Sketching for Large-Scale Learning of Mixture Models

Nicolas Keriven

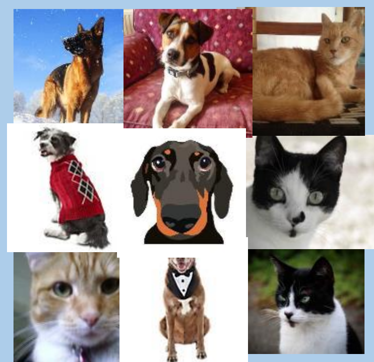
Ecole Normale Supérieure (Paris)
CFM-ENS chair in Data Science

(thesis with Rémi Gribonval at Inria Rennes)

UCL, Nov. 13th 2017

Context: machine learning

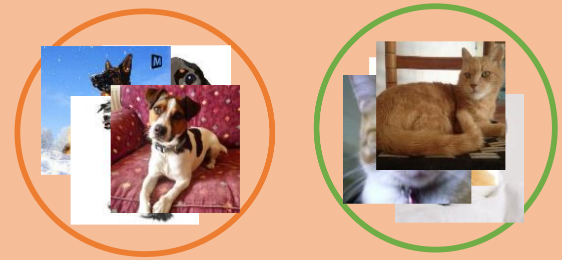
Database



Learning →

Task

- Clustering



- Classification

= cat

- etc...

Context: machine learning

Large **database**

Large elements
Billions of elements

Learning

Task

- Clustering

- Classification

= cat

- etc…

Context: machine learning

***Large* database**

Large elements
Billions of elements

Learning →

Slow, costly

Task

- Clustering

- Classification

= cat

- etc…

Context: machine learning

***Large* database**

Large elements
Billions of elements

Learning

Slow, costly

***Distributed* database**

Task

- Clustering

- Classification

 = cat

- etc…

Context: machine learning

***Large* database**

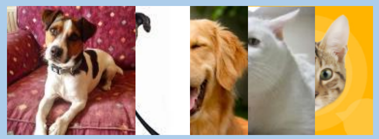
Large elements
Billions of elements

Learning

Slow, costly

***Distributed* database**

Data *Stream*

…  …

Task

- Clustering 
- Classification  = cat
- etc…

Context: machine learning

Large **database**

Large elements
Billions of elements

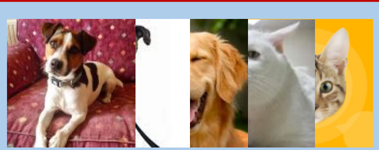
Learning

Slow, costly

Distributed **database**

Data ***Stream***

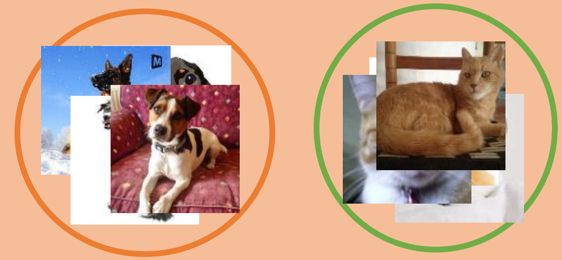
…



…

Task

- Clustering



- Classification

= cat

- etc…

Idea!

Small intermediate representation

Context: machine learning

Large **database**

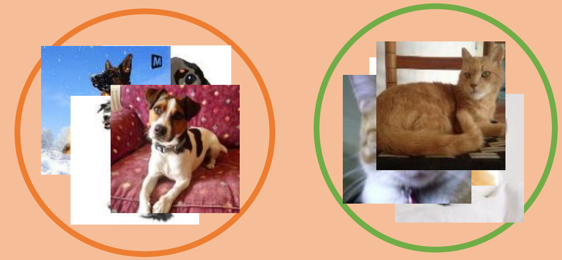
Large elements
Billions of elements

Learning

Slow, costly

Task

- Clustering



- Classification

= cat

- etc…

Distributed **database**

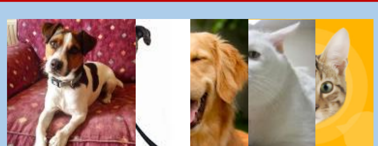
1: Compression

Idea!

Small intermediate representation

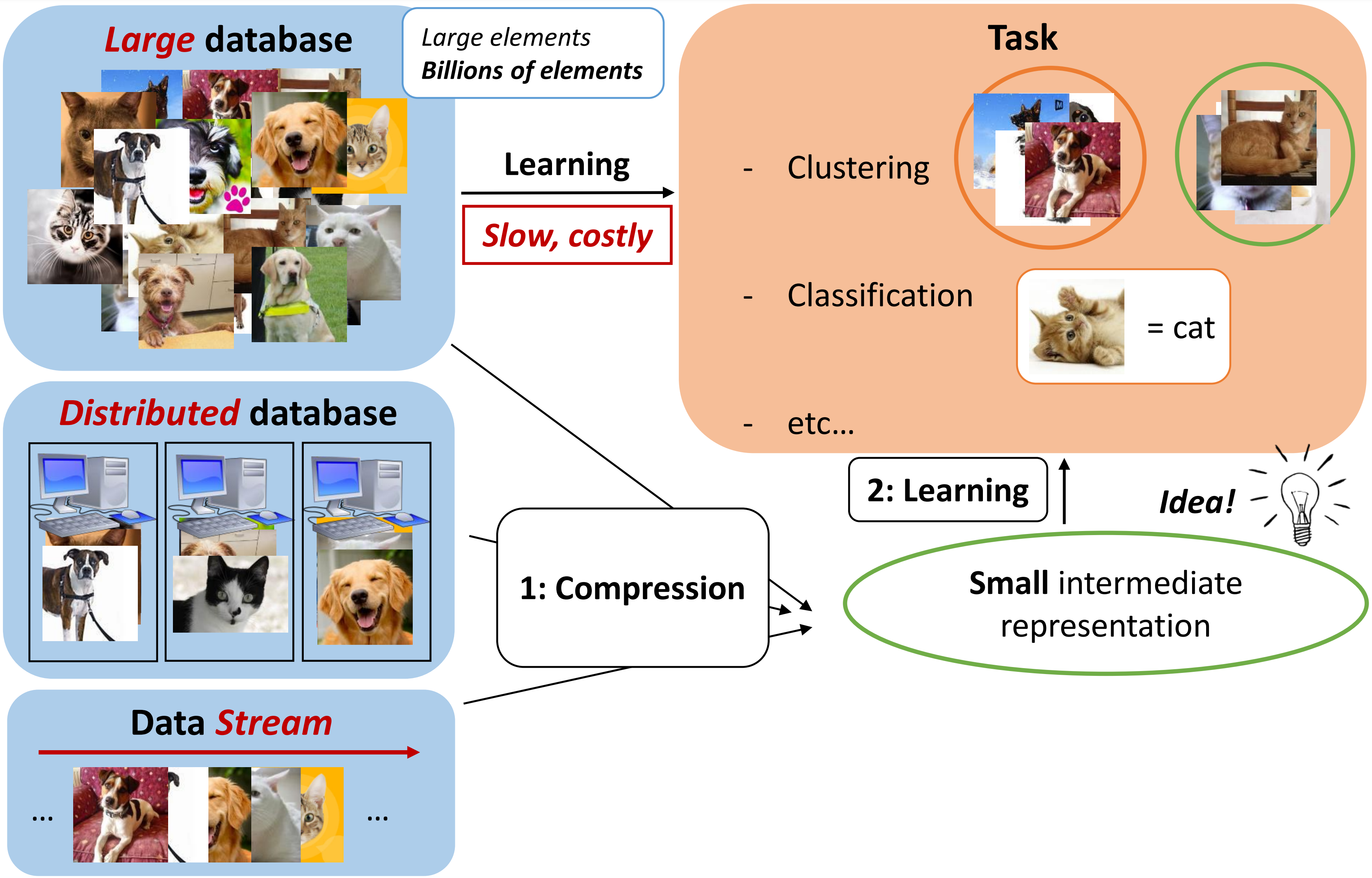
Data ***Stream***

…



…

Context: machine learning



Context: machine learning

Large **database**

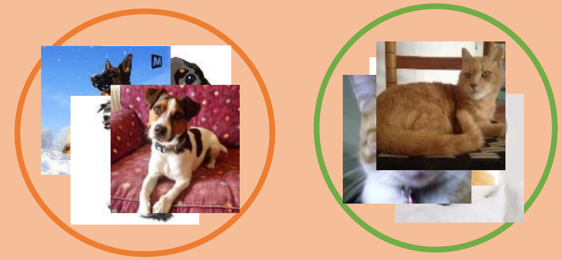
Large elements
Billions of elements

Learning

Slow, costly

Task

- Clustering



- Classification

= cat

- etc…

2: Learning

Idea!

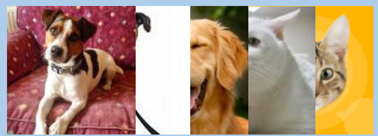
Distributed **database**

1: Compression

Small intermediate representation

Data *Stream*

…



…

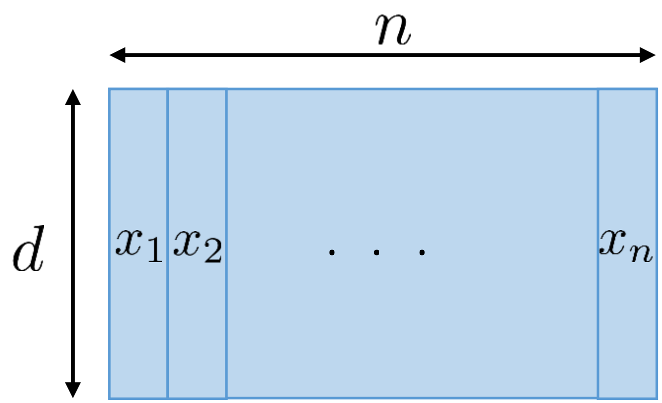
Desired properties

- **Fast** to compute (distributed, streaming, **GPU**…)
- Preserve desired **information**
- Preserve **data privacy**

Three compression schemes

Database

Feature extraction

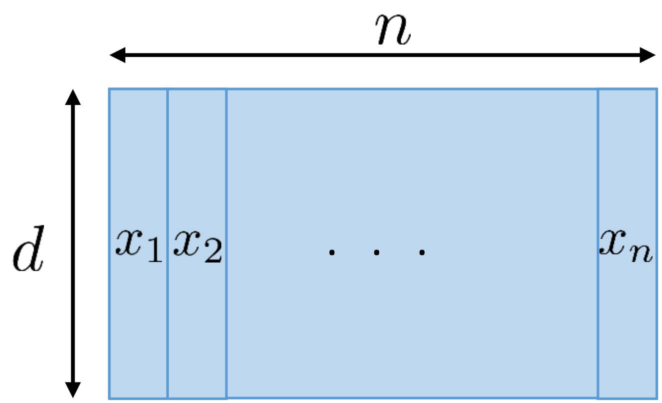


Data = Collection of vectors

Three compression schemes

Database

Feature extraction



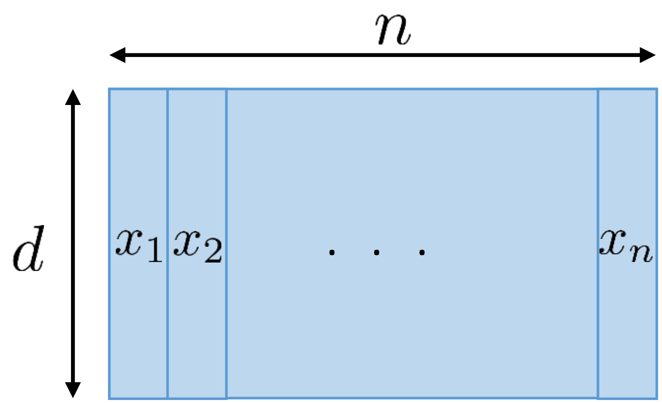
Data = Collection of vectors

Compression ?

Three compression schemes

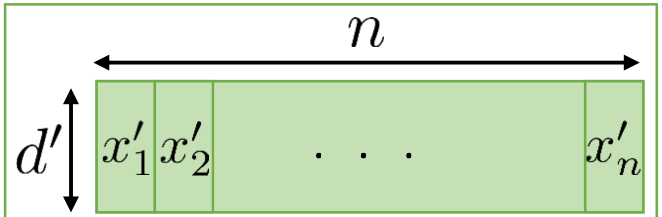
Database

Feature extraction



Data = Collection of vectors

Compression ?



Dimensionality reduction

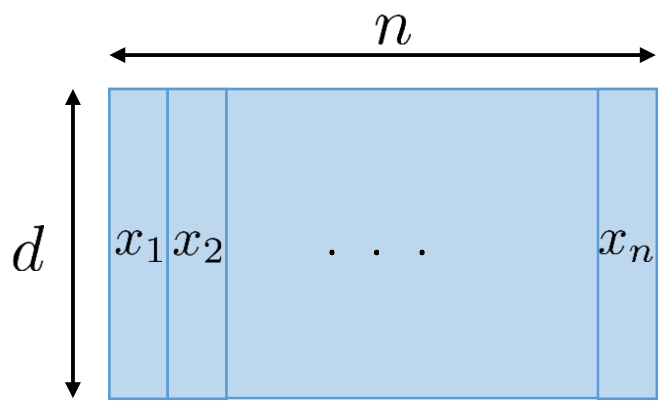
See eg *[Calderbank 2009, Boutsidis 2010]*

- Random Projection
- Feature selection

Three compression schemes

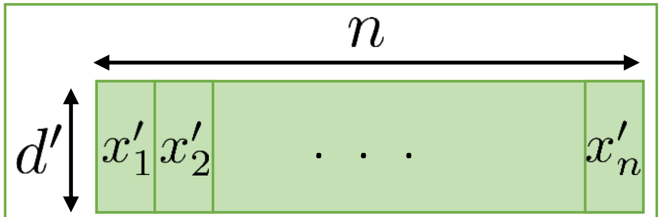
Database

Feature extraction



Data = Collection of vectors

Compression ?



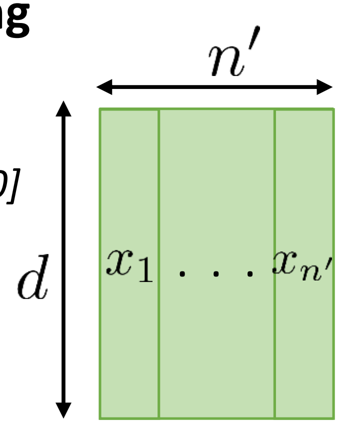
Dimensionality reduction

See eg *[Calderbank 2009, Boutsidis 2010]*

- Random Projection
- Feature selection

Subsampling coresets

See eg *[Feldman 2010]*

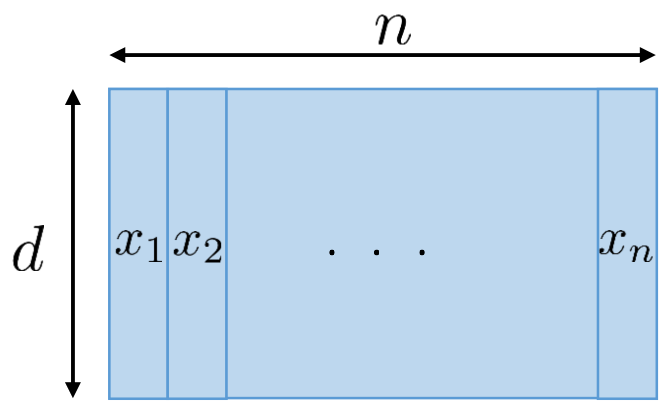


- Uniform sampling (naive)
- Adaptive sampling…

Three compression schemes

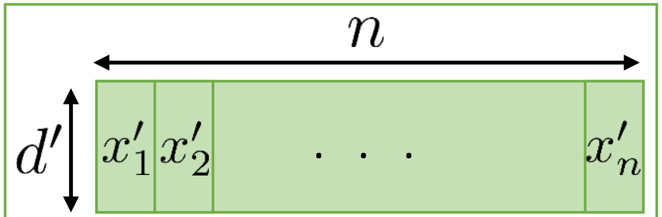
Database

Feature extraction



Data = Collection of vectors

Compression ?



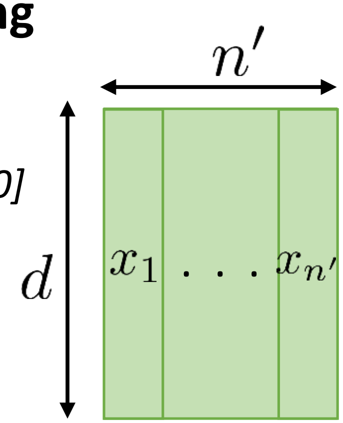
Dimensionality reduction

See eg *[Calderbank 2009, Boutsidis 2010]*

- Random Projection
- Feature selection

Subsampling coresets

See eg *[Feldman 2010]*

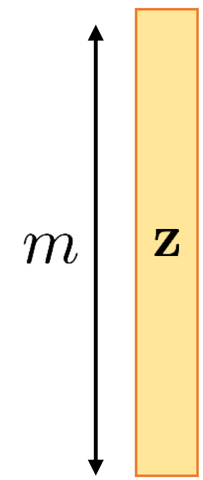


- Uniform sampling (naive)
- Adaptive sampling…

Linear sketch

See *[Thaper 2002] [Cormode 2011]*

Distributed, streaming

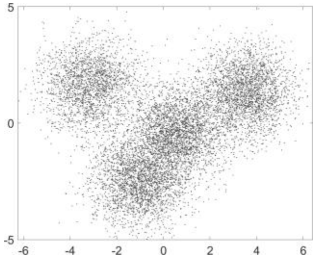


- Hash tables, histograms
- **Sketching for learning ?**

Isotropic GMM estimation *[Bourrier 2013]*

Practical illustration: sketched Gaussian Mixture Model estimation with Id cov. *[Bourrier 2013]*

Data

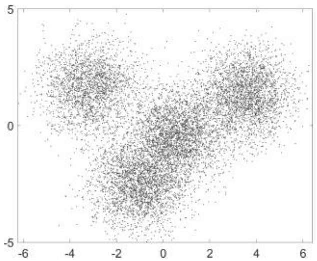


$x_1, \ldots, x_n \in \mathbb{R}^d$

Isotropic GMM estimation *[Bourrier 2013]*

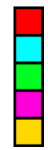
Practical illustration: sketched Gaussian Mixture Model estimation with Id cov. *[Bourrier 2013]*

