|_{\mathrm{TV}}$$

- μ : Radon measure
- $\Psi\mu = \int (\mathcal{A}\pi_\theta) d\mu(\theta)$
- $\|\cdot\|_{\mathrm{TV}}$: Total variation (« L1 norm »)

Convex:

- can be handled by eg Frank-Wolfe algorithm *[Boyd 2015]*, or in some cases as a SDP

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 ?

A bit of convex analysis

Intuition: first order conditions: μ_0 solution $\Leftrightarrow$ $\frac{1}{\lambda}\Psi^\star(\Psi\mu_0 - \hat{\mathbf{z}}) \in \partial\|\mu_0\|_{\mathrm{TV}}$

A bit of convex analysis

Intuition: first order conditions: μ_0 solution $\Leftrightarrow$ $\frac{1}{\lambda}\Psi^\star(\Psi\mu_0 - \hat{\mathbf{z}}) \in \partial\|\mu_0\|_{\mathrm{TV}}$

Def. : Dual certificate (= Lagrange multiplier in the noiseless case…)

$$\eta \in \mathrm{Im}(\Psi^\star) \cap \partial\|\mu_0\|_{\mathrm{TV}}$$

A bit of convex analysis

Intuition: first order conditions: μ_0 solution $\Leftrightarrow$ $\frac{1}{\lambda}\Psi^\star(\Psi\mu_0 - \hat{\mathbf{z}}) \in \partial\|\mu_0\|_{\mathrm{TV}}$

Def. : Dual certificate (= Lagrange multiplier in the noiseless case…)

$$\eta \in \mathrm{Im}(\Psi^\star) \cap \partial\|\mu_0\|_{\mathrm{TV}}$$

What is a dual certificate?

A bit of convex analysis

Intuition: first order conditions: μ_0 solution $\Leftrightarrow$ $\frac{1}{\lambda}\Psi^\star(\Psi\mu_0 - \hat{\mathbf{z}}) \in \partial\|\mu_0\|_{\mathrm{TV}}$

Def. : Dual certificate (= Lagrange multiplier in the noiseless case…)

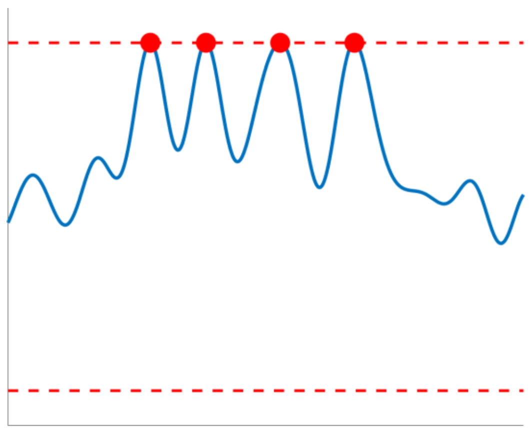
$$\eta \in \mathrm{Im}(\Psi^\star) \cap \partial\|\mu_0\|_{\mathrm{TV}}$$

What is a dual certificate?

$$\eta(\theta) = \langle \mathbf{h}, \mathcal{A}\pi_\theta \rangle$$

Such that:

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



A bit of convex analysis

Intuition: first order conditions: μ_0 solution $\Leftrightarrow$ $\frac{1}{\lambda}\Psi^\star(\Psi\mu_0 - \hat{\mathbf{z}}) \in \partial\|\mu_0\|_{\mathrm{TV}}$

Def. : Dual certificate (= Lagrange multiplier in the noiseless case…)

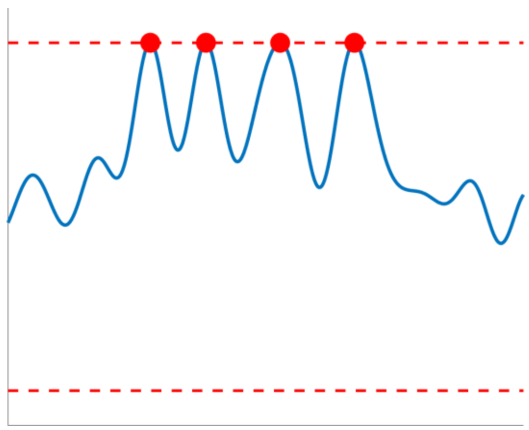
$$\eta \in \mathrm{Im}(\Psi^\star) \cap \partial\|\mu_0\|_{\mathrm{TV}}$$

