Compressive learning and sketching for large-scale machine learning

Titouan Vayer

INRIA Lyon - Ockham Team

titouan.vayer@inria.fr

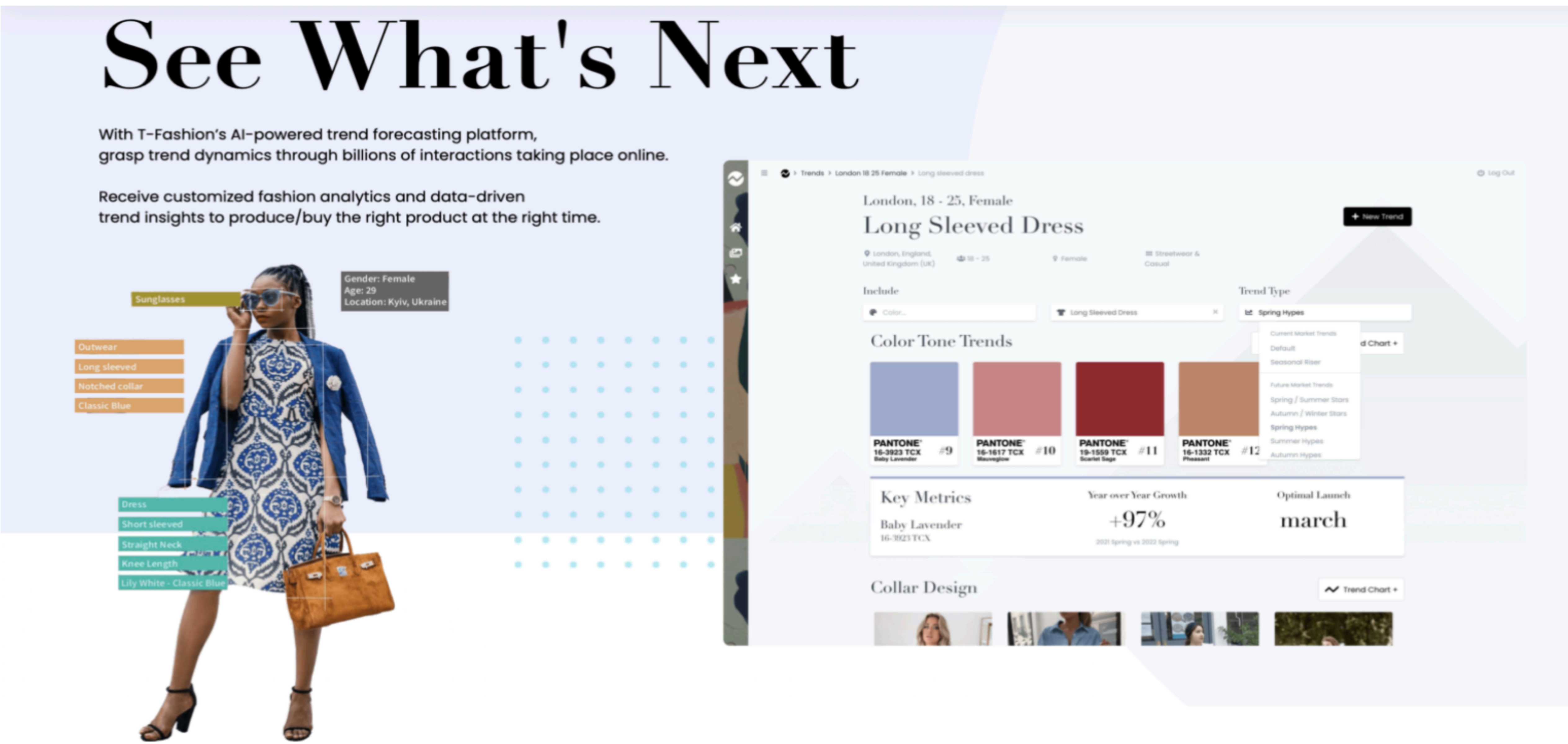
Motivations of this talk

- A modern fairy tale

Fashion Trend Forecasting with AI

JESSICA MAGALIT - NOVEMBER 9, 2021

0 0



Motivations of this talk

- A modern fairy tale

Motivations of this talk

- A modern fairy tale

The artificial intelligence revolution

18th March 2019
Electronic Specifier
Lanna Deamer

Predicting the future isn't magic, it's artificial intelligence.

DAVE WATERS

HR Technology

An AI revolution is underway, predicts Kirthiga Reddy

Hiring without bias, bringing diversity into organisations, eliminating unconscious bias: AI is game-changing, says the president of Athena SPACs.

Motivations of this talk

- A modern fairy tale

Motivations of this talk

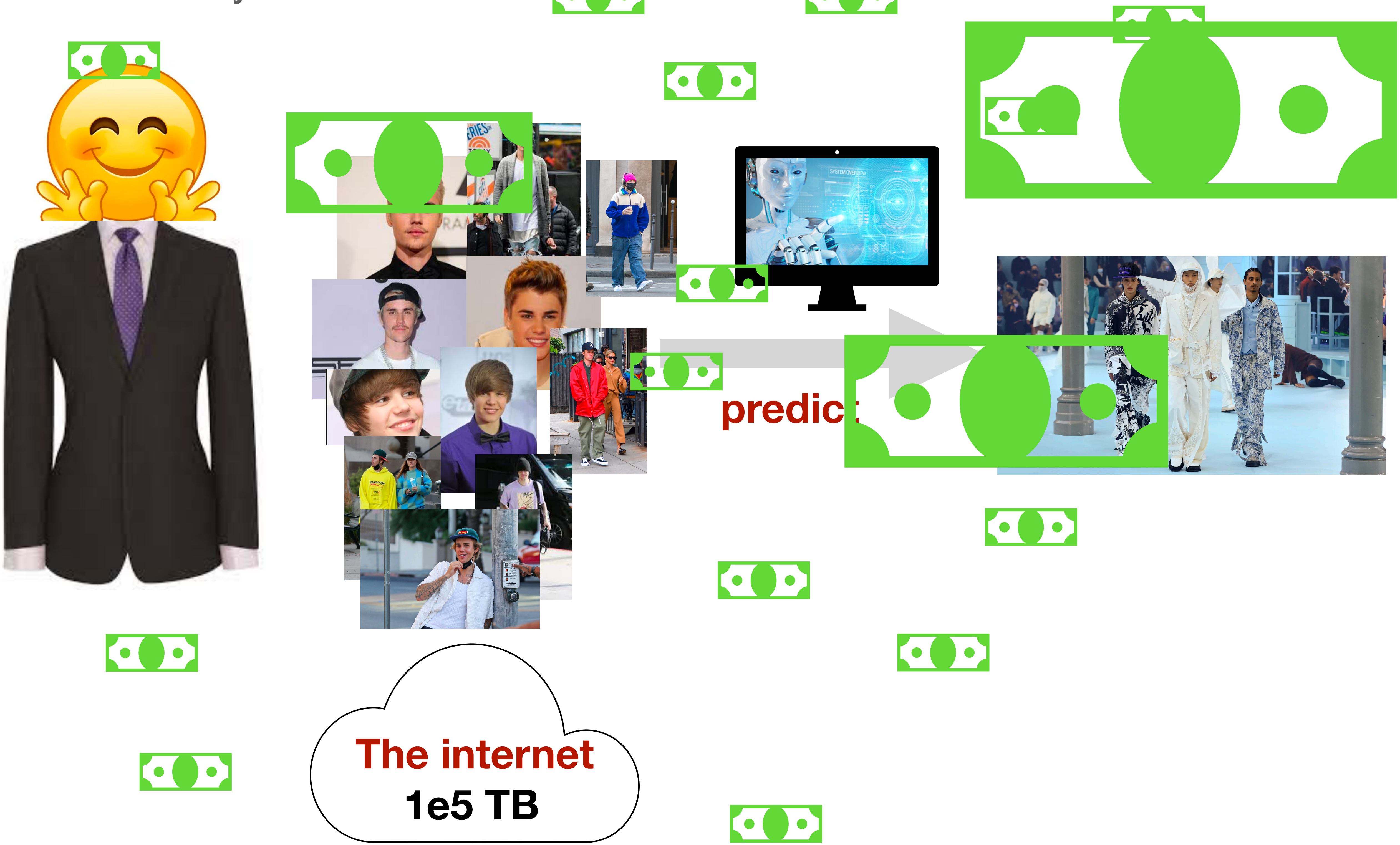
- A modern fairy tale

predict

The internet
1e5 TB

Motivations of this talk

- A modern fairy tale



Motivations of this talk

- A modern fairy tale

Motivations of this talk

- A modern fairy tale

Motivations of this talk

- A modern fairy tale

The internet

1e5 TB

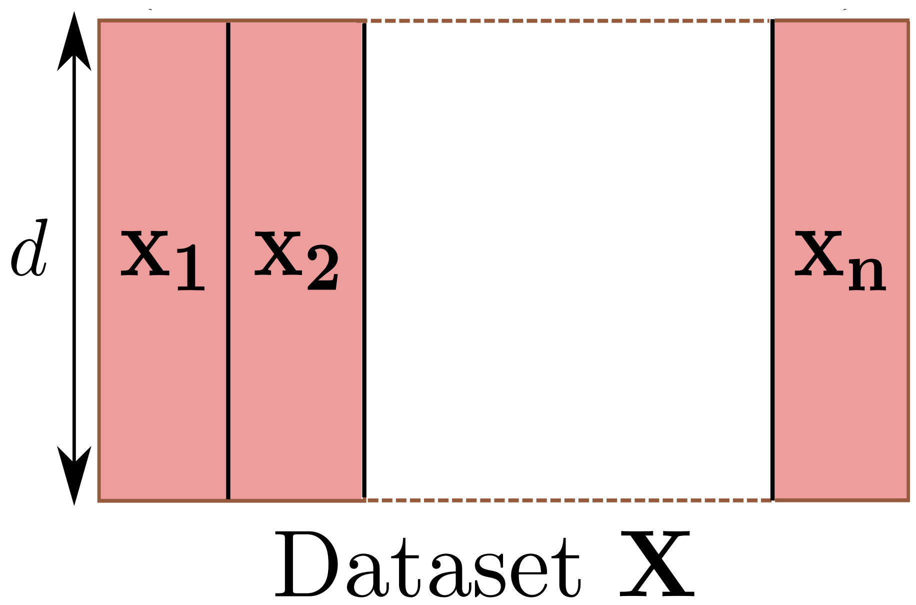
Motivations of this talk

- A modern fairy tale

Statistical correlation …

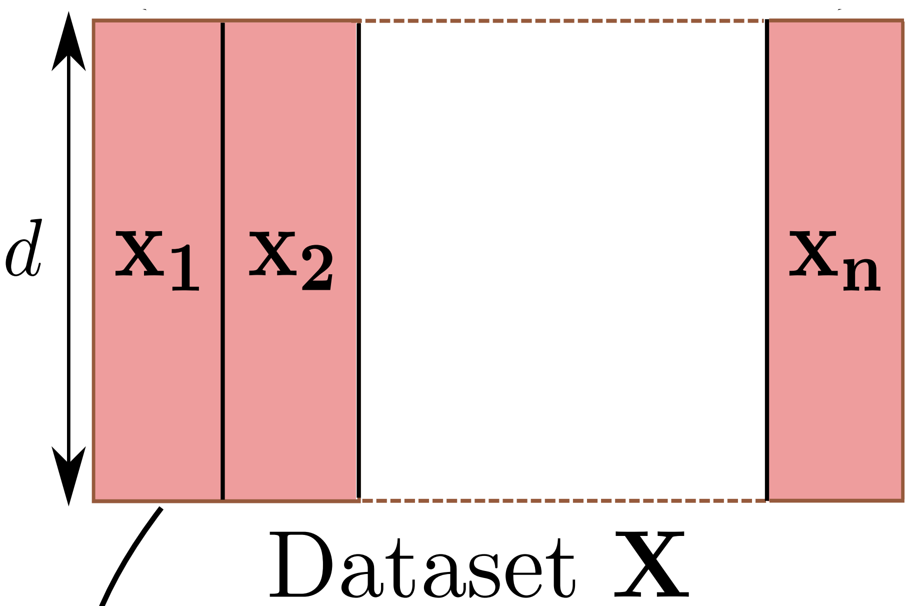
Motivations of this talk

- **Context:** this will not be a fairy tale



Motivations of this talk

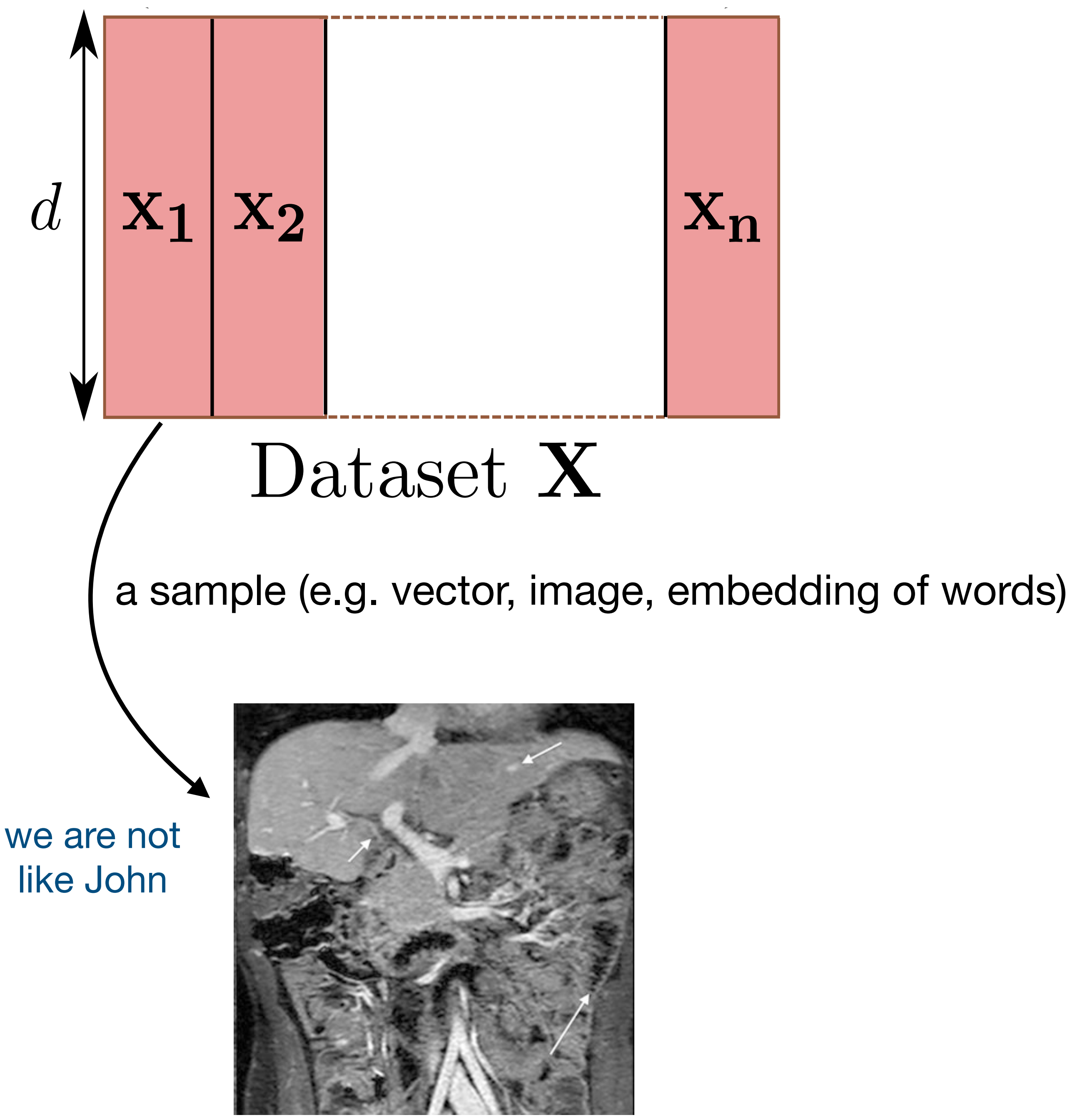
- **Context:** this will not be a fairy tale



a sample (e.g. vector, image, embedding of words)

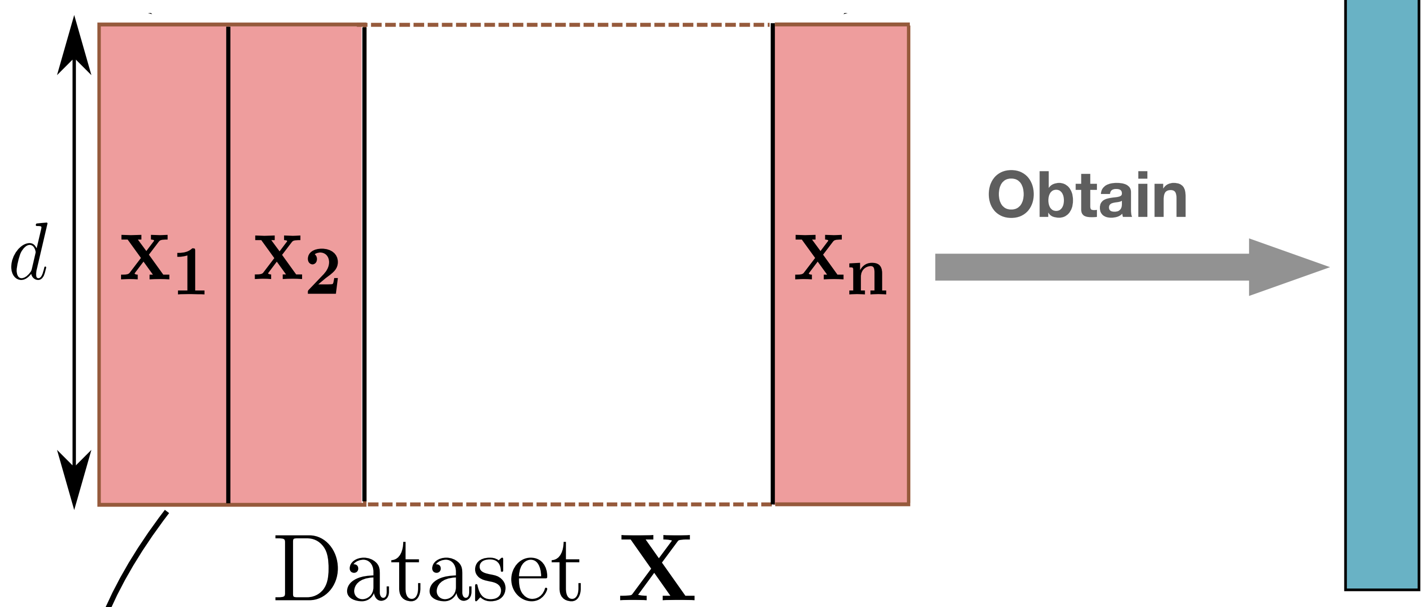
Motivations of this talk

- **Context:** this will not be a fairy tale



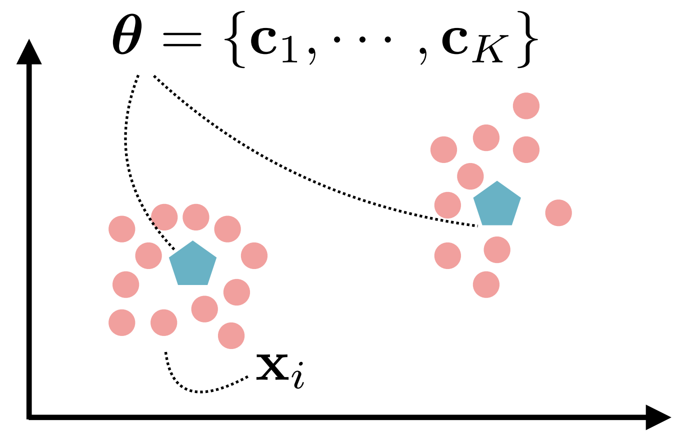
Motivations of this talk

- **Context:** this will not be a fairy tale



$\boldsymbol{\theta}$ **Parameters** that solves a specific tasks

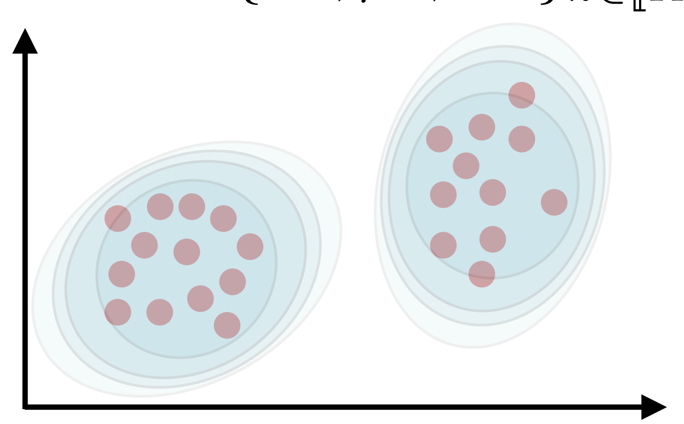
Example: K-means



a sample (e.g. vector, image, embedding of words)

Example: GMM fitting

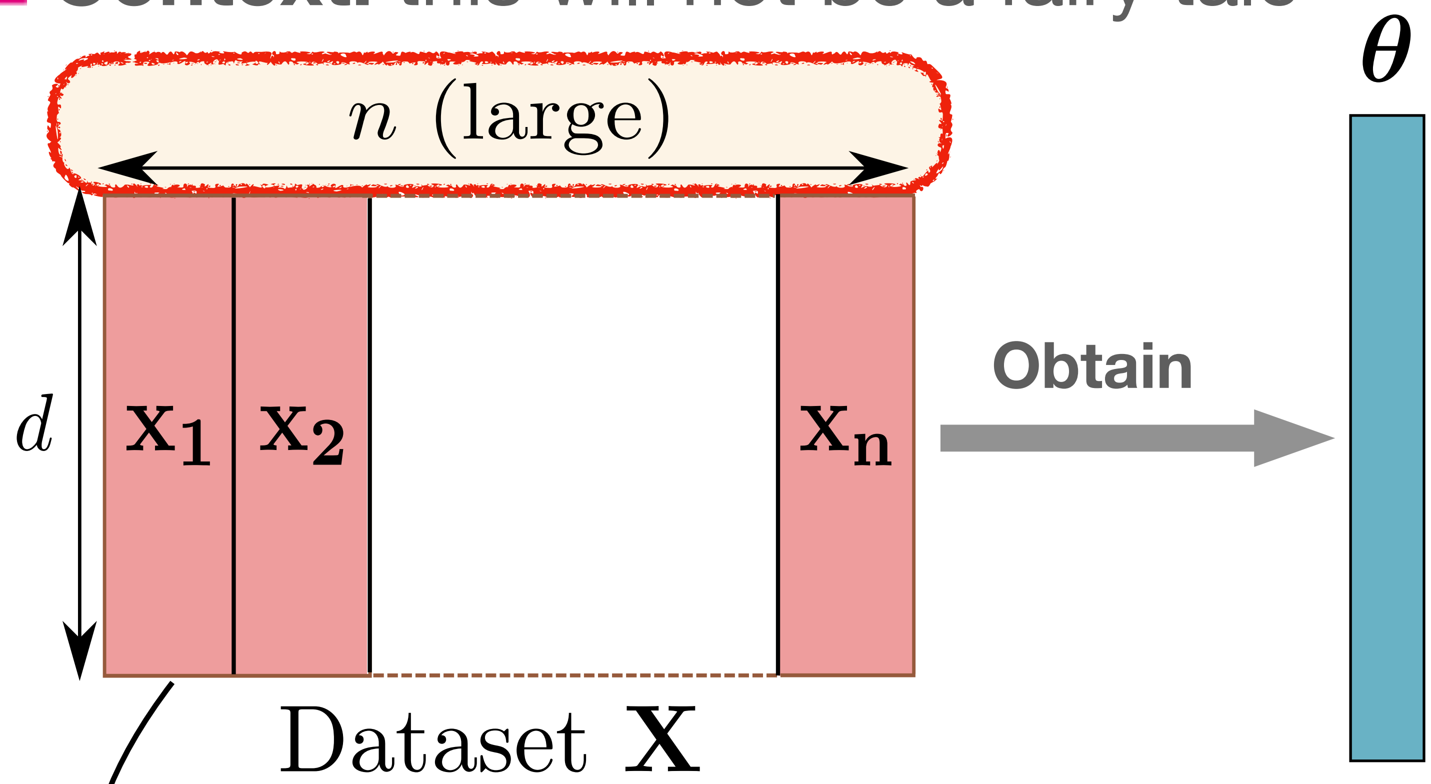
$\boldsymbol{\theta} = \{\alpha_k, \mu_k, \boldsymbol{\Sigma}_k\}_{k \in [\![K]\!]}$



we are not like John

Motivations of this talk

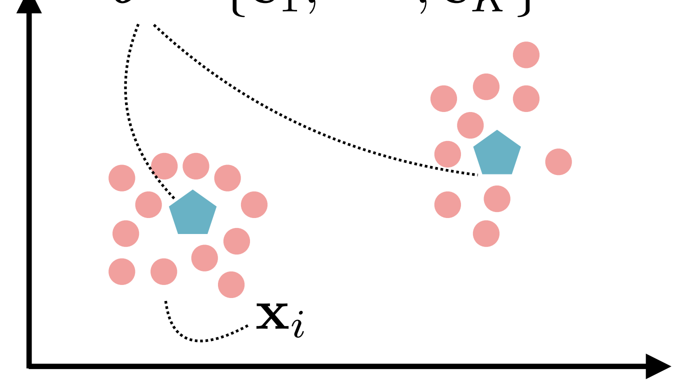
- **Context:** this will not be a fairy tale



Parameters that solves a specific tasks

Example: K-means

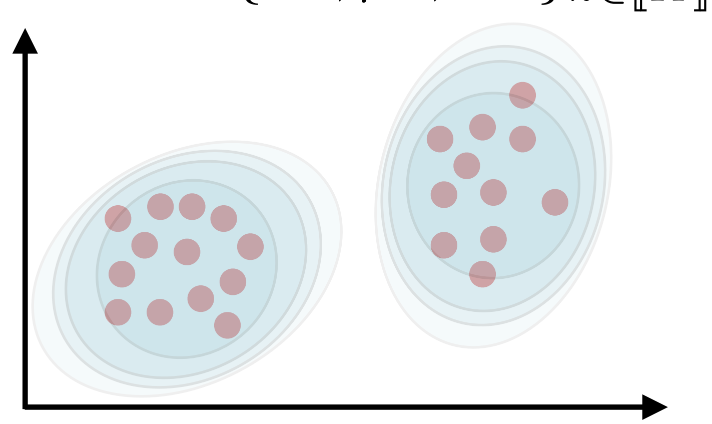
$$\boldsymbol{\theta} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$$



a sample (e.g. vector, image, embedding of words)

Example: GMM fitting

$$\boldsymbol{\theta} = \{\alpha_k, \mu_k, \boldsymbol{\Sigma}_k\}_{k \in [\![K]\!]}$$

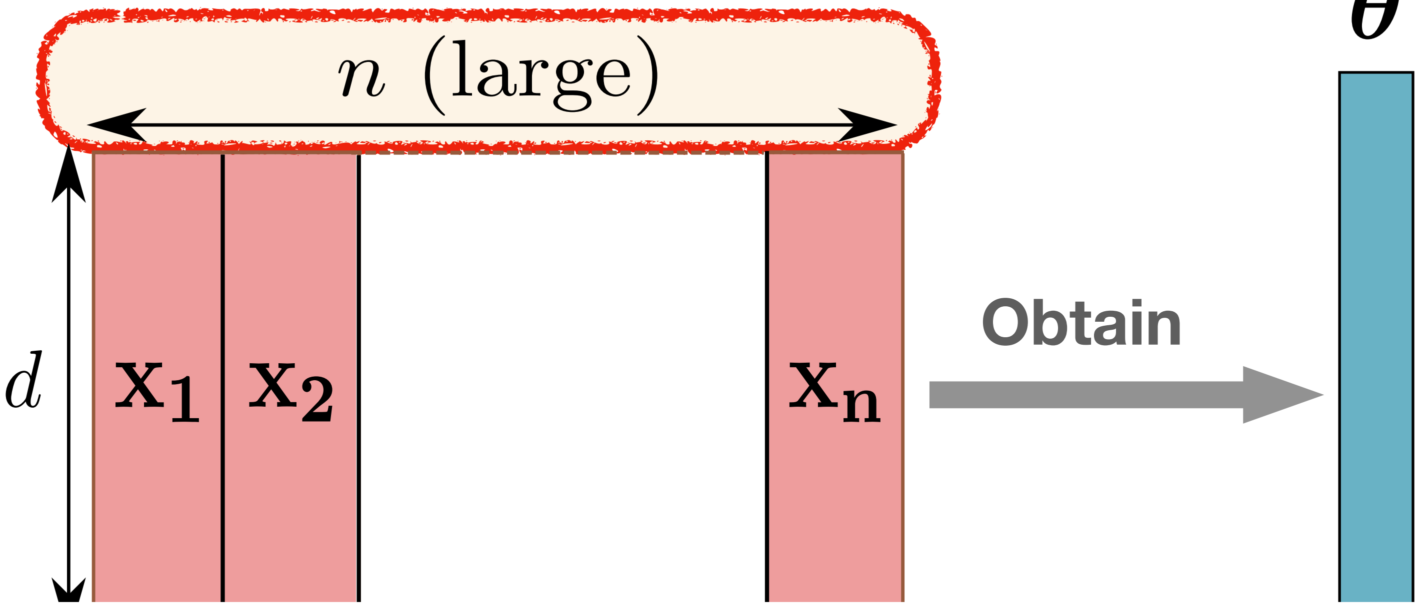


we are not like John

Large scale Machine Learning

Motivations of this talk

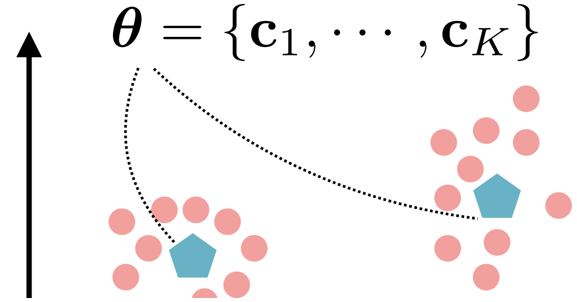
Context: this will not be a fairy tale



Parameters that solves a specific tasks

Example: K-means

$\boldsymbol{\theta} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$



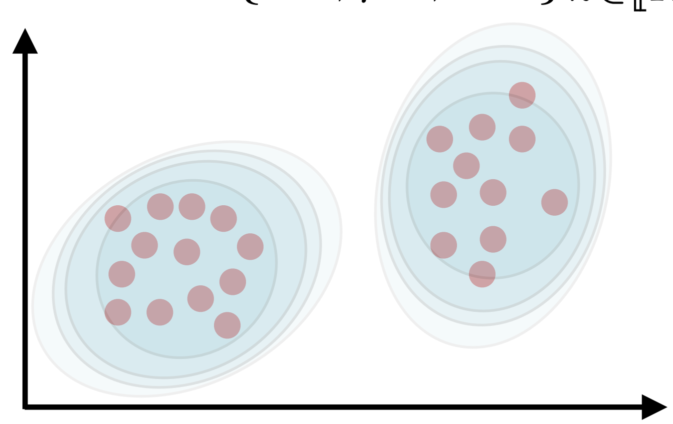
Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell, Ananya Ganesh, Andrew McCallum

we are not like John

Large scale Machine Learning

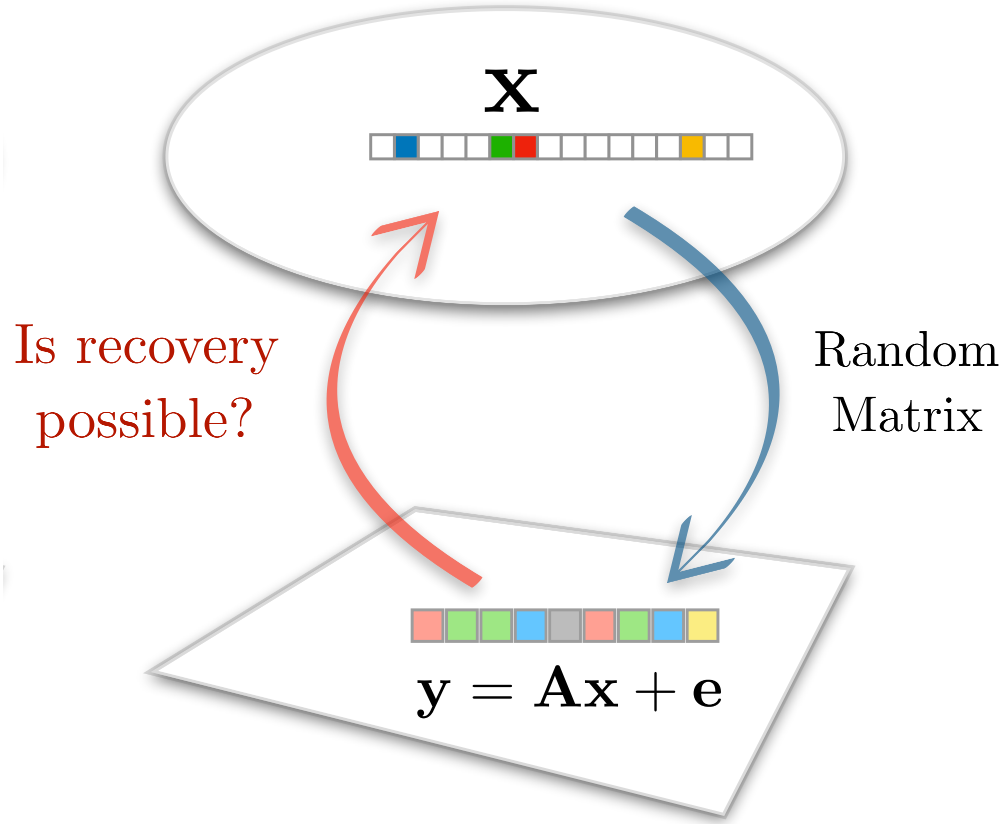
$\boldsymbol{\theta} = \{\alpha_k, \mu_k, \boldsymbol{\Sigma}_k\}_{k \in [\![K]\!]}$