Data



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

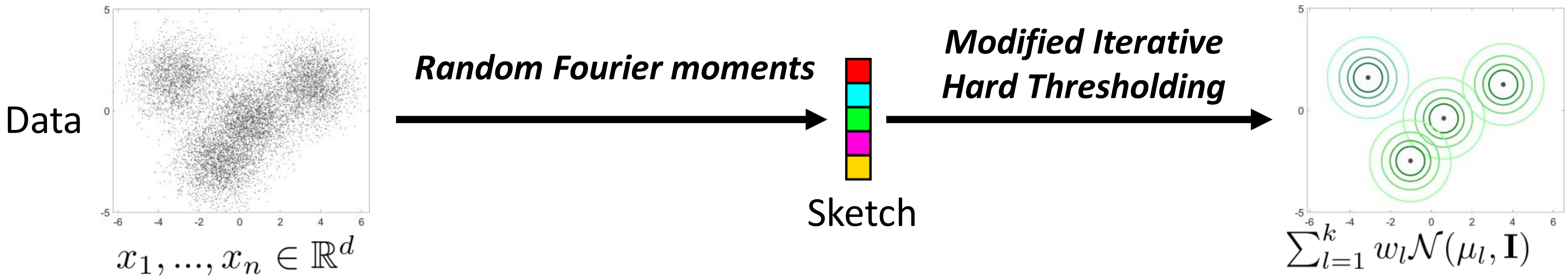
Random Fourier moments



Sketch

Isotropic GMM estimation *[Bourrier 2013]*

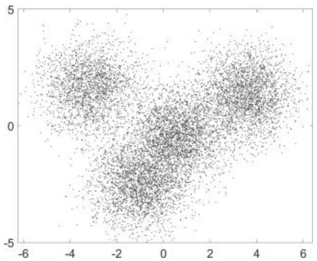
Practical illustration: sketched Gaussian Mixture Model estimation with Id cov. *[Bourrier 2013]*



Isotropic GMM estimation *[Bourrier 2013]*

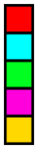
Practical illustration: sketched Gaussian Mixture Model estimation with Id cov. *[Bourrier 2013]*

Data



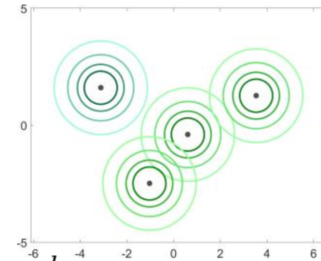
$x_1, \ldots, x_n \in \mathbb{R}^d$

Random Fourier moments



Sketch

Modified Iterative Hard Thresholding



$\sum_{l=1}^{k} w_l \mathcal{N}(\mu_l, \mathbf{I})$

Observation: *necessarily*…

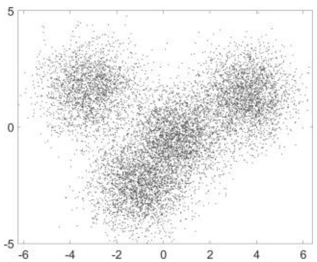
Any *linear* sketch = **empirical moments**

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

Isotropic GMM estimation *[Bourrier 2013]*

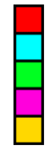
Practical illustration: sketched Gaussian Mixture Model estimation with Id cov. *[Bourrier 2013]*

Data



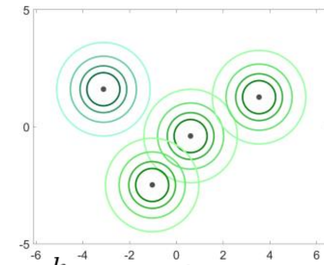
$x_1, \ldots, x_n \in \mathbb{R}^d$

Random Fourier moments



Sketch

Modified Iterative Hard Thresholding



$\sum_{l=1}^{k} w_l \mathcal{N}(\mu_l, \mathbf{I})$

Observation: *necessarily*…

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

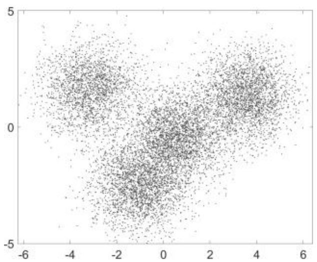
$$\boxed{\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X)} = \frac{1}{n}\sum_i \Phi(x_i)$$

$$\Phi : \mathbb{R}^d \to \mathbb{C}^m$$

Isotropic GMM estimation *[Bourrier 2013]*

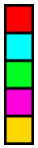
Practical illustration: sketched Gaussian Mixture Model estimation with Id cov. *[Bourrier 2013]*

Data



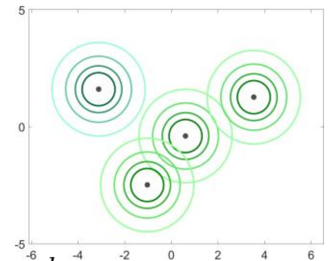
$x_1, \ldots, x_n \in \mathbb{R}^d$

Random Fourier moments



Sketch

Modified Iterative Hard Thresholding



$\sum_{l=1}^{k} w_l \mathcal{N}(\mu_l, \mathbf{I})$

Observation: *necessarily*…

Any *linear* sketch = empirical moments

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

$\Phi : \mathbb{R}^d \to \mathbb{C}^m$

… hence:

Sketch learning = moment matching

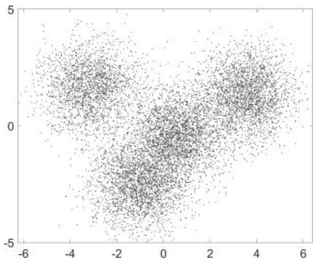
$$\boxed{\min_\theta \left\|\hat{\mathbf{z}} - \mathbb{E}_\theta \Phi(X)\right\|}$$

True moments (param. θ)

Isotropic GMM estimation *[Bourrier 2013]*

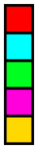
Practical illustration: sketched Gaussian Mixture Model estimation with Id cov. *[Bourrier 2013]*

Data



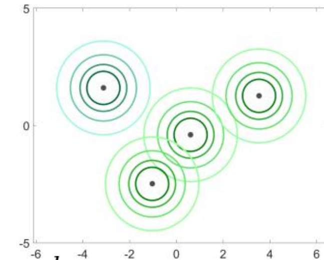
$x_1, \ldots, x_n \in \mathbb{R}^d$

Random Fourier moments



Sketch

Modified Iterative Hard Thresholding



$\sum_{l=1}^{k} w_l \mathcal{N}(\mu_l, \mathbf{I})$

Observation: *necessarily*…

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

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

$$\Phi : \mathbb{R}^d \to \mathbb{C}^m$$

… hence:

Sketch learning = moment matching

$$\boxed{\min_\theta \|\hat{\mathbf{z}} - \mathbb{E}_\theta \Phi(X)\|}$$

True moments (param. θ)

Good empirical properties of the « sketching » function Φ

- **« Sufficient » dimension m (size of the sketch)**
- **Randomly designed**

Contributions

Questions

- Generalize to other (mixture) models?
- **Theoretical guarantees?**

Contributions

Questions

- Generalize to other (mixture) models?
- **Theoretical guarantees?**

Outline

Contributions

Questions

- Generalize to other (mixture) models?
- **Theoretical guarantees?**

Outline

- **Illustration**: heuristic greedy algorithm for other sketched mixture model estimation

Contributions

Questions

- Generalize to other (mixture) models?
- **Theoretical guarantees?**

Outline

- **Illustration**: heuristic greedy algorithm for other sketched mixture model estimation
- **Theoretical analysis:** Information-preservation guarantees
  - Ideas from **Compressive Sensing**
    - Low-dimensional models (in the space of measures)
    - Random linear operators
  - Kernel mean embedding + Random features
  - Prove RIP-like conditions on sparse measures (sums of Diracs)

Outline

1. Illustration: Sketched Mixture Model Estimation

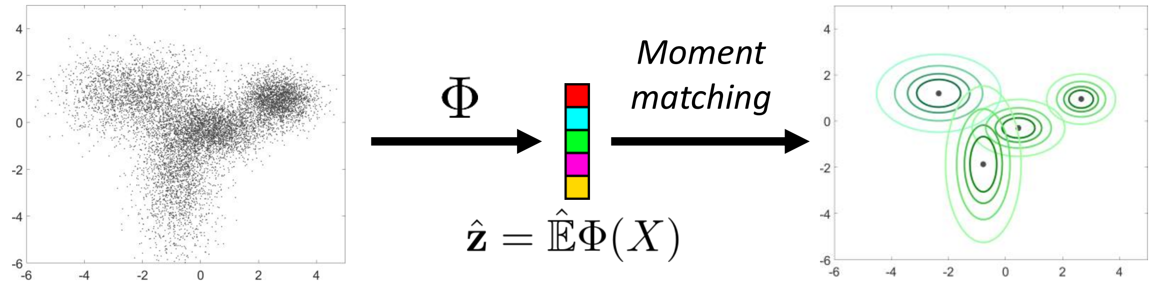
2. Information-preservation guarantees

   2.1 Restricted Isometry Property

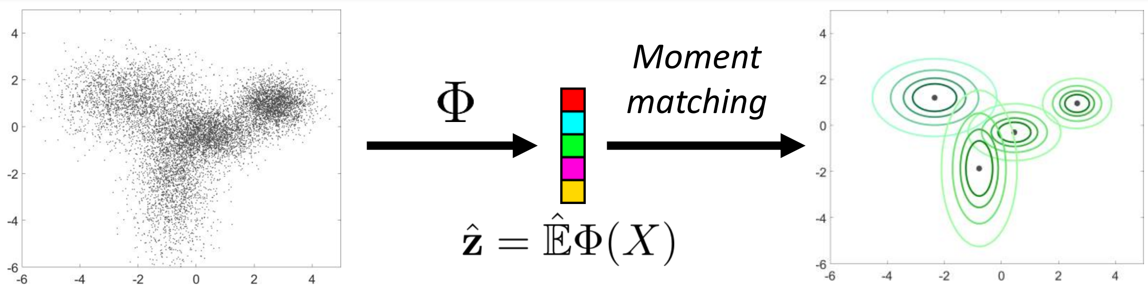
   2.2 Application: mixture model with separation assumption

3. Conclusion, outlooks

Sketched mixture model estimation [Keriven et al 2016]



Sketched mixture model estimation [Keriven et al 2016]



Goal

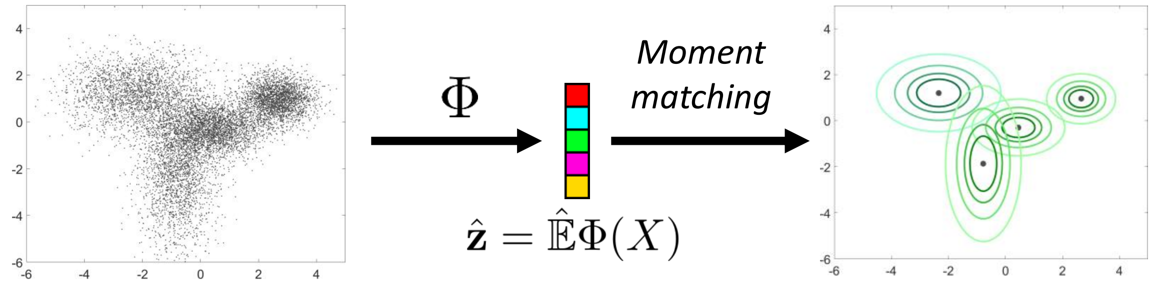
- Estimate mixture model:

$$\boxed{x_i \sim \sum_{l=1}^{k} w_l \pi_{\theta_l}}$$

$$w_l \geq 0, \quad \sum_l w_l = 1$$

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

Sketched mixture model estimation [Keriven et al 2016]



Goal

- Estimate mixture model:

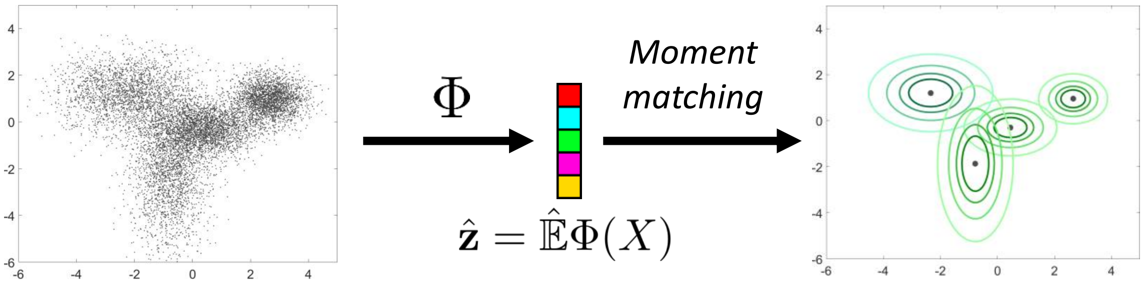
$$\boxed{x_i \sim \sum_{l=1}^{k} w_l \pi_{\theta_l}}$$

$$w_l \geq 0, \ \ \sum_l w_l = 1$$

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

Ex: $\pi_\theta = \mathcal{N}(\mu, \Sigma)$

Sketched mixture model estimation [Keriven et al 2016]



Goal

- Estimate mixture model:

$$\boxed{x_i \sim \sum_{l=1}^{k} w_l \pi_{\theta_l}}$$

$$w_l \geq 0, \quad \sum_l w_l = 1$$

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

Ex: $\pi_\theta = \mathcal{N}(\mu, \Sigma)$

Method: moment matching

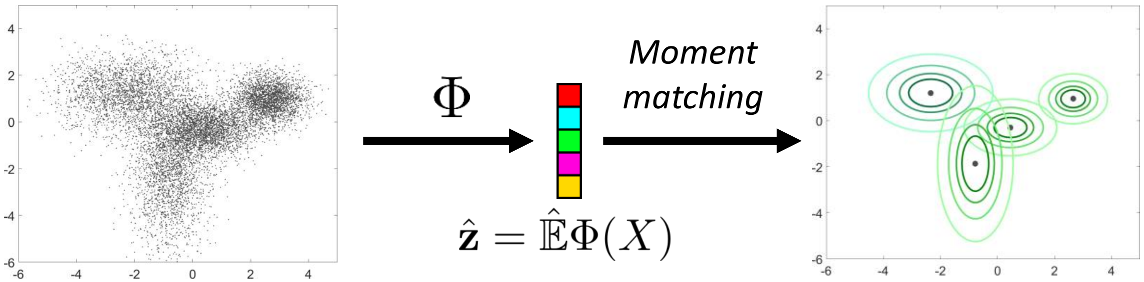
Written as

$$\boxed{\min_{\theta_l, w} \left\| \hat{\mathbf{z}} - \sum_{l=1}^{k} w_l f(\theta_l) \right\|_2}$$

where

$$\boxed{f(\theta) := \mathbb{E}_{X \sim \pi_\theta} \Phi(X)}$$

Sketched mixture model estimation [Keriven et al 2016]



Goal

- Estimate mixture model:

$$\boxed{x_i \sim \sum_{l=1}^{k} w_l \pi_{\theta_l}}$$

$$w_l \geq 0, \quad \sum_l w_l = 1$$

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

Ex: $\pi_\theta = \mathcal{N}(\mu, \Sigma)$

Method: moment matching

Written as

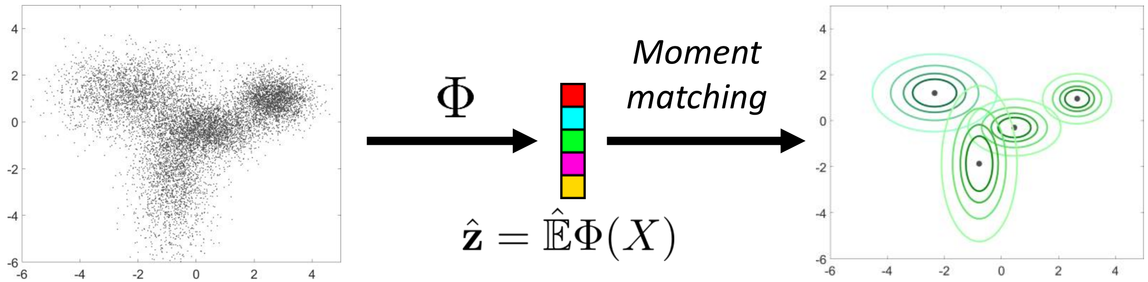
$$\boxed{\min_{\theta_l, w} \|\hat{\mathbf{z}} - \sum_{l=1}^{k} w_l f(\theta_l)\|_2}$$

where

$$\boxed{f(\theta) := \mathbb{E}_{X \sim \pi_\theta} \Phi(X)}$$

- (Highly) non-convex
- Convex relaxation? (super-resolution)
- **Proposed approach: greedy heuristic (continuous adaptation of OMP)**

Sketched mixture model estimation [Keriven et al 2016]



Goal

- Estimate mixture model:

$$\boxed{x_i \sim \sum_{l=1}^{k} w_l \pi_{\theta_l}}$$

$$w_l \geq 0, \quad \sum_l w_l = 1$$

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

Ex: $\pi_\theta = \mathcal{N}(\mu, \Sigma)$

- (Highly) non-convex
- Convex relaxation? (super-resolution)
- **Proposed approach: greedy heuristic (continuous adaptation of OMP)**

Method: moment matching

Written as

$$\boxed{\min_{\theta_l, w} \|\hat{\mathbf{z}} - \sum_{l=1}^{k} w_l f(\theta_l)\|_2}$$

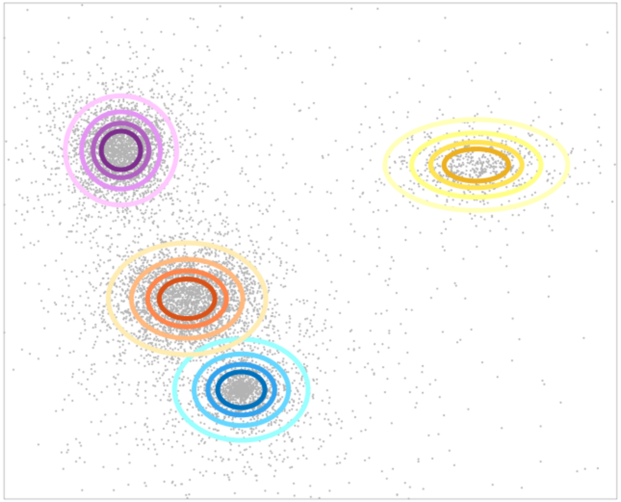
where

$$\boxed{f(\theta) := \mathbb{E}_{X \sim \pi_\theta} \Phi(X)}$$

- Can be applied as soon as $f(\theta) = \mathbb{E}_{\pi_\theta}\Phi(X)$ is computable
- **In practice: Φ random Fourier sampling** (closed-form characteristic function)

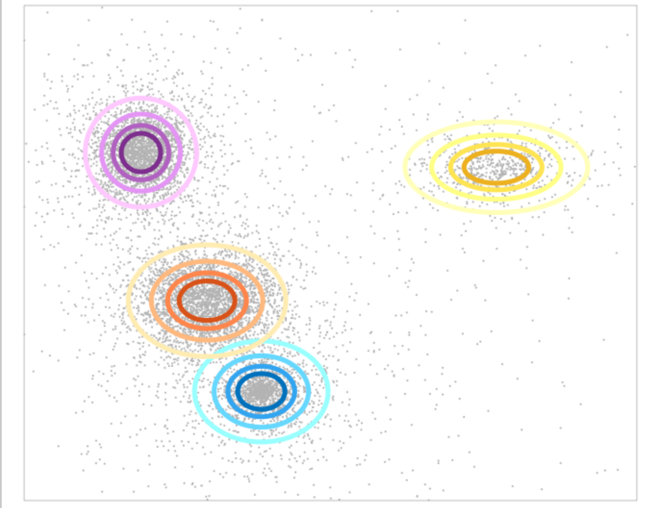
Gaussian mixture models

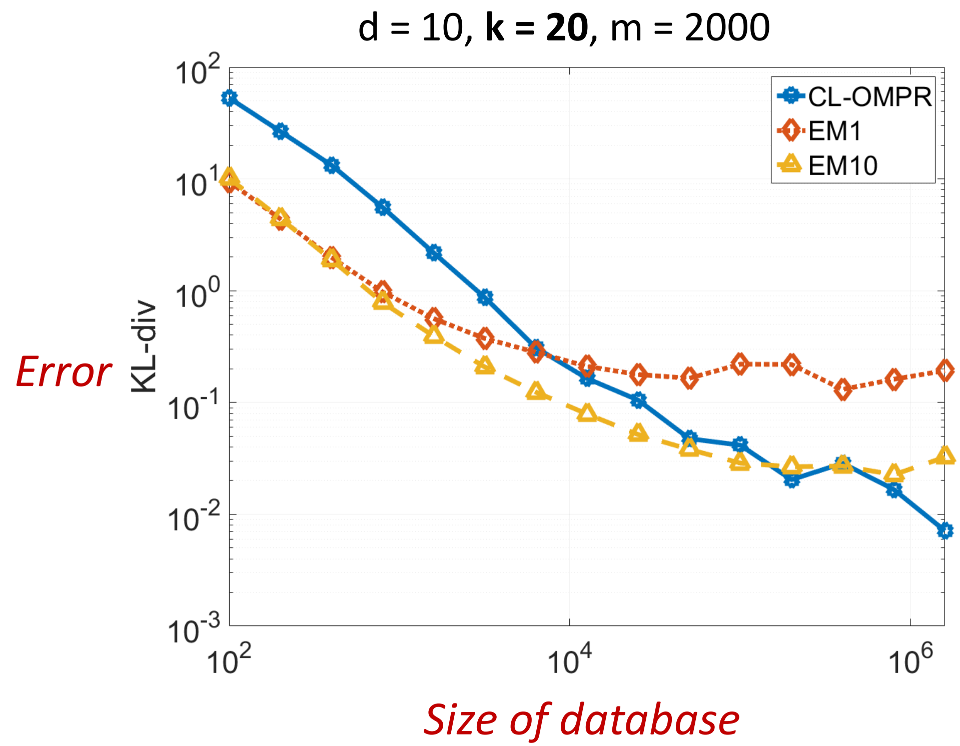
GMM with diagonal cov.