What is a dual certificate?

$$\eta(\theta) = \langle \mathbf{h}, \mathcal{A}\pi_\theta \rangle$$

Such that:

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



Ensures uniqueness and robustness…

Strategy: going back to random features

Step 1: study full kernel

Strategy: going back to random features

Step 1: study full kernel

$$\bar\eta \in \mathrm{Span}\left\{\kappa(\theta_i,\cdot), \partial_1\kappa(\theta_i,\cdot)\right\} \subset \mathrm{Im}(\mathbb{E}\Psi^\star)$$

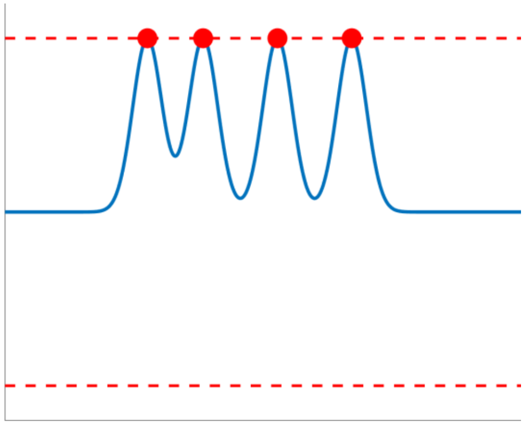
Strategy: going back to random features

Step 1: study full kernel

$$\bar{\eta} \in \operatorname{Span}\left\{\kappa(\theta_i,\cdot), \partial_1\kappa(\theta_i,\cdot)\right\} \subset \operatorname{Im}(\mathbb{E}\Psi^\star)$$

Assumptions:

- Kernel « well-behaved »
- θ_i sufficiently separated



$m = \infty$

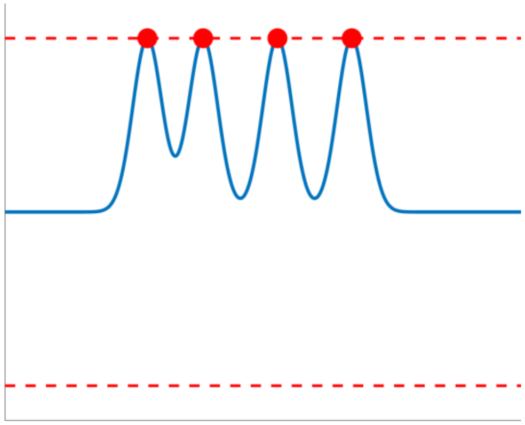
Strategy: going back to random features

Step 1: study full kernel

$$\bar{\eta} \in \mathrm{Span}\left\{\kappa(\theta_i,\cdot), \partial_1\kappa(\theta_i,\cdot)\right\} \subset \mathrm{Im}(\mathbb{E}\Psi^\star)$$

Assumptions:

- Kernel « well-behaved »
- θ_i sufficiently separated



$m = \infty$

Step 2: bounding the deviations

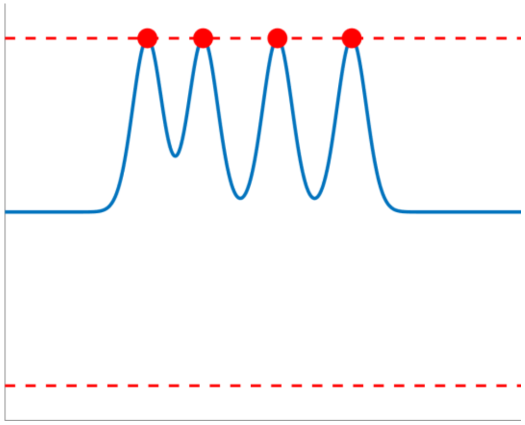
Strategy: going back to random features

Step 1: study full kernel

$$\bar{\eta} \in \mathrm{Span}\left\{\kappa(\theta_i, \cdot), \partial_1\kappa(\theta_i, \cdot)\right\} \subset \mathrm{Im}(\mathbb{E}\Psi^\star)$$