Overview of the talk

- **Part I: A journey in the compressed sensing theory**
- **Part II: A bit of machine learning theory**
- **Part III: The sketching approach**
  - Applied sketching
  - Theoretical guarantees

A journey in the compressed sensing theory



Compressed sensing theory: an invitation

- **From the basements:**
  - A signal: $\mathbf{x} \in \mathbb{R}^d$
  - An acquisition system: $\mathbf{A} \in \mathbb{R}^{m \times d}$

Observation

$$\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$$

- **Goal: recover $\mathbf{x} \in \mathbb{R}^d$ from $\mathbf{y} \in \mathbb{R}^m$**

Compressed sensing theory: an invitation

From the basements:

- A signal: $\mathbf{x} \in \mathbb{R}^d$
- An acquisition system: $\mathbf{A} \in \mathbb{R}^{m \times d}$

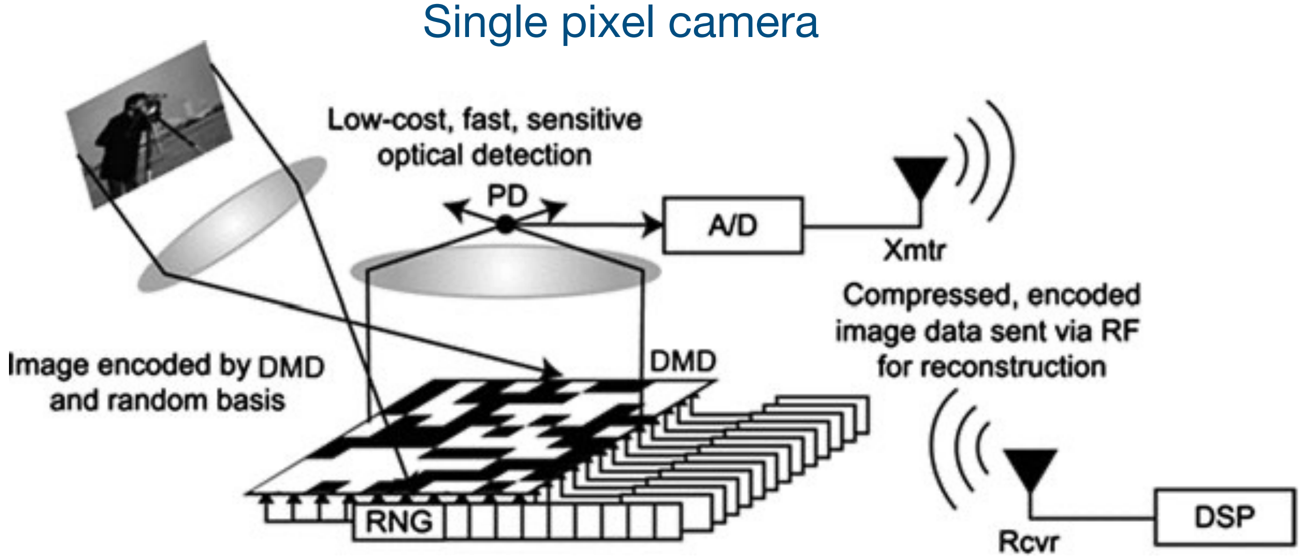
Observation

$$\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$$

Goal: recover $\mathbf{x} \in \mathbb{R}^d$ from $\mathbf{y} \in \mathbb{R}^m$

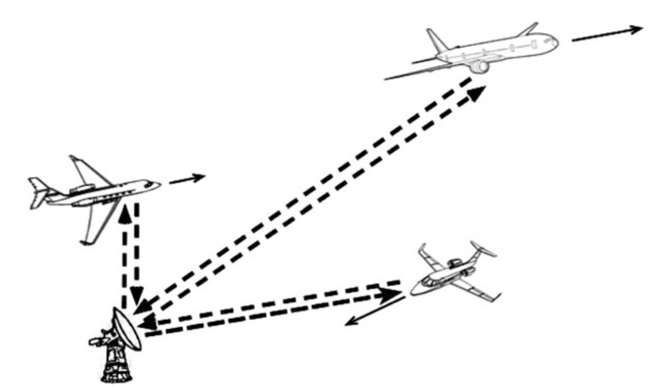
- Basis of most devices: analog-to-digital, medical imaging, radar, mobile

Single pixel camera



MRI

Radar



Compressed sensing theory: an invitation

- A signal: $\mathbf{x} \in \mathbb{R}^d$
- An acquisition system: $\mathbf{A} \in \mathbb{R}^{m \times d}$

Linear system: Find $\mathbf{x} \in \mathbb{R}^d$ s.t. $\mathbf{y} = \mathbf{A}\mathbf{x}$

- **What does the theory says ?** $\emptyset, 1, \infty$

Compressed sensing theory: an invitation

- A signal: $\mathbf{x} \in \mathbb{R}^d$
- An acquisition system: $\mathbf{A} \in \mathbb{R}^{m \times d}$

Linear system: Find $\mathbf{x} \in \mathbb{R}^d$ s.t. $\mathbf{y} = \mathbf{A}\mathbf{x}$

What does the theory says ? $\emptyset, 1, \infty$

Determined $m = d$

$$\mathbf{y} = \mathbf{A} \mathbf{x}$$

As much data as the size of the signal

If well behaved unique sol

$$\mathbf{x} := \mathbf{A}^{-1}\mathbf{y}$$

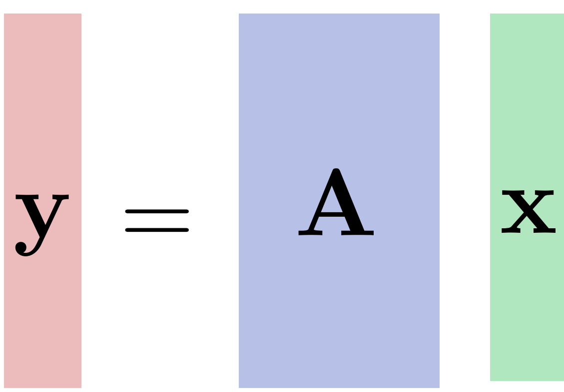
Compressed sensing theory: an invitation

- A signal: $\mathbf{x} \in \mathbb{R}^d$
- An acquisition system: $\mathbf{A} \in \mathbb{R}^{m \times d}$

Linear system: Find $\mathbf{x} \in \mathbb{R}^d$ s.t. $\mathbf{y} = \mathbf{A}\mathbf{x}$

What does the theory says ? $\emptyset, 1, \infty$

Determined $m = d$

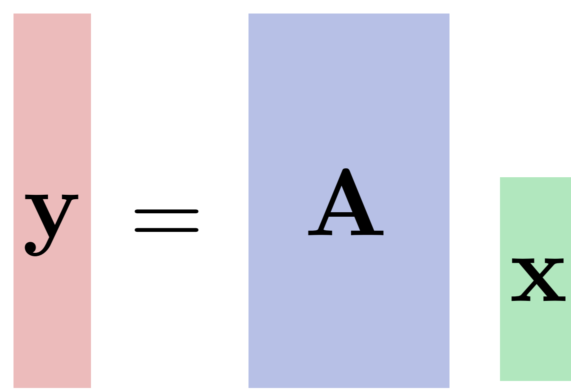


As much data as the size of the signal

If well behaved unique sol

$$\mathbf{x} := \mathbf{A}^{-1}\mathbf{y}$$

Overdetermined $m > d$



We have more data

From low dim to high dim

In general no sol:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2$$

[Gauss, 1795]

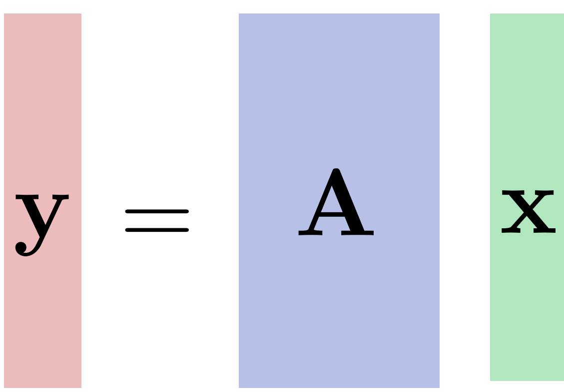
Compressed sensing theory: an invitation

- A signal: $\mathbf{x} \in \mathbb{R}^d$
- An acquisition system: $\mathbf{A} \in \mathbb{R}^{m \times d}$

Linear system: Find $\mathbf{x} \in \mathbb{R}^d$ s.t. $\mathbf{y} = \mathbf{A}\mathbf{x}$

What does the theory says ? $\emptyset, 1, \infty$

Determined $m = d$

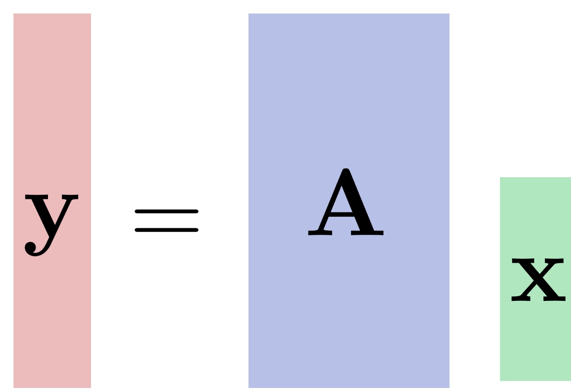


As much data as the size of the signal

If well behaved unique sol

$$\mathbf{x} := \mathbf{A}^{-1}\mathbf{y}$$

Overdetermined $m > d$



We have more data

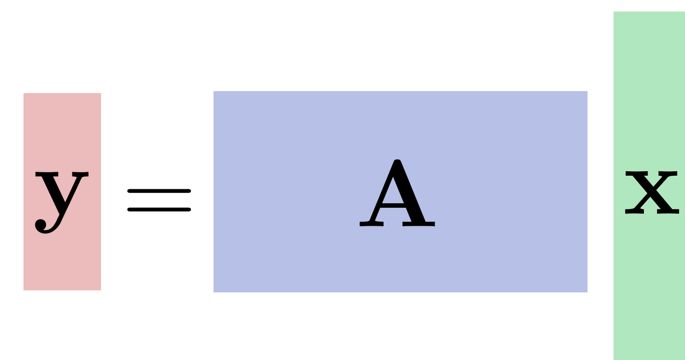
From low dim to high dim

In general no sol:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2$$

[Gauss, 1795]

Underdetermined $m < d$



We did compression

From high dim to low dim

Infinity of solutions

$$\mathbf{x}_0 + \lambda \mathbf{z}, \mathbf{z} \in \ker(\mathbf{A})$$

Compressed sensing theory: an invitation

- A signal: $\mathbf{x} \in \mathbb{R}^d$
- An acquisition system: $\mathbf{A} \in \mathbb{R}^{m \times d}$

Linear system: Find $\mathbf{x} \in \mathbb{R}^d$ s.t. $\mathbf{y} = \mathbf{A}\mathbf{x}$

What does the theory says ? $\emptyset, 1, \infty$

Determined $m = d$

$$\mathbf{y} = \mathbf{A} \mathbf{x}$$

As much data as the size of the signal

If well behaved unique sol

$$\mathbf{x} := \mathbf{A}^{-1}\mathbf{y}$$

Overdetermined $m > d$

$$\mathbf{y} = \mathbf{A} \mathbf{x}$$

We have more data

From low dim to high dim

In general no sol:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2$$

[Gauss, 1795]

Underdetermined $m < d$

$$\mathbf{y} = \mathbf{A} \mathbf{x}$$

We did compression

From high dim to low dim

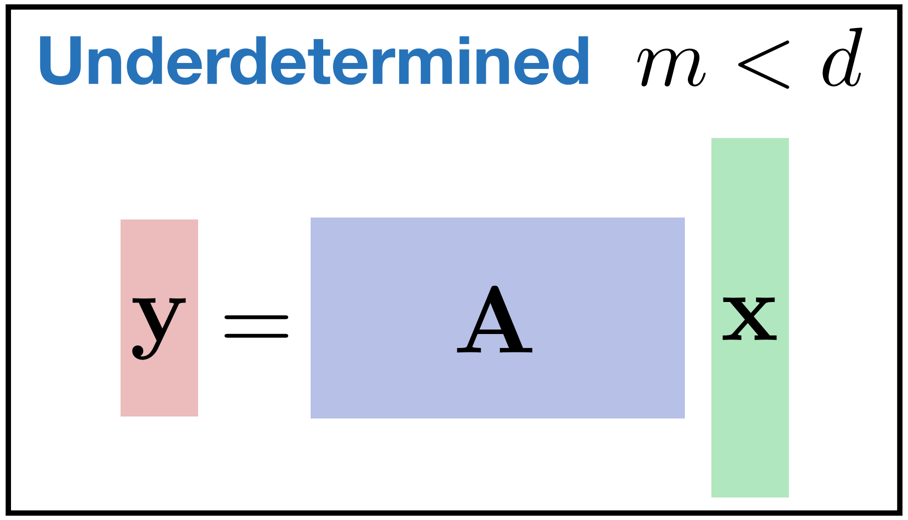
Infinity of solutions

$$\mathbf{x}_0 + \lambda \mathbf{z}, \mathbf{z} \in \ker(\mathbf{A})$$

Compressed sensing theory: an invitation

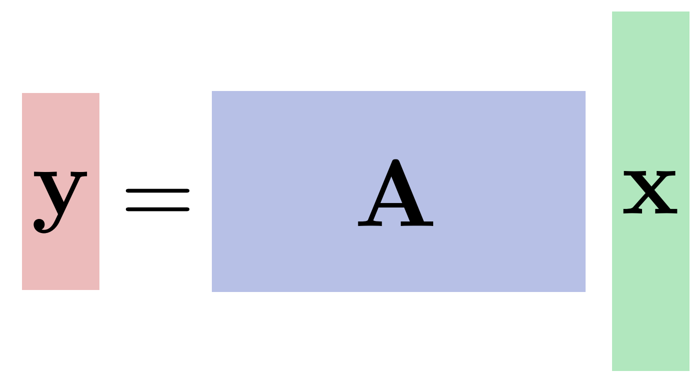
How can we recover $\mathbf{x}$?

- Ill-posed **inverse problem**
- Infinity of solutions



Compressed sensing theory: an invitation

Underdetermined $m < d$



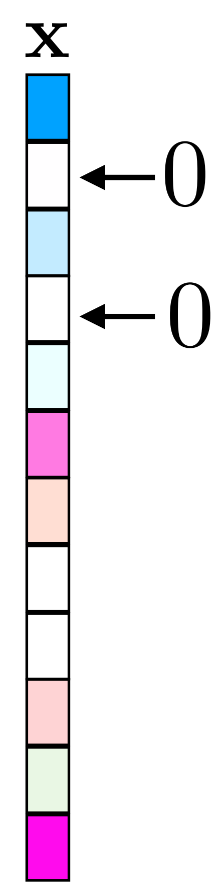
How can we recover $\mathbf{x}$?

Ill-posed **inverse problem**

Infinity of solutions

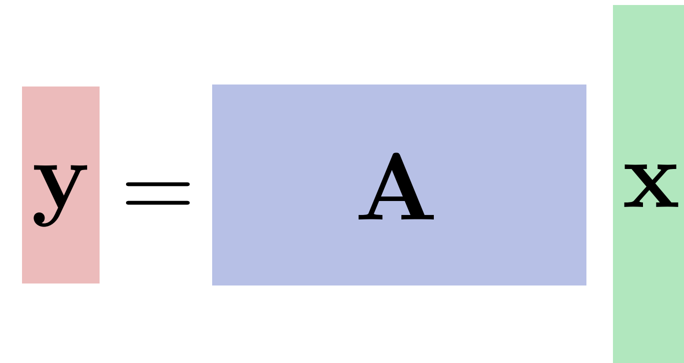
The sparse assumption

- Recovery is possible when we know something more about $\mathbf{x}$
- The signal $\mathbf{x} \in \mathbb{R}^d$ is high-dim **but it is full of zeros**



Compressed sensing theory: an invitation

Underdetermined $m < d$



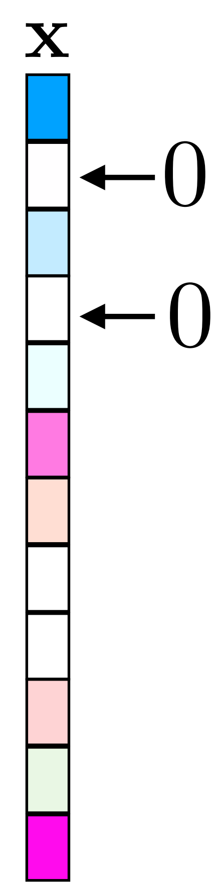
How can we recover $\mathbf{x}$?

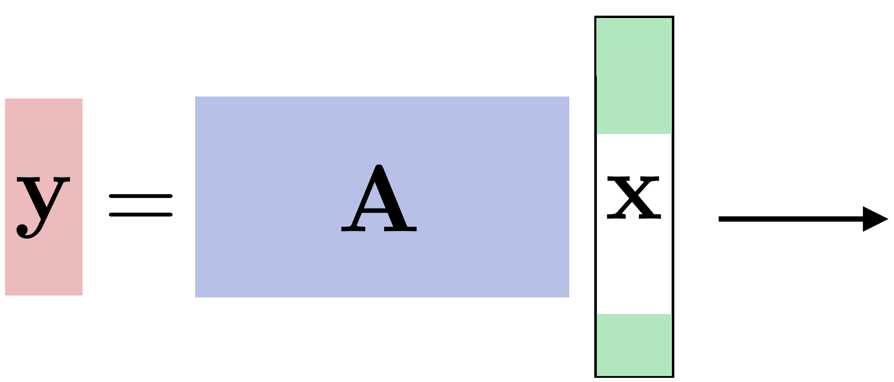
Ill-posed **inverse problem**

Infinity of solutions

The sparse assumption

- Recovery is possible when we know something more about $\mathbf{x}$
- The signal $\mathbf{x} \in \mathbb{R}^d$ is high-dim **but it is full of zeros**
- In a way $\mathbf{x}$ lives in a low-dim. space





We need much less information to recover $\mathbf{x}$

Compressed sensing theory: an invitation

Underdetermined $m < d$

$$\mathbf{y} = \mathbf{A} \mathbf{x}$$

How can we recover $\mathbf{x}$?

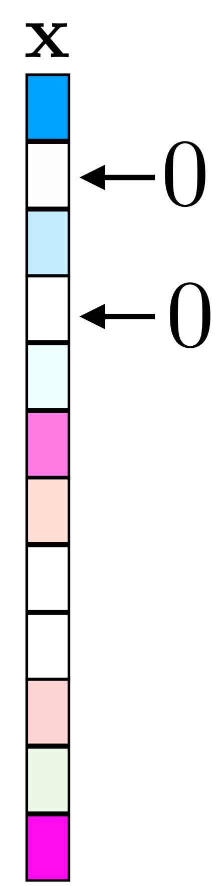
- Ill-posed **inverse problem**
- Infinity of solutions

The sparse assumption

- Recovery is possible when we know something more about $\mathbf{x}$
- The signal $\mathbf{x} \in \mathbb{R}^d$ is high-dim **but it is full of zeros**
- In a way $\mathbf{x}$ lives in a low-dim. space

$$\mathbf{y} = \mathbf{A} \mathbf{x}$$

We need much less information to recover $\mathbf{x}$



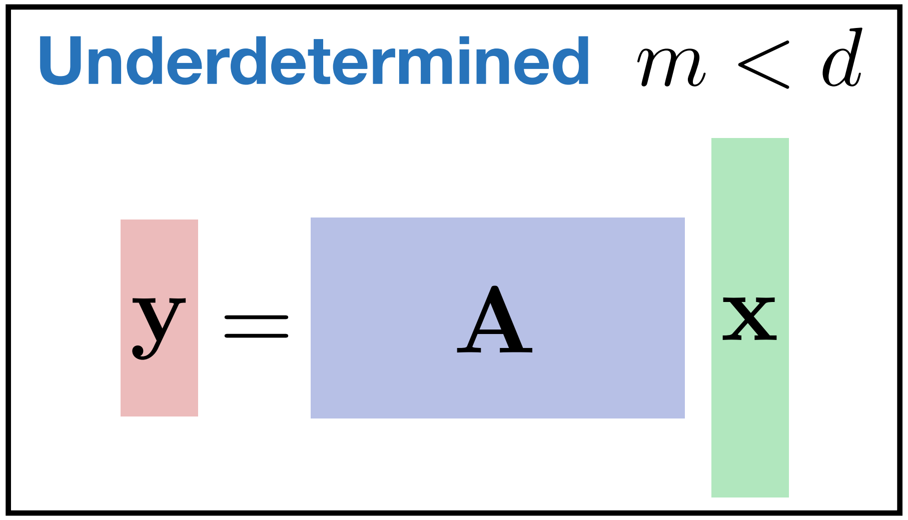
- Number of nnz components

$$\|\mathbf{x}\|_0 = \sum_{i=1}^{d} \mathbf{1}_{x_i \neq 0}$$

- E.g. $\|\mathbf{x}\|_0 \leq k$

$\boxed{\mathbf{x} \text{ is } k\text{- sparse}}$

Compressed sensing theory: an invitation



How can we recover $\mathbf{x}$?

Ill-posed **inverse problem**

Infinity of solutions

The sparse assumption

- The signal $\mathbf{x} \in \mathbb{R}^d$ is **sparse**
- We could solve:

$$(1) \quad \min_{\mathbf{x} \text{ s.t. } \mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$$

Compressed sensing theory: an invitation

Underdetermined $m < d$

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

How can we recover $\mathbf{x}$?

Ill-posed **inverse problem**

Infinity of solutions

The sparse assumption

- The signal $\mathbf{x} \in \mathbb{R}^d$ is **sparse**
- We could solve:

$$(1) \quad \min_{\mathbf{x} \text{ s.t. } \mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$$

Find the sparsest vector

Compressed sensing theory: an invitation

Underdetermined $m < d$

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

How can we recover $\mathbf{x}$?

Ill-posed **inverse problem**

Infinity of solutions

The sparse assumption

- The signal $\mathbf{x} \in \mathbb{R}^d$ is **sparse**
- We could solve:

$$(1) \quad \min_{\mathbf{x} \text{ s.t. } \mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$$

Which satisfies the constraints

Compressed sensing theory: an invitation

- **How can we recover $\mathbf{x}$?**
  - Ill-posed **inverse problem**
  - Infinity of solutions

Underdetermined $m < d$

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

- **The sparse assumption**
  - The signal $\mathbf{x} \in \mathbb{R}^d$ is **sparse**
  - We could solve:

$$(1) \quad \min_{\mathbf{x} \text{ s.t. } \mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$$

Theorem $d \geq 2k$

$\exists \mathbf{A} \in \mathbb{R}^{2k \times d}$ $(m = 2k)$

<u>every</u> k-sparse $\mathbf{x}$

can be recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$

by solving (1)

Compressed sensing theory: an invitation

Underdetermined $m < d$

$$\mathbf{y} = \mathbf{A} \mathbf{x}$$

How can we recover $\mathbf{x}$?

Ill-posed **inverse problem**

Infinity of solutions

The sparse assumption

- The signal $\mathbf{x} \in \mathbb{R}^d$ is **sparse**
- We could solve:

$$(1) \quad \min_{\mathbf{x} \text{ s.t. } \mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$$

Theorem $d \geq 2k$

$\exists \mathbf{A} \in \mathbb{R}^{2k \times d}\, (m = 2k)$

<u>every</u> k-sparse $\mathbf{x}$

can be recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$

by solving (1)

Is it done ? No …

- Not robust to noise $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$
- (1) is NP-hard
- Not robust w.r.t. sparsity level

Compressed sensing theory: an invitation

- **How can we recover $\mathbf{x}$?**
  - Ill-posed inverse problem
  - Infinity of solutions

Underdetermined $m < d$

$$\mathbf{y} = \mathbf{A} \mathbf{x}$$

- **The sp**
  - Th
  - W

$= \mathbf{A}\mathbf{x}$

by solving (1)

$$(1) \quad \min_{\mathbf{x} \text{ s.t. } \mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$$

Is $\mathbf{x}$ sparse in practice ??

- **Is it done ? No …**
  - Not robust to noise $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$
  - (1) is NP-hard
  - Not robust w.r.t. sparsity level

Compressed sensing theory: an invitation

- Most data encountered are sparse in another representation

$$\text{Original signal} \longrightarrow \mathbf{x} = \mathbf{D}\mathbf{x}_0 \longleftarrow \text{sparse vector}$$

- $\mathbf{D} \in \mathbb{R}^{d \times d}$ is another ortho. basis than the canonical one $\mathbf{D}^{-1} = \mathbf{D}^\top$

$$\mathbf{x} = \mathbf{D}\mathbf{x}_0 \iff \mathbf{x}_0 = \mathbf{D}^\top \mathbf{x}$$

Compressed sensing theory: an invitation

- Most data encountered are sparse in another representation

$$\text{Original signal} \longrightarrow \mathbf{x} = \mathbf{D}\mathbf{x}_0 \longleftarrow \text{sparse vector}$$

- $\mathbf{D} \in \mathbb{R}^{d\times d}$ is another ortho. basis than the canonical one $\mathbf{D}^{-1} = \mathbf{D}^\top$
  - Discrete Fourier basis (FFT)
  - Discrete Cosinus Transform
  - Wavelet

$$D_{pq} = \exp(-2i\tfrac{\pi}{d}pq) \qquad D_{pq} = \cos(\tfrac{\pi}{d}(p+\tfrac{1}{2})q)$$

[Haar, 1909]
[Gabor, 1946]
[Morlet & Grossmann, 1984]
[Mallat, 1986]
[Daubechies,1987]

Compressed sensing theory: an invitation

- Most data encountered are sparse in another representation

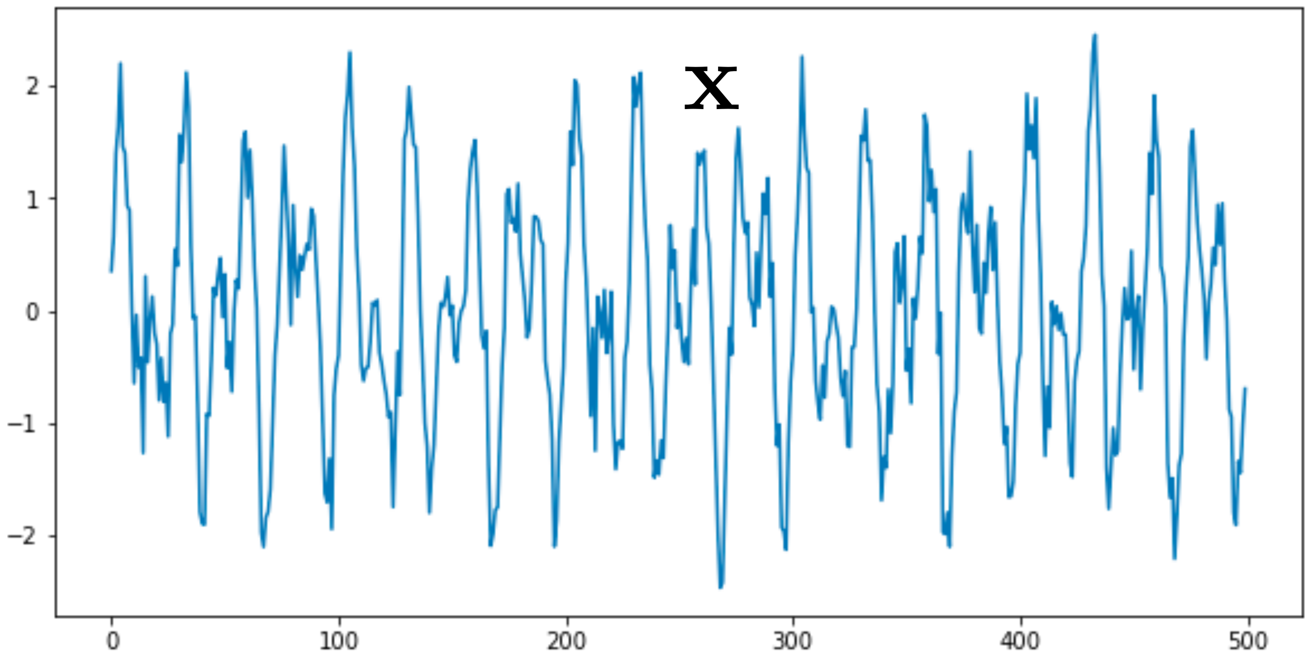
$$\text{Original signal} \longrightarrow \mathbf{x} = \mathbf{D}\mathbf{x}_0 \longleftarrow \text{sparse vector}$$