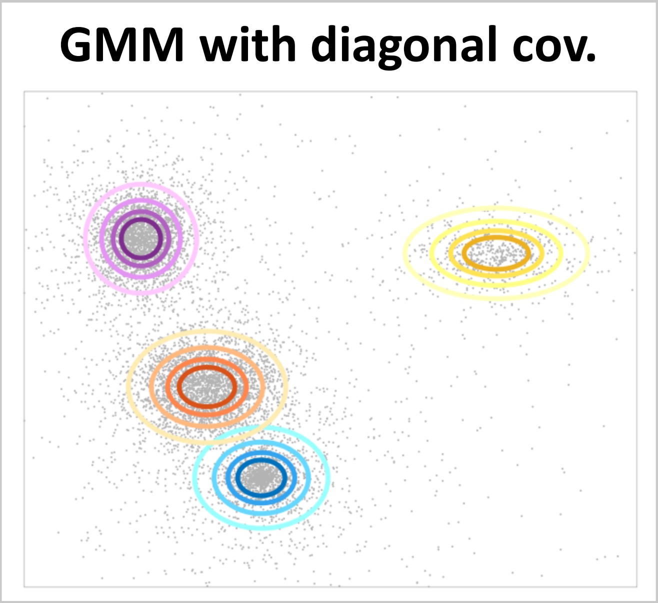
Gaussian mixture models

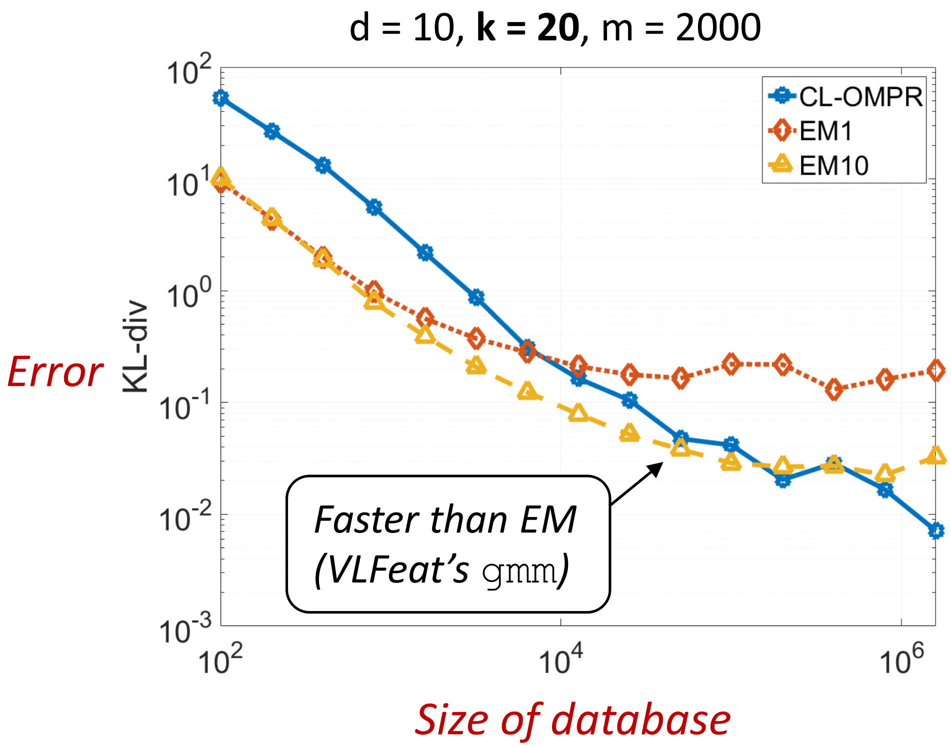
GMM with diagonal cov.





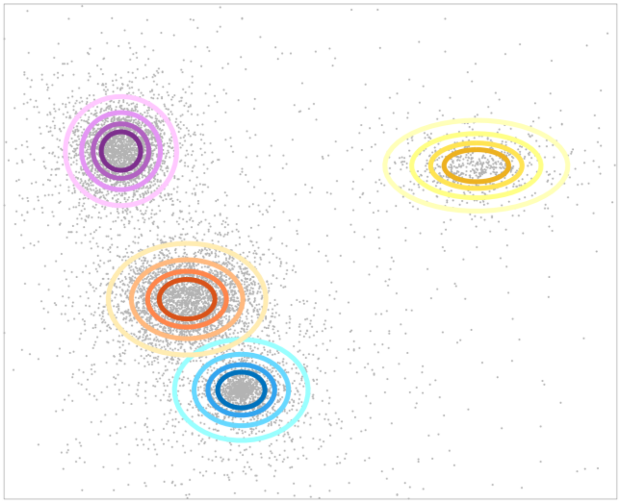
Gaussian mixture models

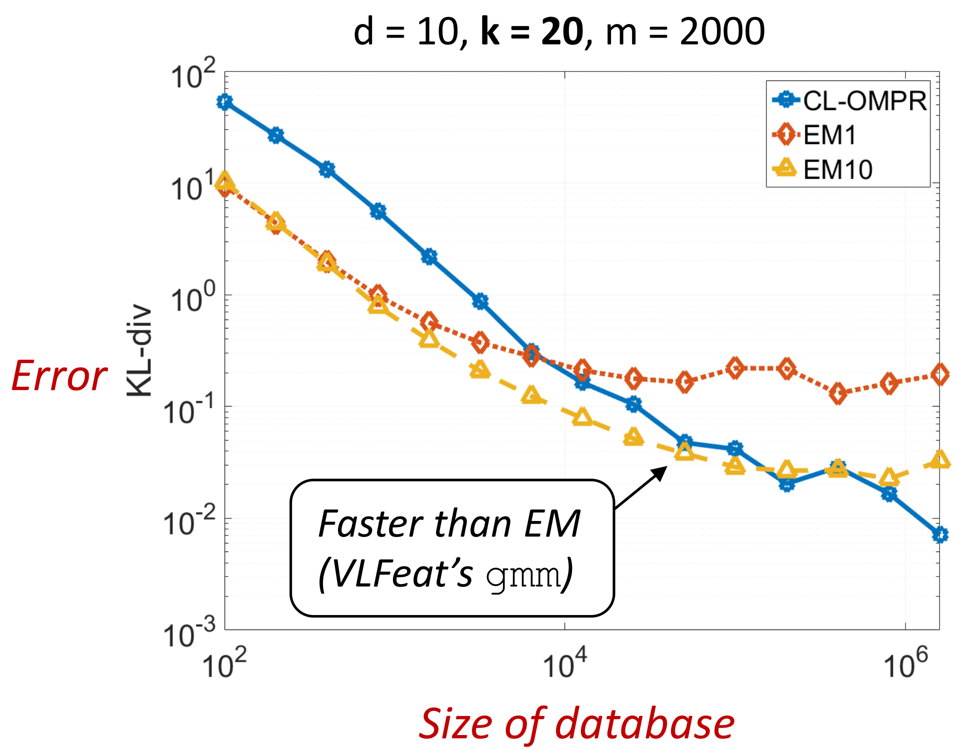




Gaussian mixture models

GMM with diagonal cov.



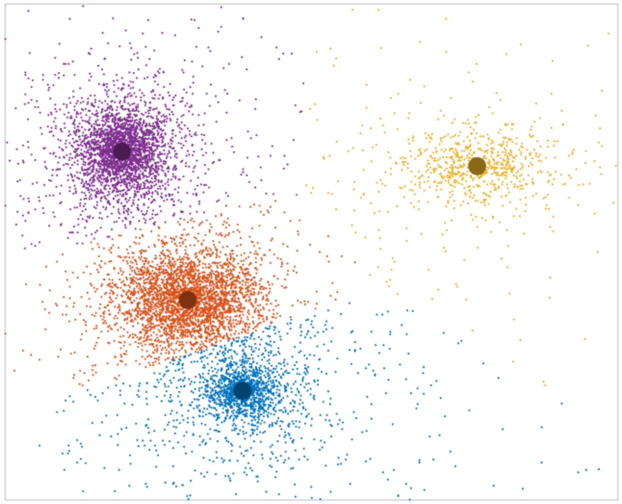


Application: ***speaker verification*** *[Reynolds 2000]* (d=12, k=64)

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

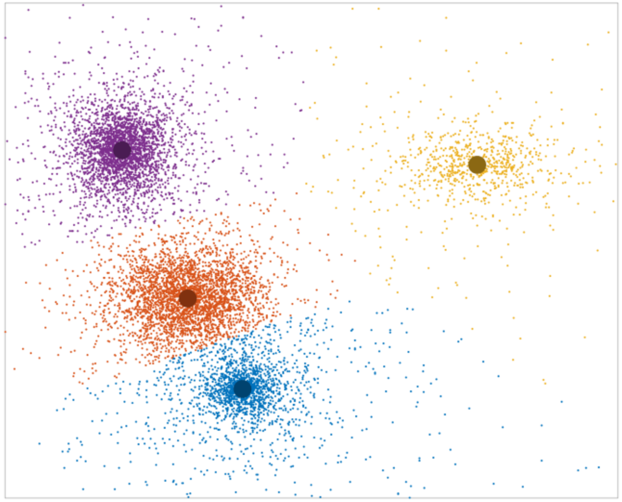
Compressive k-means [Keriven et al 2017]

Mixture of Diracs



Compressive k-means [Keriven et al 2017]

Mixture of Diracs

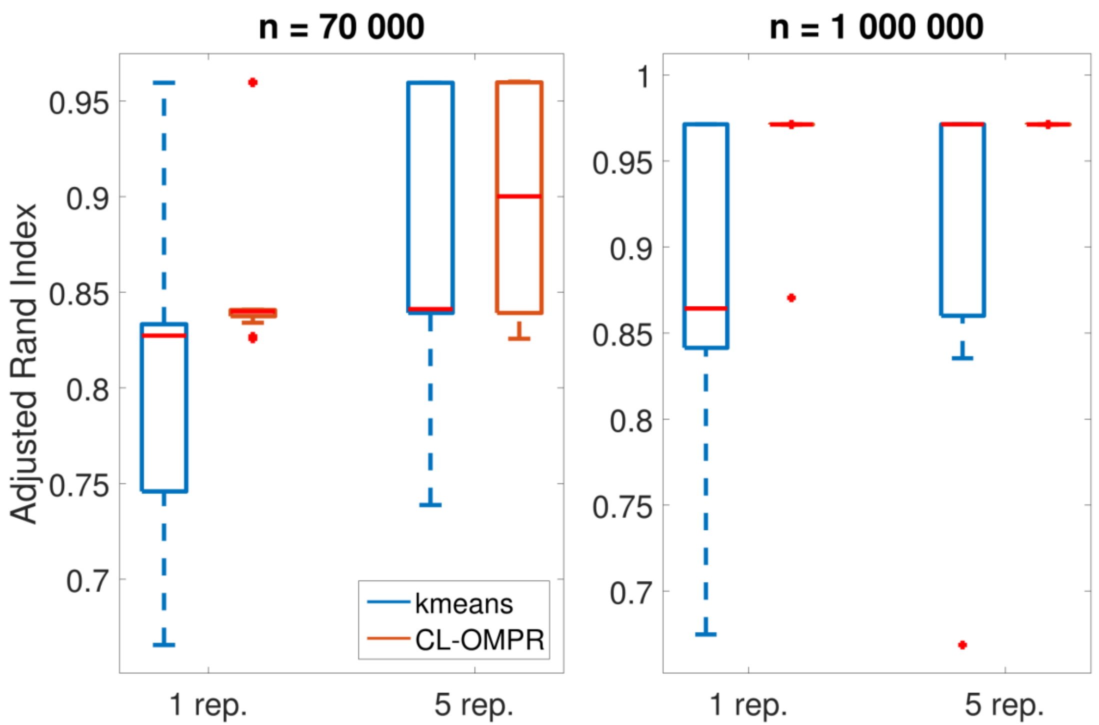


Application: Spectral clustering

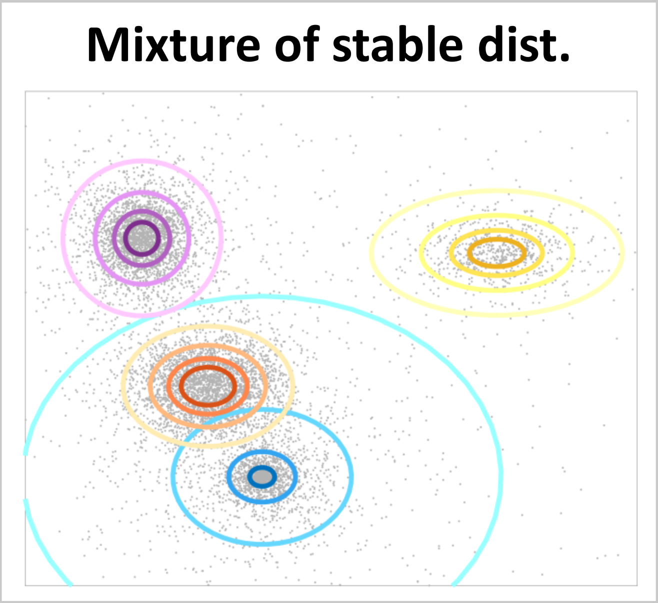
for MNIST classification [Uw 2001]

(d=10, k=10, m=1000)

Classif. Perf.

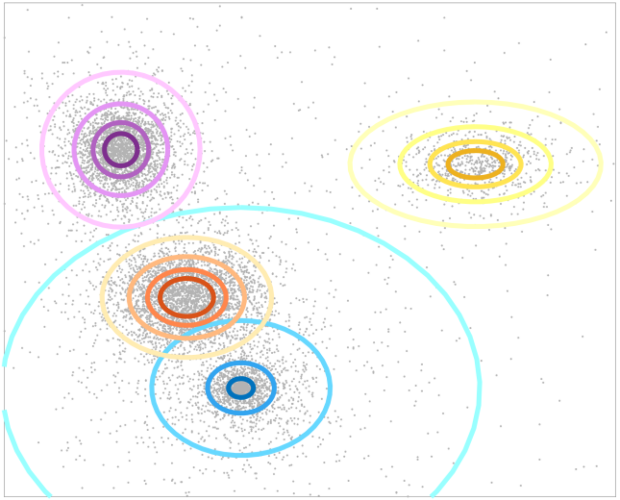


Mixtures of alpha-stable distribution



Mixtures of alpha-stable distribution

Mixture of stable dist.

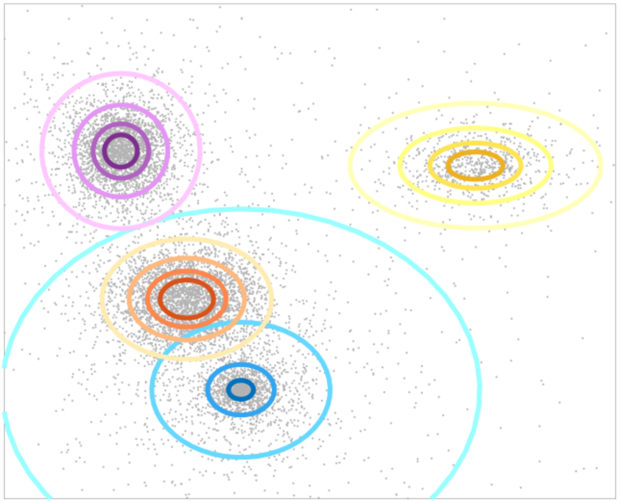


Toy example

- CL-OMPR with $d = 10$, $k = 3$
  - **10^{-2} precision** in **80 sec**
- MCMC with $d = 1$, $k = 3$
  - **10^{-1} precision** in **1.5 hours**

Mixtures of alpha-stable distribution

Mixture of stable dist.