Assumptions:

- Kernel « well-behaved »
- θ_i sufficiently separated



$m = \infty$

Step 2: bounding the deviations

- Pointwise deviation (concentration ineq.)
- Covering numbers

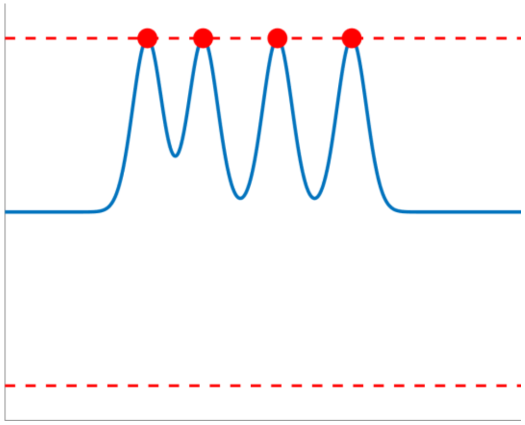
Strategy: going back to random features

Step 1: study full kernel

$$\bar{\eta} \in \mathrm{Span}\left\{\kappa(\theta_i, \cdot), \partial_1 \kappa(\theta_i, \cdot)\right\} \subset \mathrm{Im}(\mathbb{E}\Psi^\star)$$

Assumptions:

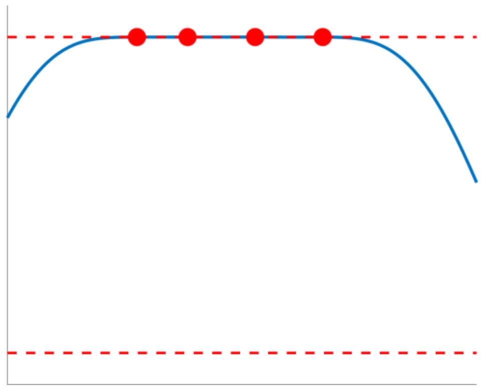
- Kernel « well-behaved »
- θ_i sufficiently separated



$m = \infty$

Step 2: bounding the deviations

- Pointwise deviation (concentration ineq.)
- Covering numbers



m=10

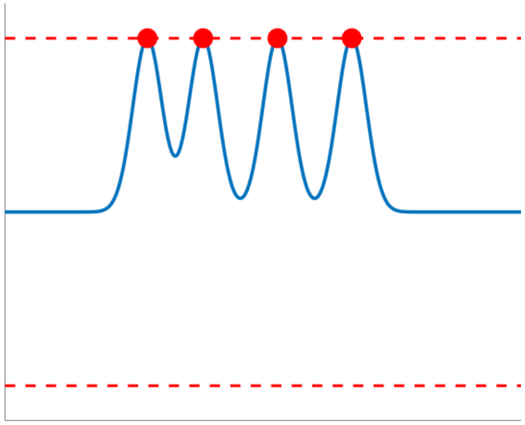
Strategy: going back to random features

Step 1: study full kernel

$$\bar{\eta} \in \text{Span}\left\{\kappa(\theta_i, \cdot), \partial_1 \kappa(\theta_i, \cdot)\right\} \subset \text{Im}(\mathbb{E}\Psi^\star)$$

Assumptions:

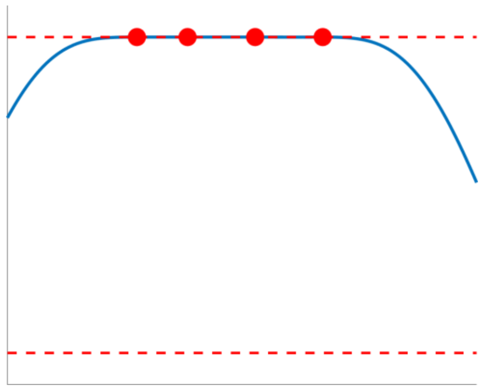
- Kernel « well-behaved »
- θ_i sufficiently separated