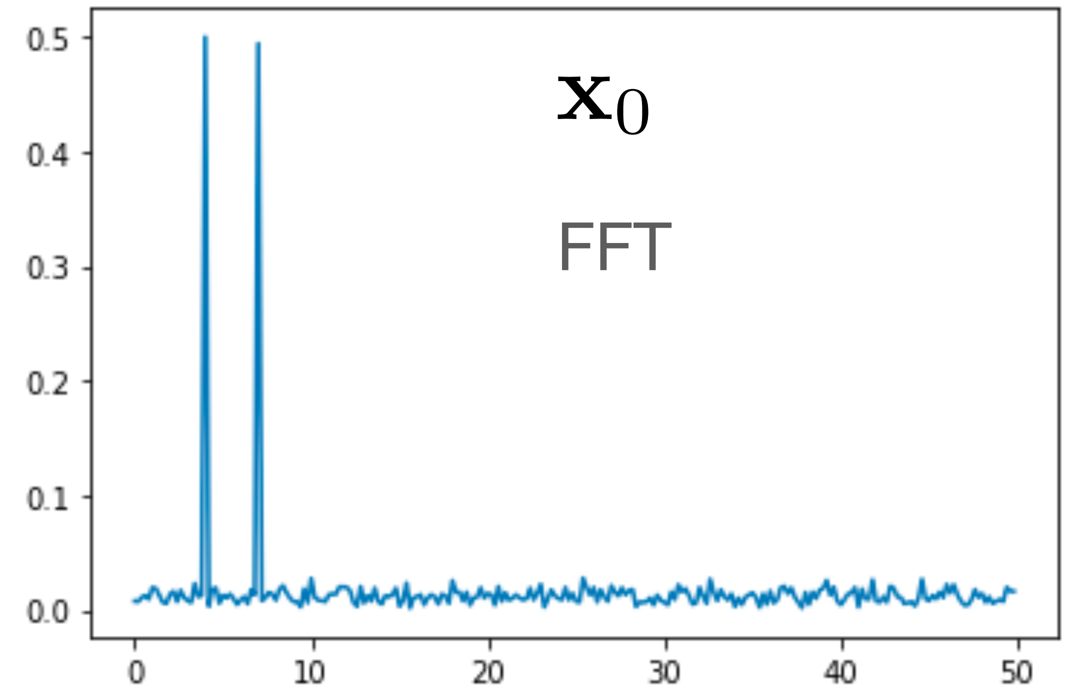
- $\mathbf{D} \in \mathbb{R}^{d\times d}$ is another ortho. basis than the canonical one $\mathbf{D}^{-1} = \mathbf{D}^\top$

  - Discrete Fourier basis (FFT)
  - Discrete Cosinus Transform
  - Wavelet

$$D_{pq} = \exp(-2i\frac{\pi}{d}pq) \qquad D_{pq} = \cos(\frac{\pi}{d}(p+\frac{1}{2})q)$$

- Quite magical …





Compressed sensing theory: an invitation

- Most data encountered are sparse in another representation

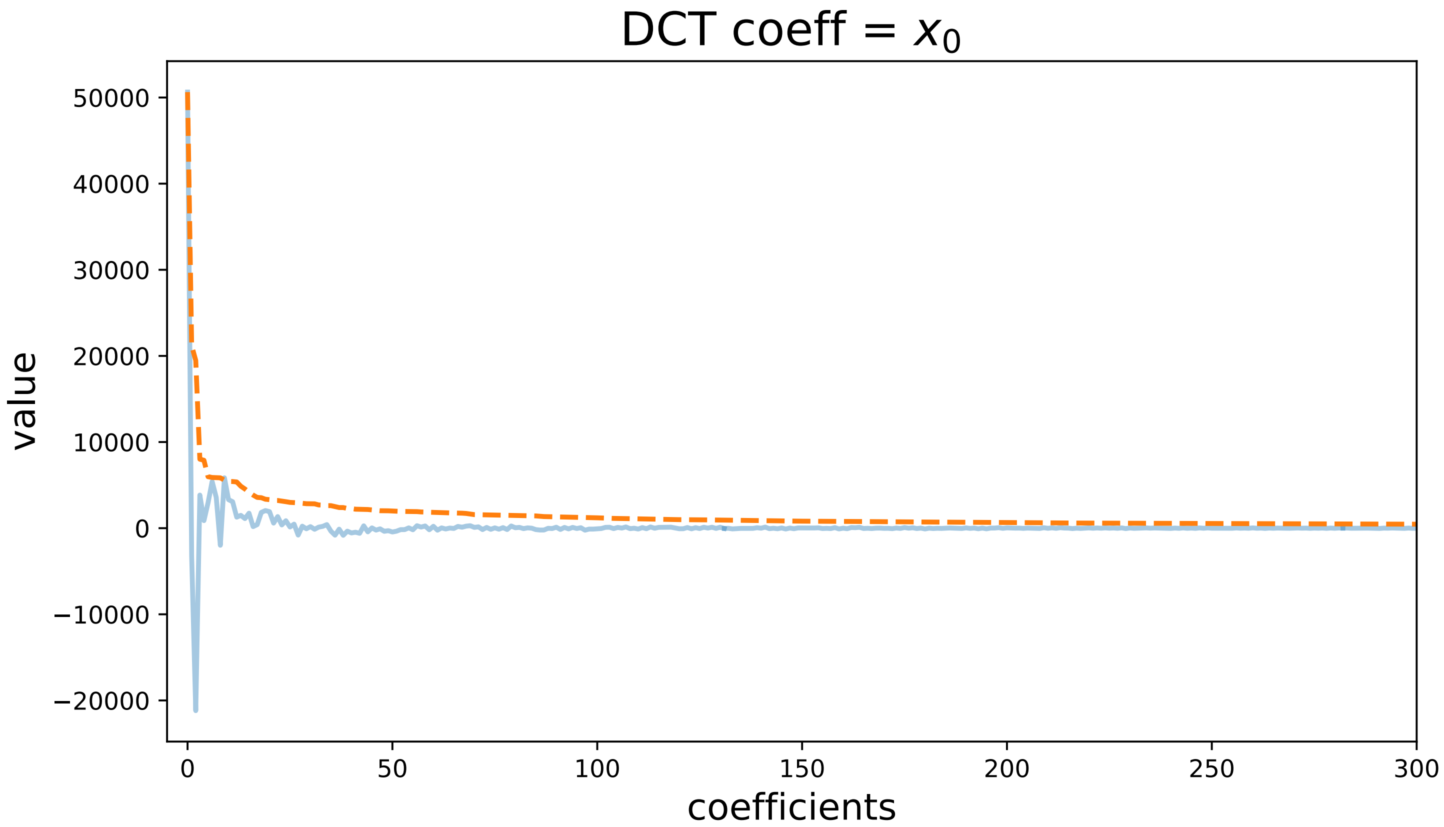
$$\text{Original signal} \longrightarrow \mathbf{x} = \mathbf{D}\mathbf{x}_0 \longleftarrow \text{sparse vector}$$

- $\mathbf{D} \in \mathbb{R}^{d \times d}$ is another ortho. basis than the canonical one $\mathbf{D}^{-1} = \mathbf{D}^\top$
  - Discrete Fourier basis (FFT)
  - Discrete Cosinus Transform
  - Wavelet

$$D_{pq} = \exp(-2i\tfrac{\pi}{d}pq) \qquad D_{pq} = \cos(\tfrac{\pi}{d}(p + \tfrac{1}{2})q)$$

- Quite magical …

$d = 512 * 512$



Compressed sensing theory: an invitation

- Most data encountered are sparse in another representation

$$\text{Original signal} \longrightarrow \mathbf{x} = \mathbf{D}\mathbf{x}_0 \longleftarrow \text{sparse vector}$$

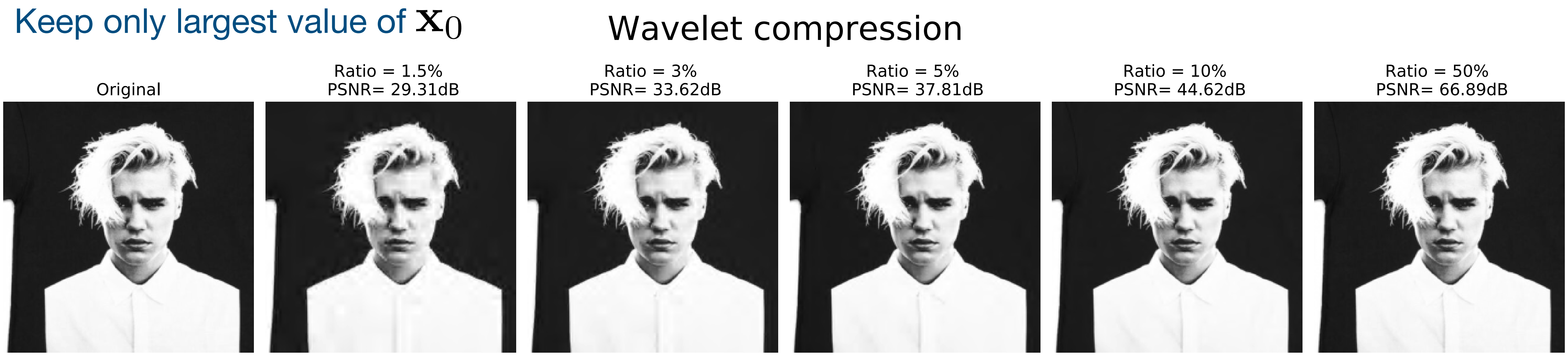
- $\mathbf{D} \in \mathbb{R}^{d\times d}$ is another ortho. basis than the canonical one $\mathbf{D}^{-1} = \mathbf{D}^\top$

  - Discrete Fourier basis (FFT)
  - Discrete Cosinus Transform
  - Wavelet

$$D_{pq} = \exp(-2i\tfrac{\pi}{d}pq) \qquad D_{pq} = \cos(\tfrac{\pi}{d}(p+\tfrac{1}{2})q)$$

- At the core of many compression schemes (JPEG, MPEG, MP3 ...)

Keep only largest value of $\mathbf{x}_0$



Compressed sensing theory: an invitation

- Most data encountered are sparse in another representation

$$\text{Original signal} \longrightarrow \mathbf{x} = \mathbf{D}\mathbf{x}_0 \longleftarrow \text{sparse vector}$$

- $\mathbf{D} \in \mathbb{R}^{n \times n}$ is another ortho. basis than the canonical one $\mathbf{D}^{-1} = \mathbf{D}^{\top}$
  - Discrete … avelet

$D_{pq} = e$ …

- At the c … …)

$\mathbf{x}$ is reasonably sparse if we look right

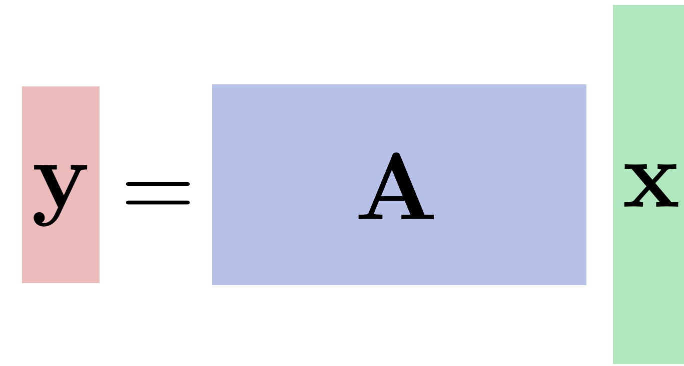
Keep only largest value of $\mathbf{x}_0$

Wavelet compression

Original	Ratio = 1.5% PSNR= 29.31dB	Ratio = 3% PSNR= 33.62dB	Ratio = 5% PSNR= 37.81dB	Ratio = 10% PSNR= 44.62dB	Ratio = 50% PSNR= 66.89dB

Compressed sensing theory: an invitation

Underdetermined $m < d$



How can we recover $\mathbf{x}$?

- Ill-posed inverse problem
- Infinity of solutions

The sparse assumption

- The signal $\mathbf{x} \in \mathbb{R}^d$ is **sparse**

- We could solve: $$(1) \quad \min_{\mathbf{x} \text{ s.t. } \mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$$

Is it done ? No …

- Not robust to noise $$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$$
- (1) is NP-hard
- Not robust w.r.t. sparsity level

Compressed sensing theory: an invitation

- **How can we recover $\mathbf{x}$?**
  - The signal $\mathbf{x} \in \mathbb{R}^d$ is **sparse**

Underdetermined $m < d$

$$\mathbf{y} = \mathbf{A} \mathbf{x}$$

- **The LASSO**

[Thibshirani, 1996]
[Chen & Donoho, 1995]

$$\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

Compressed sensing theory: an invitation

Underdetermined $m < d$

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

How can we recover $\mathbf{x}$?

- The signal $\mathbf{x} \in \mathbb{R}^d$ is **sparse**

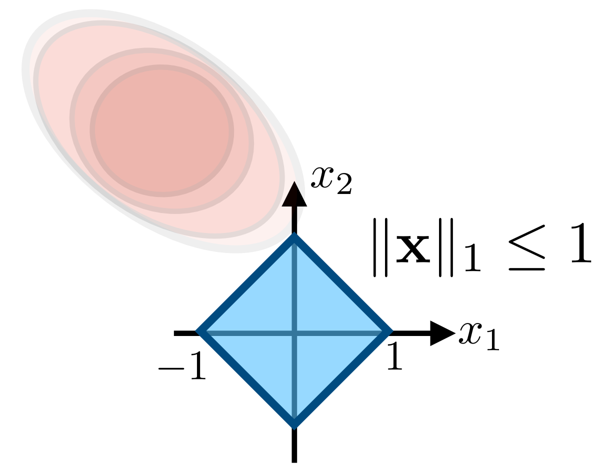
The LASSO

[Thibshirani, 1996]
[Chen & Donoho, 1995]

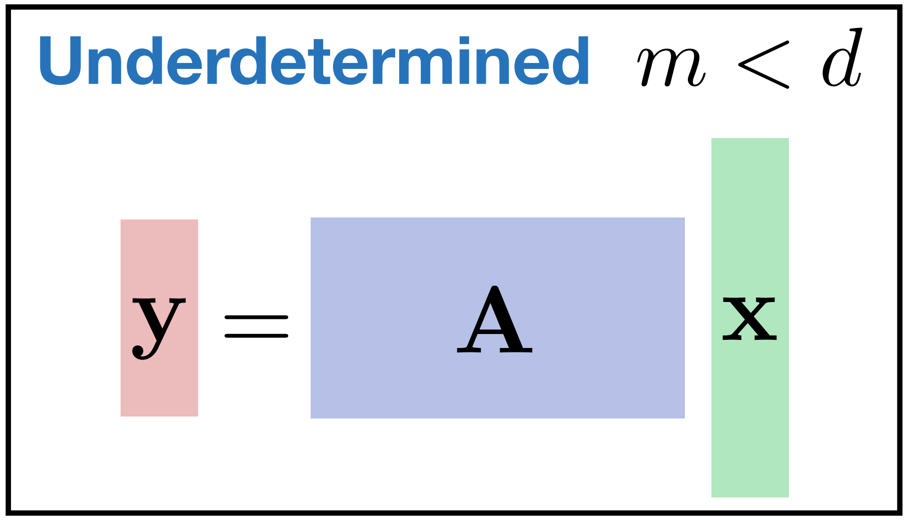
$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

$\mathbf{A}\mathbf{x}$ is closed to $\mathbf{y}$

$\mathbf{x}$ is penalized so as to have small L1 norm promotes sparsity



Compressed sensing theory: an invitation



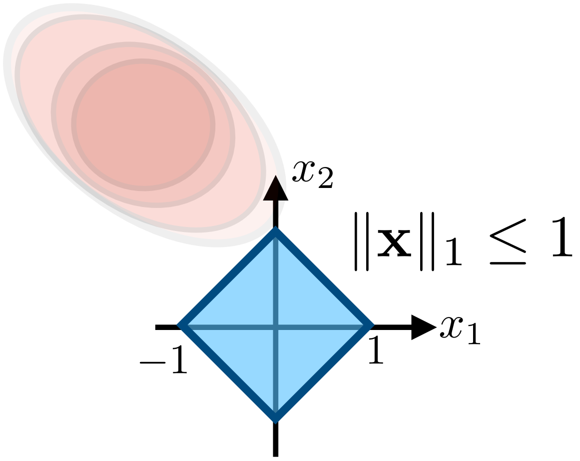
How can we recover $\mathbf{x}$?

- The signal $\mathbf{x} \in \mathbb{R}^d$ is **sparse**

The LASSO

$$\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

- Major impact on ML
  [Bach & al, 2012]
- Strictly convex



Can be solved using many algorithms…

- Proximal algorithms (iterative thresholding): ISTA, FISTA
  [Beck & Teboulle, 2009]
- (Block) Coordinate descent algorithms [Friedman & al, 2007]
- Here we will use CELER [Massias & al, 2018]

Compressed sensing theory: an invitation

The LASSO in practice for CS

Compressed sensing theory: an invitation

The LASSO in practice for CS

$$d = 128 * 128$$

$\mathbf{x}$ 

Compressed sensing theory: an invitation

The LASSO in practice for CS

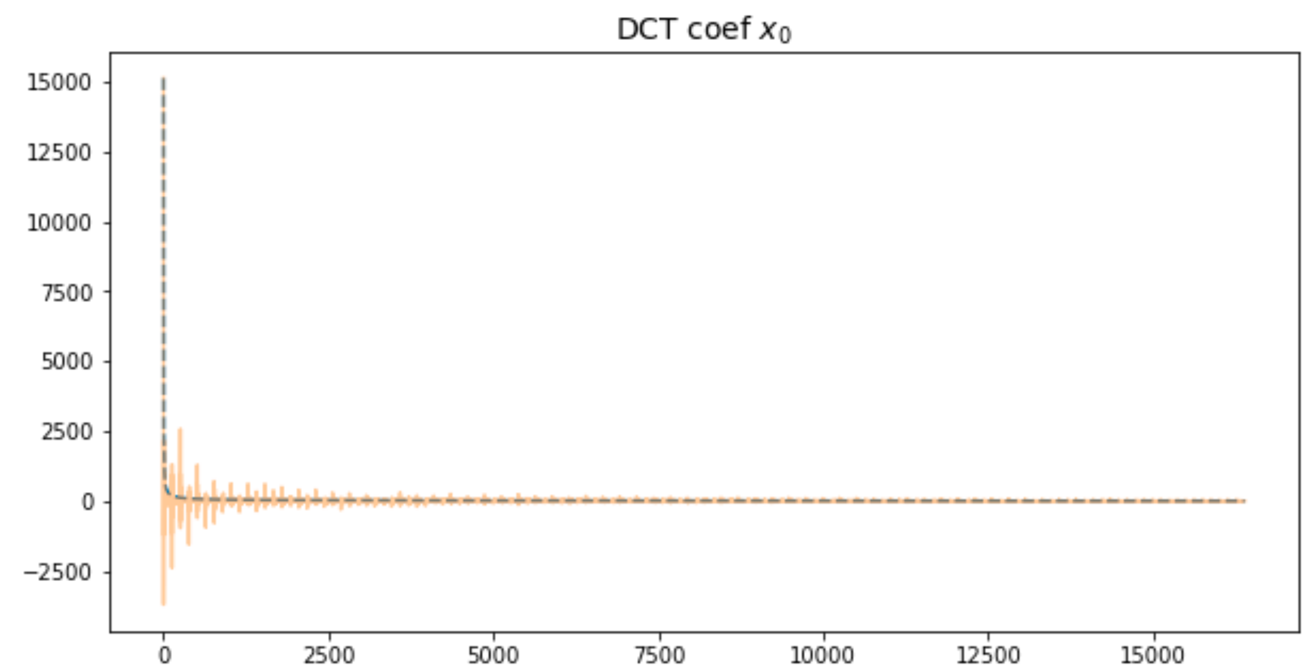
$$d = 128 * 128$$

$\mathbf{x}$

- An acquisition system: $\mathbf{A} \in \mathbb{R}^{m \times d}$
  - We do compression $m < d$
  - We take $A_{ij} \sim \frac{1}{\sqrt{m}} \mathcal{N}(0,1)$

Observation $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$

- $\mathbf{x}$ is sparse in another basis (DCT)



Compressed sensing theory: an invitation

The LASSO in practice for CS

$d = 128 * 128$

$\mathbf{x}$

- An acquisition system: $\mathbf{A} \in \mathbb{R}^{m \times d}$
  - We do compression $m < d$
  - We take $A_{ij} \sim \frac{1}{\sqrt{m}} \mathcal{N}(0,1)$

Observation $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$

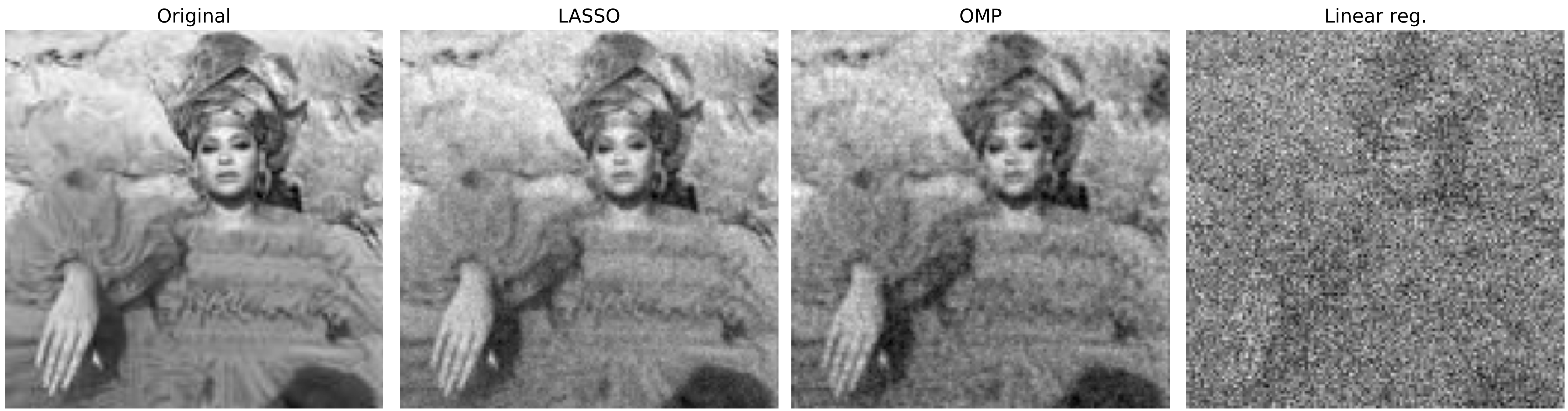
- $\mathbf{x}$ is sparse in another basis (DCT)
- We solve the LASSO

$$\min_{\mathbf{x}_0 \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{A}\mathbf{D}\mathbf{x}_0\|_2^2 + \lambda \|\mathbf{x}_0\|_1$$

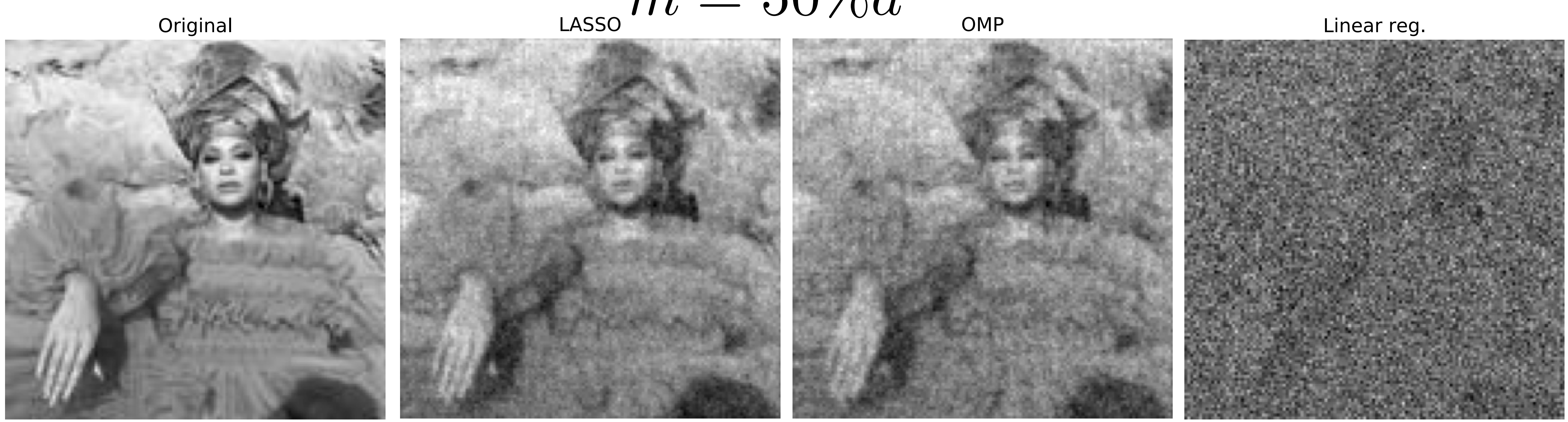
Compressed sensing theory: an invitation

The LASSO in practice for CS

$$m = 70\% d$$

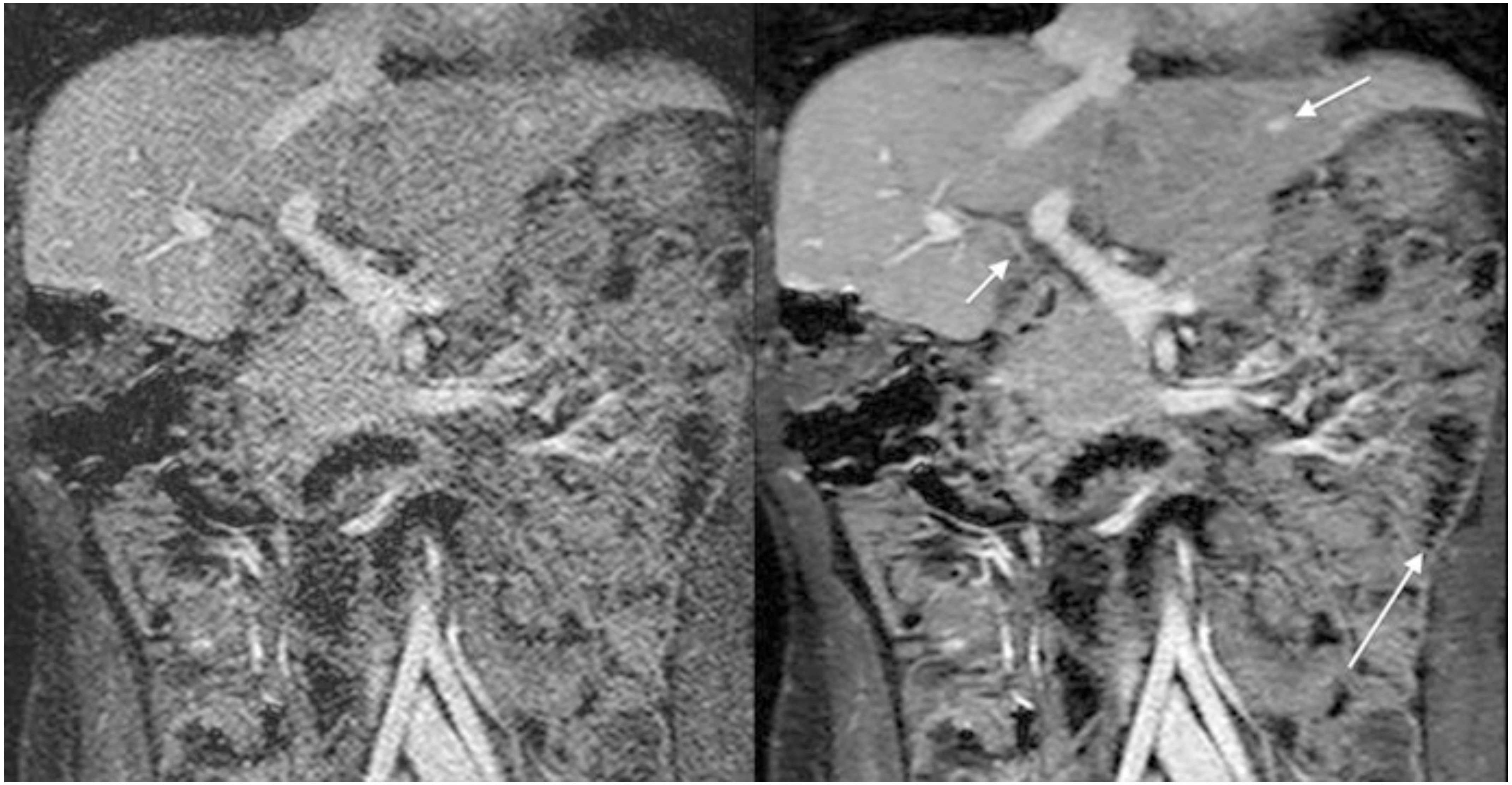


$$m = 50\% d$$



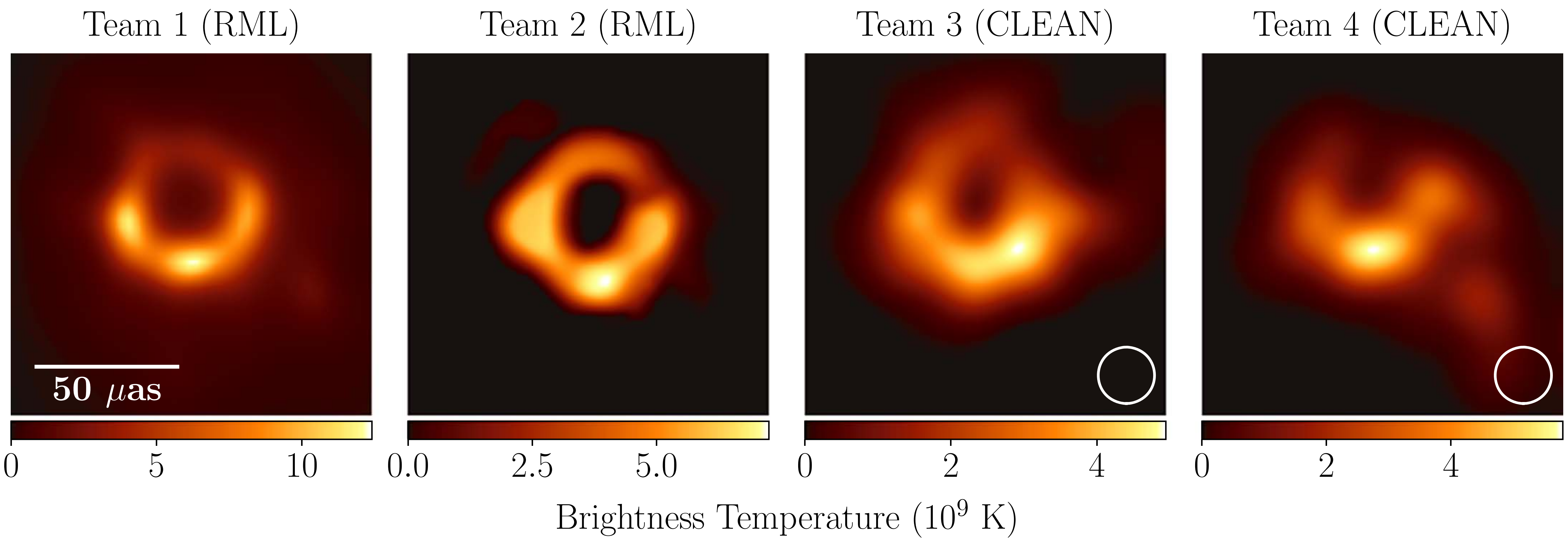
Compressed sensing theory: an invitation

Other CS results



Compressed sensing theory: an invitation

Other CS results



Compressed sensing theory: an invitation

Guarantees for the LASSO

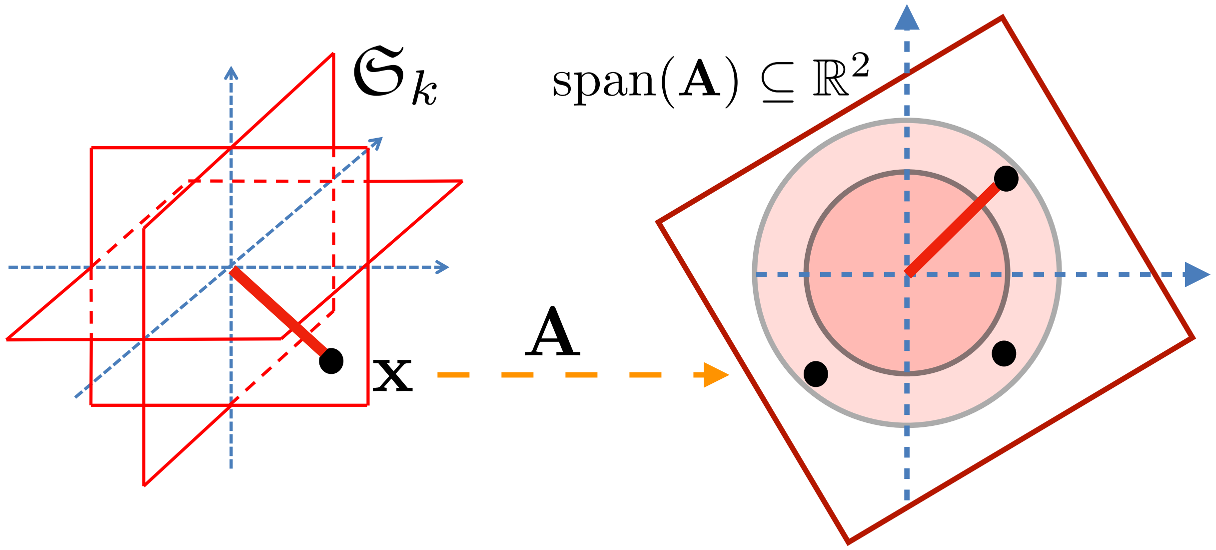
- Does the LASSO truly recovers $\mathbf{x} \in \mathbb{R}^d$?