Toy example

- CL-OMPR with $d = 10$, $k = 3$
  - **10^{-2} precision** in **80 sec**
- MCMC with $d = 1$, $k = 3$
  - **10^{-1} precision** in **1.5 hours**

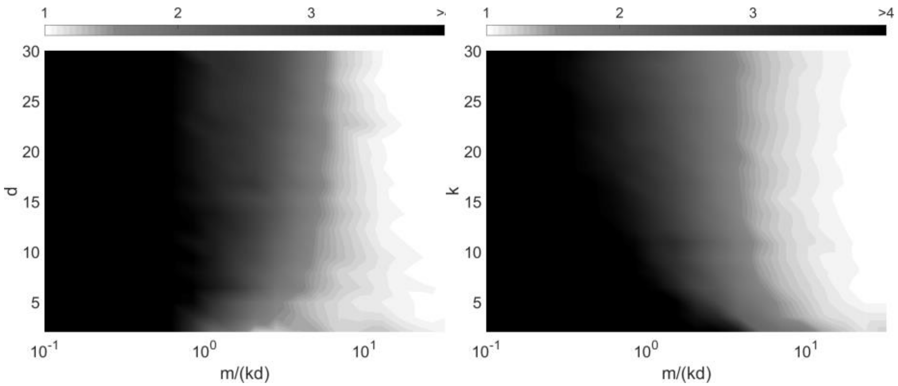
***Application**: audio source separation* *[submitted]*

Model: hybrid between rank-1 alpha-stable and Gaussian noise…

	SDR (dB)	SIR (dB)	MER (dB)
Oracle	8.33 ± 3.16	18.3 ± 4.13	N/A
Gaussian (EM)	3.50 ± 2.87	9.04 ± 4.92	12.3 ± 11.0
CF-α	$\mathbf{4.11 \pm 2.59}$	$\mathbf{9.17 \pm 3.51}$	$\mathbf{12.65 \pm 9.73}$

How big a sketch ?

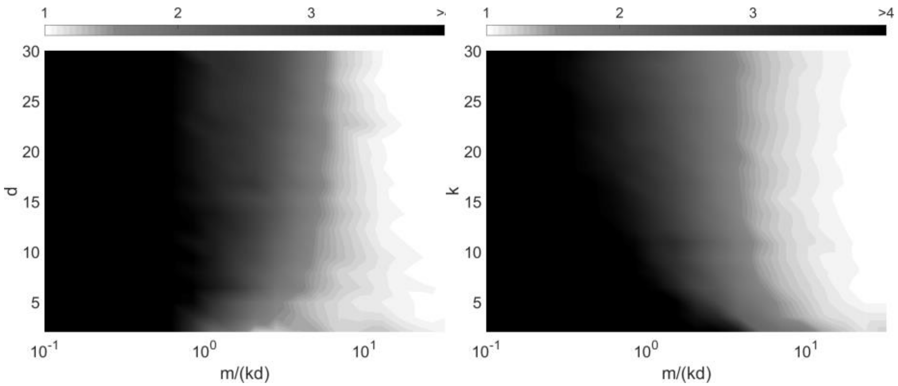
k-means



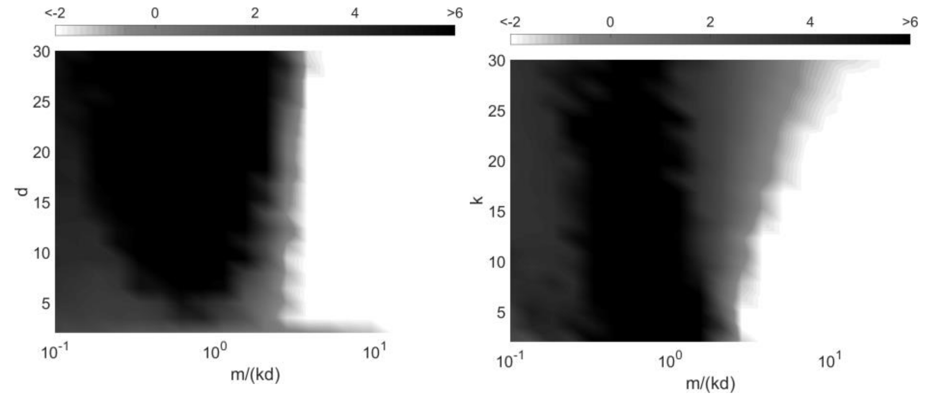
Relative sketch size m/(kd)

How big a sketch ?

k-means



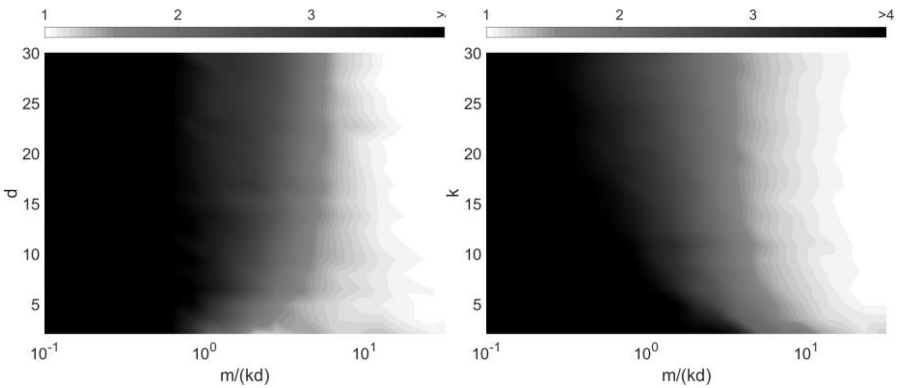
GMM



Relative sketch size m/(kd)

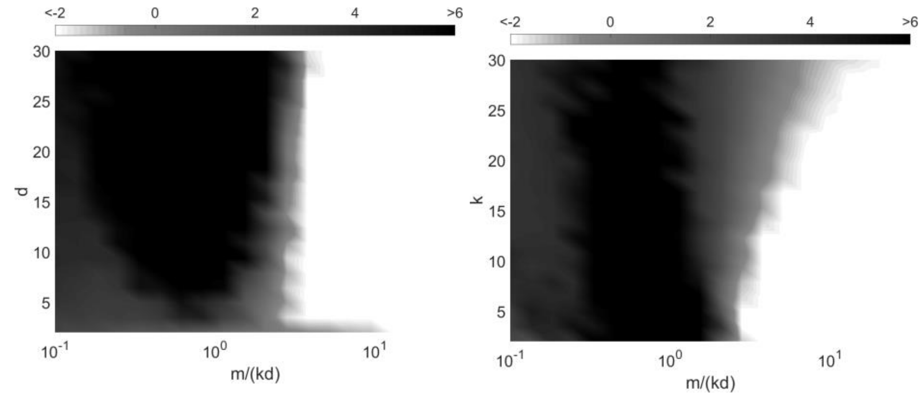
How big a sketch ?

k-means

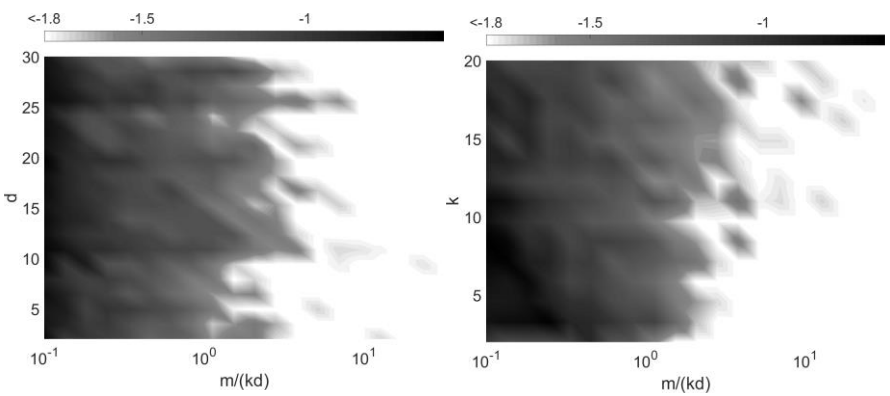


Relative sketch size m/(kd)

GMM

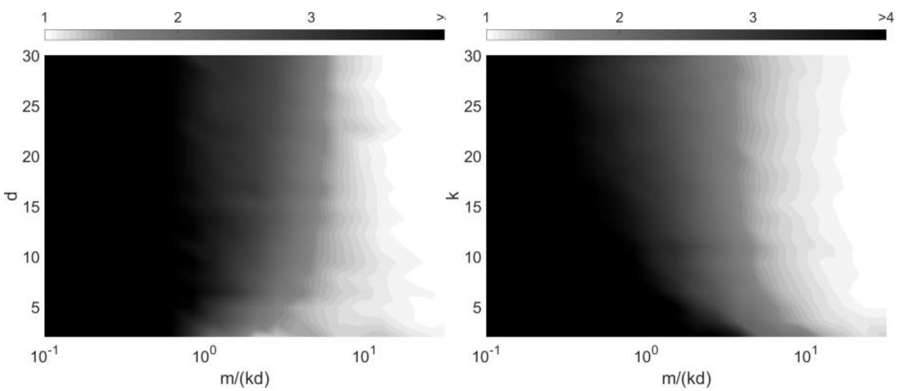


Stable distributions

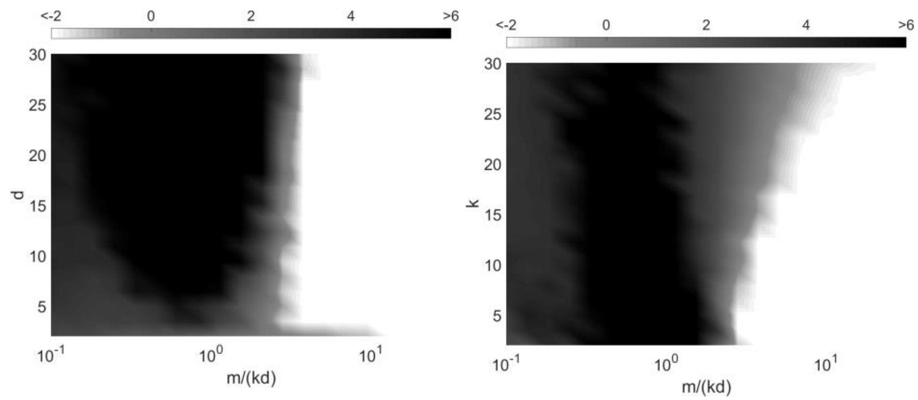


How big a sketch ?

k-means

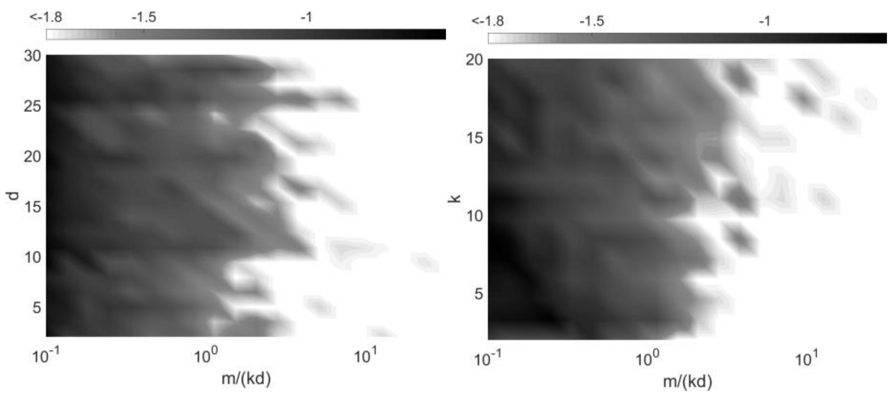


GMM



Relative sketch size m/(kd)

Stable distributions



Sufficient sketch size

$$m \approx \mathcal{O}(kd)$$

Outline

1. Illustration: Sketched Mixture Model Estimation

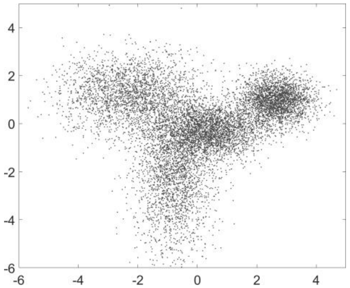
2. Information-preservation guarantees

   2.1 Restricted Isometry Property

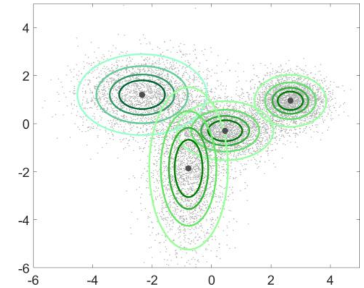
   2.2 Application: mixture model with separation assumption

3. Conclusion, outlooks

Linear inverse problem

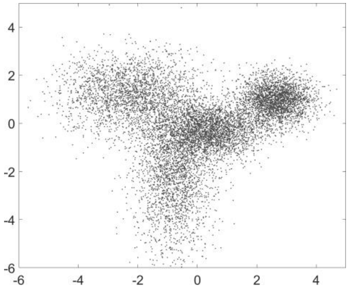


$\xrightarrow{\Phi}$ $\hat{\mathbf{z}}$ $\xrightarrow{\textit{Moment matching}}$

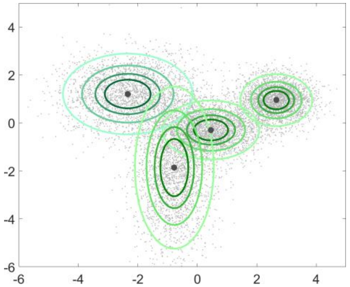


$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X)$$

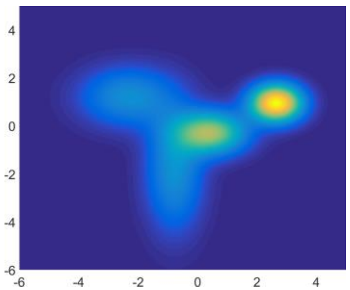
Linear inverse problem



$$\xrightarrow{\Phi} \quad \hat{\mathbf{z}} \quad \xrightarrow{\textit{Moment matching}}$$



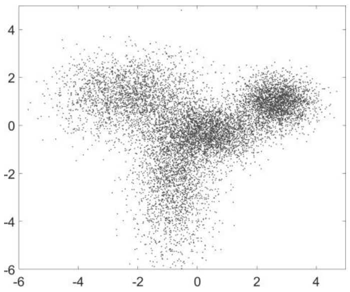
$$\hat{\mathbf{z}} = \hat{\mathbb{E}}\Phi(X)$$



$\pi^\star$

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

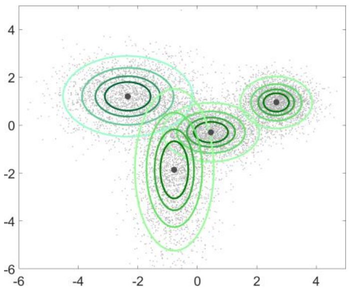
Linear inverse problem



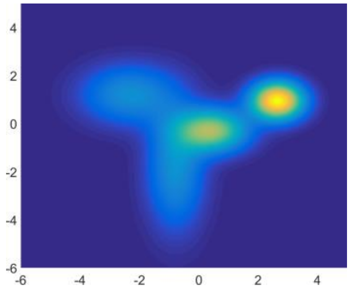
Φ

$\hat{\mathbf{z}}$

Moment matching



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

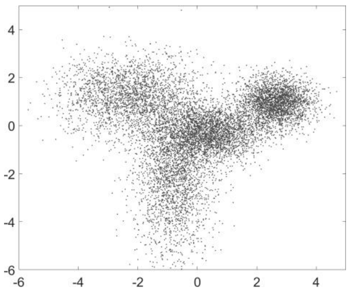


$\pi^\star$

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

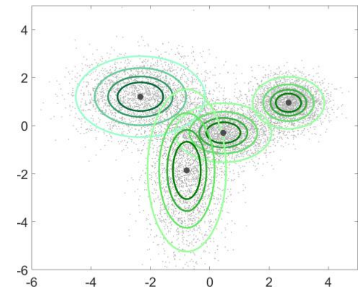
Reformulation of the sketching

Linear inverse problem

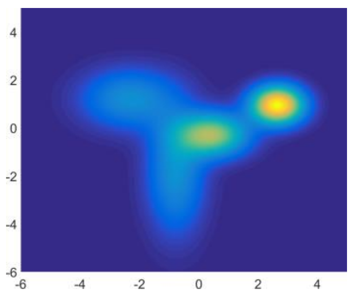


Φ

Moment matching



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



$\pi^\star$

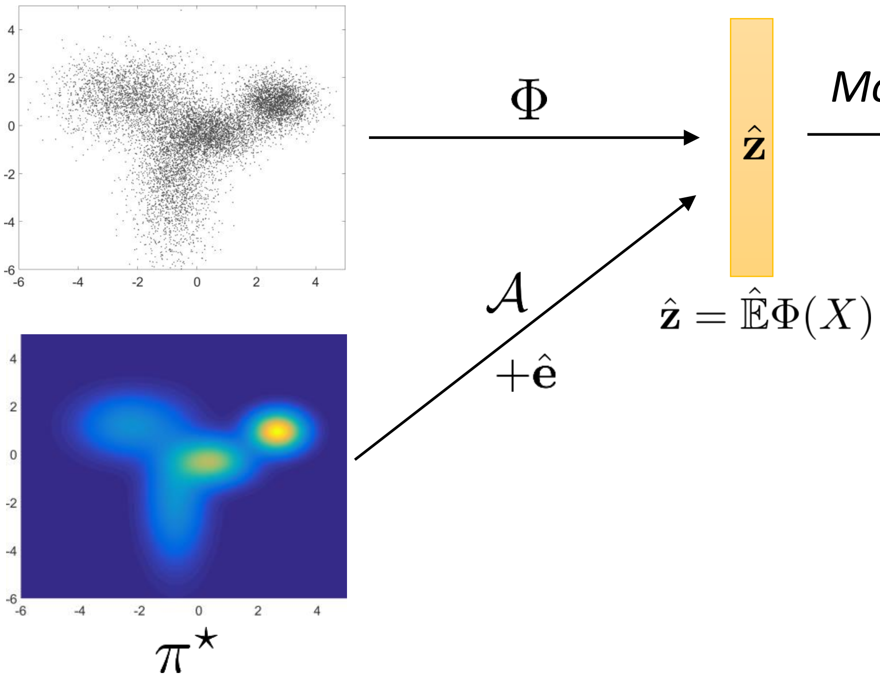
True distribution: $\boxed{x_1, ..., x_n \overset{i.i.d.}{\sim} \pi^\star}$

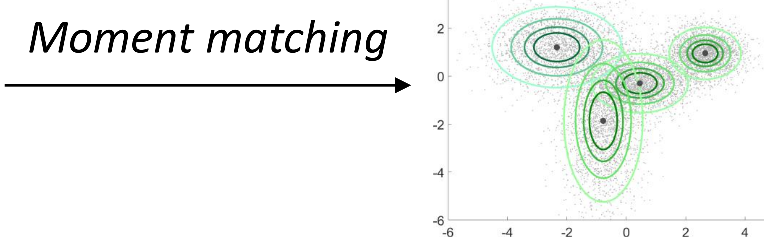
Reformulation of the sketching

- Linear operator:

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

Linear inverse problem





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

- Estimation problem = **linear inverse problem** on measures

Reformulation of the sketching

- Linear operator:

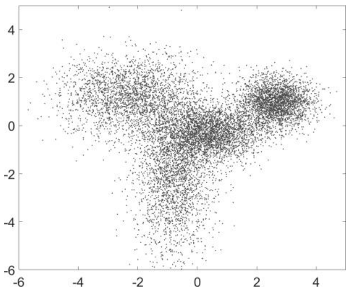
$$\boxed{\mathcal{A}\pi = \mathbb{E}_{X\sim\pi}\Phi(X)}$$

- « Noisy » linear measurement:

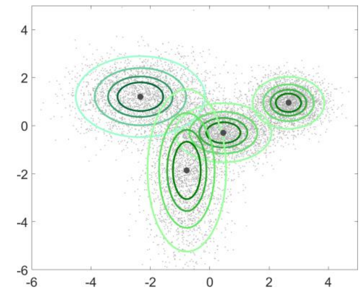
$$\boxed{\hat{\mathbf{z}} = \mathcal{A}\pi^\star + \hat{\mathbf{e}}}$$

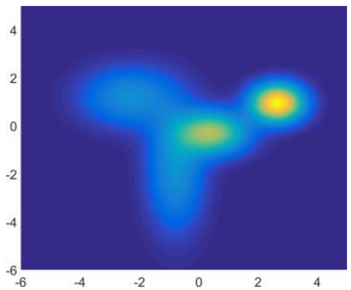
Noise $\hat{\mathbf{e}} = \hat{\mathbb{E}}\Phi(X) - \mathbb{E}_{\pi^\star}\Phi(X)$ ***small***

Linear inverse problem



$\xrightarrow{\Phi}$ $\hat{\mathbf{z}}$ $\xrightarrow{\text{Moment matching}}$





$\pi^\star$

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

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

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

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

Reformulation of the sketching

- Linear operator:

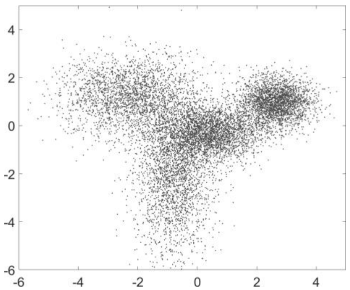
$$\mathcal{A}\pi = \mathbb{E}_{X\sim\pi}\Phi(X)$$

- « 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***

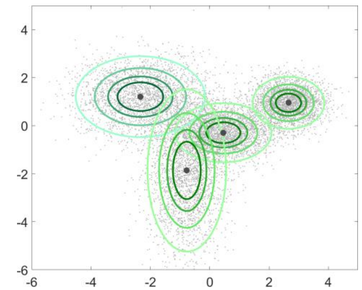
Linear inverse problem



Φ $\hat{\mathbf{z}}$

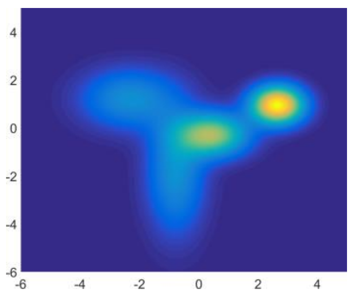
Solve

Moment matching



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

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



$\pi^\star$

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

Reformulation of the sketching

- Linear operator:

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

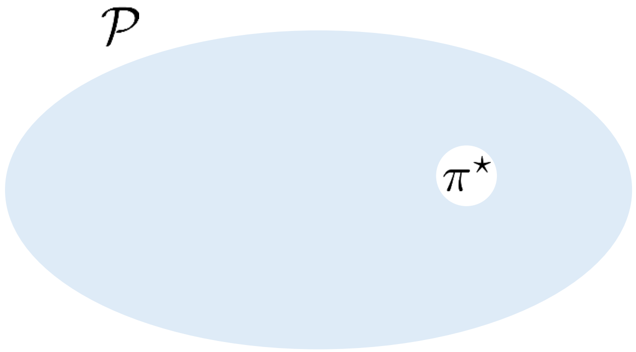
- « Noisy » linear measurement:

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

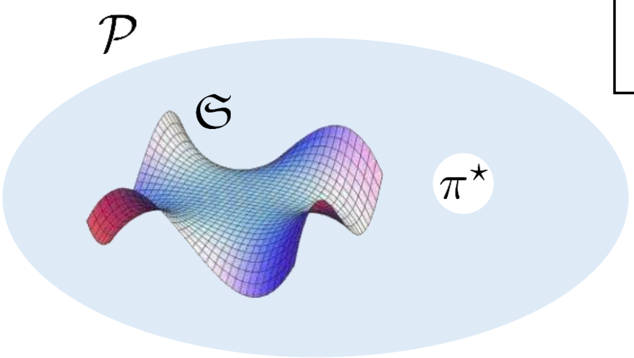
Noise $\hat{\mathbf{e}} = \hat{\mathbb{E}}\Phi(X) - \mathbb{E}_{\pi^\star}\Phi(X)$ ***small***

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

Information preservation guarantees

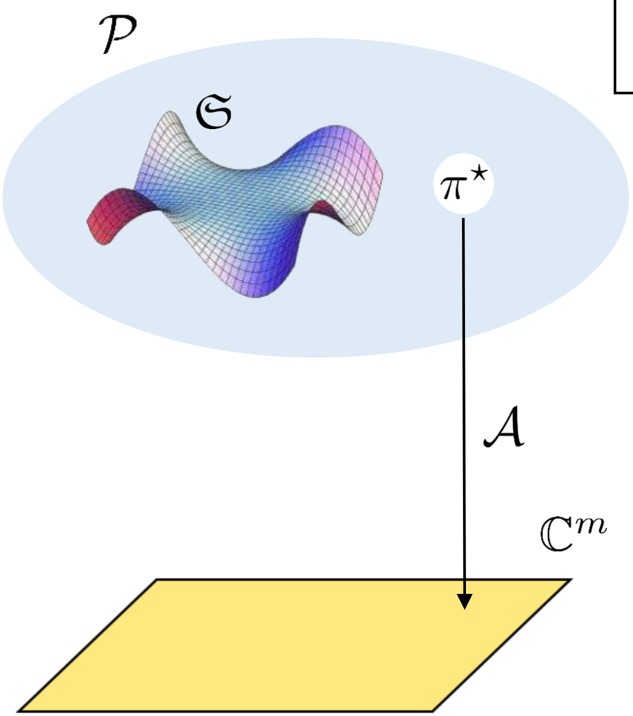


Information preservation guarantees



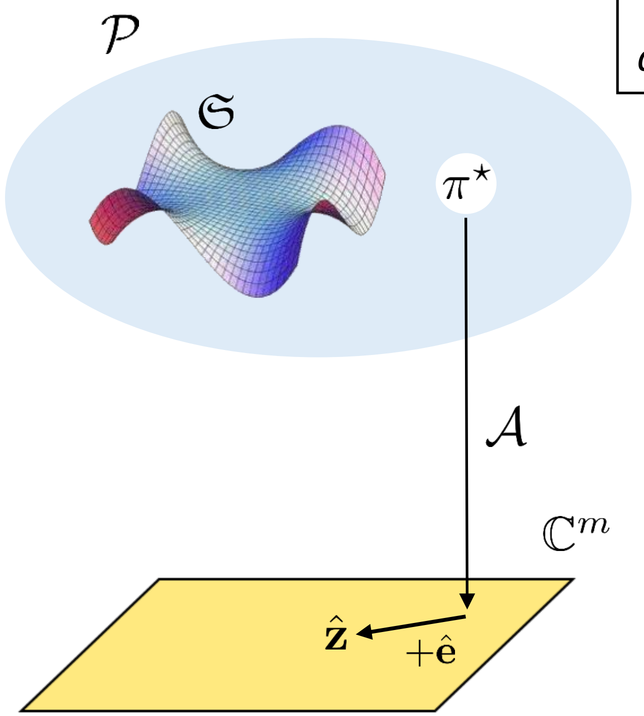
$\mathfrak{S}$ *: Model set of « simple » distributions (eg. GMMs)*

Information preservation guarantees



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

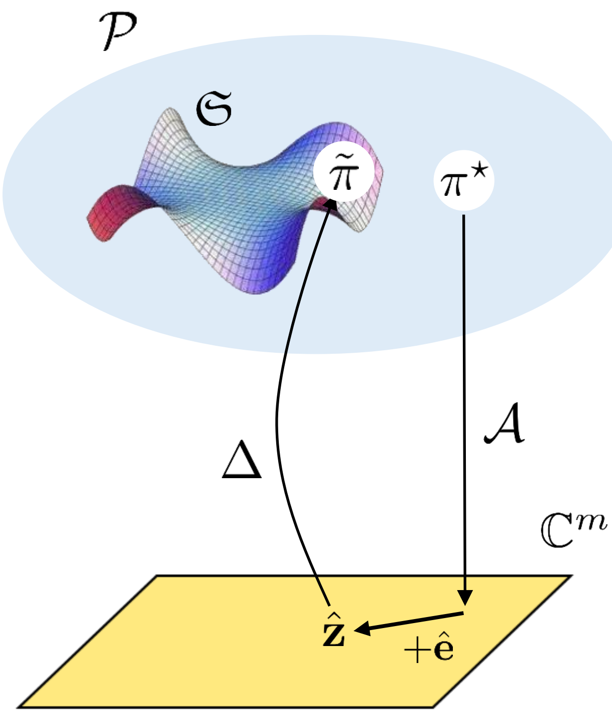
Information preservation guarantees



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

Information preservation guarantees

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



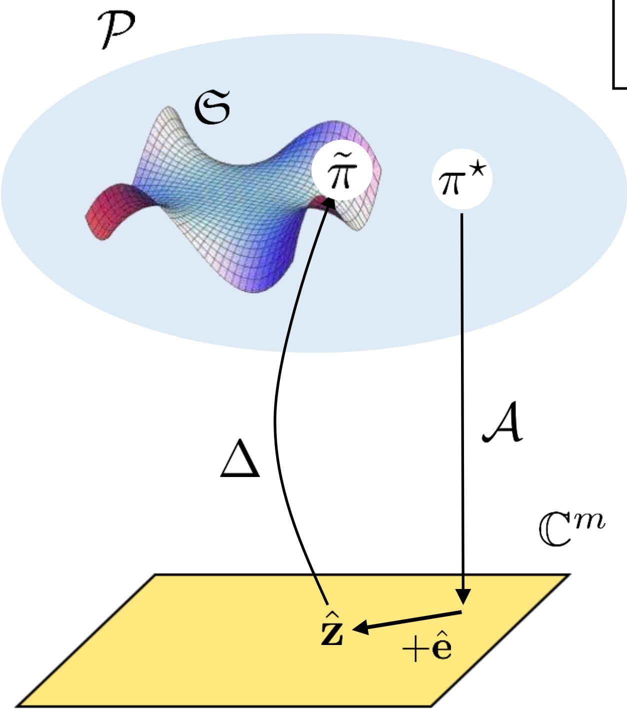
Goal

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

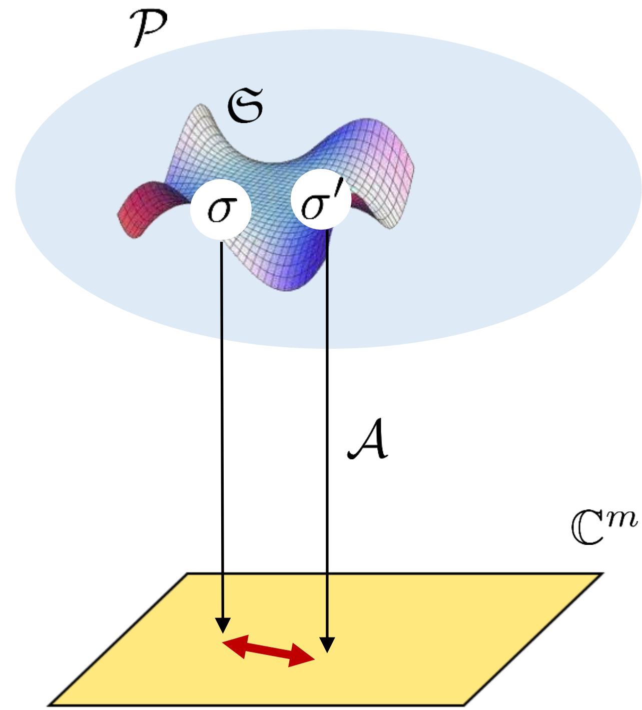
« Instance-optimal » decoder

Information preservation guarantees

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



$\Longrightarrow$



Goal

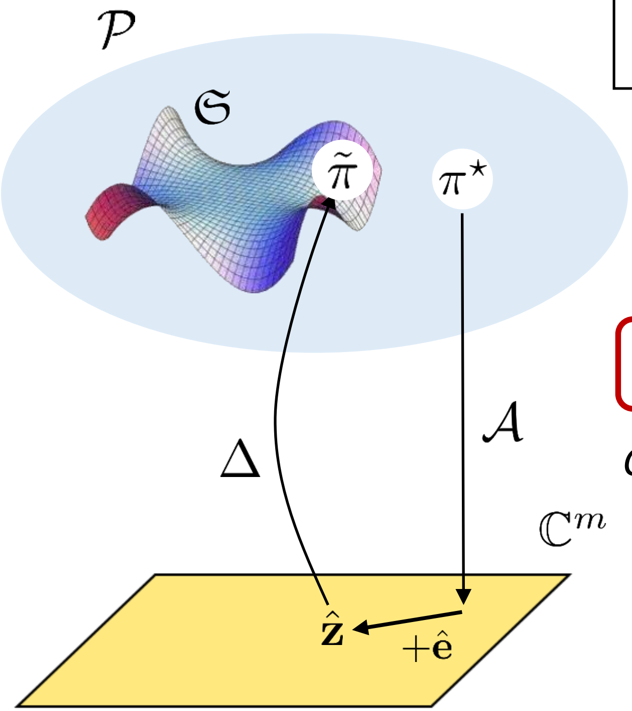
Prove the existence of a *decoder* Δ robust to noise and stable to modeling error.

« Instance-optimal » decoder

Lower Restricted Isometry Property

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

Information preservation guarantees

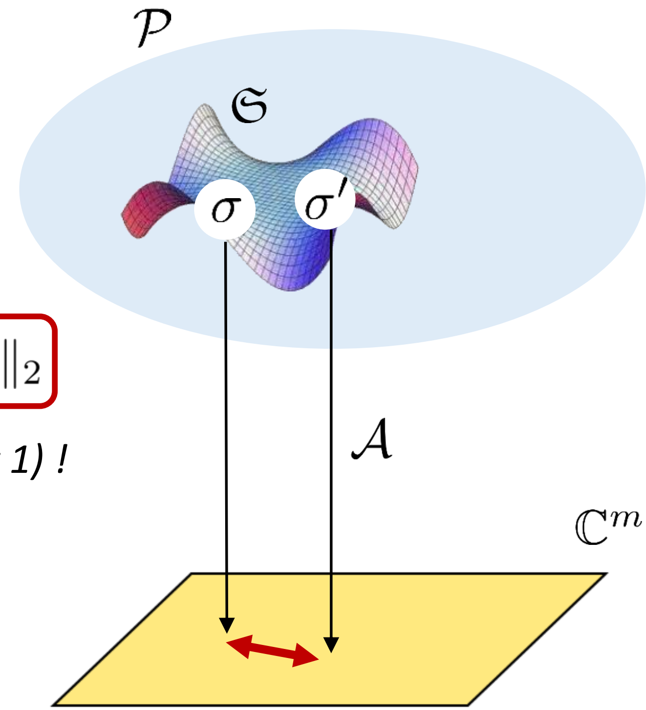


$\mathfrak{S}$: Model set of « simple » distributions (eg. GMMs)

$$\Longrightarrow$$
$$\Longleftarrow$$

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

Cost function used in practice (Part 1) !



Goal

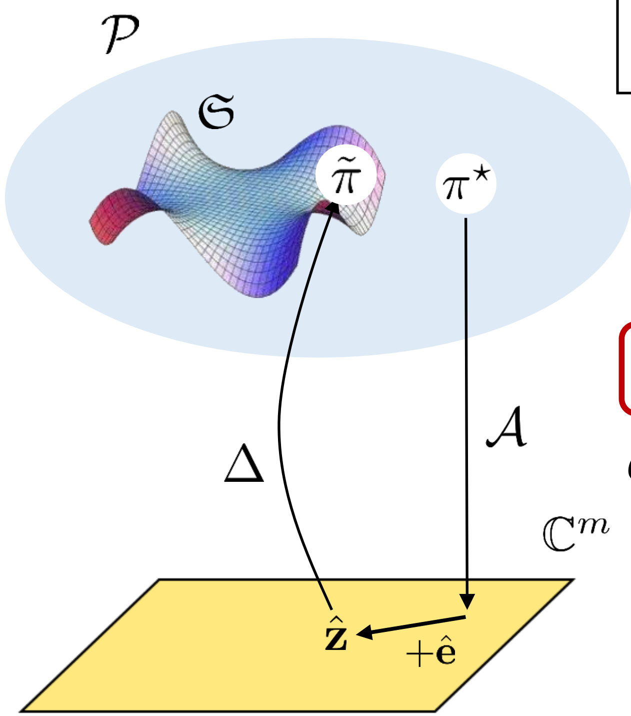
Prove the existence of a *decoder* Δ robust to noise and stable to modeling error.

« Instance-optimal » decoder

Lower Restricted Isometry Property

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

Information preservation guarantees



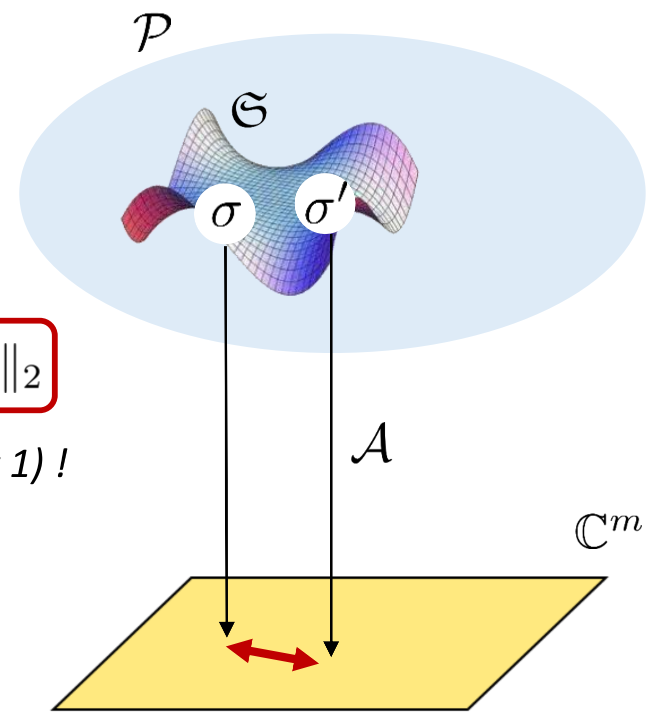
$\mathfrak{S}$: Model set of « simple » distributions (eg. GMMs)

$\Longrightarrow$

$\Longleftarrow$

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

Cost function used in practice (Part 1) !



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.)

Mathematical framework: mean kernel, random features

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

Mathematical framework: mean kernel, random features

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

Metric: mean kernel

[Gretton 2006, Borgwardt 2006]

Mathematical framework: mean kernel, random features

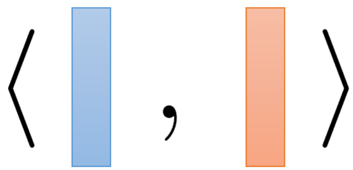
Goal: LRIP w.h.p. on $\mathcal{A}$, $\forall \sigma, \sigma' \in \mathfrak{S}$, $\|\sigma - \sigma'\| \lesssim \|\mathcal{A}\sigma - \mathcal{A}\sigma'\|_2$.