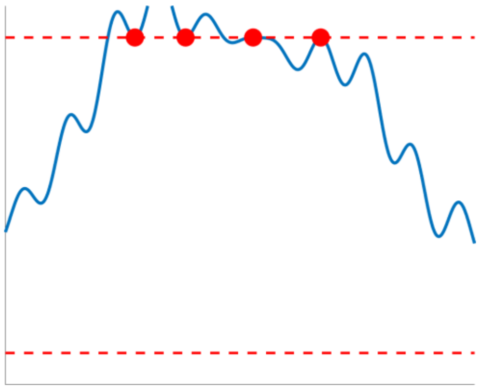
$m = \infty$

Step 2: bounding the deviations

- Pointwise deviation (concentration ineq.)
- Covering numbers



m=10



m=20

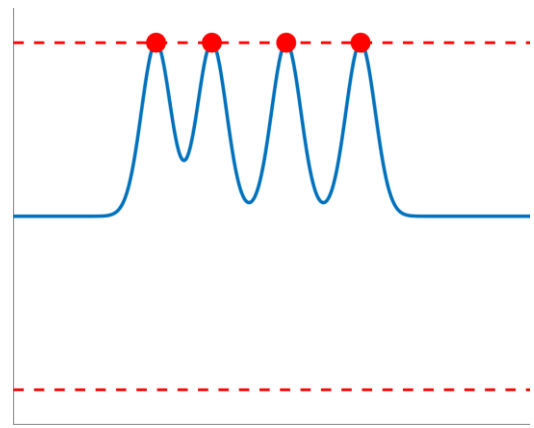
Strategy: going back to random features

Step 1: study full kernel

$$\bar\eta \in \mathrm{Span}\left\{\kappa(\theta_i,\cdot), \partial_1\kappa(\theta_i,\cdot)\right\} \subset \mathrm{Im}(\mathbb{E}\Psi^\star)$$

Assumptions:

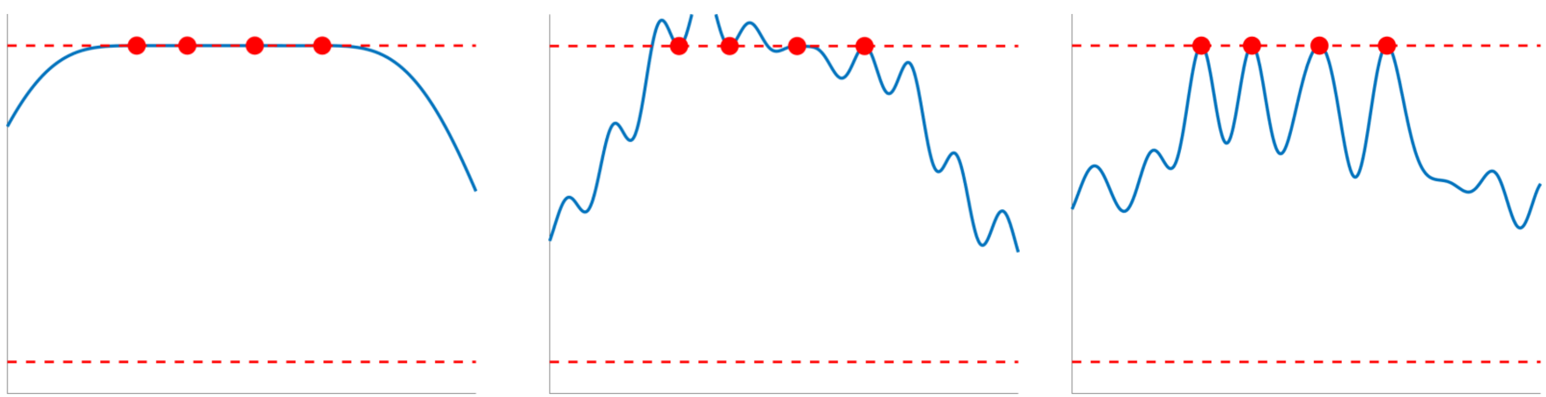
- Kernel « well-behaved »
- θ_i sufficiently separated