Compressed sensing theory: an invitation

- **Guarantees for the LASSO**
  - Does the LASSO truly recovers $\mathbf{x} \in \mathbb{R}^d$?

- **The restricted isometric property (RIP)** [Candes & Tao, 2005]

$$\boxed{\begin{array}{l} \exists \delta_k \in [0,1[\quad \forall \mathbf{x} \ k\text{-sparse} \\ (1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2 \end{array}}$$



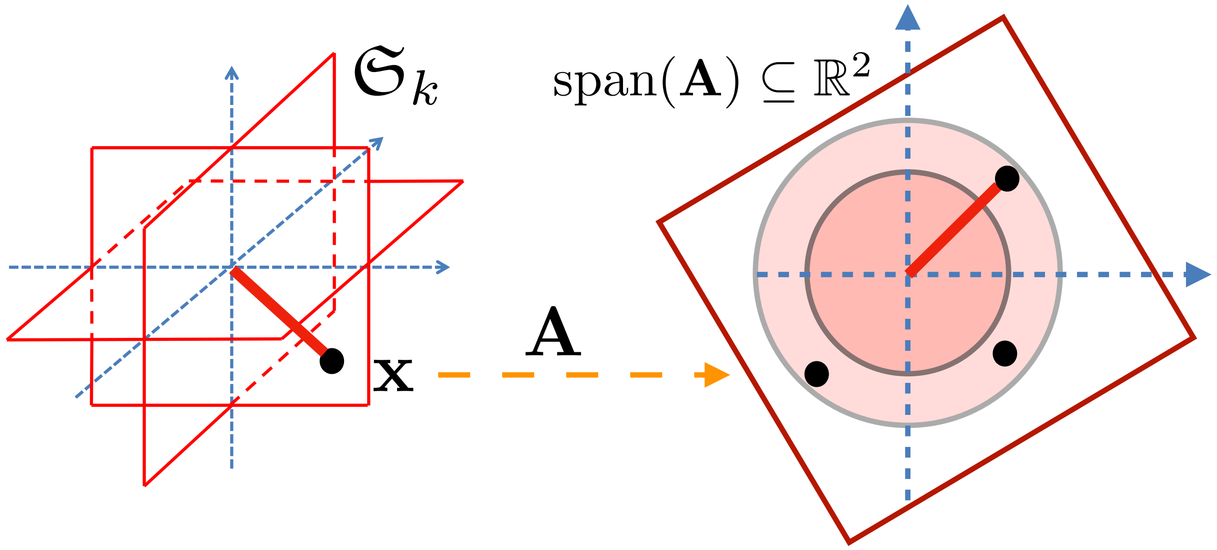
Compressed sensing theory: an invitation

Guarantees for the LASSO

- Does the LASSO truly recovers $\mathbf{x} \in \mathbb{R}^d$?

The restricted isometric property (RIP) [Candes & Tao, 2005]

$$\boxed{\begin{array}{l} \exists \delta_k \in [0,1[\quad \forall \mathbf{x} \ k\text{-sparse} \\ (1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2 \end{array}}$$



- E.g. **Gaussian matrices**

$$\mathbf{A} \in \mathbb{R}^{m \times d} \text{ with } A_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0,1)$$

$$\boxed{m \gtrsim \delta^{-2}\, k \ln(e\tfrac{d}{k})}$$ with overwhelming prob. we have the RIP with δ

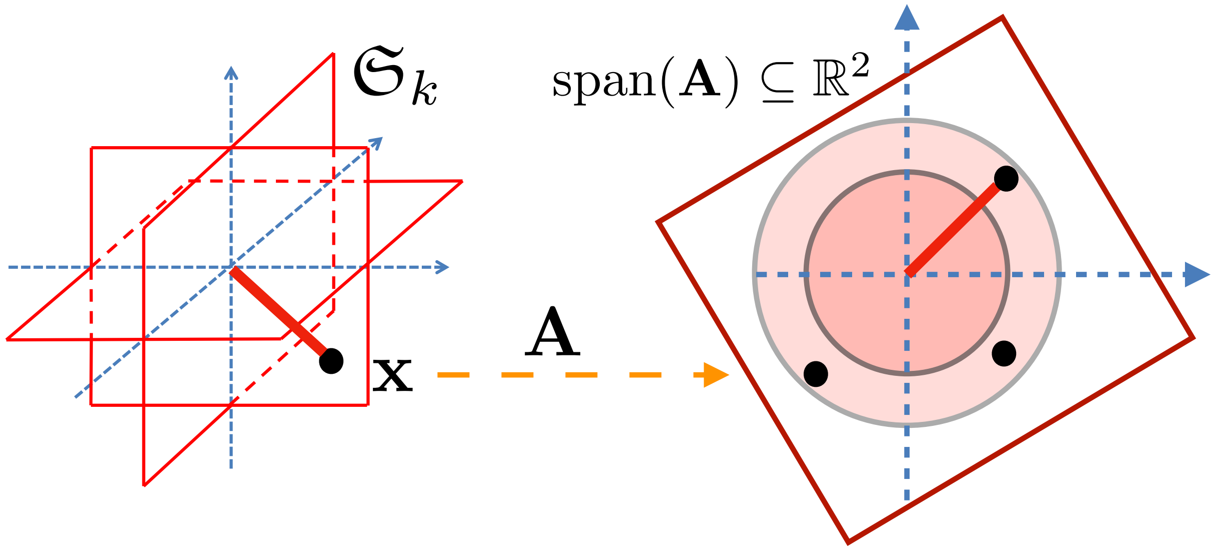
Compressed sensing theory: an invitation

Guarantees for the LASSO

- Does the LASSO truly recovers $\mathbf{x} \in \mathbb{R}^d$?

The restricted isometric property (RIP) [Candes & Tao, 2005]

$$\boxed{\begin{array}{l} \exists \delta_k \in [0,1[\quad \forall \mathbf{x} \ k\text{-sparse} \\ (1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2 \end{array}}$$



- E.g. **Gaussian matrices**

$$\mathbf{A} \in \mathbb{R}^{m \times d} \text{ with } A_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0,1)$$

$$\boxed{m \gtrsim \delta^{-2}\, k \ln(e\tfrac{d}{k})}$$ ▶ with overwhelming prob. we have the RIP with δ

- Same holds for $\mathbf{AD}$ where $\mathbf{A}$ is **Gaussian** and $\mathbf{D}$ **orthogonal**

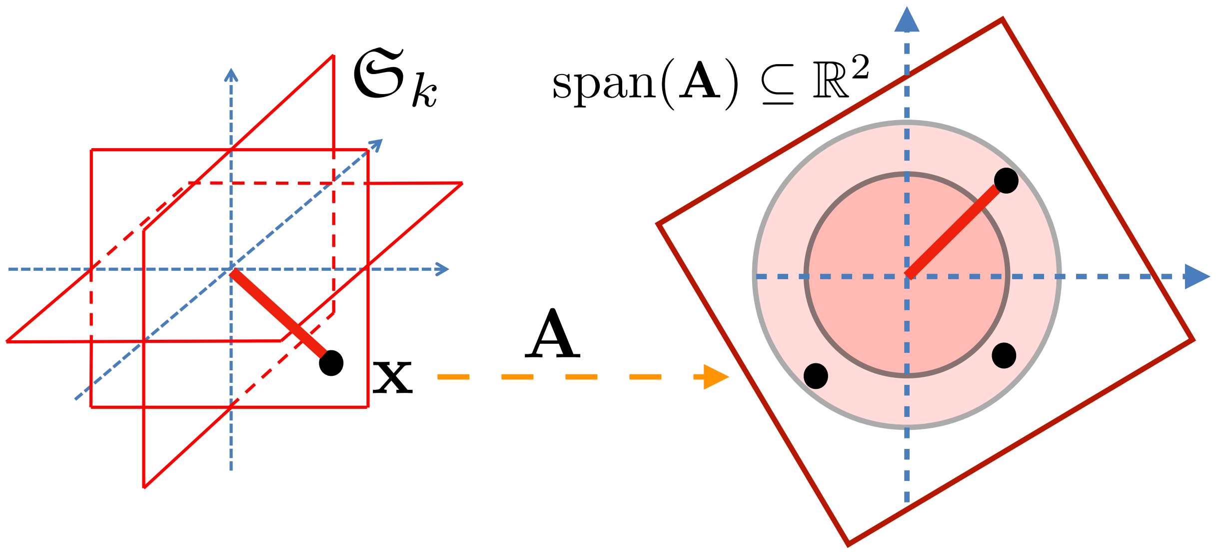
Compressed sensing theory: an invitation

Guarantees for the LASSO


- Does the LASSO truly recovers $\mathbf{x} \in \mathbb{R}^d$?

The restricted isometric property (RIP) [Candes & Tao, 2005]

$$\exists \delta_k \in [0,1[\quad \forall \mathbf{x} \ k\text{-sparse}$$
$$(1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2$$



- E.g. **Gaussian matrices**

$$m \gtrsim \delta^{-2} \, k \ln(e\tfrac{d}{k})$$

- Related with the **magical Johnson-Lindenstrauss Lemma**

Every $\mathbf{x}_1, \cdots, \mathbf{x}_k \in \mathbb{R}^d$ can be linearly embedded in $\mathbb{R}^m$ with $\delta > 0$ distortion provided $m \gtrsim \delta^{-2} \ln(k)$

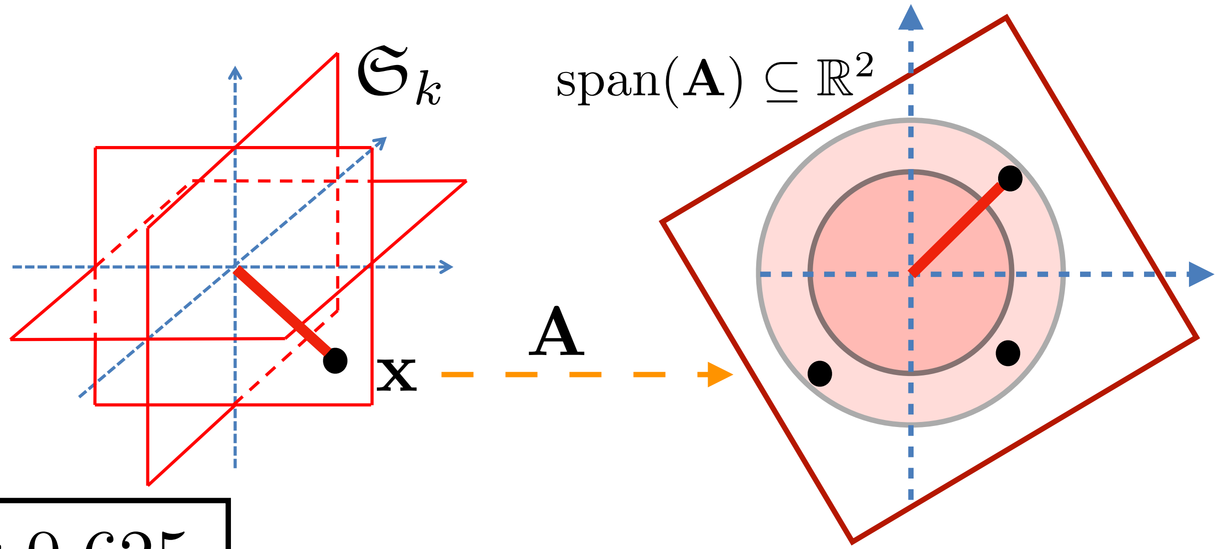
- **Does not depend on d !**

Compressed sensing theory: an invitation

- **Guarantees for the LASSO**
  - Does the LASSO truly recovers $\mathbf{x} \in \mathbb{R}^d$?


- **The restricted isometric property (RIP)** [Candes & Tao, 2005]

$$\boxed{\begin{array}{l} \exists \delta_k \in [0,1[\quad \forall \mathbf{x} \ k\text{-sparse} \\ (1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2 \end{array}}$$



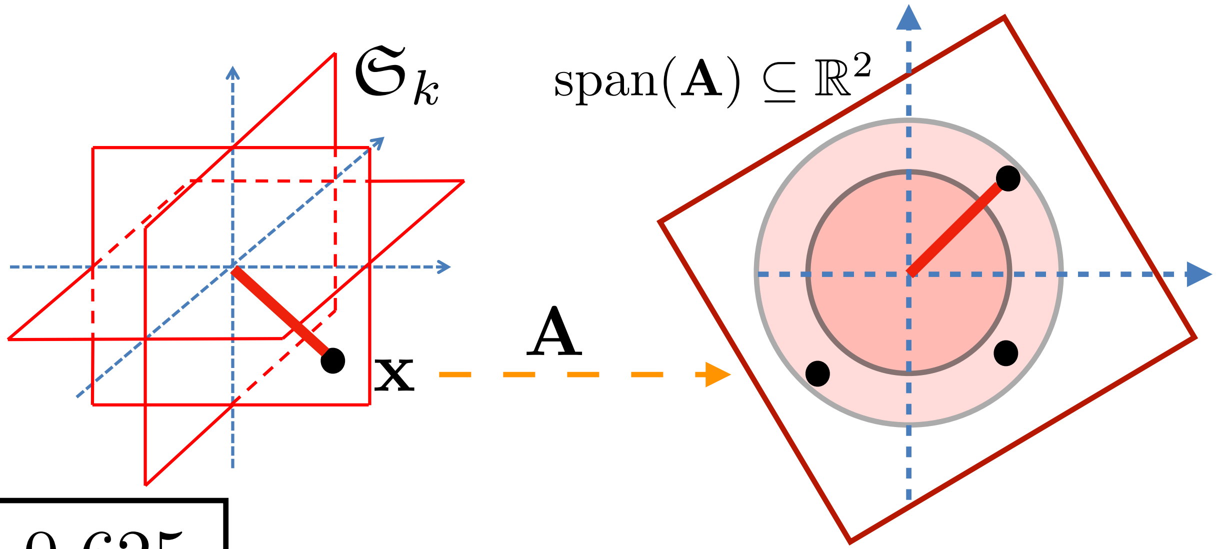
- **The result:** $\boxed{\text{Suppose } \delta_{2k} < 4/\sqrt{41} \approx 0.625}$

Compressed sensing theory: an invitation

- **Guarantees for the LASSO**
  - Does the LASSO truly recovers $\mathbf{x} \in \mathbb{R}^d$?

- **The restricted isometric property (RIP)** [Candes & Tao, 2005]

$$\boxed{\begin{array}{l} \exists \delta_k \in [0,1[\quad \forall \mathbf{x}\ k\text{-sparse} \\ (1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2 \end{array}}$$



- **The result:** $\boxed{\text{Suppose } \delta_{2k} < 4/\sqrt{41} \approx 0.625}$

Take $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$
with $\|\mathbf{e}\|_2 \leq \eta$

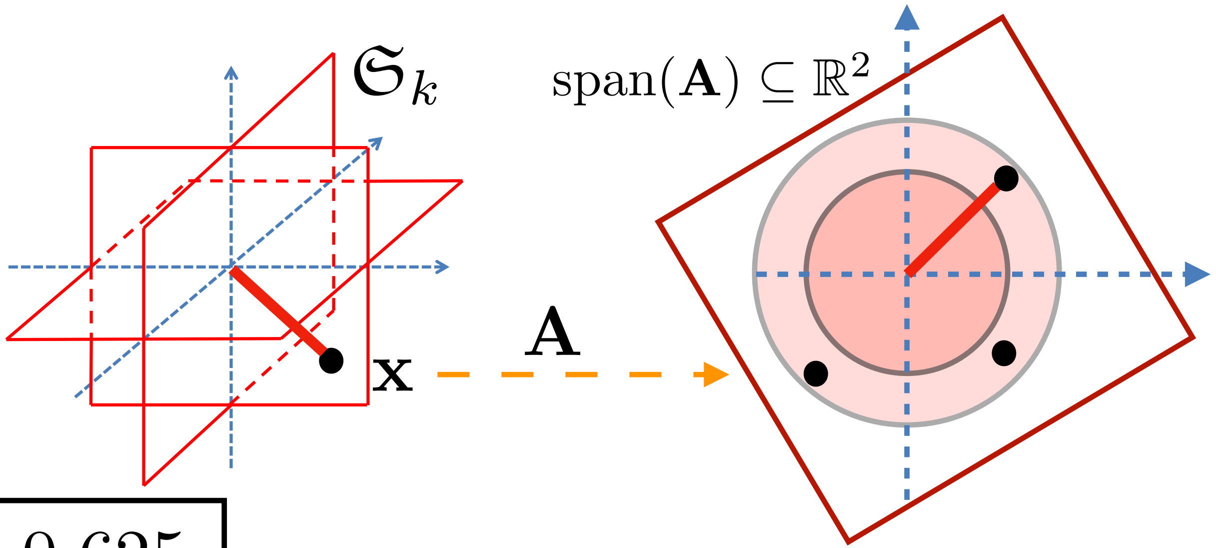
Compressed sensing theory: an invitation

Guarantees for the LASSO


- Does the LASSO truly recovers $\mathbf{x} \in \mathbb{R}^d$?

The restricted isometric property (RIP) [Candes & Tao, 2005]

$$\boxed{\begin{array}{l} \exists \delta_k \in [0,1[\quad \forall \mathbf{x}\ k\text{-sparse} \\ (1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2 \end{array}}$$



The result: $\boxed{\text{Suppose } \delta_{2k} < 4/\sqrt{41} \approx 0.625}$

Take $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$
with $\|\mathbf{e}\|_2 \leq \eta$

$\widehat{\mathbf{x}}$ sol. of $\begin{cases} \min \|\mathbf{z}\|_1 \\ \text{s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2 \leq \eta \end{cases}$

Compressed sensing theory: an invitation

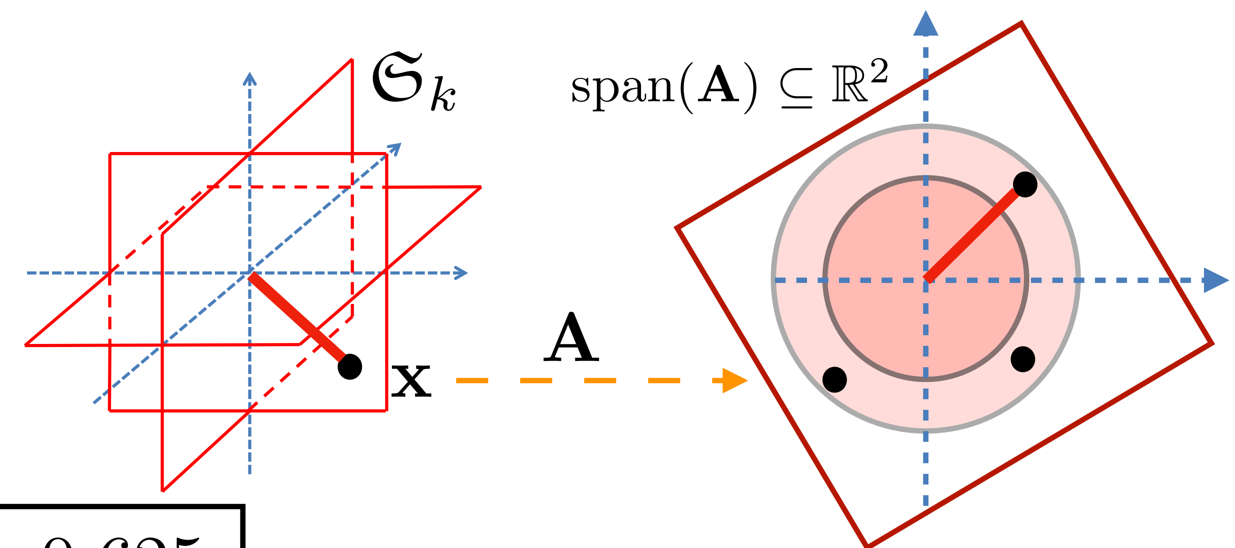
Guarantees for the LASSO

- Does the LASSO truly recovers $\mathbf{x} \in \mathbb{R}^d$?

The restricted isometric property (RIP) [Candes & Tao, 2005]

$$\exists \delta_k \in [0,1[\quad \forall \mathbf{x} \ k\text{-sparse}$$

$$(1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2$$



The result:

$$\boxed{\text{Suppose } \delta_{2k} < 4/\sqrt{41} \approx 0.625}$$

Take $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$
with $\|\mathbf{e}\|_2 \leq \eta$

$\widehat{\mathbf{x}}$ sol. of $\begin{cases} \min \|\mathbf{z}\|_1 \\ \text{s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2 \leq \eta \end{cases}$

$$\Longrightarrow \quad \boxed{\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 \leq \frac{C_1}{\sqrt{k}} d^\circ(\mathbf{x}, \mathfrak{S}_k) + C_2 \eta}$$

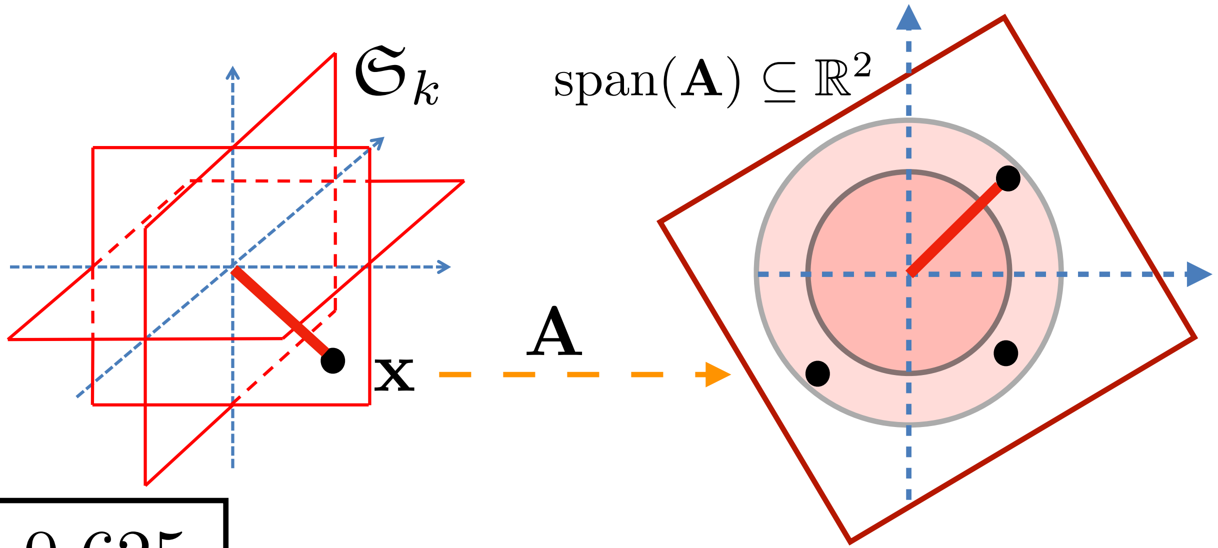
Compressed sensing theory: an invitation

Guarantees for the LASSO

- Does the LASSO truly recovers $\mathbf{x} \in \mathbb{R}^d$?

The restricted isometric property (RIP) [Candes & Tao, 2005]

$$\exists \delta_k \in [0,1[\quad \forall \mathbf{x} \ k\text{-sparse}$$
$$(1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2$$



The result: $\boxed{\text{Suppose } \delta_{2k} < 4/\sqrt{41} \approx 0.625}$

Take $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$
with $\|\mathbf{e}\|_2 \leq \eta$

$\hat{\mathbf{x}}$ sol. of $\begin{cases} \min \|\mathbf{z}\|_1 \\ \text{s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2 \leq \eta \end{cases}$

$$\implies \|\mathbf{x} - \hat{\mathbf{x}}\|_2$$

Error between the estimation and the ground truth

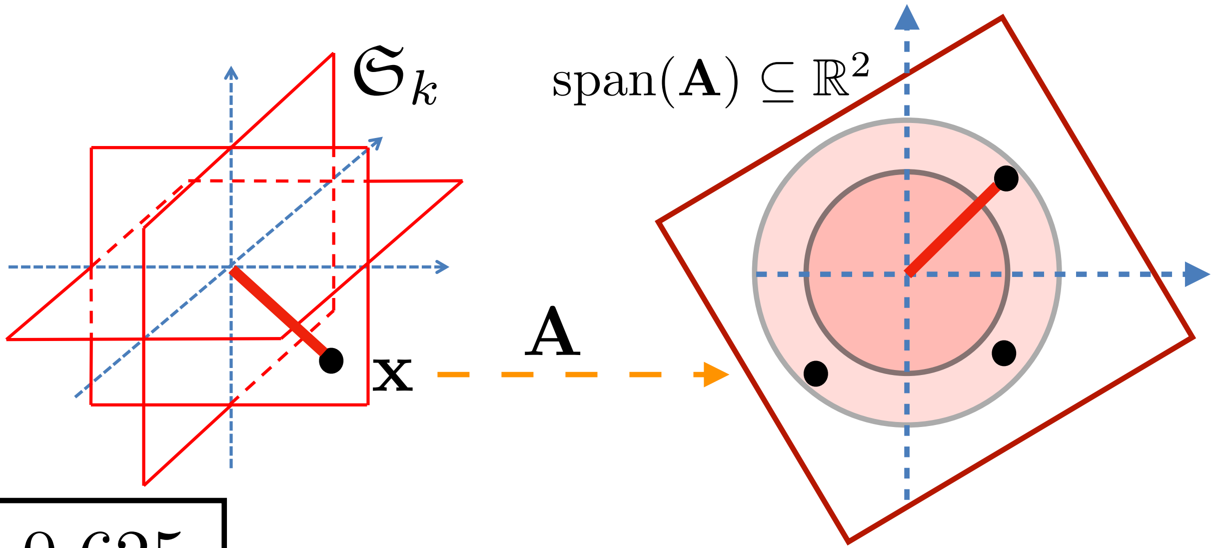
Compressed sensing theory: an invitation

Guarantees for the LASSO

- Does the LASSO truly recovers $\mathbf{x} \in \mathbb{R}^d$?

The restricted isometric property (RIP) [Candes & Tao, 2005]

$$\boxed{\begin{aligned} &\exists \delta_k \in [0,1[\quad \forall \mathbf{x} \ k\text{-sparse} \\ &(1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2 \end{aligned}}$$



The result: Suppose $\delta_{2k} < 4/\sqrt{41} \approx 0.625$

Take $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \leq \eta$

$\hat{\mathbf{x}}$ sol. of $\begin{cases} \min \|\mathbf{z}\|_1 \\ \text{s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2 \leq \eta \end{cases}$

$$\Longrightarrow \quad \boxed{\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq \frac{C_1}{\sqrt{k}} d^\circ(\mathbf{x}, \mathfrak{S}_k)}$$

Zero if the vector is exactly k-sparse

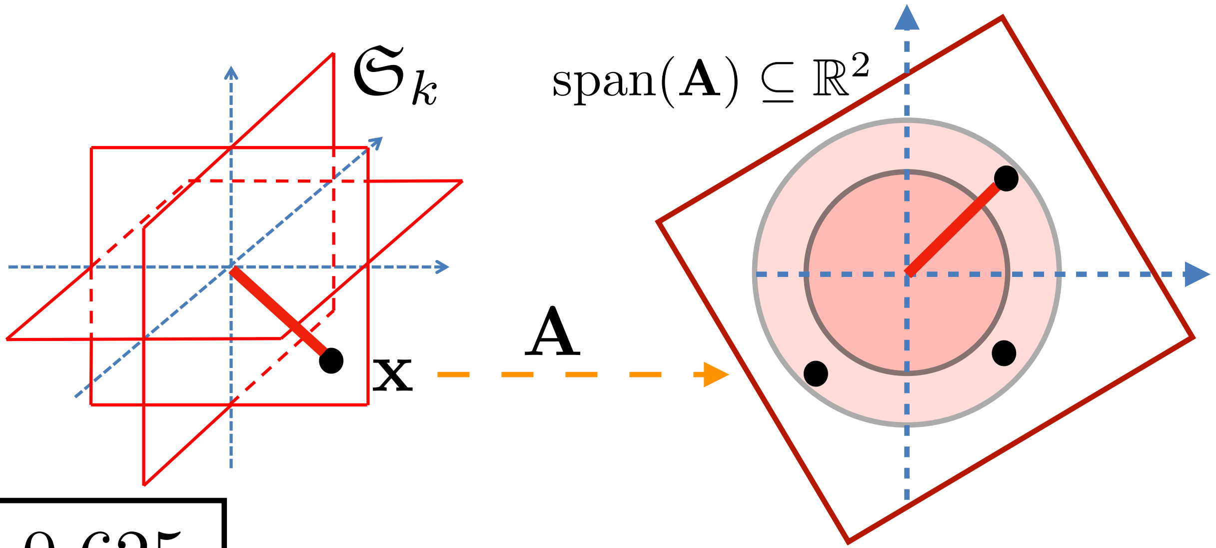
Compressed sensing theory: an invitation

Guarantees for the LASSO

- Does the LASSO truly recovers $\mathbf{x} \in \mathbb{R}^d$?