Metric: mean kernel

[Gretton 2006, Borgwardt 2006]

Reproducing kernel:
adjustable geometry on **any** set of objects

$\kappa(x, x')$

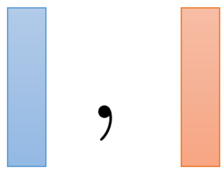


Goal: LRIP w.h.p. on $\mathcal{A}$, $\forall \sigma, \sigma' \in \mathfrak{S}$, $\|\sigma - \sigma'\| \lesssim \|\mathcal{A}\sigma - \mathcal{A}\sigma'\|_2$.

Metric: mean kernel

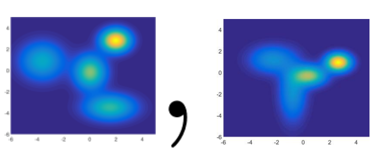
[Gretton 2006, Borgwardt 2006]

Reproducing kernel:
adjustable geometry on **any** set of objects

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

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

Kernel **between** **distributions** **of objects**

$\langle$  $\rangle$

Adjustable
$\|\pi - \pi'\|_\kappa$

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

Metric: mean kernel

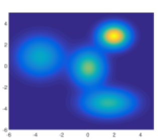
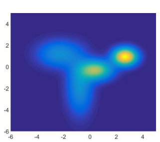
[Gretton 2006, Borgwardt 2006]

Reproducing kernel:
adjustable geometry on **any** set of objects

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

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

Kernel **between** **distributions** **of objects**

$\langle$  ,  $\rangle$

Adjustable
$\|\pi - \pi'\|_\kappa$

Sketching operator: Random Features

[Rahimi 2007]

Mathematical framework: mean kernel, random features

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

Metric: mean kernel

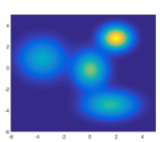
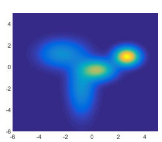
[Gretton 2006, Borgwardt 2006]

Reproducing kernel:
adjustable geometry on **any** set of objects

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

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

Kernel **between** **distributions** **of objects**

$\langle$  ,  $\rangle$

Adjustable
$\|\pi - \pi'\|_\kappa$

Sketching operator: Random Features

[Rahimi 2007]

Random $\Phi : \mathbb{R}^d \to \mathbb{C}^m$ such that:

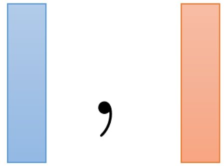
$$\kappa(x, x') \approx \Phi(x)^*\Phi(x')$$

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

Metric: mean kernel

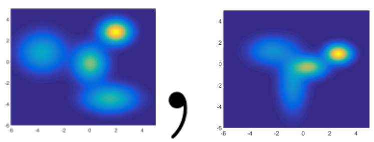
[Gretton 2006, Borgwardt 2006]

Reproducing kernel:
adjustable geometry on **any** set of objects

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

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

Kernel **between distributions of objects**

$\langle$  $\rangle$

Adjustable
$\|\pi - \pi'\|_\kappa$

Sketching operator: Random Features

[Rahimi 2007]

Random $\Phi : \mathbb{R}^d \to \mathbb{C}^m$ such that:

$$\kappa(x, x') \approx \Phi(x)^* \Phi(x')$$

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

Basis for the RIP

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

Mathematical framework: mean kernel, random features

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

Metric: mean kernel

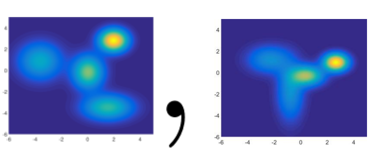
[Gretton 2006, Borgwardt 2006]

Reproducing kernel:
adjustable geometry on **any** set of objects

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

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

Kernel **between** **distributions** **of objects**

$\langle$  $\rangle$

Adjustable

$\|\pi - \pi'\|_\kappa$

Sketching operator: Random Features

[Rahimi 2007]

Random $\Phi : \mathbb{R}^d \to \mathbb{C}^m$ such that:

$$\kappa(x, x') \approx \Phi(x)^* \Phi(x')$$

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

Basis for the RIP

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

Bernstein...

Mathematical framework: mean kernel, random features

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

Metric: mean kernel

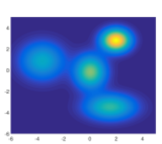
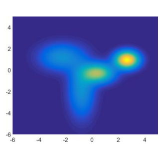
[Gretton 2006, Borgwardt 2006]

Reproducing kernel:
adjustable geometry on **any** set of objects

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

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

Kernel **between** **distributions** **of objects**

$\Big\langle$  ,  $\Big\rangle$

Adjustable
$\|\pi - \pi'\|_\kappa$

Sketching operator: Random Features

[Rahimi 2007]

Random $\Phi : \mathbb{R}^d \to \mathbb{C}^m$ such that:

$$\kappa(x, x') \approx \Phi(x)^* \Phi(x')$$

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

Basis for the RIP

$$\color{blue}{\|\pi - \pi'\|_\kappa^2} \approx \|\color{red}{\mathcal{A}}(\pi - \pi')\|_2^2$$

Bernstein…

Ideally…

Number of random features =
intrinsic dimensionality of the problem

Normalized secant set

Reformulation of the LRIP

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

Normalized secant set

Reformulation of the LRIP

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

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

Normalized secant set

Reformulation of the LRIP

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

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

Definition: Normalized Secant set

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

Normalized secant set

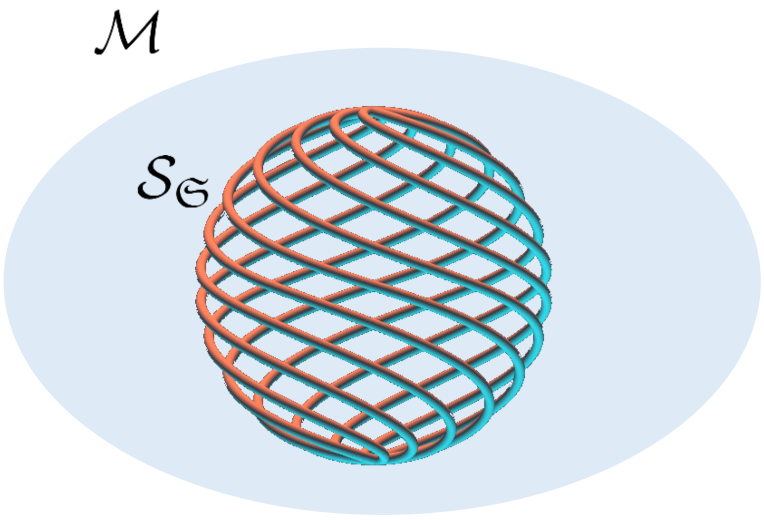
Reformulation of the LRIP

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

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

Definition: Normalized Secant set

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



Normalized secant set

Reformulation of the LRIP

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

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

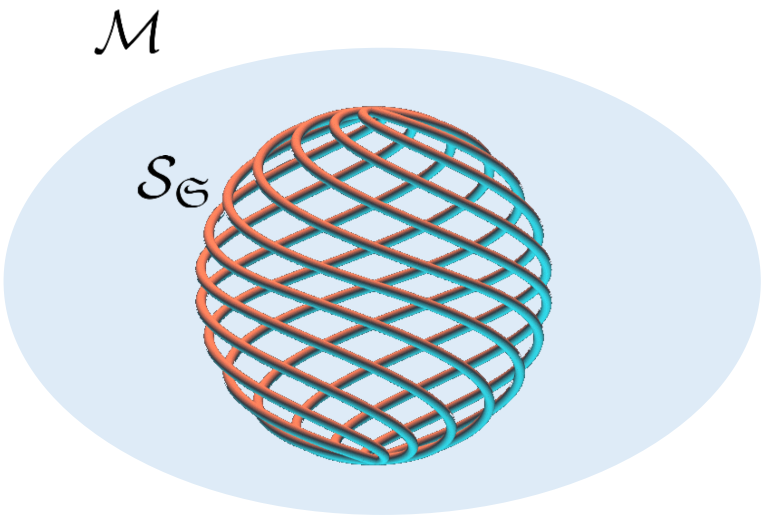
Definition: Normalized Secant set

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

New goal

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

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



Normalized secant set

Reformulation of the LRIP

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

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

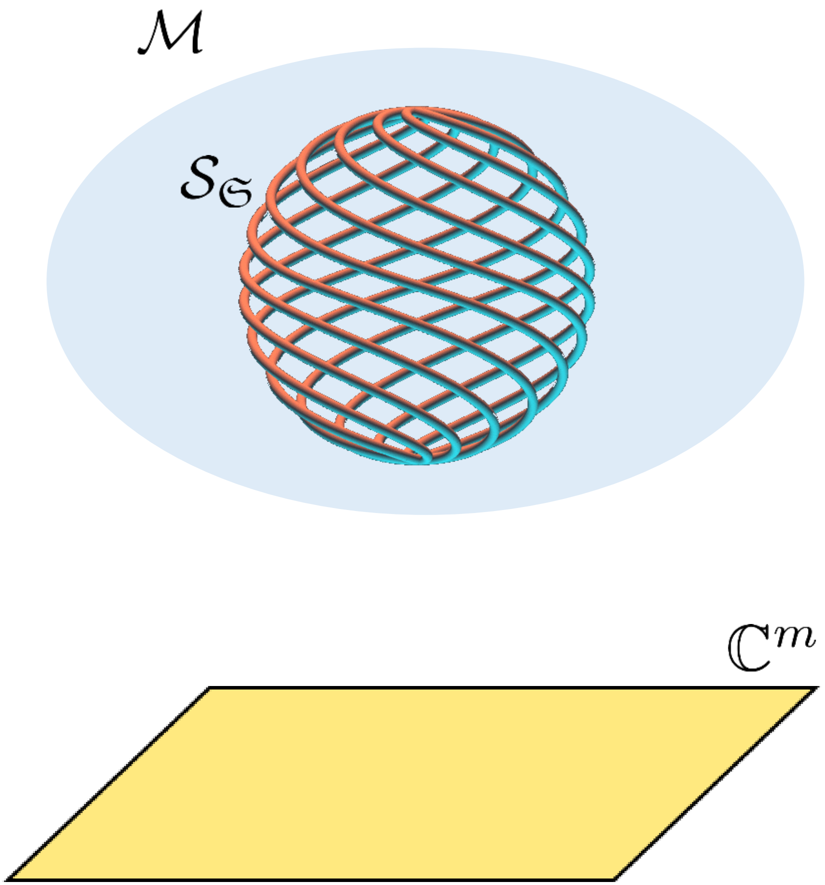
Definition: Normalized Secant set

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

New goal

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

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



Normalized secant set

Reformulation of the LRIP

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

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

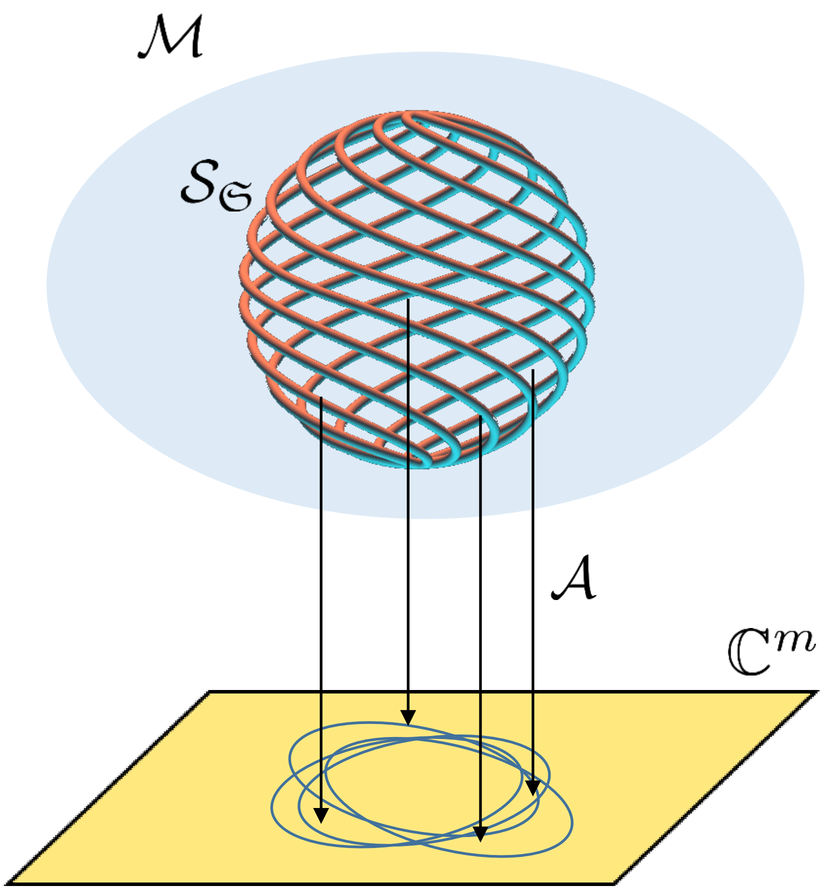
Definition: Normalized Secant set

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

New goal

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

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



Normalized secant set

Reformulation of the LRIP

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

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

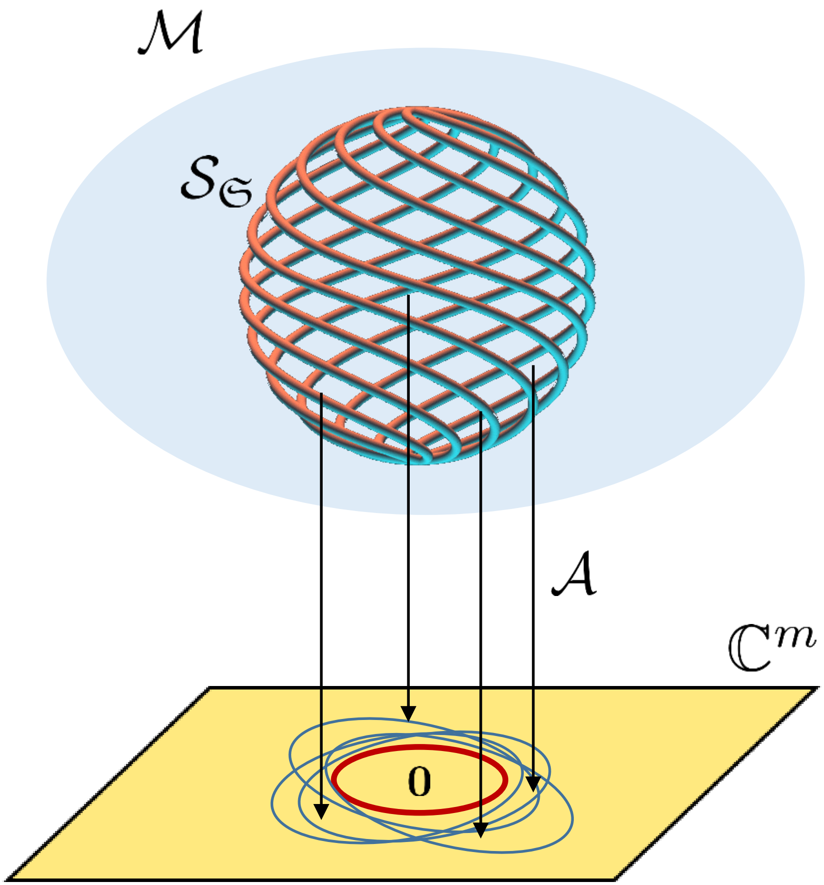
Definition: Normalized Secant set

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

New goal

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

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



Proving the LRIP

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

Proving the LRIP

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

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

Proving the LRIP

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

① **Pointwise LRIP** $\forall s,\ \text{w.h.p. on } \mathcal{A},\ \text{LRIP}.$

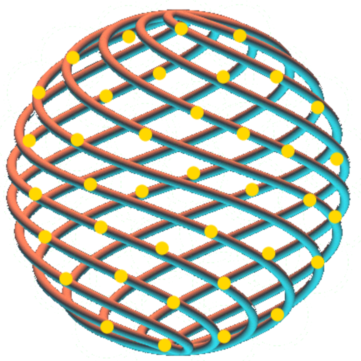
② **Extension to LRIP: covering numbers**

Proving the LRIP

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

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

2 **Extension to LRIP: covering numbers**

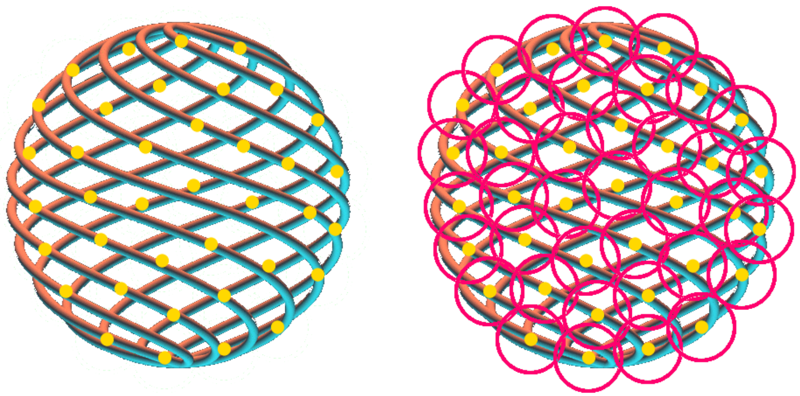


Proving the LRIP

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

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

2 **Extension to LRIP: covering numbers**



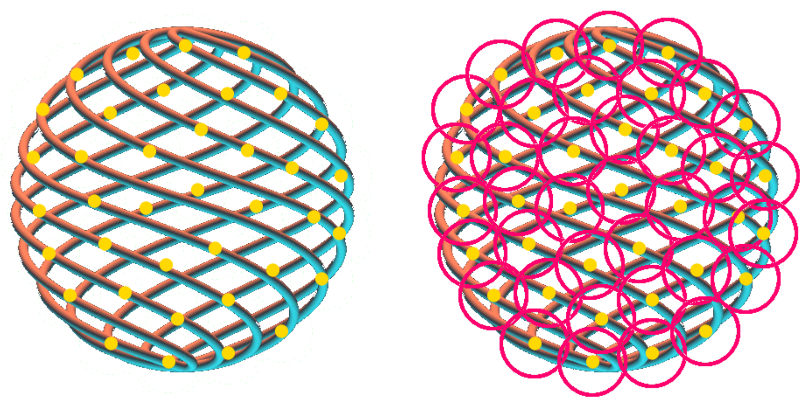
Proving the LRIP

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

① **Pointwise LRIP**

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

② **Extension to LRIP: covering numbers**



w.h.p. on $\mathcal{A}$, $\forall s$, 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 $\boxed{m \geq C \;\times\; \log(\text{cov. num.})}$,

Main result [Keriven 2016]

Main hypothesis

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

Result

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

Quality of pointwise LRIP

Dimensionality of the model

Main result [Keriven 2016]

Main hypothesis

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

Result

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

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 [Keriven 2016]

Main hypothesis

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

Result

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

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 [Keriven 2016]

Main hypothesis

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

Result

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

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**

First results

Simplified hyp.: the **model itself** $\mathfrak{S}$ is compact (instead of $\mathcal{S}(\mathfrak{S})$)

First results

Simplified hyp.: the **model itself** $\mathfrak{S}$ is compact (instead of $\mathcal{S}(\mathfrak{S})$)

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

First results

Simplified hyp.: the **model itself** $\mathfrak{S}$ is compact (instead of $\mathcal{S}(\mathfrak{S})$)

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

First results

Simplified hyp.: the **model itself** $\mathfrak{S}$ is compact (instead of $\mathcal{S}(\mathfrak{S})$)

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

MMD

Sub-optimal !