$m=\infty$

Step 2: bounding the deviations

- Pointwise deviation (concentration ineq.)
- Covering numbers



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

Results for separated GMM

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

1: Ideal scaling in sparsity

Results for separated GMM

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

1: Ideal scaling in sparsity

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

Results for separated GMM

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

1: Ideal scaling in sparsity

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

In progress…

Results for separated GMM

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

1: Ideal scaling in sparsity

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

↑ *In progress…*

- $\tilde{\mu}$ ***not necessarily sparse***, but:
- Mass of $\tilde{\mu}$ concentrated around true θ_i
- *Proof*: infinite-dimensional golfing scheme (new)

Results for separated GMM

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

1: Ideal scaling in sparsity

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

↑ *In progress…*

- $\tilde{\mu}$ ***not necessarily sparse***, but:
- Mass of $\tilde{\mu}$ concentrated around true θ_i
- *Proof*: infinite-dimensional golfing scheme (new)

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

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

1: Ideal scaling in sparsity

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

↑ *In progress…*

- $\tilde{\mu}$ ***not necessarily sparse***, but:
- Mass of $\tilde{\mu}$ concentrated around true θ_i
- *Proof*: infinite-dimensional golfing scheme (new)

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] (constructive!)*

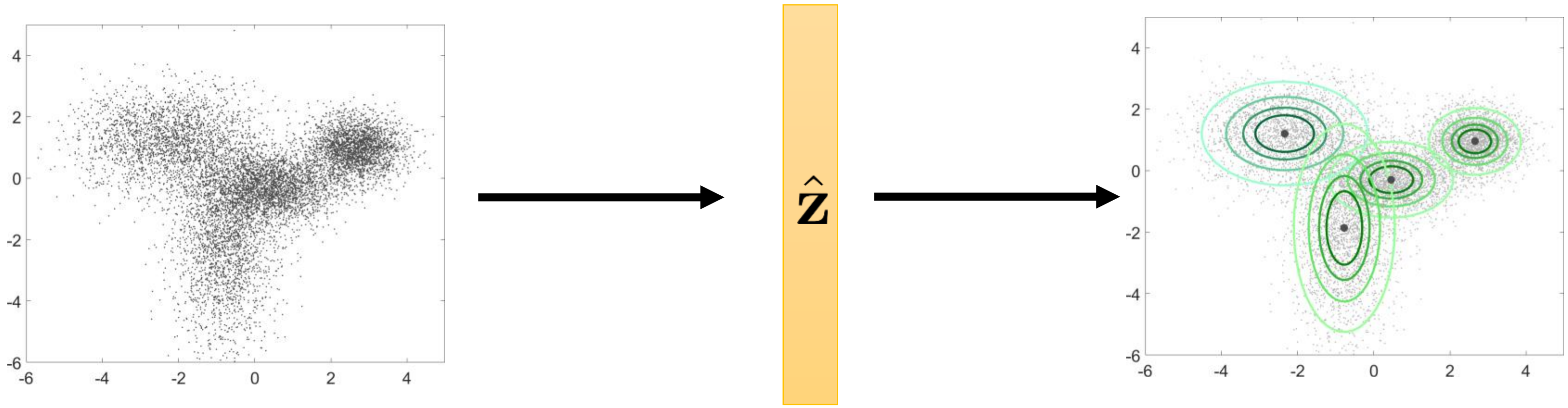
Outline

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
  - But: Non-convex minimization

Summary, outlooks

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

- **Dual certificate analysis**
  - Convex minimization
  - Does not handle modelling error
  - 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
  - But: Non-convex minimization

- **Dual certificate analysis**
  - Convex minimization
  - Does not handle modelling error
  - 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 !

- Keriven, Bourrier, Gribonval, Pérez. **Sketching for Large-Scale Learning of Mixture Models** *Information & Inference: a Journal of the IMA,* 2017. <arXiv:1606.02838>

- Keriven, Tremblay, Traonmilin, Gribonval. **Compressive k-means** *ICASSP,* 2017.

- Gribonval, Blanchard, Keriven, Traonmilin. **Compressive Statistical Learning with Random Feature Moments**. *Preprint* 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.** *Submitted*

- **Code**: sketchml.gforge.inria.fr,
  github: nkeriven