The restricted isometric property (RIP) [Candes & Tao, 2005]

$$\boxed{\begin{array}{l} \exists \delta_k \in [0,1[\quad \forall \mathbf{x} \ k\text{-sparse} \\ (1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2 \end{array}}$$



The result:

$\boxed{\text{Suppose } \delta_{2k} < 4/\sqrt{41} \approx 0.625}$

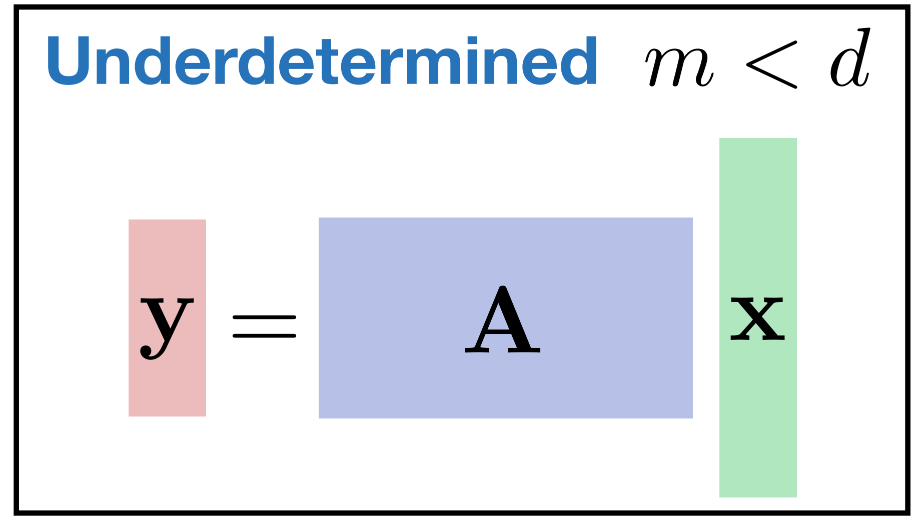
Take $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \leq \eta$

$\widehat{\mathbf{x}}$ sol. of $\begin{cases} \min \|\mathbf{z}\|_1 \\ \text{s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2 \leq \eta \end{cases}$

$$\Longrightarrow \boxed{\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 \leq \frac{C_1}{\sqrt{k}} d^\circ(\mathbf{x}, \mathfrak{S}_k) + C_2\eta}$$

Zero if there is no noise

Compressed sensing theory: an invitation



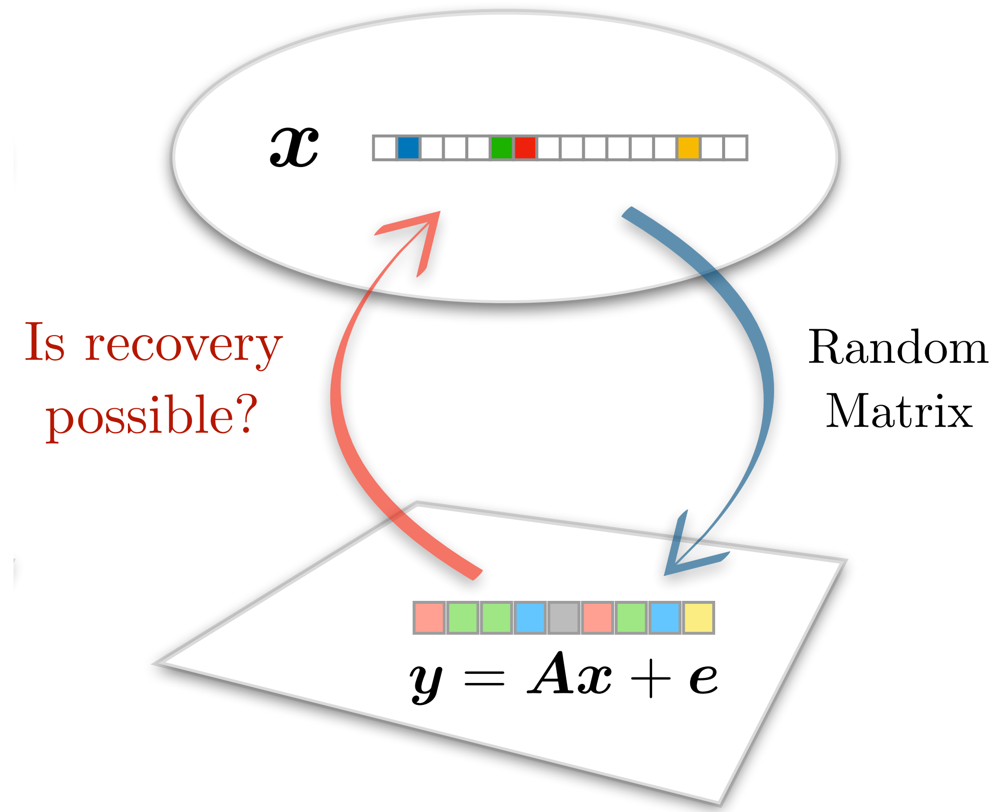
How can we recover $\mathbf{x}$?

Ill-posed inverse problem

Infinity of solutions

The sparse assumption

The true vector lives in a low-dim space



Algorithmic solutions

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

Theoretical guarantees

- E.g. RIP with Gaussian matrices

$$m \gtrsim \delta^{-2}\, k \ln(e\tfrac{d}{k})$$

Overview of the talk

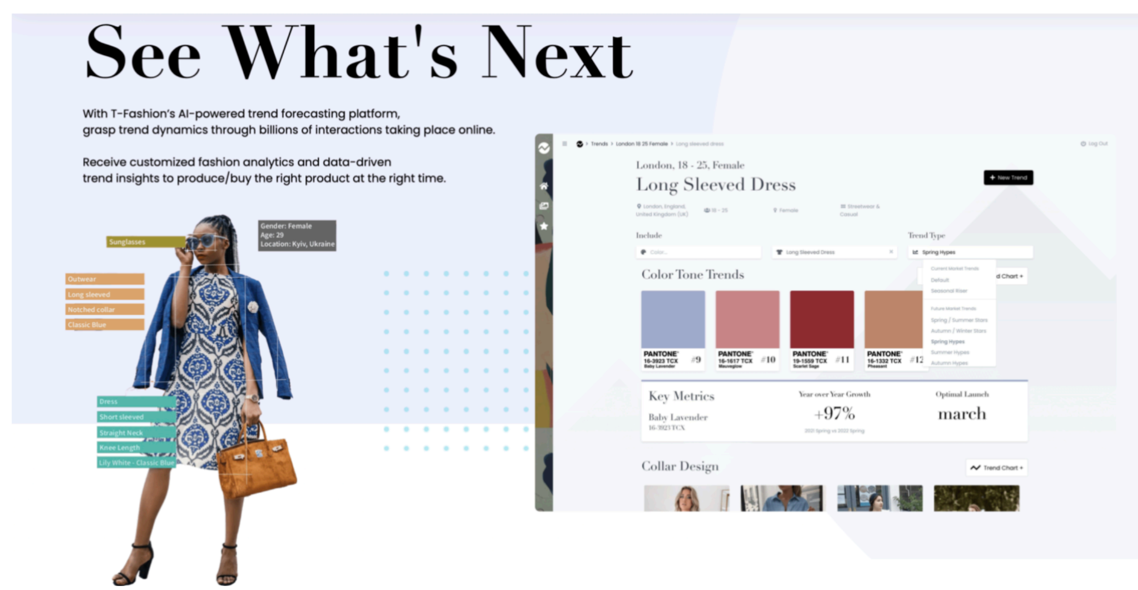
- **Part I: A journey in the compressed sensing theory**
- **Part II: A bit of machine learning theory**
- **Part III: The sketching approach**
  - Applied sketching
  - Theoretical guarantees

A bit of machine learning

Fashion Trend Forecasting with AI

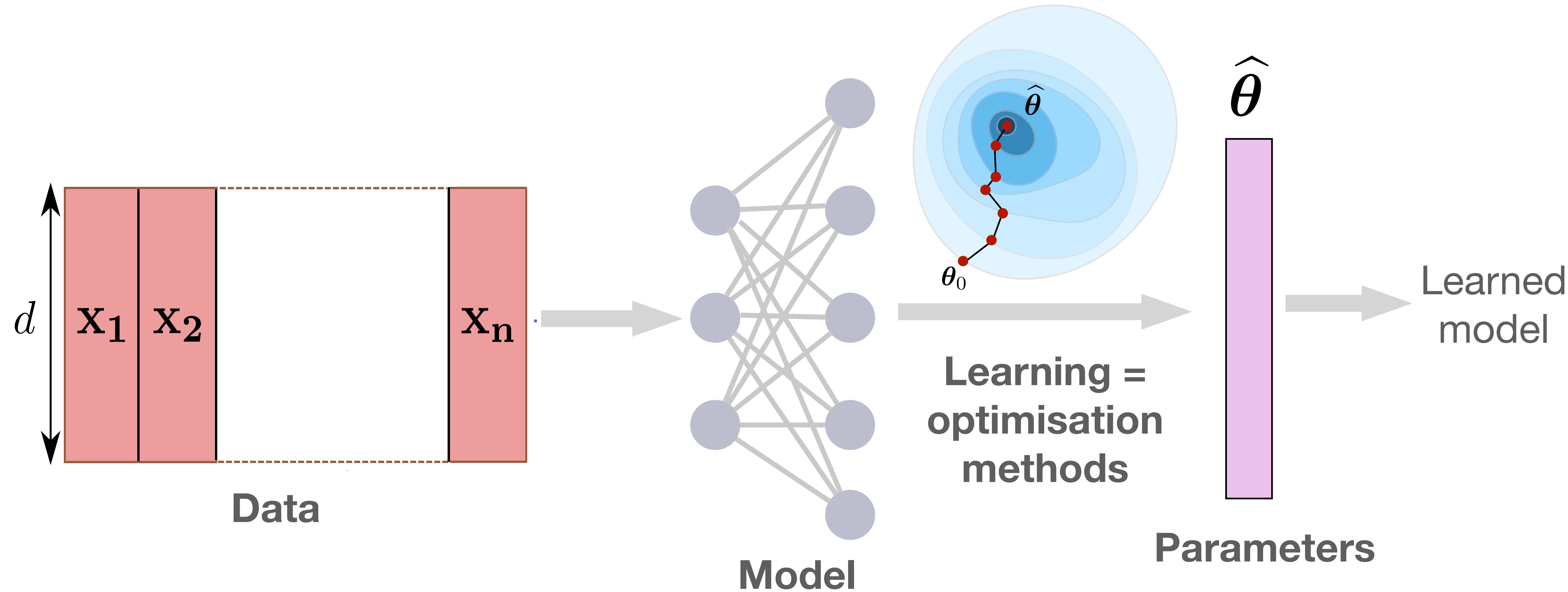
JESSICA MAGALIT - NOVEMBER 9, 2021

0 0



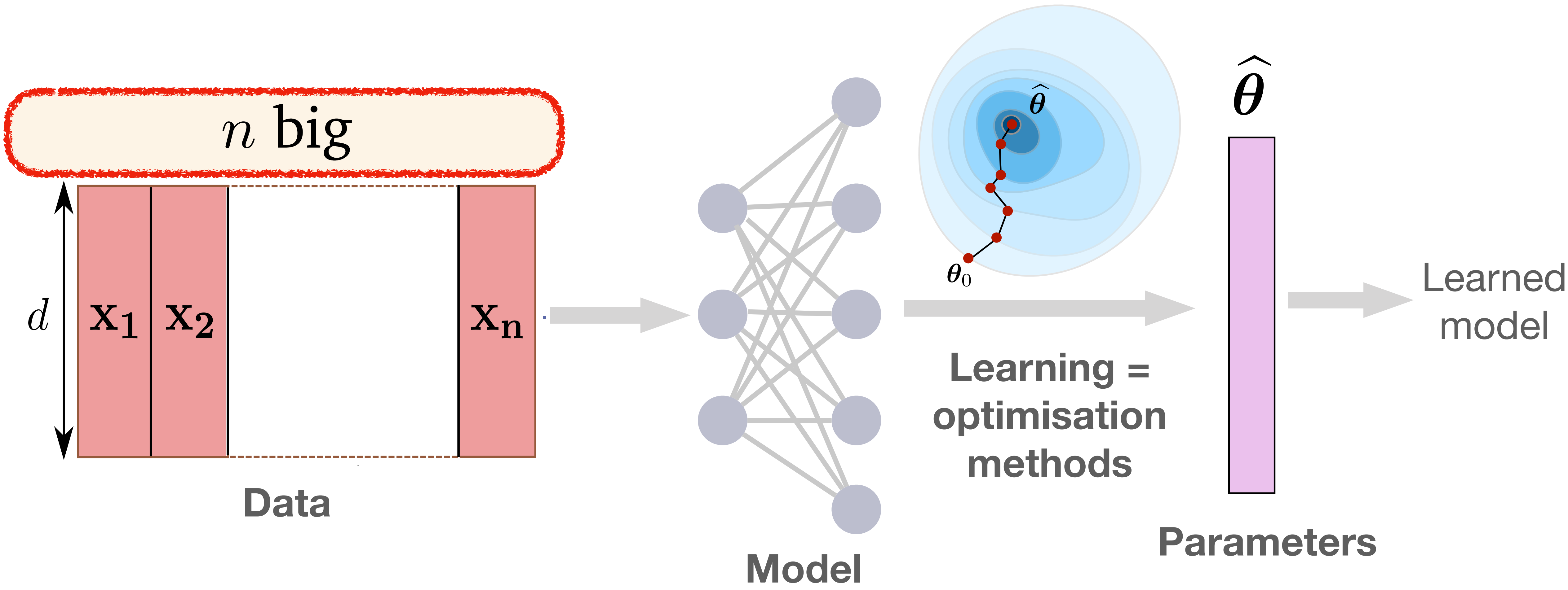
Machine learning theory

The big picture:



Machine learning theory

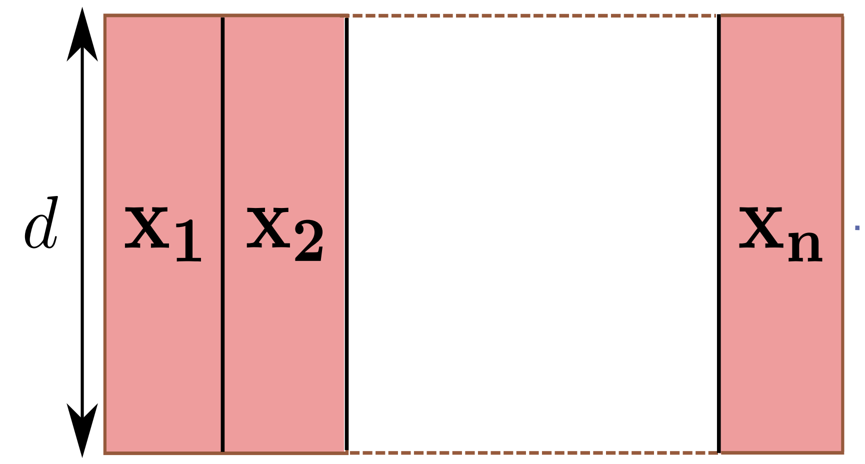
- **The big picture:**



- **Large scale machine learning**

Machine learning theory

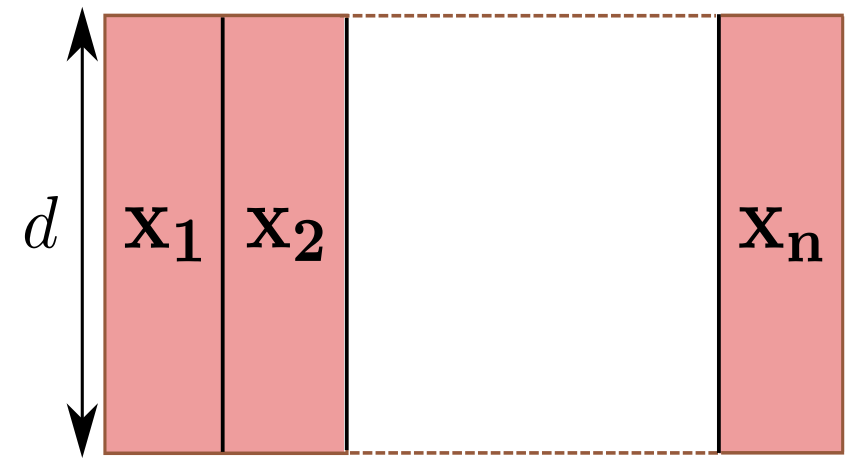
- **The ML setting:**
  - Data points $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$
  - Each point $\mathbf{x}_i \sim \pi$
  - π is unknown and generates the data



Machine learning theory

The ML setting:

- Data points $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$
- Each point $\mathbf{x}_i \sim \pi$
- π is unknown and generates the data



Empirical risk minimization:

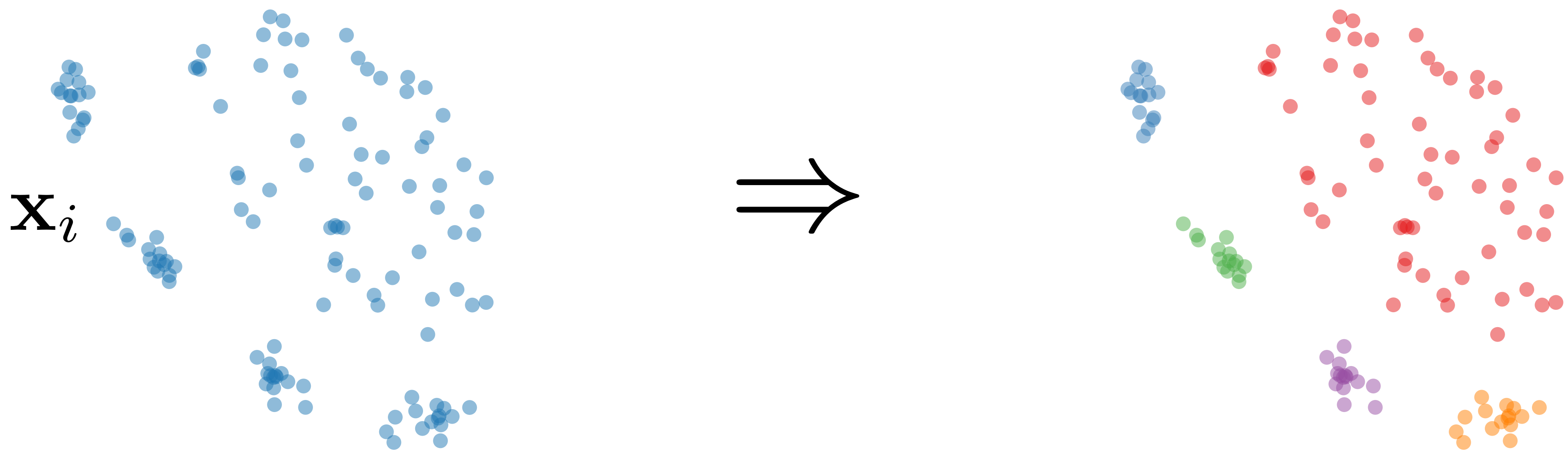
- Find parameters: $\widehat{\boldsymbol{\theta}} \in \Theta$
- That minimizes:

$$\boxed{\min_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda \, \mathrm{Reg}(\boldsymbol{\theta})}$$

- Where $\ell : \mathbb{R}^d \times \Theta \to \mathbb{R}$ is a loss and Reg a regularization term

Machine learning theory

Unsupervised learning: K-means clustering



- Organize training samples in groups (say K)
- Find K clusters that best represent our data
- We look for $\boldsymbol{\theta} = (\mathbf{c}_1, \cdots, \mathbf{c}_K), \mathbf{c}_k \in \mathbb{R}^d$
- The loss is $\ell(\mathbf{x}_i, \boldsymbol{\theta}) = \min_{k \in [\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$

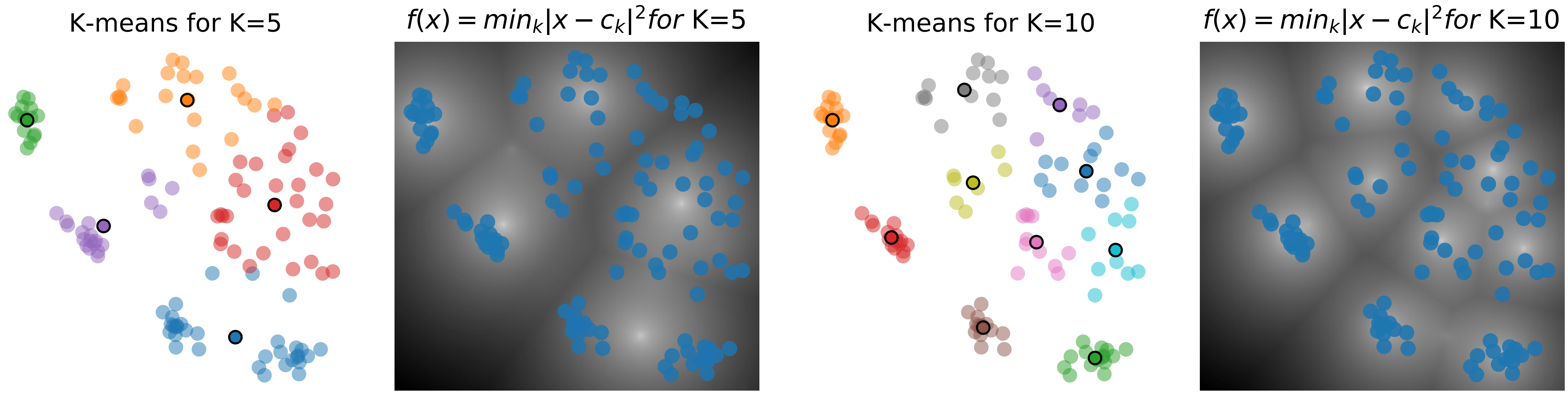
Squared distance between the point and its closest cluster

Machine learning theory

Unsupervised learning: K-means clustering

- We aim at solving: [Steinhaus & al, 1956, McQuenn & al, 1967]

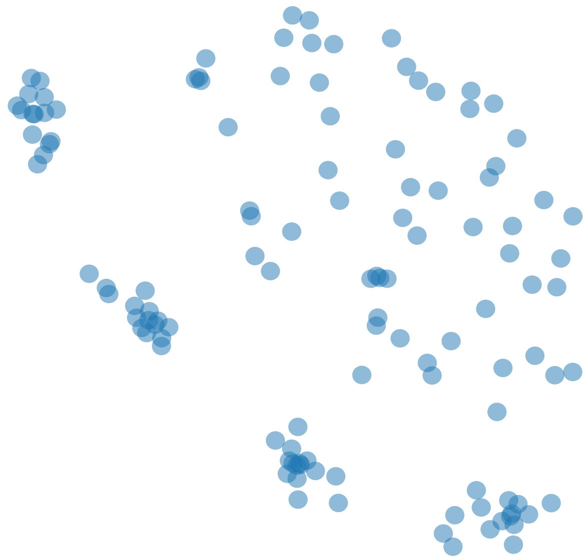
$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \frac{1}{n}\sum_{i=1}^{n} \min_{k\in[\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$$



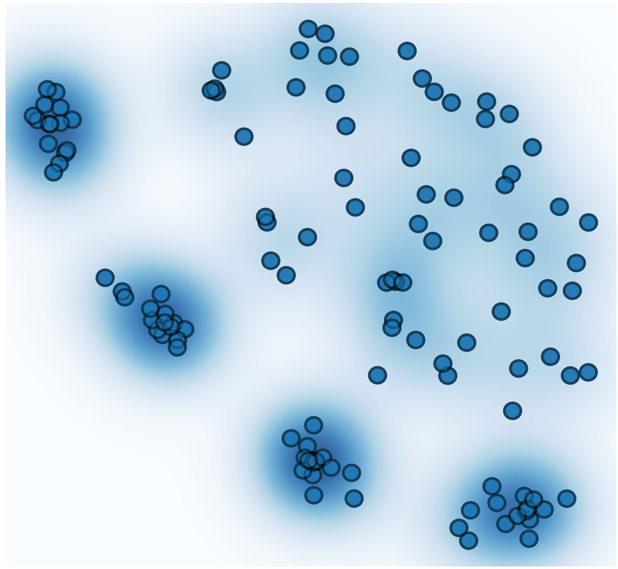
- It is a NP-Hard problem: can be tackled by **Lloyd's algorithm**
- Complexity (in time): $\mathcal{O}(nKd)$

Machine learning theory

Unsupervised learning: GMM



$\Longrightarrow$

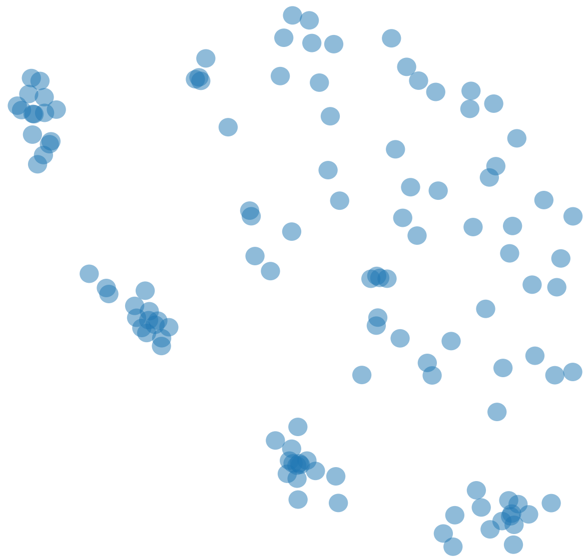


- Estimate a probability density $\widehat{\pi}$ from the samples
- $\forall \mathbf{x}, \widehat{\pi}(\mathbf{x}) \geq 0, \int \widehat{\pi}(\mathbf{x}) \mathrm{d}\mathbf{x} = 1$

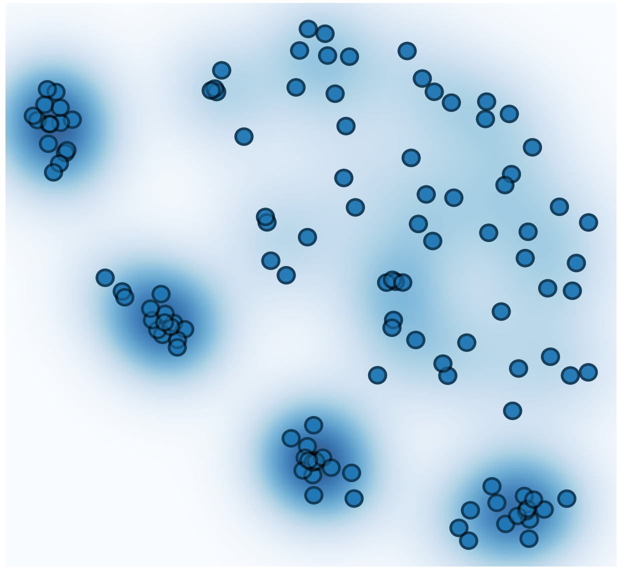
} Density estimation

Machine learning theory

Unsupervised learning: GMM



$\Longrightarrow$



- Estimate a probability density $\widehat{\pi}$ from the samples
- $\forall \mathbf{x}, \widehat{\pi}(\mathbf{x}) \geq 0, \int \widehat{\pi}(\mathbf{x}) \mathrm{d}\mathbf{x} = 1$

} Density estimation

Parametrized probability distributions

- Look only for a family of distributions given by param. $\boldsymbol{\theta}$
- E.g. Gaussian $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$

$$\pi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) := (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Machine learning theory

- **Unsupervised learning: GMM**
  - The model is a mixture of Gaussian

$$\boldsymbol{\theta} = \{\alpha_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$$

$$\boxed{\pi_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k \pi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}(\mathbf{x})}$$

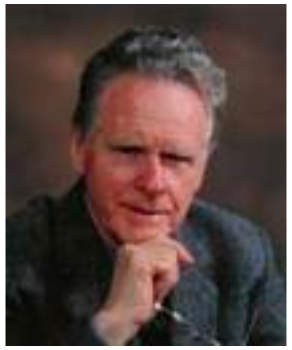
Machine learning theory

Unsupervised learning: GMM

- The model is a mixture of Gaussian

$$\boldsymbol{\theta} = \{\alpha_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$$

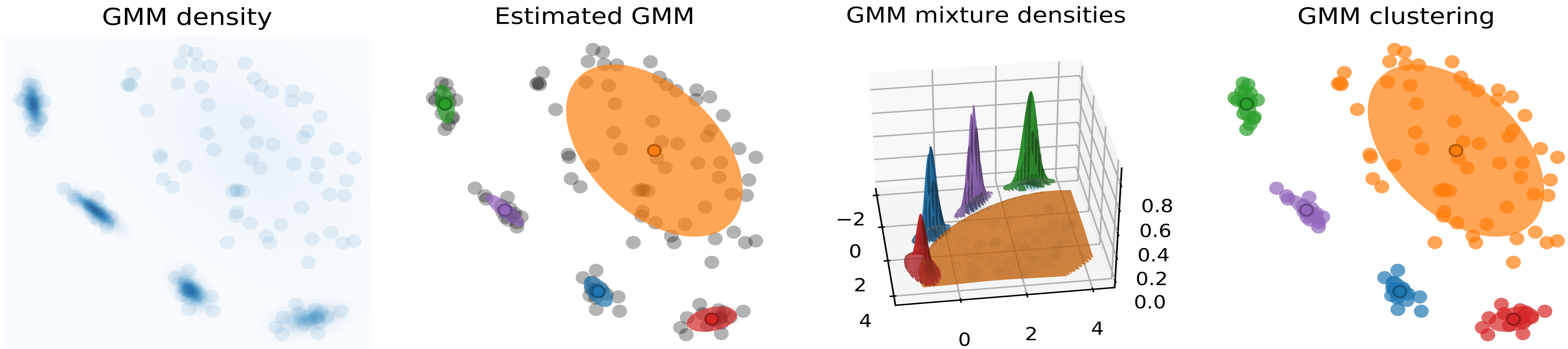
$$\pi_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k \pi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}(\mathbf{x})$$



- We aim at solving the MLE problem: [Dempster & al, 1977]

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} -\log(\pi_{\boldsymbol{\theta}}(\mathbf{x}_i))$$

- Here $\ell(\mathbf{x}_i, \boldsymbol{\theta}) = -\log(\pi_{\boldsymbol{\theta}}(\mathbf{x}_i))$ (neg log likehood)



Machine learning theory

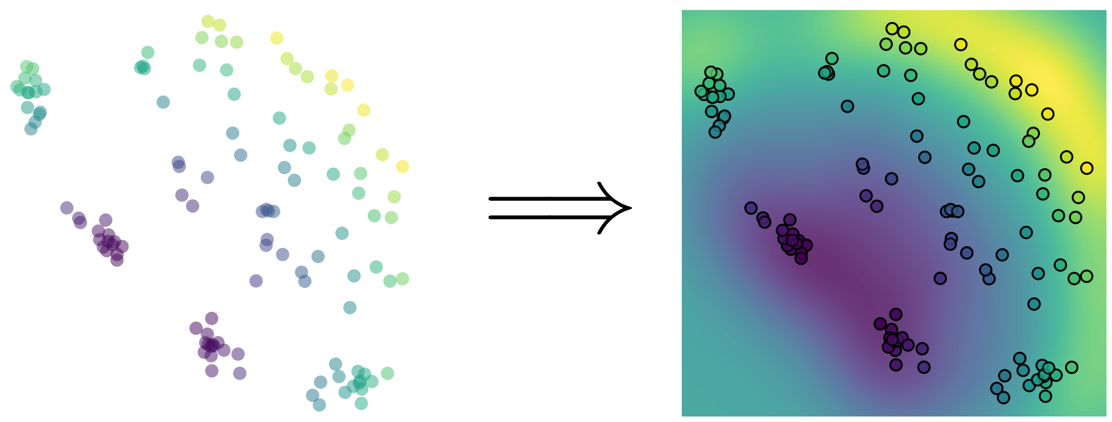
- **Supervised learning:**
  - In the supervised setting: $\mathbf{x}_i = (\mathbf{z}_i, y_i)$

Machine learning theory

Supervised learning:

- In the supervised setting: $\mathbf{x}_i = (\mathbf{z}_i, y_i)$

Regression: $y_i \in \mathbb{R}$ $h(\mathbf{z}_i) \in \mathbb{R}$



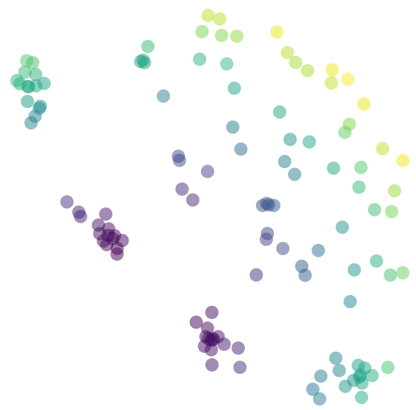
$$\ell(\mathbf{x}_i, h) = \|y_i - h(\mathbf{z}_i)\|_2^2$$

Machine learning theory

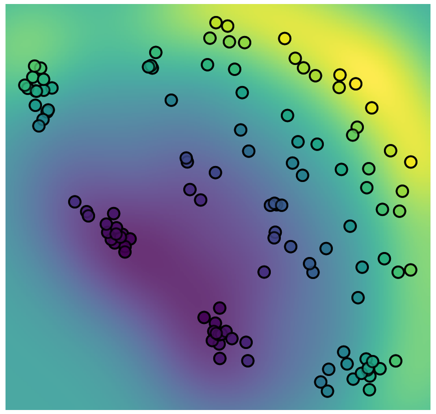
Supervised learning:

- In the supervised setting: $\mathbf{x}_i = (\mathbf{z}_i, y_i)$

Regression: $y_i \in \mathbb{R}$ $h(\mathbf{z}_i) \in \mathbb{R}$

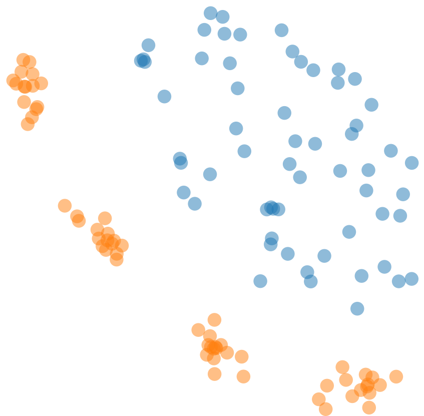


$\Longrightarrow$

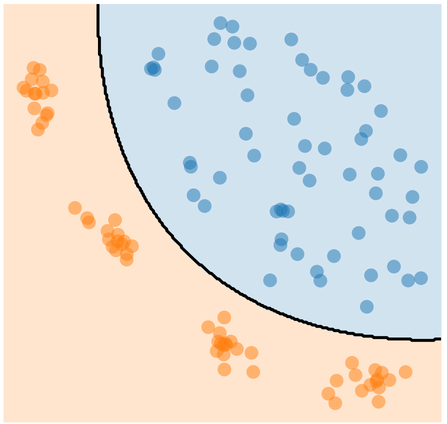


$$\ell(\mathbf{x}_i, h) = \|y_i - h(\mathbf{z}_i)\|_2^2$$

Classification: $y_i, h(\mathbf{z}_i) \in \{-1, 1\}$



$\Longrightarrow$

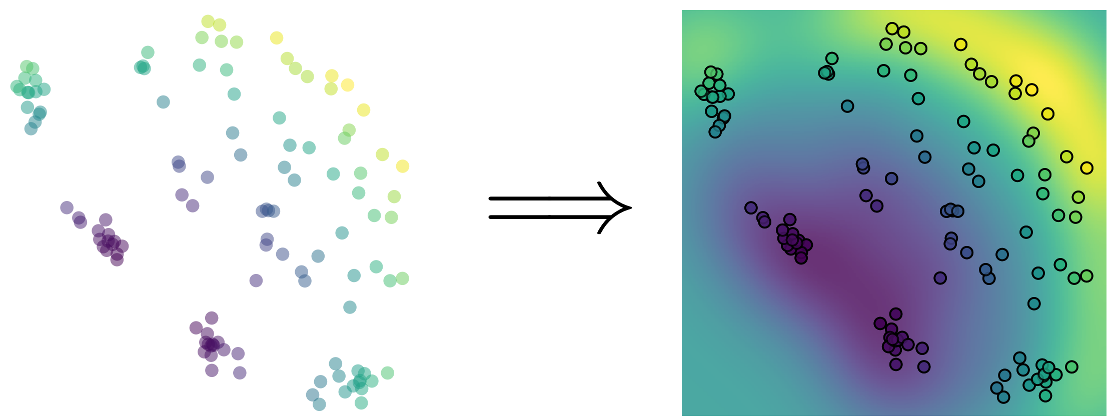


Machine learning theory

Supervised learning:

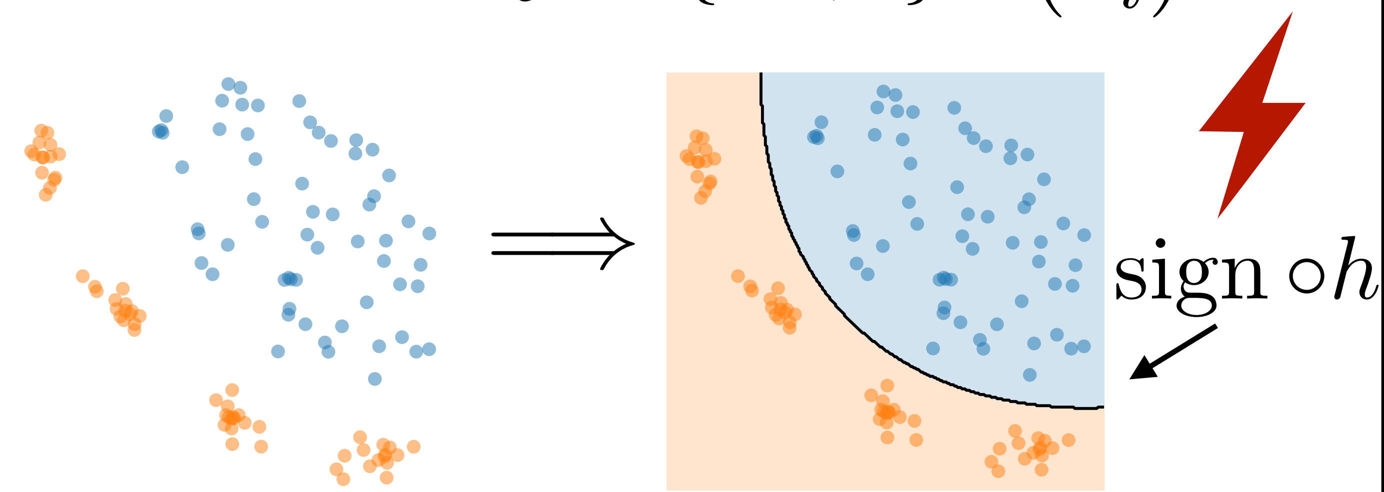
- In the supervised setting: $\mathbf{x}_i = (\mathbf{z}_i, y_i)$

Regression: $y_i \in \mathbb{R}$ $h(\mathbf{z}_i) \in \mathbb{R}$



$$\ell(\mathbf{x}_i, h) = \|y_i - h(\mathbf{z}_i)\|_2^2$$

Classification: $y_i \in \{-1, 1\}$ $h(\mathbf{z}_i) \in \mathbb{R}$

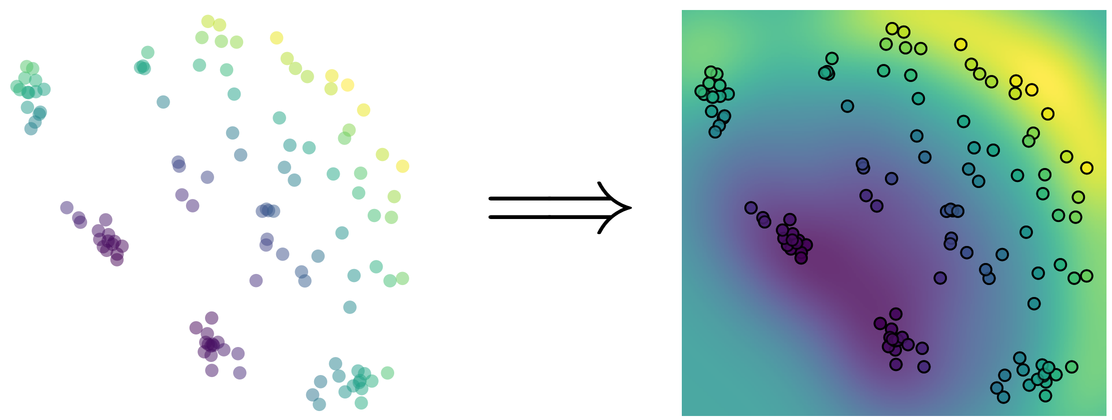


Machine learning theory

Supervised learning:

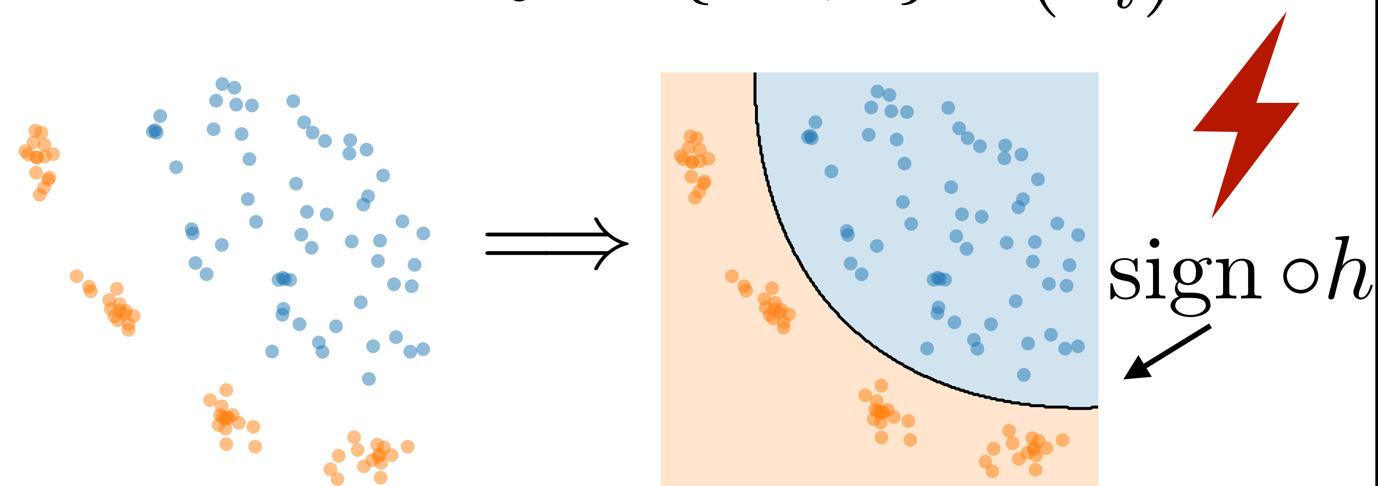
- In the supervised setting: $\mathbf{x}_i = (\mathbf{z}_i, y_i)$

Regression: $y_i \in \mathbb{R}$ $h(\mathbf{z}_i) \in \mathbb{R}$

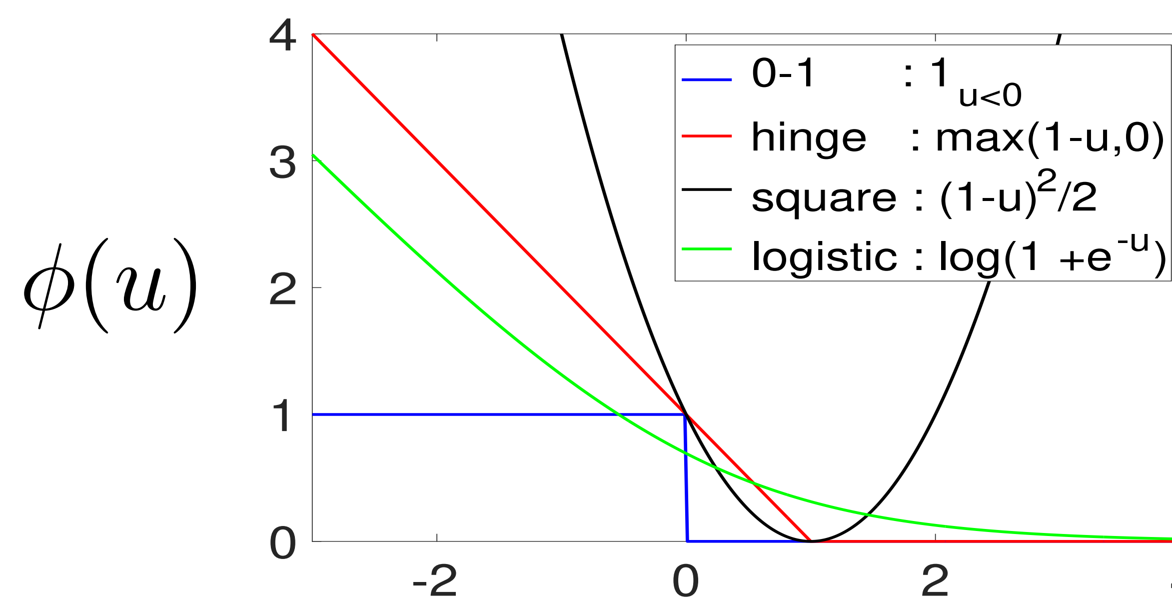


$$\ell(\mathbf{x}_i, h) = \|y_i - h(\mathbf{z}_i)\|_2^2$$

Classification: $y_i \in \{-1, 1\}$ $h(\mathbf{z}_i) \in \mathbb{R}$



$$\ell(\mathbf{x}_i, h) = \phi(y_i h(\mathbf{x}_i))$$

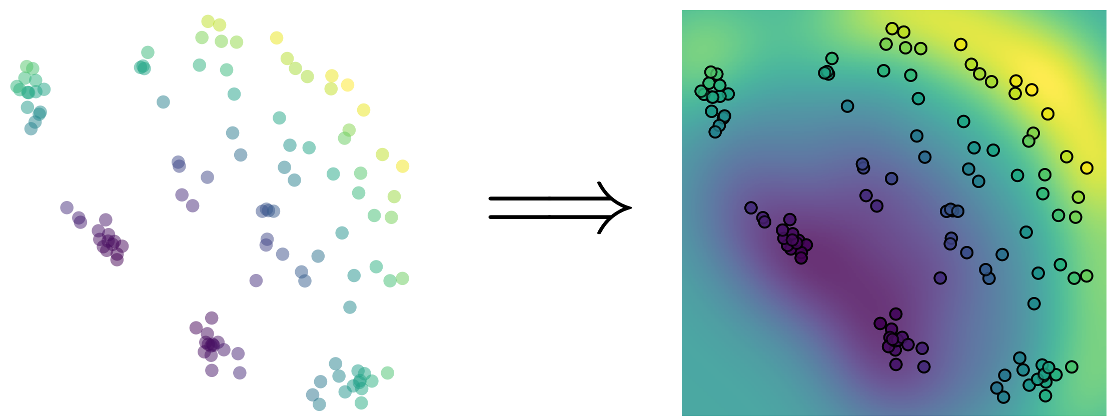


Machine learning theory

Supervised learning:

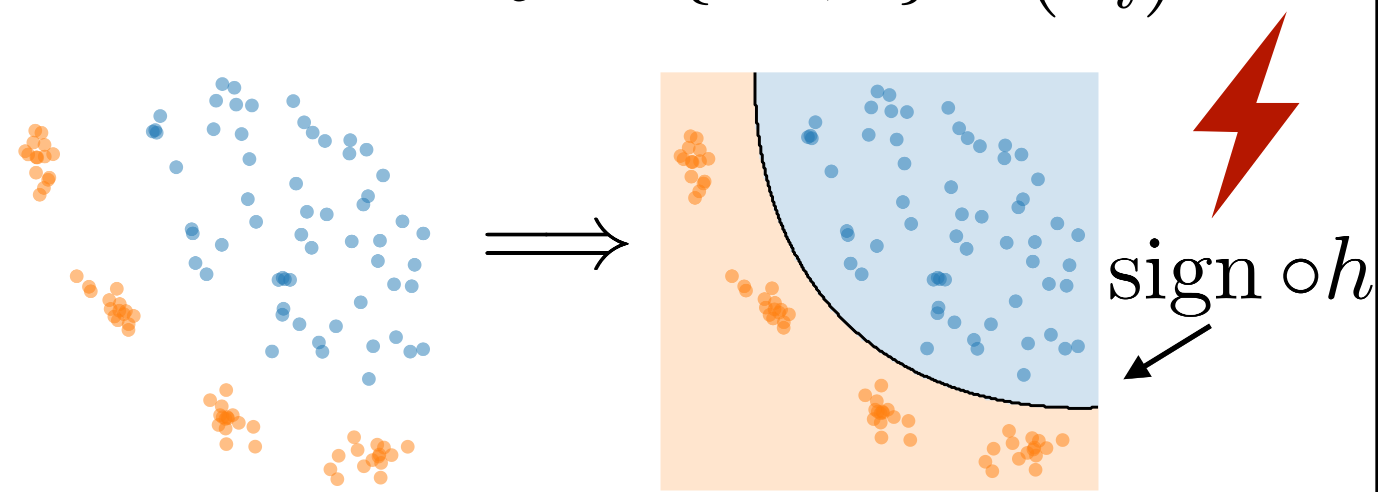
- In the supervised setting: $\mathbf{x}_i = (\mathbf{z}_i, y_i)$

Regression: $y_i \in \mathbb{R}$ $h(\mathbf{z}_i) \in \mathbb{R}$

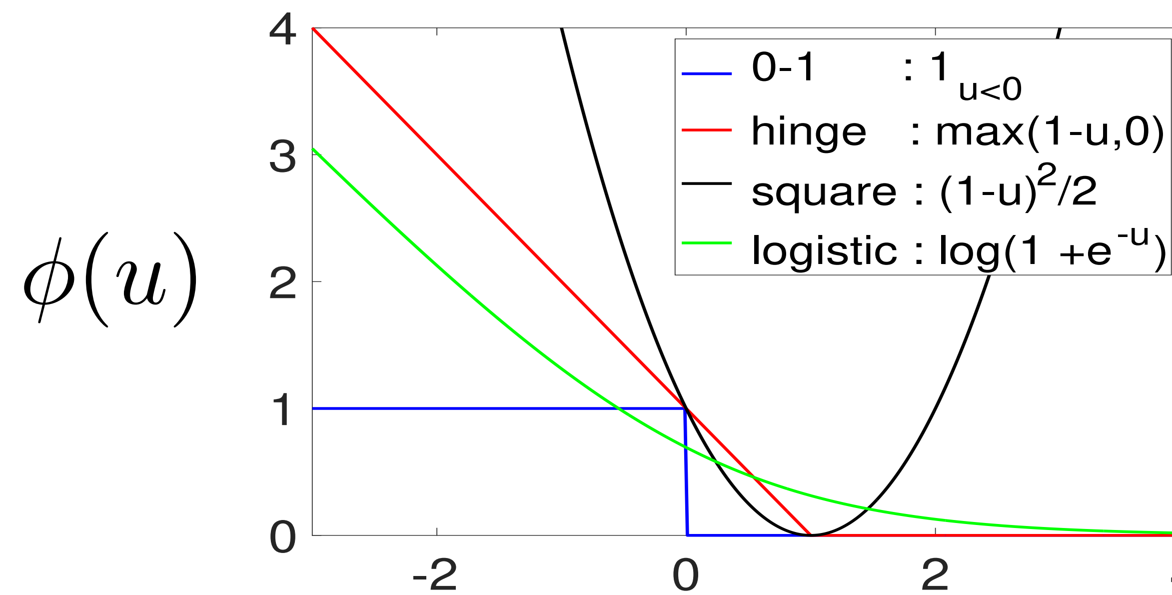


$$\ell(\mathbf{x}_i, h) = \|y_i - h(\mathbf{z}_i)\|_2^2$$

Classification: $y_i \in \{-1, 1\}$ $h(\mathbf{z}_i) \in \mathbb{R}$

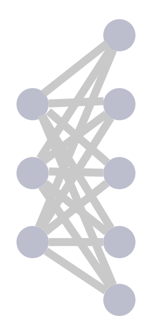


$$\ell(\mathbf{x}_i, h) = \phi(y_i h(\mathbf{x}_i))$$



Parametrized model

$$\boxed{h = f_{\boldsymbol{\theta}}}$$



Linear reg: $f_{\boldsymbol{\theta}}(\mathbf{z}) = \boldsymbol{\theta}^\top \mathbf{z}$

Neural networks: $f_{\boldsymbol{\theta}}(\mathbf{z}) = \mathrm{NN}_{\boldsymbol{\theta}}(\mathbf{z})$

Machine learning theory

ML in practice

- ERM $$\boxed{\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda\,\mathrm{Reg}(\boldsymbol{\theta})}$$

Machine learning theory

ML in practice

- ERM $\boxed{\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda\,\mathrm{Reg}(\boldsymbol{\theta})}$

everything differentiable

Machine learning theory

ML in practice:

- ERM $\boxed{\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda\,\mathrm{Reg}(\boldsymbol{\theta})}$
- Empirical risk: $\mathcal{R}_n(\boldsymbol{\theta})$ - - - - ▶ you **want** to minimize it

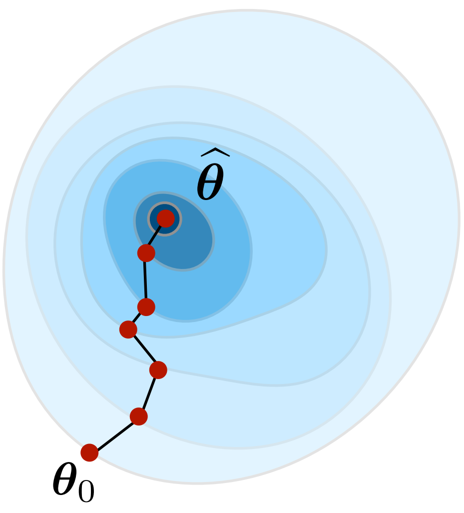
Machine learning theory

ML in practice:

- ERM $$\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda\,\mathrm{Reg}(\boldsymbol{\theta})$$
- Empirical risk: $\mathcal{R}_n(\boldsymbol{\theta})$ ----▶ you **want** to minimize it

Gradient descent:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} \mathcal{R}_n(\boldsymbol{\theta})$$



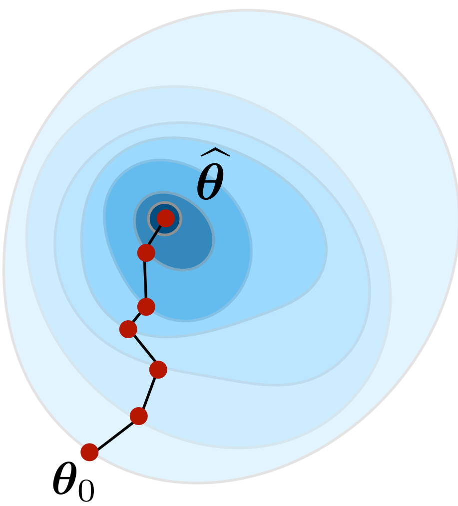
Machine learning theory

ML in practice:

- ERM $$\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda\,\mathrm{Reg}(\boldsymbol{\theta})$$
- Empirical risk: $\mathcal{R}_n(\boldsymbol{\theta})$ ▶ you **want** to minimize it

Gradient descent:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta\nabla_{\boldsymbol{\theta}}\mathcal{R}_n(\boldsymbol{\theta})$$



We are lazy: PyTorch

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta\,\mathrm{autodiff}[\mathcal{R}_n(\boldsymbol{\theta})]$$

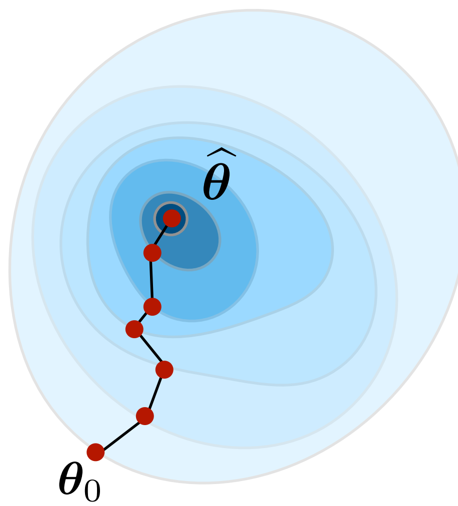
Machine learning theory

ML in practice:

- ERM $$\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda\,\mathrm{Reg}(\boldsymbol{\theta})$$
- Empirical risk: $\mathcal{R}_n(\boldsymbol{\theta})$ - - - - ▶ you **want** to minimize it

Gradient descent:

$$\boxed{\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta\nabla_{\boldsymbol{\theta}}\mathcal{R}_n(\boldsymbol{\theta})}$$



We are lazy: PyTorch

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta\,\mathrm{autodiff}[\mathcal{R}_n(\boldsymbol{\theta})]$$

Many many many variants

- Fix step-size, or change it $(\eta_k)_k$
- Momentum, averaging, adaptative step-size strategies:
  - Momentum and Accelerated gradients [Nesterov, 1983]
  - RMSPROP [Tieleman & Hinton, 2012]
  - Adam [Kingma & Ba, 2014]

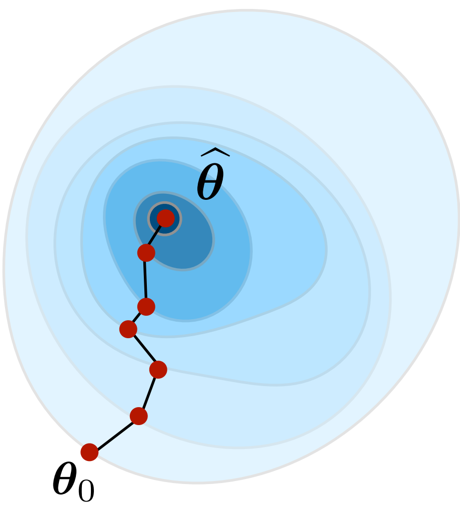
Machine learning theory

ML in practice:

- ERM $$\boxed{\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda\,\mathrm{Reg}(\boldsymbol{\theta})}$$
- Empirical risk: $\mathcal{R}_n(\boldsymbol{\theta})$ - - - - ▶ you **want** to minimize it

Gradient descent:

$$\boxed{\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} \mathcal{R}_n(\boldsymbol{\theta})}$$



We are lazy: PyTorch

$$\boxed{\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta\,\mathrm{autodiff}[\mathcal{R}_n(\boldsymbol{\theta})]}$$

Many many many variants

Hope that $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_\infty$ is « good »

- Fix step-size, or change it $(\eta_k)_k$
- Momentum, averaging, adaptative step-size strategies:

  Momentum and Accelerated gradients [Nesterov, 1983]

  RMSPROP [Tieleman & Hinton, 2012] | Adam [Kingma & Ba, 2014]

Machine learning theory

Hope that $\widehat{\theta} = \theta_{\infty}$ is « good » ??

Machine learning theory

Hope that $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_\infty$ is « good » ??

- **Approximate the true risk**

  - **ERM:** $\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, \mathcal{R}_n(\boldsymbol{\theta})$ $= \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i, f_{\boldsymbol{\theta}})$

Machine learning theory

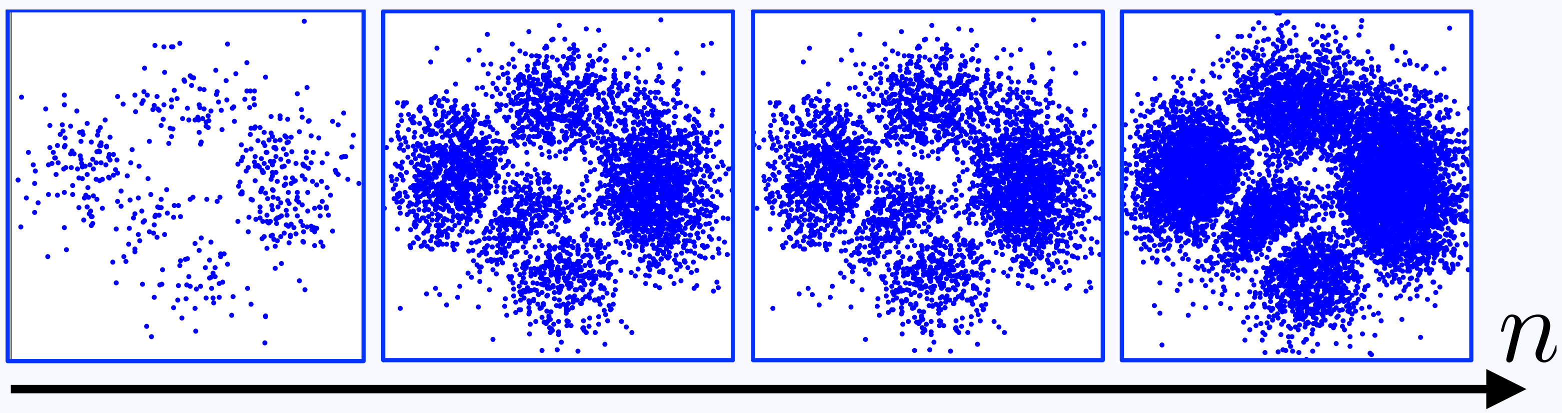
Hope that $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_\infty$ is « good » ??

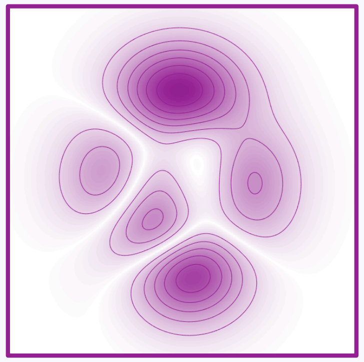
- **Approximate the true risk**
  - **ERM:** $\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_n(\boldsymbol{\theta})$ $= \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{x}_i, \boldsymbol{\theta})$ $= \underset{\mathbf{x}\sim\pi_n}{\mathbb{E}}[\ell(\mathbf{x}, \boldsymbol{\theta})]$

Side note

Empirical distribution $\pi_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$

True distrib. π





$n = \infty$

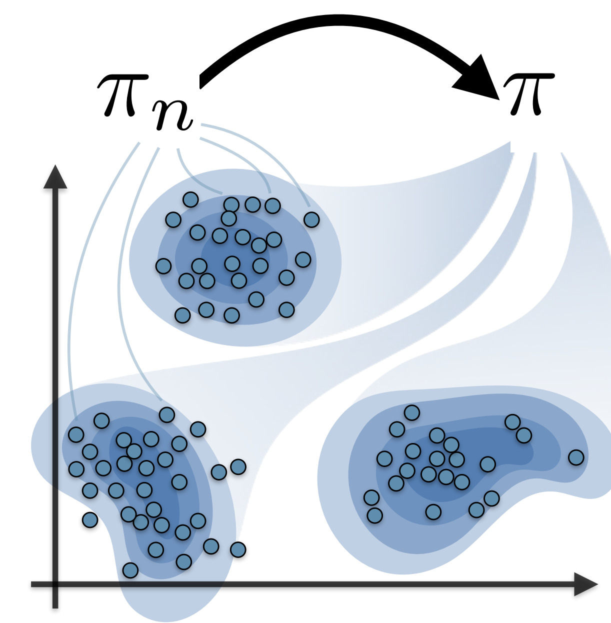
Machine learning theory

Hope that $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_\infty$ is « good » ??

Approximate the true risk

- **ERM:** $\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{R}_n(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x} \sim \pi_n} [\ell(\mathbf{x}, \boldsymbol{\theta})]$
- **True risk:** $\boxed{\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x} \sim \pi} [\ell(\mathbf{x}, \boldsymbol{\theta})]}$
- Best param. (if we have all the data in the world)

$$\boldsymbol{\theta}^\star \in \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{R}(\boldsymbol{\theta})$$



Machine learning theory

Hope that $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_\infty$ is « good » ??