First results

Simplified hyp.: the **model itself** $\mathfrak{S}$ is compact (instead of $\mathcal{S}(\mathfrak{S})$)

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

MMD

Sub-optimal !

Application to:
- **GMM with diagonal covariance**
- **Mixture of elliptic stable distributions** *(no existing estimator)*

First results

Simplified hyp.: the **model itself** $\mathfrak{S}$ is compact (instead of $\mathcal{S}(\mathfrak{S})$)

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

MMD

Sub-optimal !

Application to:

- **GMM with diagonal covariance**
- **Mixture of elliptic stable distributions** *(no existing estimator)*

Questions:

- Get rid of the $\mathcal{O}(1/\sqrt{m})$?
- Replace $\|\cdot\|_\kappa$ with another metric for learning?

Outline

1 Illustration: Sketched Mixture Model Estimation

2 Information-preservation guarantees

2.1 Restricted Isometry Property

2.2 Application: mixture model with separation assumption

3 Conclusion, outlooks

Fine control for mixture models

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

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

Main difficulty

*Controlling metrics between **distributions in the model** that get **close to each other** in infinite-dimensional space*

Fine control for mixture models

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

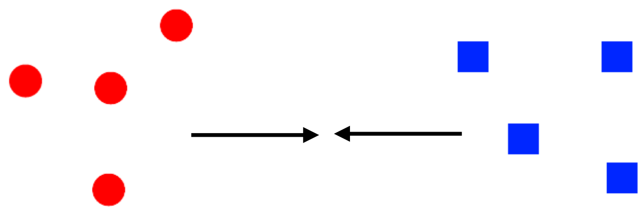
$$\mathcal{S}_{\mathfrak{S}} = \left\{ \frac{\sigma - \sigma'}{\|\sigma - \sigma'\|_\kappa} \right\}$$

Main difficulty

*Controlling metrics between **distributions in the model** that get **close to each other** in infinite-dimensional space*

Case of mixture models

$$\|\textstyle\sum_l w_l \pi_{\theta_l} - \sum_l w'_l \pi_{\theta'_l}\|_\kappa \to 0 \text{ : what happens ?}$$



Fine control for mixture models

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

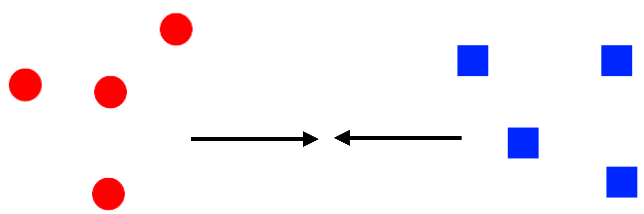
$$\mathcal{S}_{\mathfrak{S}} = \left\{ \frac{\sigma - \sigma'}{\|\sigma - \sigma'\|_\kappa} \right\}$$

Main difficulty

Controlling metrics between ***distributions in the model*** *that get* ***close to each other*** *in infinite-dimensional space*

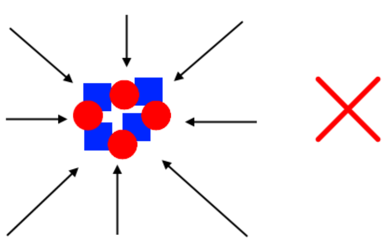
Case of mixture models

$$\|\textstyle\sum_l w_l \pi_{\theta_l} - \sum_l w'_l \pi_{\theta'_l}\|_\kappa \to 0 : \text{ what happens ?}$$

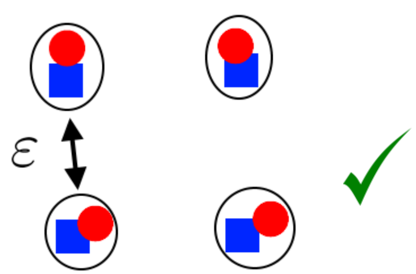


Separation hypothesis

No hypothesis



Separation hypothesis



Compressive statistical learning

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

Compressive statistical learning

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

Hypotheses

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

Compressive statistical learning

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

Hypotheses

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

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

Compressive statistical learning
[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

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

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

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

Compressive statistical learning

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

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

Hypotheses

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

Sketch

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

Result

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

Compressive statistical learning
[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

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

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

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

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

Sketch size

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

Compressive statistical learning

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

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

Hypotheses

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

Sketch

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

Result

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

Sketch size

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

GMM with known covariance

Compressive statistical learning

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

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

Hypotheses

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

Sketch

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

Result

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

Sketch size

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

GMM with known covariance

Hypotheses

- Sufficiently separated means
- Bounded domain for means

Compressive statistical learning

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

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

Hypotheses

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

Sketch

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

Result

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

Sketch size

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

GMM with known covariance

Hypotheses

- Sufficiently separated means
- Bounded domain for means

Sketch

- Fourier features

Compressive statistical learning

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

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

Hypotheses

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

Sketch

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

Result

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

Sketch size

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

GMM with known covariance

Hypotheses

- Sufficiently separated means
- Bounded domain for means

Sketch

- Fourier features

Result

- With respect to **log-likelihood**

Compressive statistical learning

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

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

Hypotheses

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

Sketch

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

Result

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

Sketch size

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

GMM with known covariance

Hypotheses

- Sufficiently separated means
- Bounded domain for means

Sketch

- Fourier features

Result

- With respect to **log-likelihood**

Sketch size

$$m \geq \mathcal{O}(k^2 d^2 \texttt{polylog}(k,d) \varphi(\text{sep.}))$$

Compressive statistical learning

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

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

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

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

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

Sketch size

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

GMM with known covariance

Hypotheses
- Sufficiently separated means
- Bounded domain for means

Sketch
- Fourier features

Result
- With respect to **log-likelihood**

Sketch size

$$m \geq \mathcal{O}(k^2 d^2 \texttt{polylog}(k,d) \varphi(\text{sep.}))$$

Recently (not published yet)

$$d^2 \rightarrow d$$

Compressive statistical learning

[Gribonval, Blanchard, Keriven, Traonmilin 2017]

k-means with mixtures of Diracs

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

Hypotheses

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

Sketch

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

Result

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

Sketch size

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

GMM with known covariance

Hypotheses

- Sufficiently separated means
- Bounded domain for means

Sketch

- Fourier features

Result

- With respect to **log-likelihood**

Sketch size

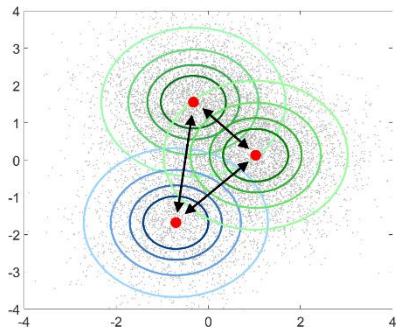
$$m \geq \mathcal{O}(k^2 d^2 \texttt{polylog}(k,d) \varphi(\text{sep.}))$$

Recently (not published yet)

$$d^2 \rightarrow d$$

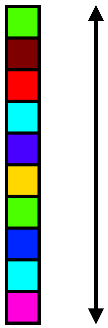
$$m \geq C \times \log(\text{cov. num.})$$

GMM trade-off



Separation of means

Trade-off



Size of sketch

More
High
Freq.

Separation of means	Number of measurements
$\mathcal{O}(\sqrt{d \log k})$	$m \geq \mathcal{O}(k^2 d \cdot \texttt{polylog}(k,d))$
$\mathcal{O}(\sqrt{d + \log k})$	$m \geq \mathcal{O}(k^3 d \cdot \texttt{polylog}(k,d))$
$\mathcal{O}(\sqrt{\log k})$	$m \geq \mathcal{O}(k^2 d e^d \cdot \texttt{polylog}(k,d))$

Outline

1 Illustration: Sketched Mixture Model Estimation

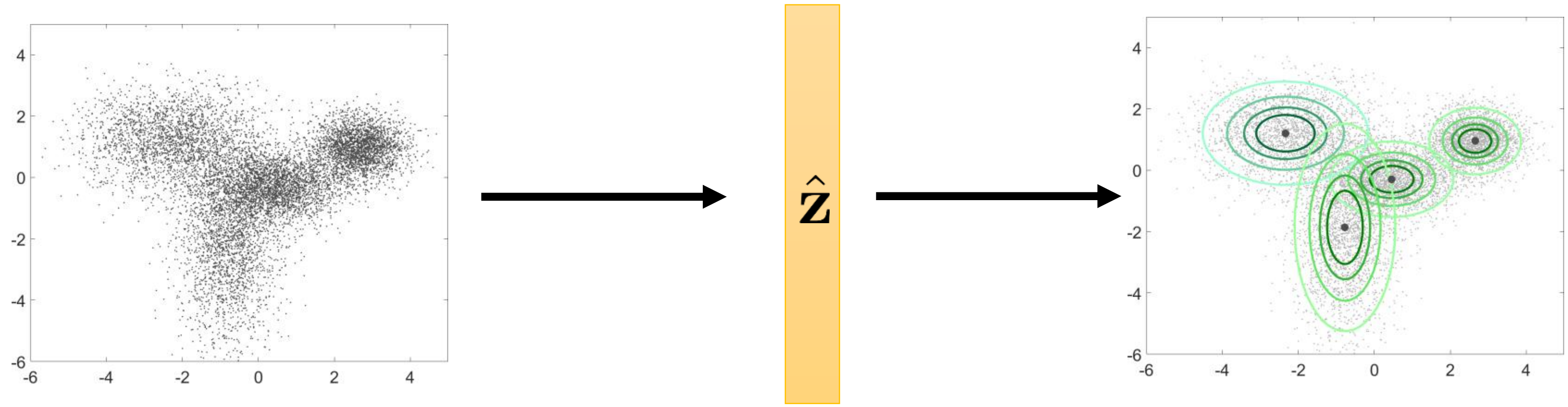
2 Information-preservation guarantees

2.1 Restricted Isometry Property

2.2 Application: mixture model with separation assumption

3 Conclusion, outlooks

Sketch learning



- Sketching method for **large-scale density estimation**
  - Well-adapted to **distributed** or **streaming** context
  - Focus on **mixture model estimation**

Summary of contributions

- Practical illustration: **flexible heuristic algorithm for any sketched mixture model estimation**
  - GMM with diagonal covariance
  - k-means (mixture of Diracs)
  - *Mixture of multivariate elliptic stable distributions*

Summary of contributions

- Practical illustration: **flexible heuristic algorithm for any sketched mixture model estimation**
  - GMM with diagonal covariance
  - k-means (mixture of Diracs)
  - *Mixture of multivariate elliptic stable distributions*

- Information-preservation guarantees
  - Infinite dimensional **Compressive Sensing** (Restricted isometry property)
  - **Kernel methods** on distributions (Kernel mean, Random features)

Summary of contributions

- Practical illustration: **flexible heuristic algorithm for any sketched mixture model estimation**
  - GMM with diagonal covariance
  - k-means (mixture of Diracs)
  - *Mixture of multivariate elliptic stable distributions*

- Information-preservation guarantees
  - Infinite dimensional **Compressive Sensing** (Restricted isometry property)
  - **Kernel methods** on distributions (Kernel mean, Random features)

- Generic assumptions of *low-dimensionality* of the model set

Summary of contributions

- Practical illustration: **flexible heuristic algorithm for any sketched mixture model estimation**
  - GMM with diagonal covariance
  - k-means (mixture of Diracs)
  - *Mixture of multivariate elliptic stable distributions*

- Information-preservation guarantees
  - Infinite dimensional **Compressive Sensing** (Restricted isometry property)
  - **Kernel methods** on distributions (Kernel mean, Random features)

- Generic assumptions of *low-dimensionality* of the model set

- Focus on mixture models
  - Estimator of mixture of multivariate elliptic stable distributions
  - **Statistical learning** with **controlled sketch size** for k-means, sketched GMM with known covariance

Outlook : convex relaxation *[with G. Peyré, C. Poon]*

Algorithm with guarantees?

Outlook : convex relaxation *[with G. Peyré, C. Poon]*

Algorithm with guarantees?

- Convex relaxation: ***super-resolution***

$$\min_\mu \frac{1}{2}\|\mathbf{z} - \mathcal{A}\mu\|^2 + \lambda\|\mu\|_{\mathrm{TV}}$$

Outlook : convex relaxation *[with G. Peyré, C. Poon]*

Algorithm with guarantees?

- Convex relaxation: ***super-resolution***

$$\min_\mu \frac{1}{2}\|\mathbf{z}-\mathcal{A}\mu\|^2+\lambda\|\mu\|_{\mathrm{TV}}$$

  - Dual formulation: **SDP…**

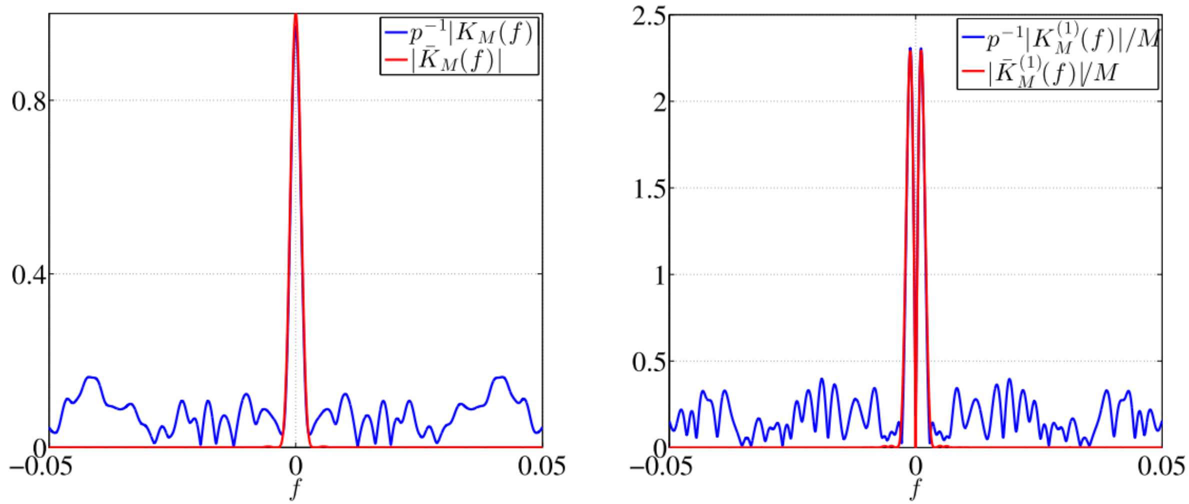
Outlook : convex relaxation *[with G. Peyré, C. Poon]*

Algorithm with guarantees?

- Convex relaxation: ***super-resolution***

$$\min_\mu \frac{1}{2}\|\mathbf{z} - \mathcal{A}\mu\|^2 + \lambda\|\mu\|_{\mathrm{TV}}$$

  - Dual formulation: **SDP…**



Tang2015

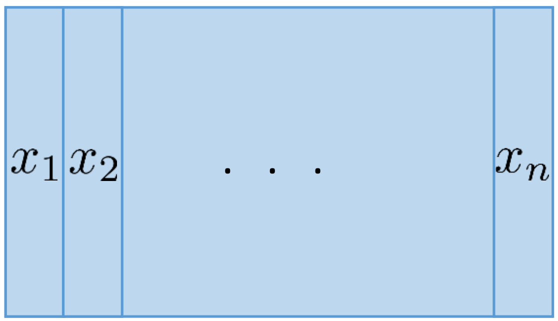
> ➢ Extend to any kernels with random features
> ➢ Application in machine learning…

Outlook : dimensionality

- Combine with **dimension reduction**? *[A. Chatalic]*
  - First map in low-dimension, then sketch

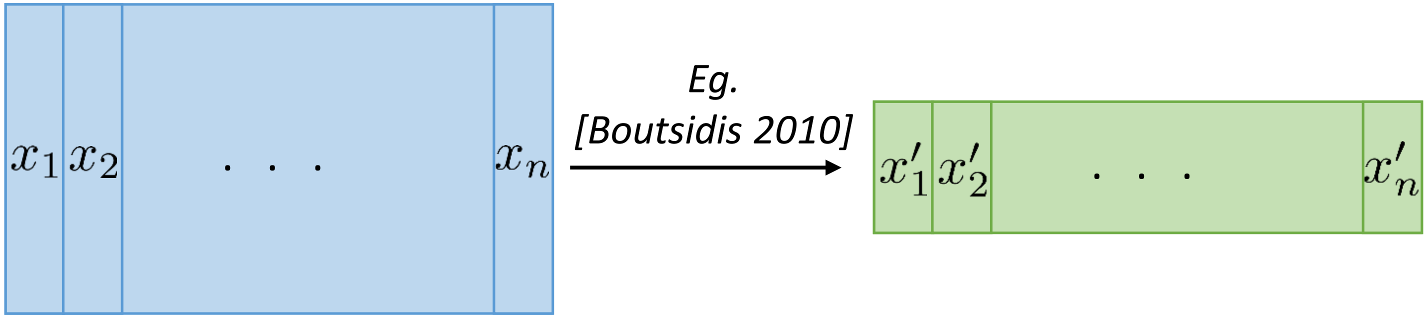
Outlook : dimensionality

- Combine with **dimension reduction**? *[A. Chatalic]*
  - First map in low-dimension, then sketch



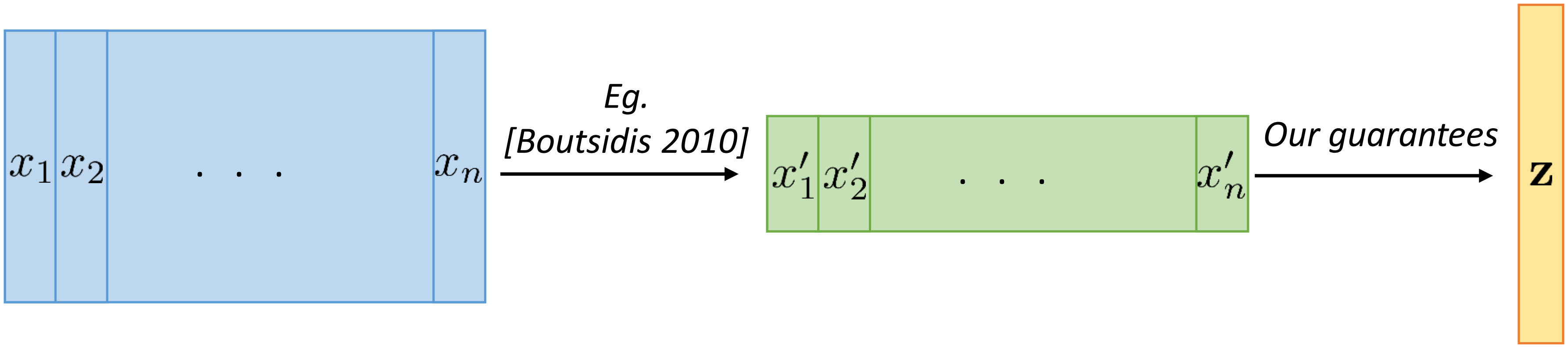
Outlook : dimensionality

- Combine with **dimension reduction**? *[A. Chatalic]*
  - First map in low-dimension, then sketch



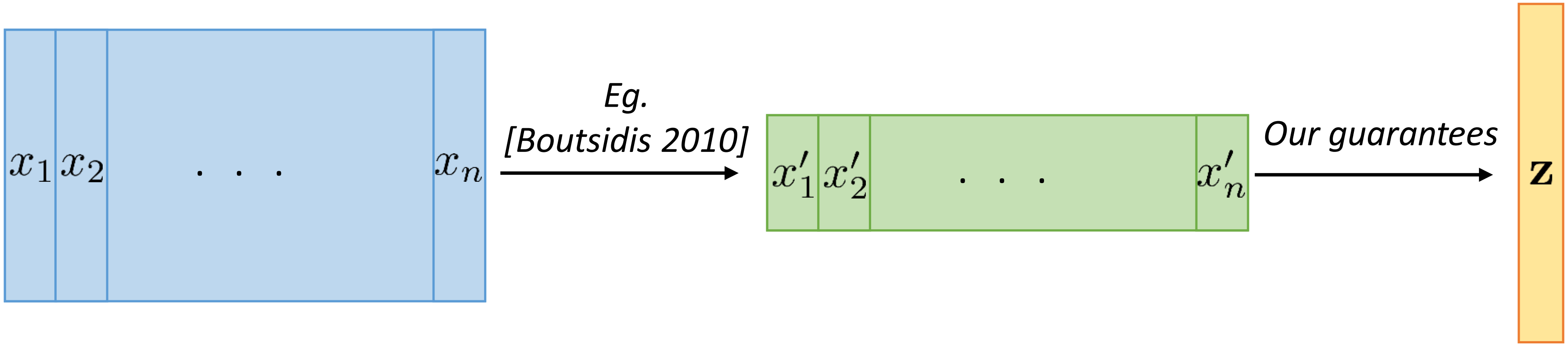
Outlook : dimensionality

- Combine with **dimension reduction**? *[A. Chatalic]*
  - First map in low-dimension, then sketch



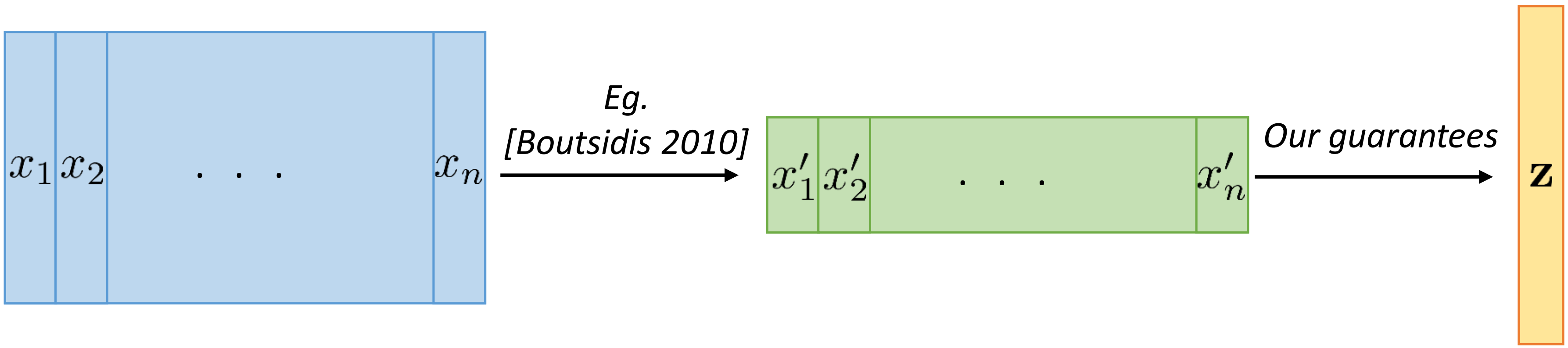
Outlook : dimensionality

- Combine with **dimension reduction**? *[A. Chatalic]*
  - First map in low-dimension, then sketch
  - Use fast transforms *[eg. Le 2013]*



Outlook : dimensionality

- Combine with **dimension reduction**? *[A. Chatalic]*
  - First map in low-dimension, then sketch
  - Use fast transforms *[eg. Le 2013]*



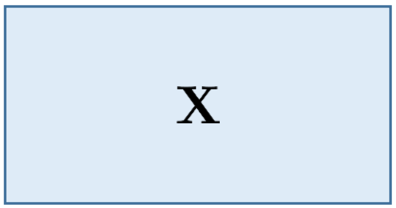
- More generally, extend the idea

Random sampling = intrinsic dimensionality

Oliva2016

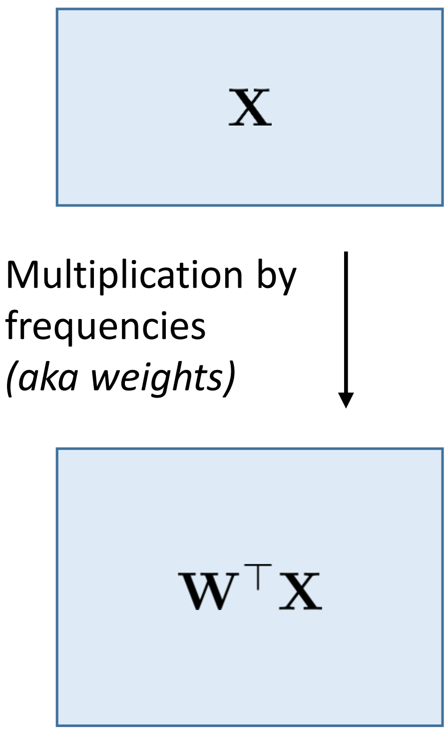
Outlook : neural networks

- Extension to multi-layer sketches ?



Outlook : neural networks

- Extension to multi-layer sketches ?



Outlook : neural networks

- Extension to multi-layer sketches ?

$$\mathbf{X}$$

Multiplication by frequencies *(aka weights)*

$$\mathbf{W}^\top \mathbf{X}$$

Complex exponential *(aka pointwise non-linearity)*

$$\rho\left(\mathbf{W}^\top \mathbf{X}\right)$$

Outlook : neural networks

- Extension to multi-layer sketches ?

$\mathbf{X}$

Multiplication by frequencies *(aka weights)*

$\mathbf{W}^\top \mathbf{X}$

Complex exponential *(aka pointwise non-linearity)*

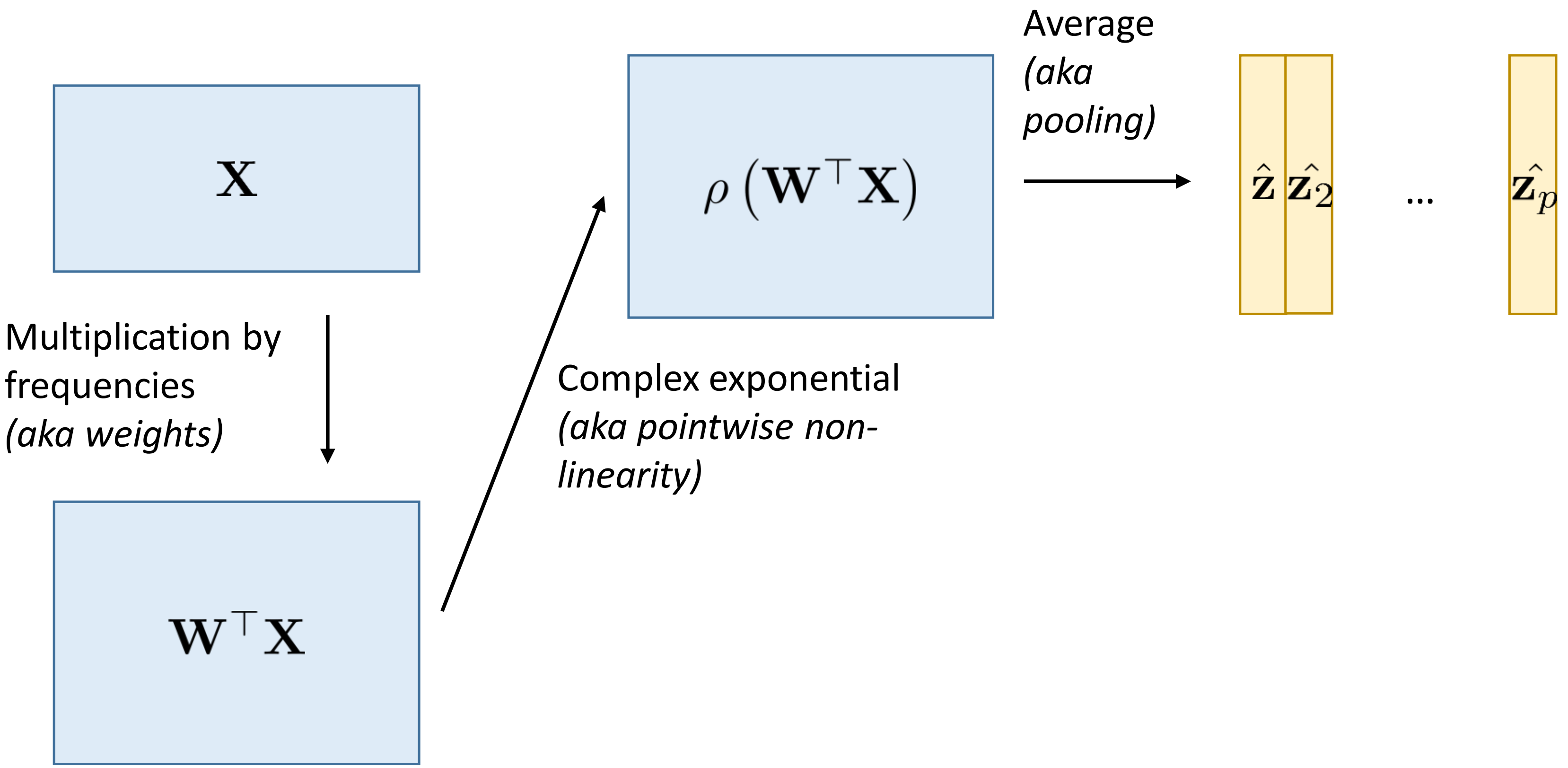
$\rho\left(\mathbf{W}^\top \mathbf{X}\right)$

Average *(aka pooling)*

$\hat{\mathbf{z}}$

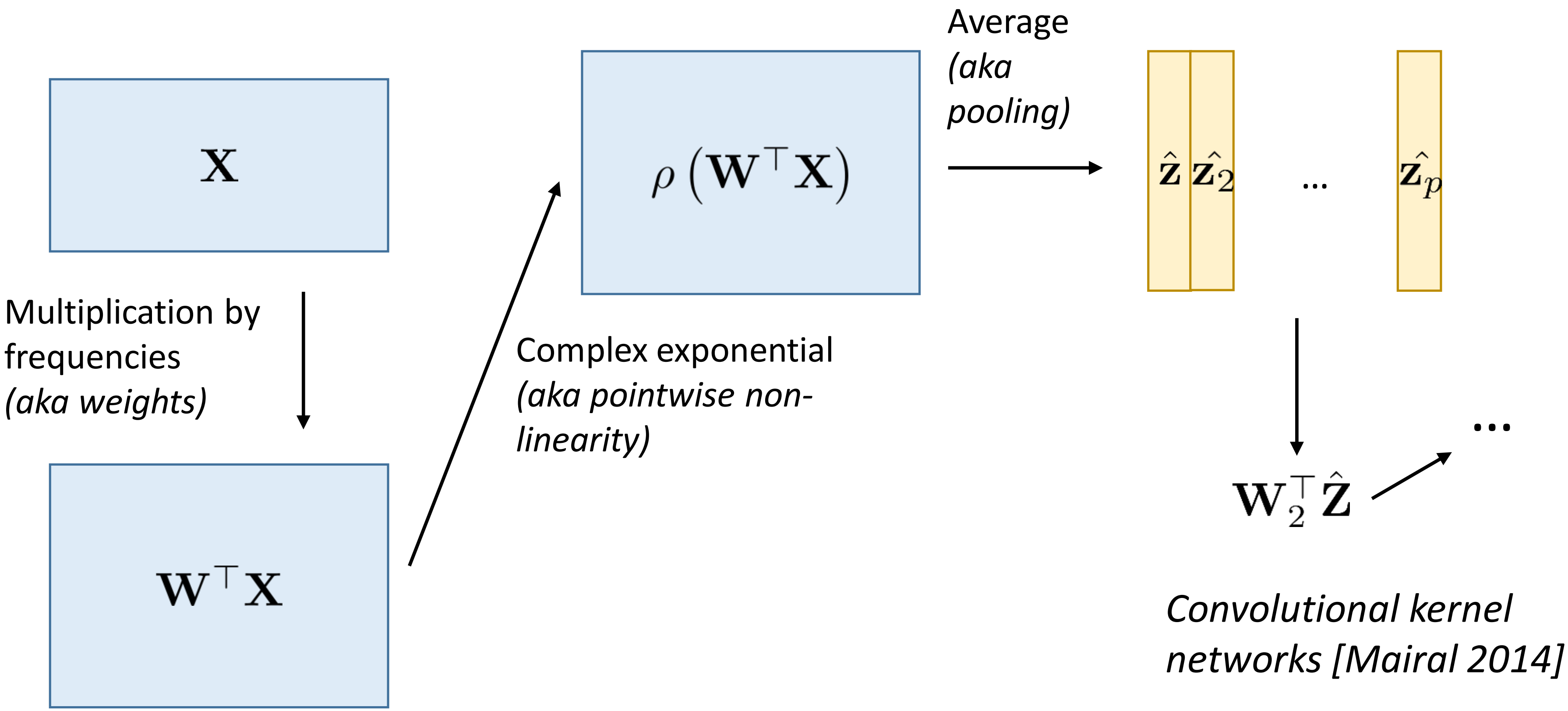
Outlook : neural networks

- Extension to multi-layer sketches ?



Outlook : neural networks

- Extension to multi-layer sketches ?



Outlook : neural networks

- Extension to multi-layer sketches ?

$\mathbf{X}$

Multiplication by frequencies *(aka weights)*

$\mathbf{W}^\top \mathbf{X}$

Complex exponential *(aka pointwise non-linearity)*

$\rho\left(\mathbf{W}^\top \mathbf{X}\right)$

Average *(aka pooling)*

$\hat{\mathbf{z}}$ $\hat{\mathbf{z}}_2$ … $\hat{\mathbf{z}}_p$

$\mathbf{W}_2^\top \hat{\mathbf{Z}}$ …

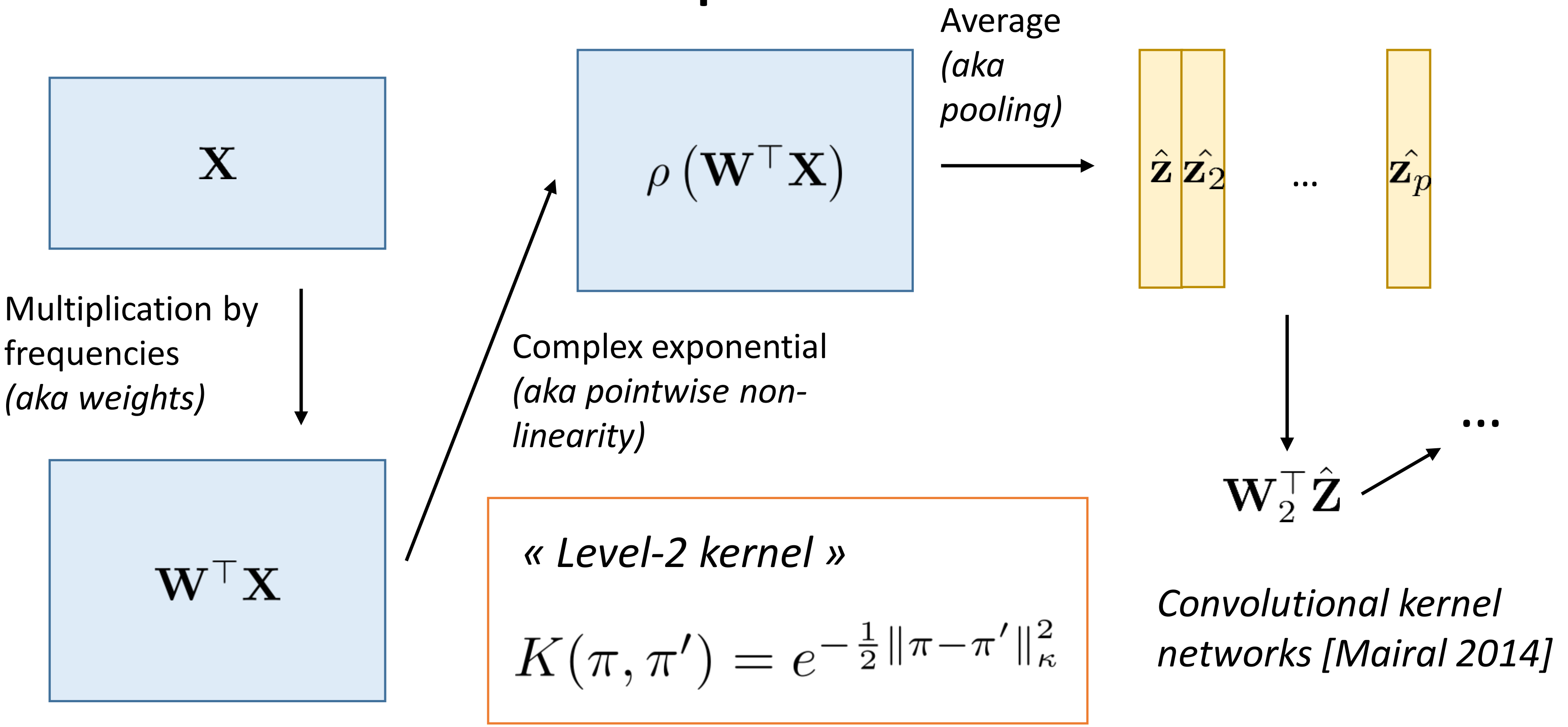
« Level-2 kernel »

$$K(\pi, \pi') = e^{-\frac{1}{2}\|\pi - \pi'\|_\kappa^2}$$

Convolutional kernel networks [Mairal 2014]

Outlook : neural networks

- Extension to multi-layer sketches ?
  - Equivalence between LRIP and robust information-preservation still valid for **non-linear operators**



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>

- **Code**: sketchml.gforge.inria.fr