Approximate the true risk

- **ERM:** $\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_n(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim\pi_n}[\ell(\mathbf{x},\boldsymbol{\theta})]$
- **True risk:** $\boxed{\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim\pi}[\ell(\mathbf{x},\boldsymbol{\theta})]}$
- Best param. (if we have all the data in the world)

$$\boldsymbol{\theta}^\star \in \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{R}(\boldsymbol{\theta})$$

Wished guarantees:

$$0 \leq \mathcal{R}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^\star) \leq \lambda_n \underset{n\to+\infty}{\to} 0$$

Machine learning theory

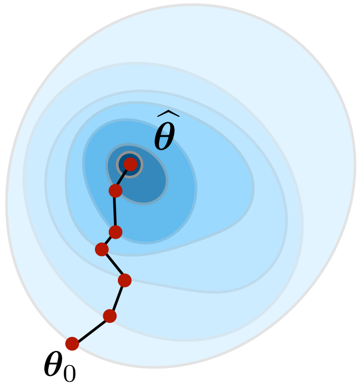
ML in practice:

- ERM $$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda \,\mathrm{Reg}(\boldsymbol{\theta})$$

- Empirical risk: $\mathcal{R}_n(\boldsymbol{\theta})$ ----▶ you **want** to minimize it

Gradient descent:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} \mathcal{R}_n(\boldsymbol{\theta})$$



Guarantees:

Hope that $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_\infty$ is « good »

Machine learning theory

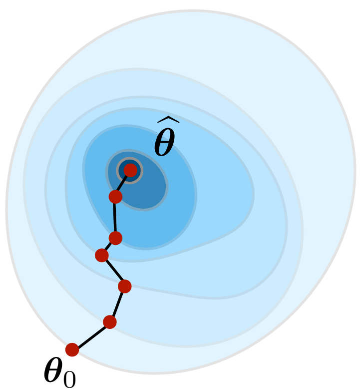
ML in practice:

- ERM $\boxed{\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda \, \mathrm{Reg}(\boldsymbol{\theta})}$
- Empirical risk: $\mathcal{R}_n(\boldsymbol{\theta})$ - - - - ▶ you **want** to minimize it

Gradient descent:

$$\boxed{\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} \mathcal{R}_n(\boldsymbol{\theta})}$$

$\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ell(\mathbf{x}_i, \boldsymbol{\theta})$



Guarantees:

Hope that $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_\infty$ is « good »

Expensive when large scale

- Complexity: $\mathcal{O}(n \times C_\nabla \times n_{it})$ may not even fit in memory
- Alternative: SGD, mini-batches [Bottou, 2010]
- **BUT** requires multiple passes (epochs)

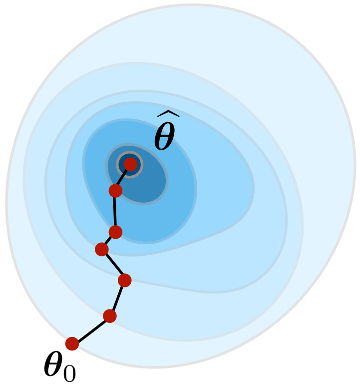
Machine learning theory

ML in practice:

- ERM $$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda\,\mathrm{Reg}(\boldsymbol{\theta})$$
- Empirical risk: $\mathcal{R}_n(\boldsymbol{\theta})$ ----▶ you **want** to minimize it

Gradient descent:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} \mathcal{R}_n(\boldsymbol{\theta})$$



$$\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ell(\mathbf{x}_i, \boldsymbol{\theta})$$

Expensive when large scale

- Complexity: $\mathcal{O}(n \times C_{\nabla} \times n_{it})$
- Alternative: SGD, mini-batches [Bottou, 2010]
- **BUT** requires multiple passes (epochs)

Guarantees:

Hope that $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_\infty$ is « good »

Difficult to prove

- Non-convexity/ high-dim
- Stats/Optim/Approx theory

Machine learning theory

ML in practice:

- ERM $$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda\,\mathrm{Reg}(\boldsymbol{\theta})$$

Gradie

$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}$

« good »

Compressive learning theory

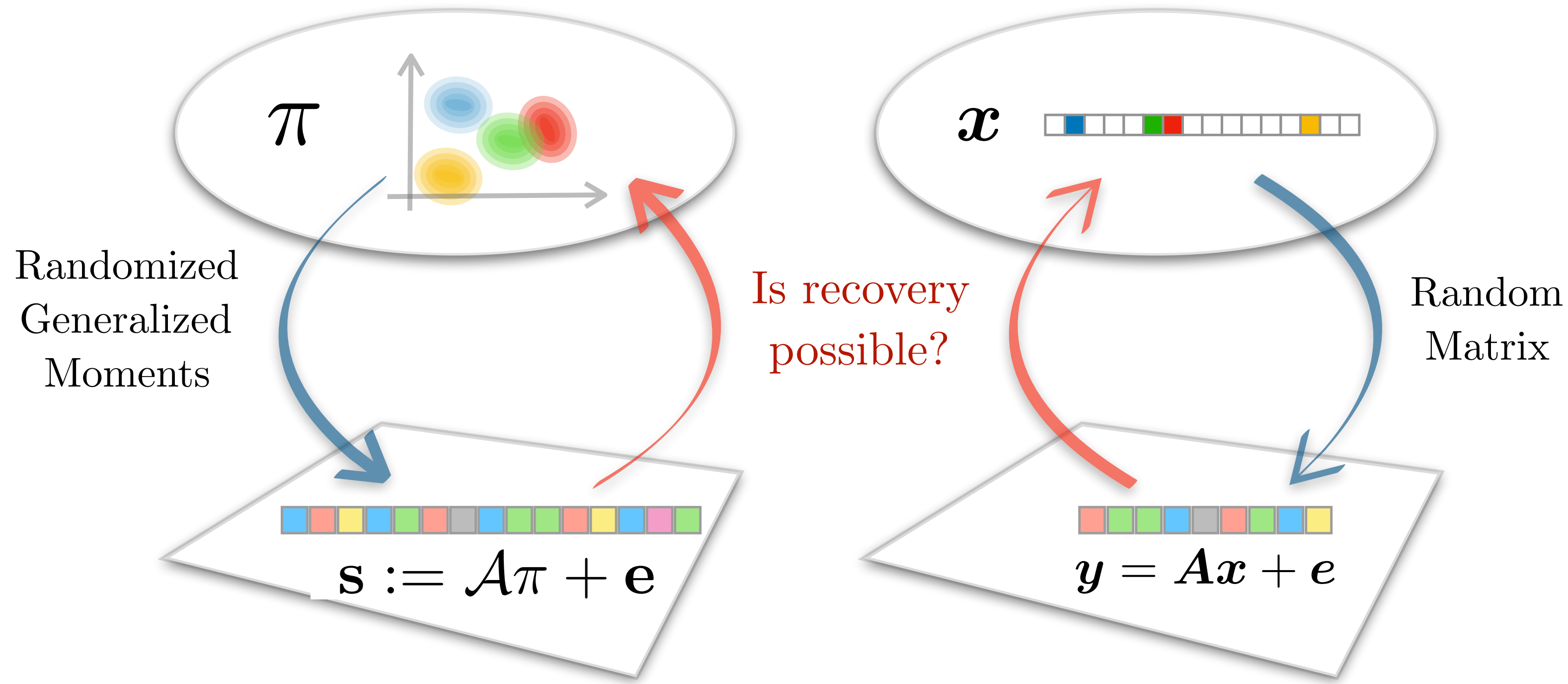
$\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ell(\mathbf{x}_i, \boldsymbol{\theta})$

Expensive when large scale

- Complexity: $\mathcal{O}(n \times C_{\nabla} \times n_{it})$
- Alternative: SGD, mini-batches [Bottou, 2010]
- **BUT** requires multiple passes (epochs)

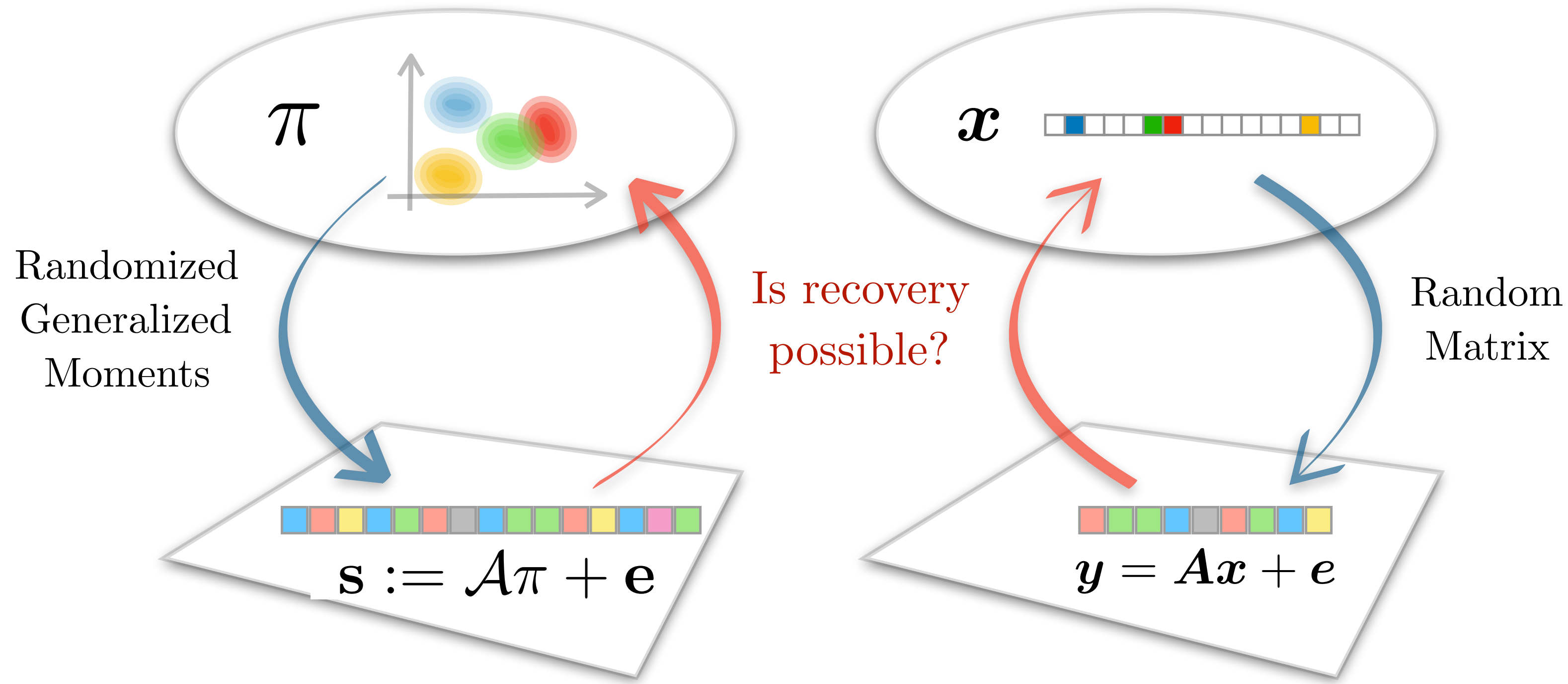
Difficult to prove

- Non-convexity/ high-dim
- Stats/Optim/Approx theory



Compressive Learning

- Theory of sketching
- Sketching in practice
- Theoretical guarantees
- Limitations & perspectives

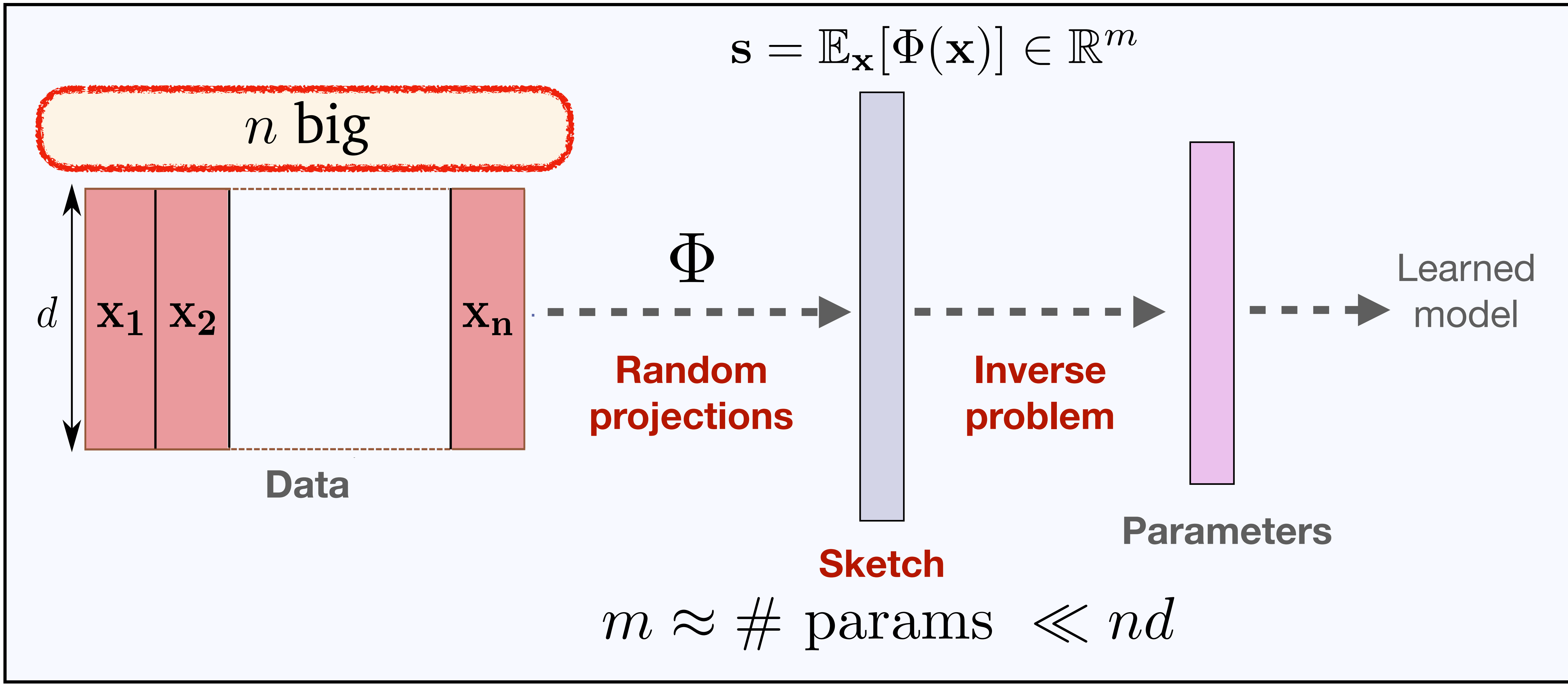


Compressive Learning

- Theory of sketching
- Sketching in practice
- Theoretical guarantees
- Limitations & perspectives

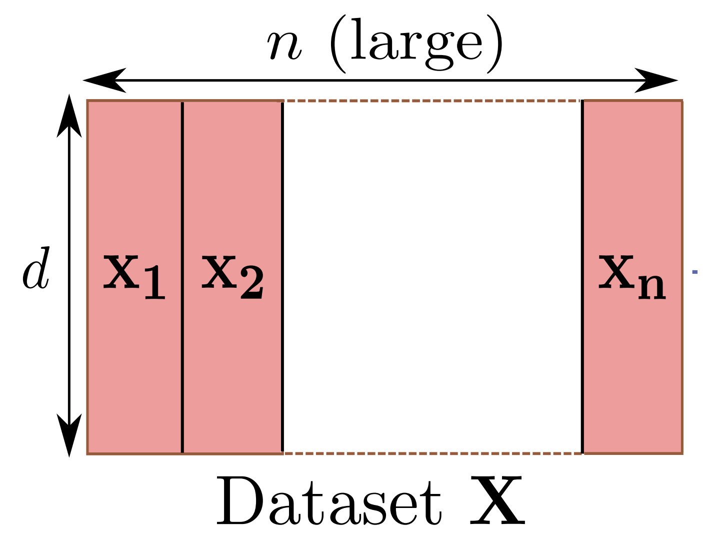
Theory of sketching

The big picture



Theory of sketching

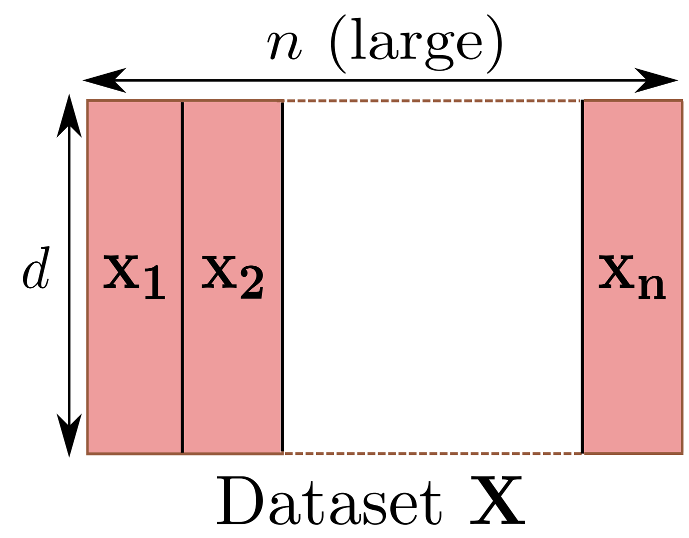
- « Dimension » reduction
  - « Low-dim » representation of a dataset



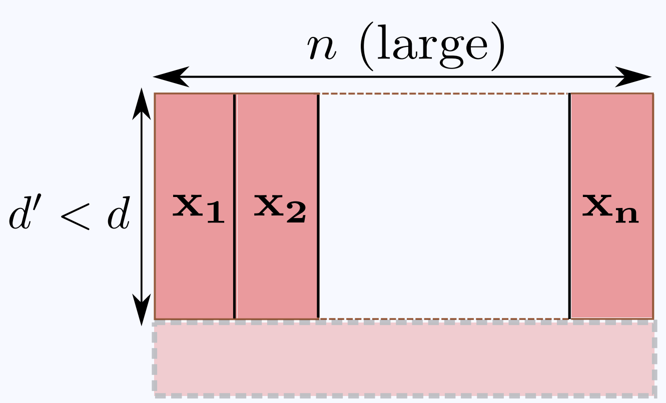
Theory of sketching

« Dimension » reduction

- « Low-dim » representation of a dataset



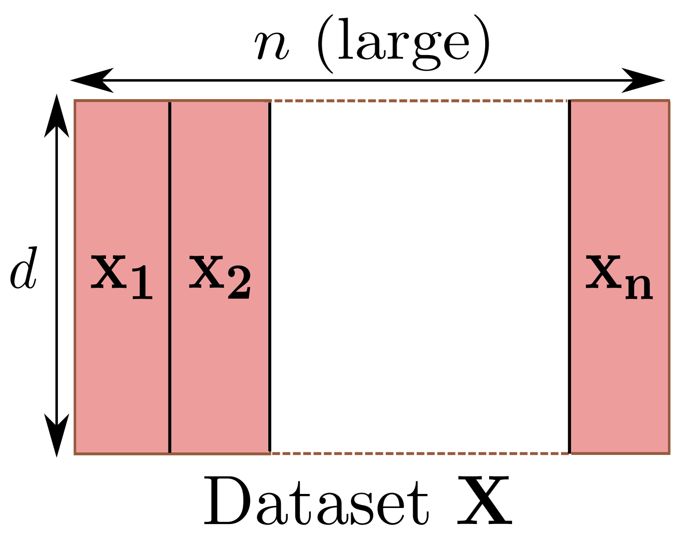
Dimension reduction



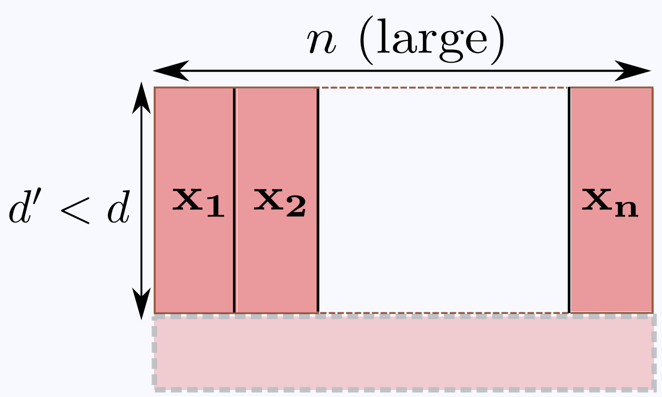
- Random projections (JL lemma)
- Feature selection
- Minimum distorsion embedding, PCA

Theory of sketching

- **« Dimension » reduction**
  - « Low-dim » representation of a dataset

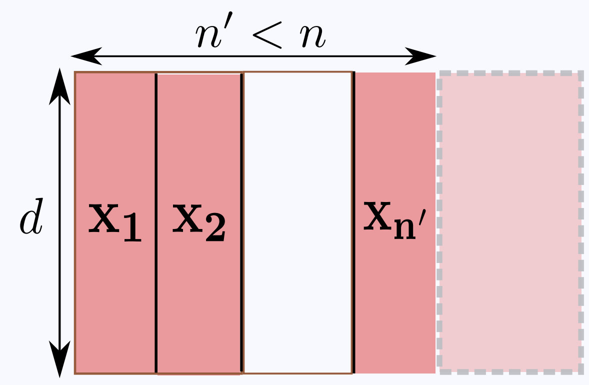


Dimension reduction



- Random projections (JL lemma)
- Feature selection
- Minimum distortion embedding, PCA

Subsampling

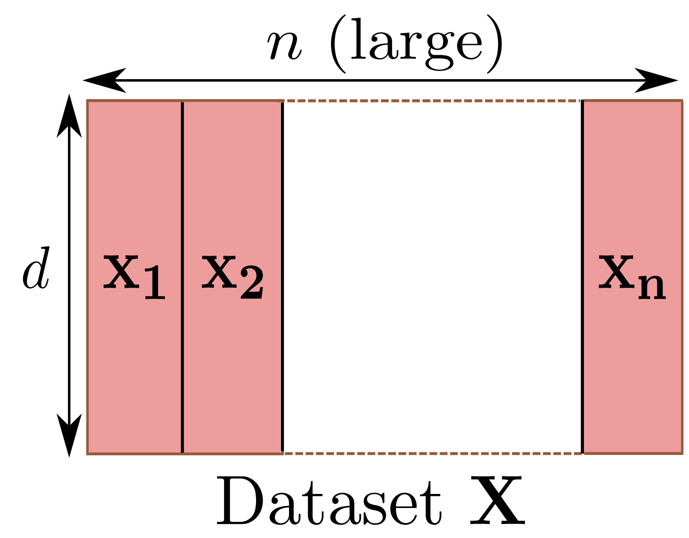


- Coresets
- Importance sampling

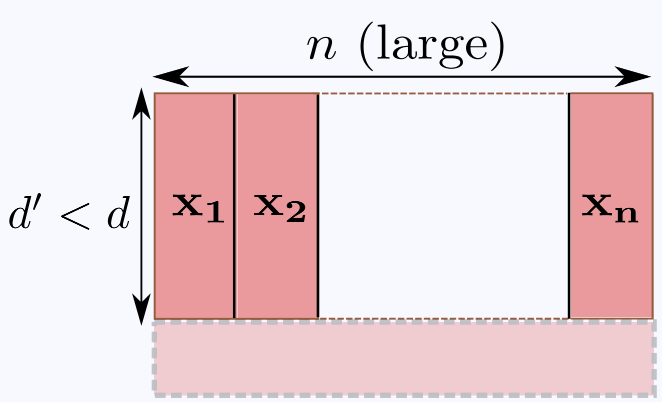
Theory of sketching

« Dimension » reduction

- « Low-dim » representation of a dataset



Dimension reduction

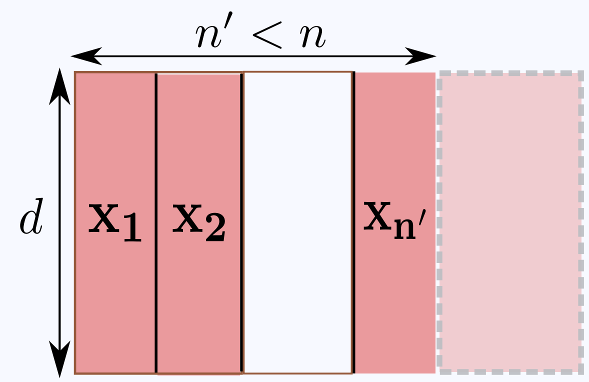


Random projections (JL lemma)

Feature selection

Minimum distortion embedding, PCA

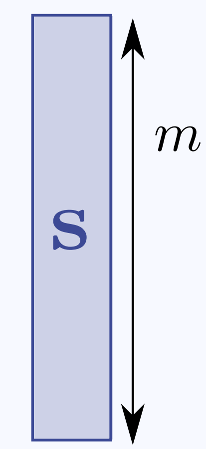
Subsampling



Coresets

Importance sampling

Here: linear « sketch »

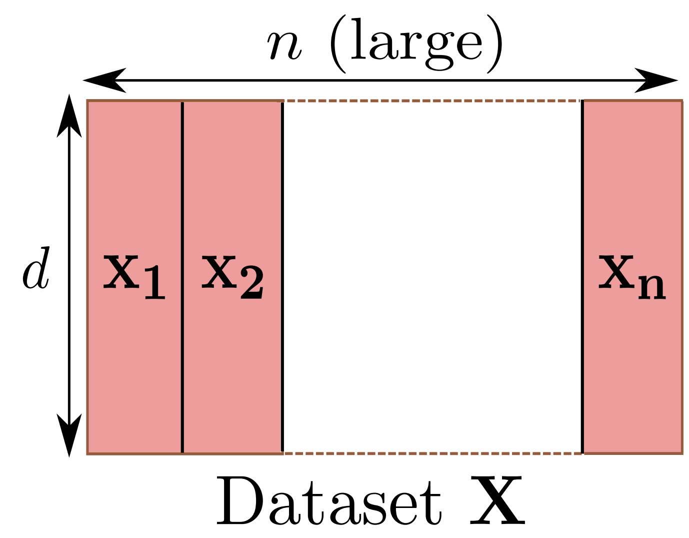


Only one vector

[Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, Yann Traonmilin, Antoine Chatalic, Vincent Schellekens, Laurent Jacques…]

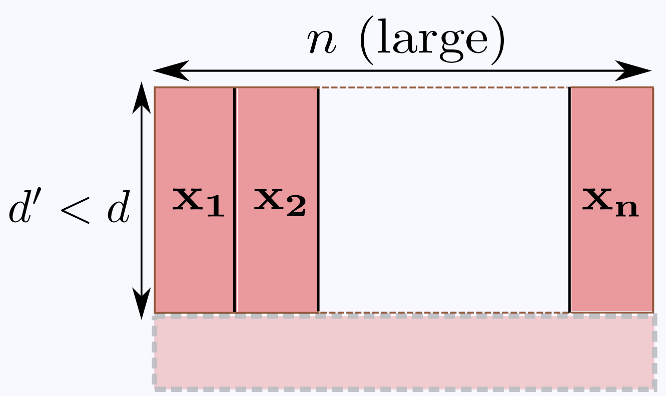
Theory of sketching

- « Dimension » reduction
  - « Low-dim » representation of a dataset



How do we sketch ? How do we learn from sketch ?

Dimension reduction

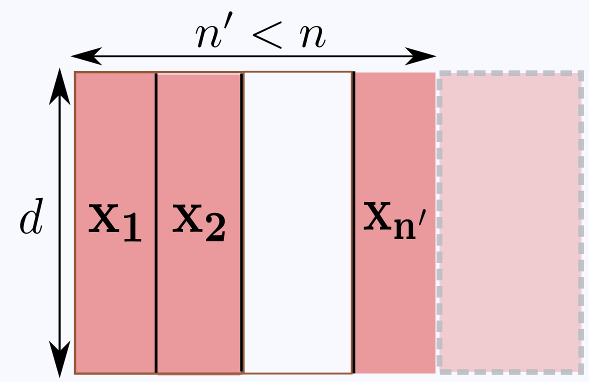


Random projections (JL lemma)

Feature selection

Minimum distortion embedding, PCA

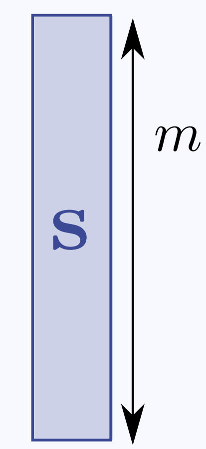
Subsampling



Coresets

Importance sampling

Here: linear « sketch »



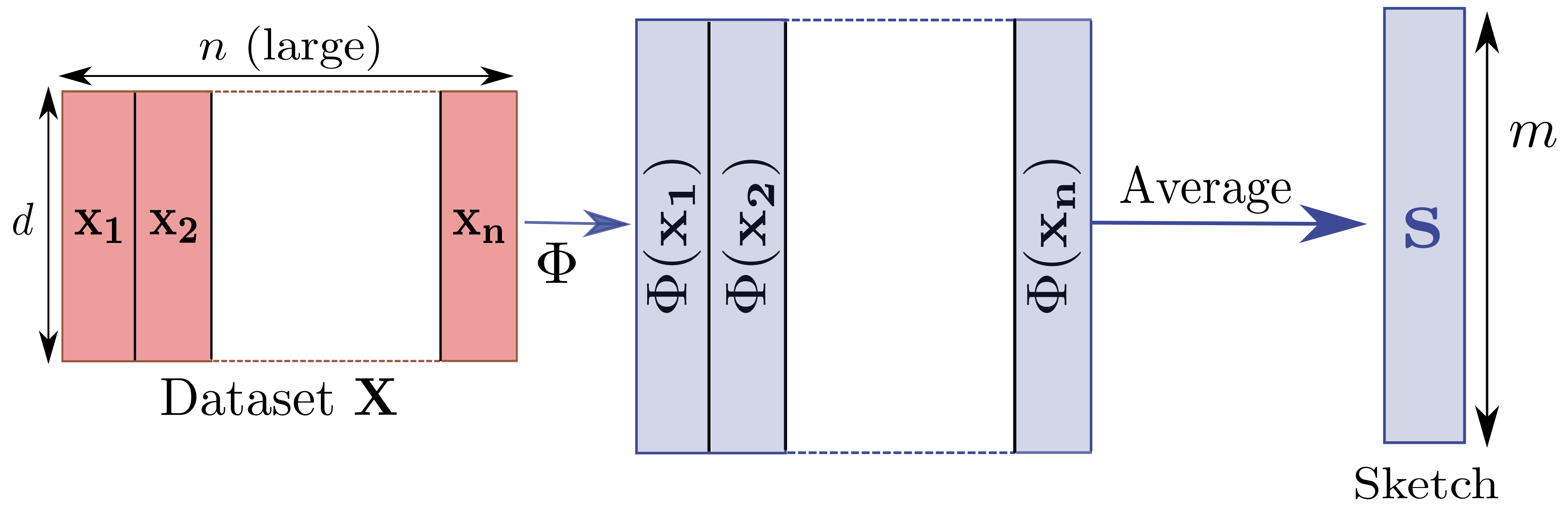
Only one vector

[Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, Yann Traonmilin, Antoine Chatalic, Vincent Schellekens, Laurent Jacques…]

Theory of sketching

Obtaining the sketch

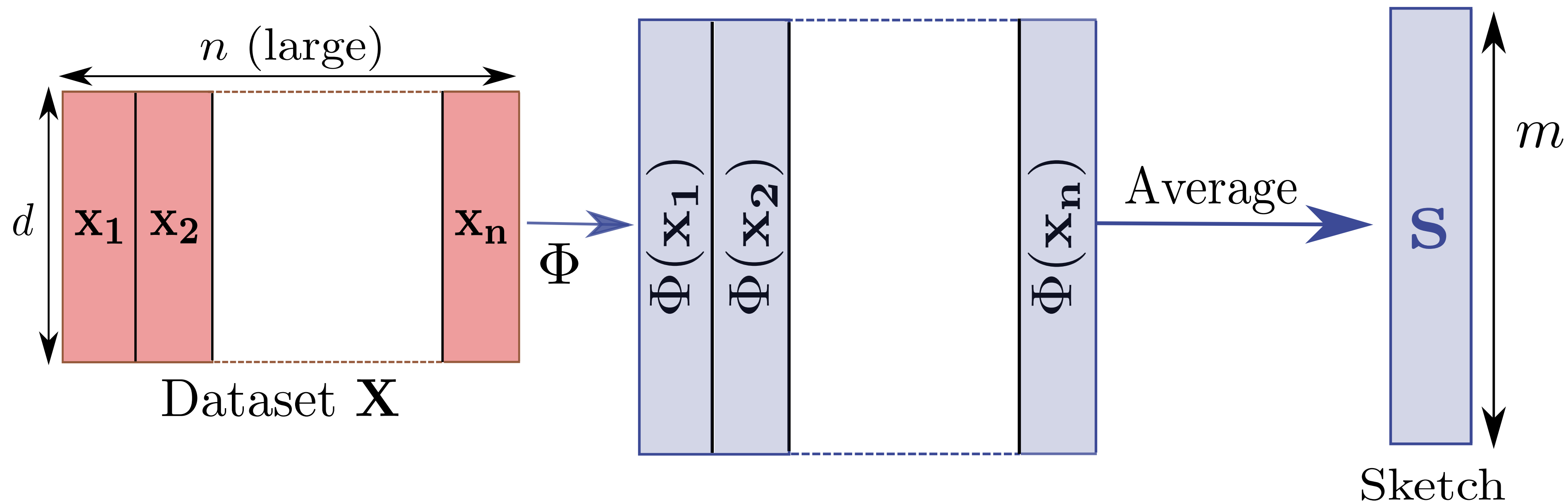
- A function called feature operator $\boxed{\Phi : \mathbb{R}^d \to \mathbb{R}^m}$
- Averaging **n points ->** $\mathbf{s} := \frac{1}{n} \sum_{i=1}^{n} \Phi(\mathbf{x}_i)$



Theory of sketching

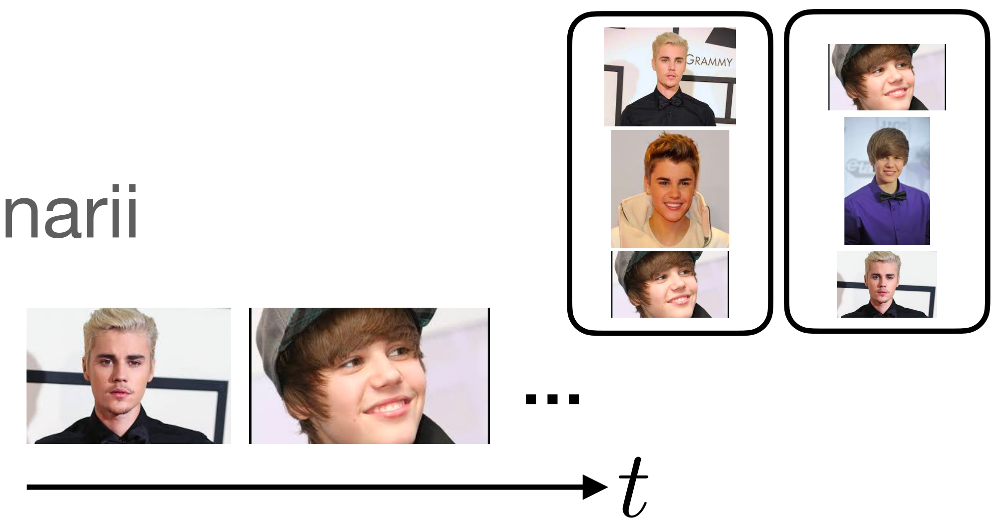
Obtaining the sketch

- A function called feature operator $\boxed{\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^m}$
- Averaging **n points ->** $\mathbf{s} := \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i)$



Average is a simple idea but …

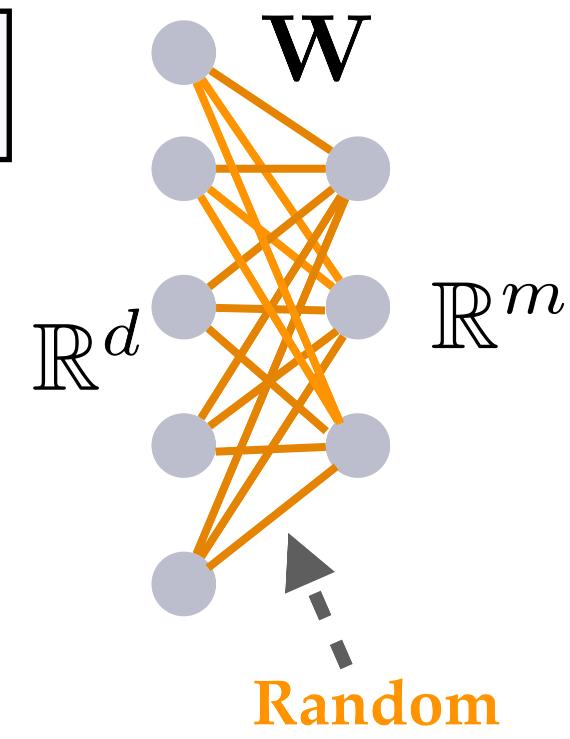
- Suitable for distributed /streaming scenarii
- It can be calculated in parallel



Theory of sketching

Randomization: the core of sketching

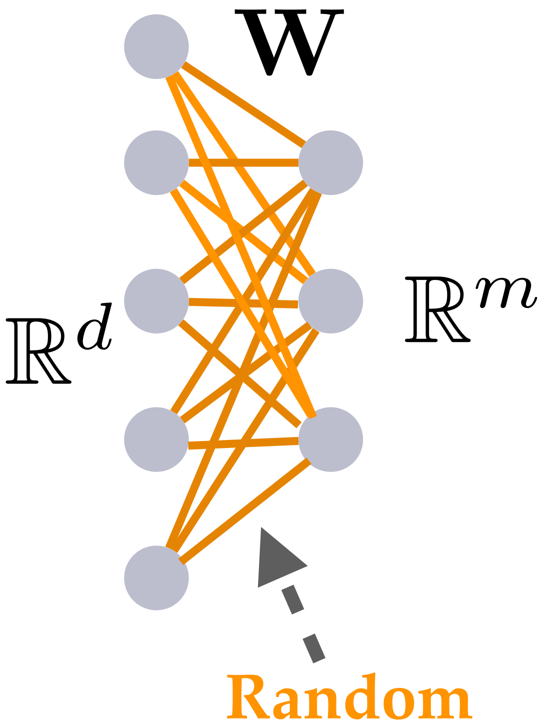
- A function called feature operator $\boxed{\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^m}$
- In practice $\Phi(\mathbf{x}) = \rho(\mathbf{W}\mathbf{x})$
- $\mathbf{W} \in \mathbb{R}^{m \times d}$ is a **random matrix**
- $\rho : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is a non-linear activation



Theory of sketching

Randomization: the core of sketching

- A function called feature operator $\boxed{\Phi : \mathbb{R}^d \to \mathbb{R}^m}$
- In practice $\Phi(\mathbf{x}) = \rho(\mathbf{W}\mathbf{x})$
- $\mathbf{W} \in \mathbb{R}^{m \times d}$ is a **random matrix**
- $\rho : \mathbb{R}^m \to \mathbb{R}^m$ is a non-linear activation



Random Fourier Features (RFF)

- $\mathbf{W} \in \mathbb{R}^{m \times d}$ is **Gaussian** $W_{ij} \sim \mathcal{N}(0, \sigma^2)$
- $\rho(\mathbf{y}) = \frac{1}{\sqrt{m}}(\exp(-iy_1), \cdots, \exp(-iy_m))$

[Rahimi & Recht, 2008]

$$\boxed{\Phi(\mathbf{x}) = \frac{1}{\sqrt{m}}(\exp(-i\boldsymbol{\omega}_1^\top \mathbf{x}), \cdots, \exp(-i\boldsymbol{\omega}_m^\top \mathbf{x}))}$$

$\mathbf{W} = [\boldsymbol{\omega}_1^\top, \cdots, \boldsymbol{\omega}_m^\top]$

Theory of sketching

A linear operator on distributions

- The **sketching operator** $\boxed{\mathcal{A} : \pi \to \int_{\mathbb{R}^d} \Phi(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) \in \mathbb{R}^m}$
- This is a **linear operator** on measures/distributions

Theory of sketching

A linear operator on distributions

- The **sketching operator** $\boxed{\mathcal{A} : \pi \to \int_{\mathbb{R}^d} \Phi(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) \in \mathbb{R}^m}$
- This is a **linear operator** on measures/distributions
- Generalized moments $\mathcal{A}\pi = \mathop{\mathbb{E}}_{\mathbf{x}\sim\pi}[\Phi(\mathbf{x})]$

Theory of sketching

A linear operator on distributions

- The **sketching operator** $$\boxed{\mathcal{A} : \pi \to \int_{\mathbb{R}^d} \Phi(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) \in \mathbb{R}^m}$$
- This is a **linear operator** on measures/distributions
- Generalized moments $\mathcal{A}\pi = \mathbb{E}_{\mathbf{x}\sim\pi}[\Phi(\mathbf{x})]$

$\Phi(\mathbf{x}) = \mathbf{x}$

Mean (moment 1)

$$\mathcal{A}\pi = \mathbb{E}_{\mathbf{x}\sim\pi}[\mathbf{x}]$$

$\Phi(\mathbf{x}) = \mathbf{x}\mathbf{x}^\top$

Variance (moment 2)

$$\mathcal{A}\pi = \mathbb{E}_{\mathbf{x}\sim\pi}[\mathbf{x}\mathbf{x}^\top]$$

Theory of sketching

A linear operator on distributions

- The **sketching operator** $\boxed{\mathcal{A} : \pi \to \int_{\mathbb{R}^d} \Phi(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) \in \mathbb{R}^m}$
- This is a **linear operator** on measures/distributions
- Generalized moments $\mathcal{A}\pi = \mathbb{E}_{\mathbf{x}\sim\pi}[\Phi(\mathbf{x})]$

$$\mathcal{A}\pi = \mathbb{E}_{\mathbf{x}\sim\pi}[\Phi(\mathbf{x})] \xrightarrow{\Phi(\mathbf{x}) = \mathbf{x}} \mathcal{A}\pi = \mathbb{E}_{\mathbf{x}\sim\pi}[\mathbf{x}]$$

Mean (moment 1)

$$\mathcal{A}\pi = \mathbb{E}_{\mathbf{x}\sim\pi}[\Phi(\mathbf{x})] \xrightarrow{\Phi(\mathbf{x}) = \mathbf{x}\mathbf{x}^\top} \mathcal{A}\pi = \mathbb{E}_{\mathbf{x}\sim\pi}[\mathbf{x}\mathbf{x}^\top]$$

Variance (moment 2)

$$\mathcal{A}\pi = \mathbb{E}_{\mathbf{x}\sim\pi}[\Phi(\mathbf{x})] \xrightarrow{\Phi = \mathrm{RFF}} \mathcal{A}\pi = (\mathcal{F}[\pi](\boldsymbol{\omega}_1), \cdots, \mathcal{F}[\pi](\boldsymbol{\omega}_m))$$

Sampling of the Fourier transform of the distrib.

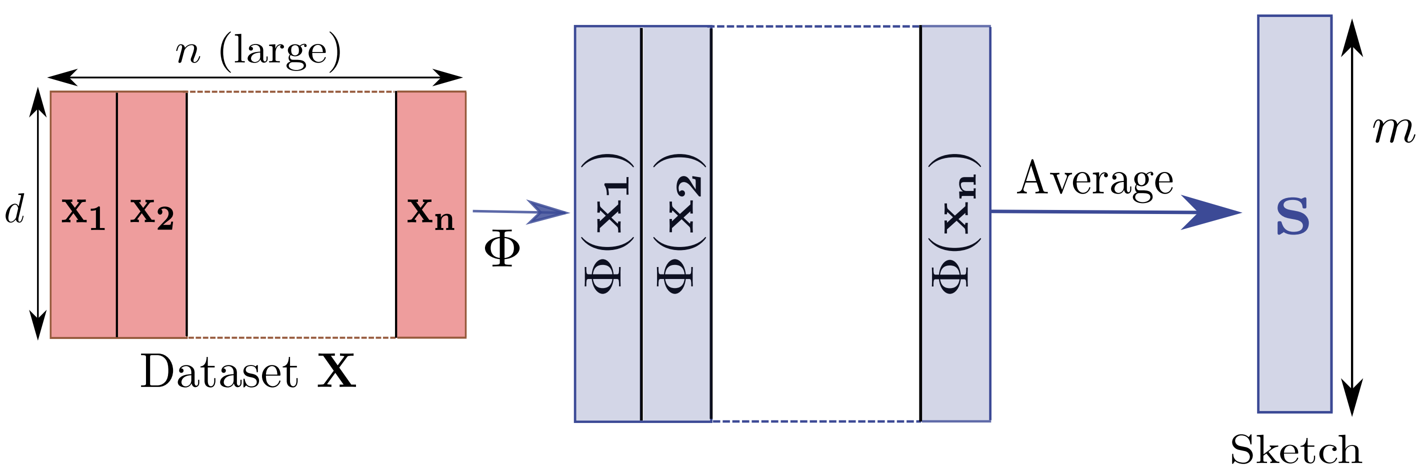
Theory of sketching

A linear operator on distributions

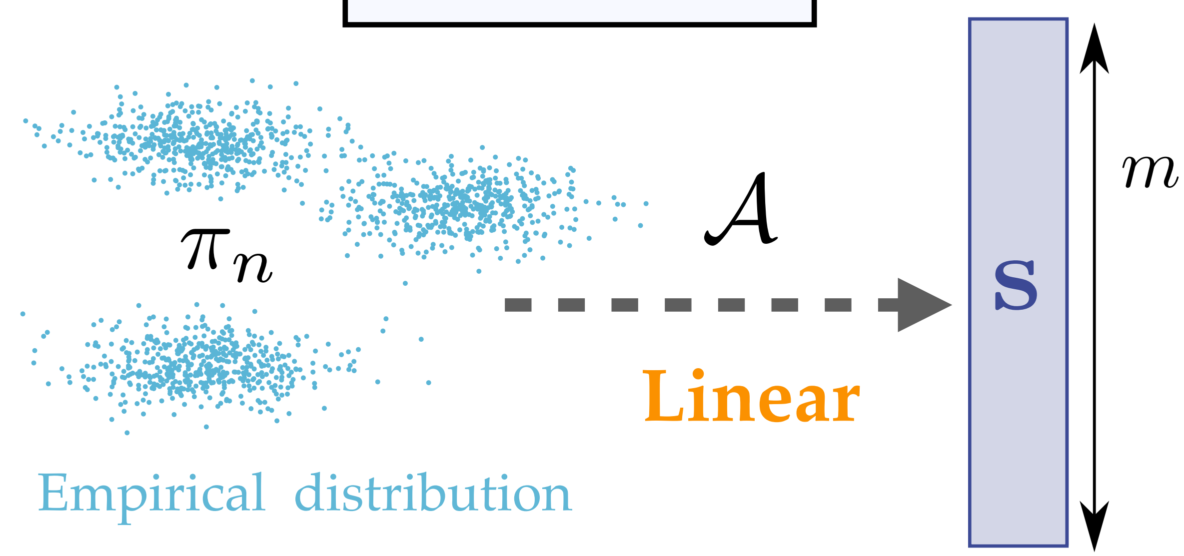
- The **sketching operator** $$\boxed{\mathcal{A} : \pi \rightarrow \int_{\mathbb{R}^d} \Phi(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) \in \mathbb{R}^m}$$
- This is a **linear operator** on measures/distributions
- Generalized moments $$\mathcal{A}\pi = \mathbb{E}_{\mathbf{x}\sim\pi}[\Phi(\mathbf{x})]$$

The two « ways » of sketching

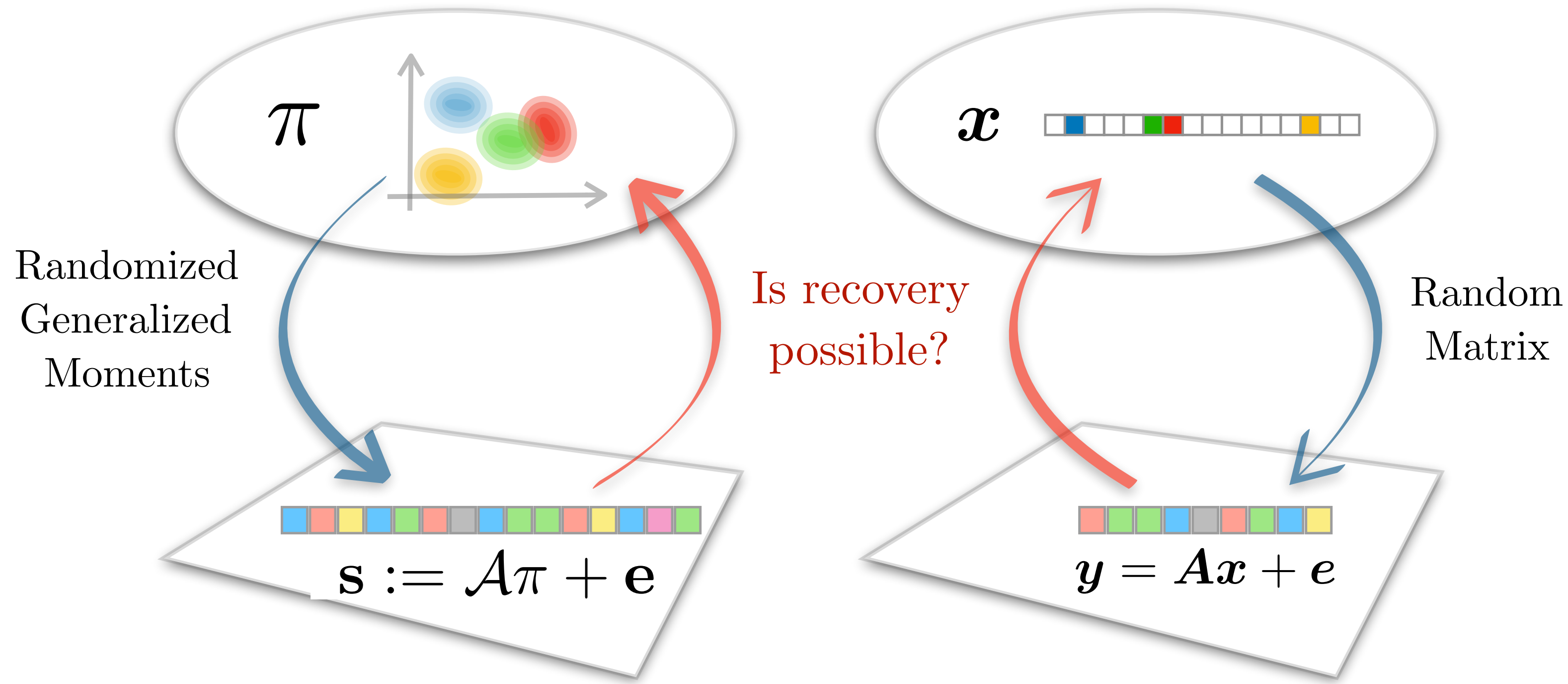
$$\boxed{\mathbf{s} := \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i)}$$



$$\boxed{\mathbf{s} := \mathcal{A}\pi_n}$$



$$\pi_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i}$$



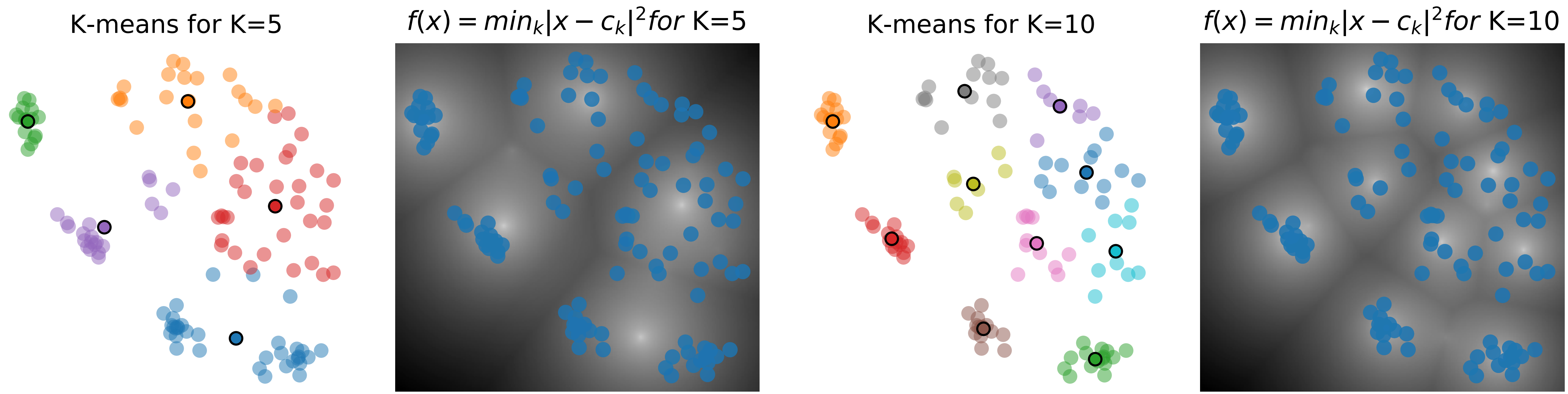
Compressive Learning

- Theory of sketching
- Sketching in practice
- RIP for theoretical guarantees
- Limitations & perspectives

Sketching in practice

Come back to K-means:

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \frac{1}{n}\sum_{i=1}^{n} \min_{k\in[\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$$

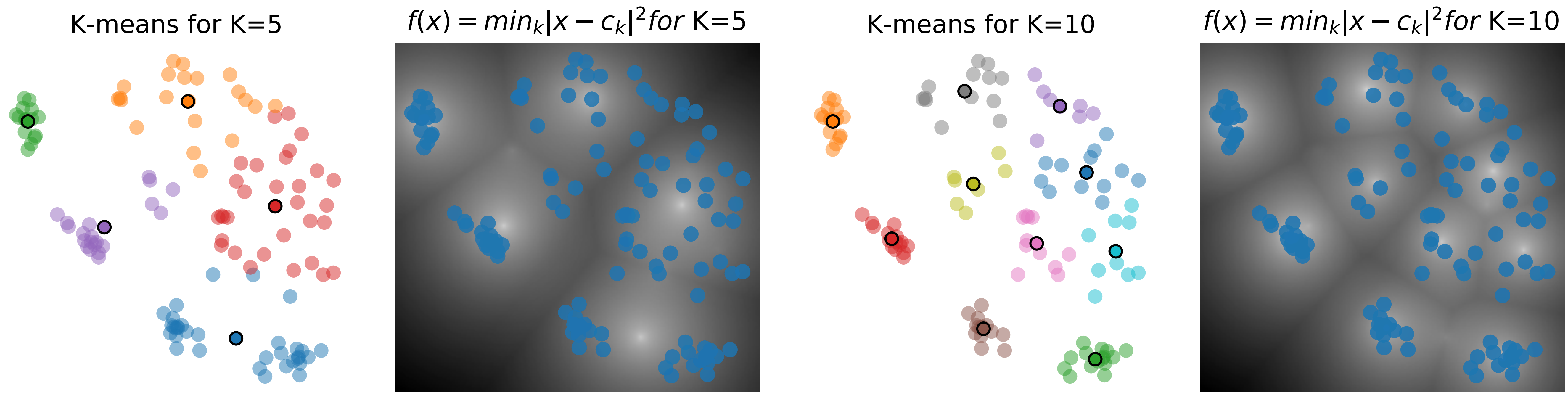


Sketching in practice

We want to find Kd params.

- **Come back to K-means:**

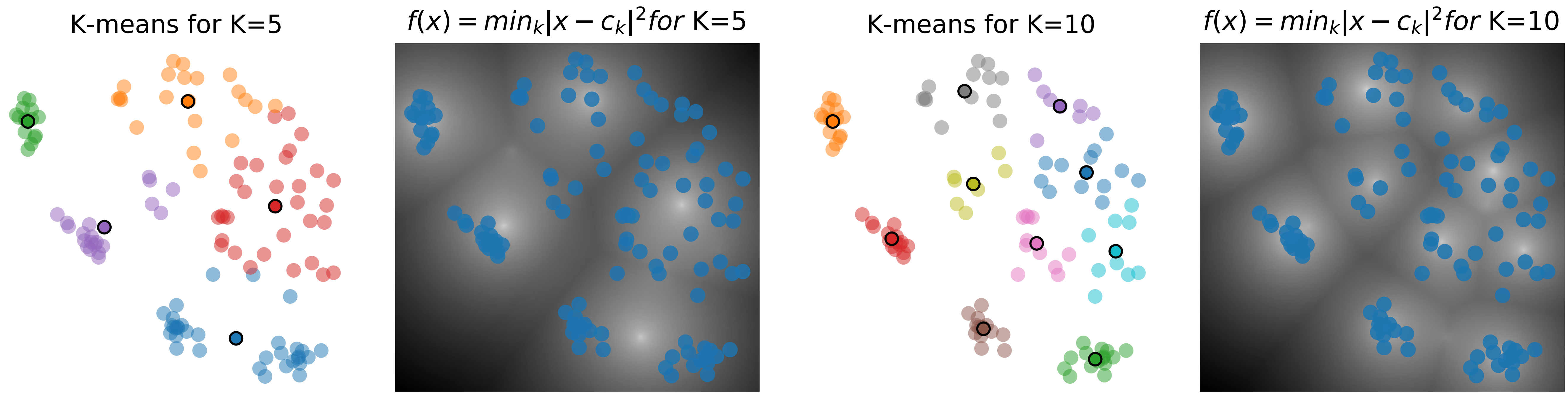
$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \frac{1}{n}\sum_{i=1}^n \min_{k\in[\![K]\!]} \|\mathbf{x}_i-\mathbf{c}_k\|_2^2$$



Sketching in practice

Come back to K-means:

$$\boxed{\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \frac{1}{n}\sum_{i=1}^{n} \min_{k\in[\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2}$$



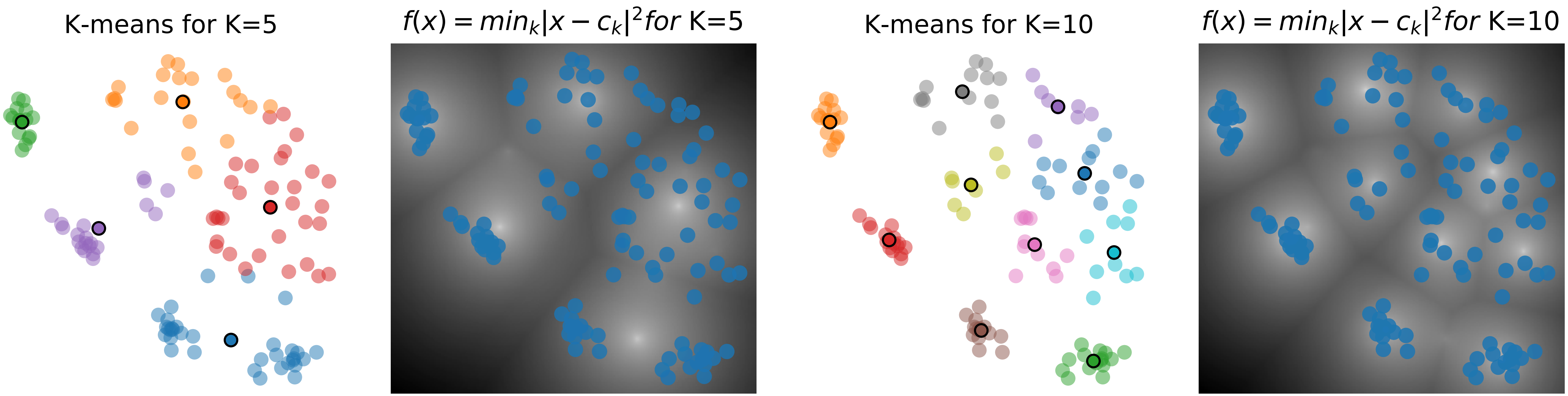
Another point of view:

- Find a distribution $\widehat{\pi} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\mathbf{c}_k}$ that bests approximate π
- We have access to the whole dataset i.e. π_n

Sketching in practice

- **Come back to K-means:**

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \frac{1}{n}\sum_{i=1}^{n} \min_{k\in[\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$$



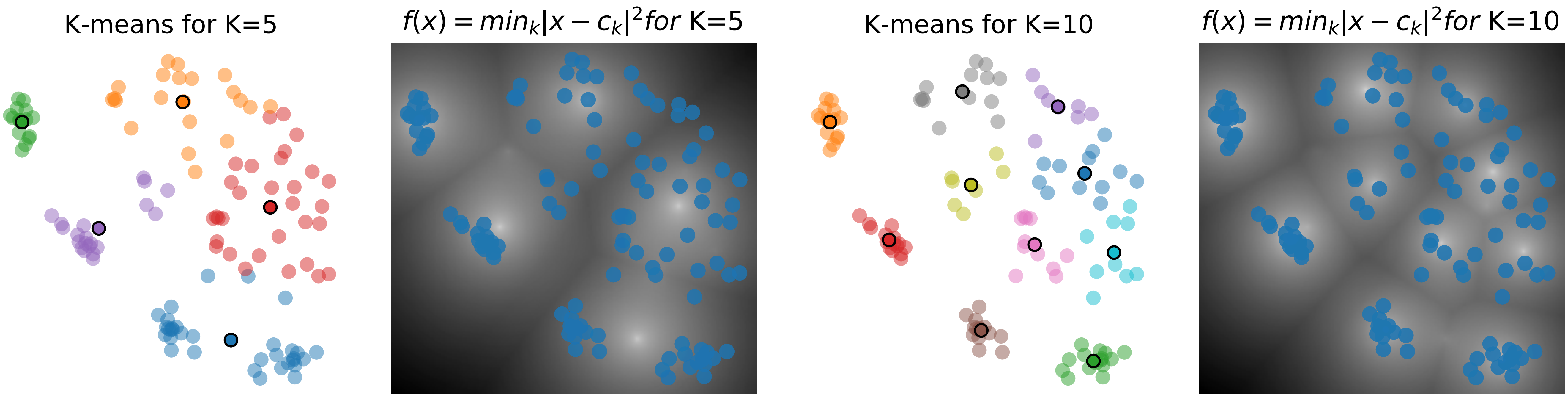
- **Another point of view:**
  - Find a distribution $\widehat{\pi} = \frac{1}{K}\sum_{k=1}^{K}\delta_{\mathbf{c}_k}$ that bests approximate π
  - We have access to the whole dataset i.e. π_n
- **In sketching:**
  - We only have access to $\boxed{\mathbf{s} := \mathcal{A}\pi_n}$

How do we do ?

Sketching in practice

Come back to K-means:

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \frac{1}{n}\sum_{i=1}^{n} \min_{k\in[\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$$



Another point of view:

- Find a distribution $\widehat{\pi} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\mathbf{c}_k}$ that **bests approximate** π

we will change that

- We have access to the whole dataset i.e. π_n

In sketching:

- We only have access to $\boxed{\mathbf{s} := \mathcal{A}\pi_n}$

How do we do ?

Sketching in practice

Sketching for large scale K-means:

- We aim at solving:

$$\boxed{\min_{\widehat{\pi} \text{ s.t. } \widehat{\pi} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\mathbf{c}_k}} \|\mathbf{s} - \mathcal{A}\widehat{\pi}\|_2}$$

Sketching in practice

Sketching for large scale K-means:

- We aim at solving:

$$\min_{\widehat{\pi} \text{ s.t. } \widehat{\pi} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\mathbf{c}_k}} \|\mathbf{s} - \mathcal{A}\widehat{\pi}\|_2$$

Find a distrib. of K diracs

Sketching in practice

Sketching for large scale K-means:

- We aim at solving:

$$\boxed{\min_{\widehat{\pi} \text{ s.t. } \widehat{\pi} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\mathbf{c}_k}} \|\mathbf{s} - \mathcal{A}\widehat{\pi}\|_2}$$

Sketch of the distrib

Sketching in practice

Sketching for large scale K-means:

- We aim at solving:

$$\min_{\widehat{\pi} \text{ s.t. } \widehat{\pi} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\mathbf{c}_k}} \|\mathbf{s} - \mathcal{A}\widehat{\pi}\|_2$$

Sketch of the data

Sketching in practice

Sketching for large scale K-means:

- We aim at solving:

$$\boxed{\min_{\widehat{\pi} \text{ s.t. } \widehat{\pi}=\frac{1}{K}\sum_{k=1}^{K}\delta_{\mathbf{c}_k}} \|\mathbf{s} - \mathcal{A}\widehat{\pi}\|_2}$$

- Find a distribution of K diracs whose sketch is the closest to the sketch of the dataset

- Different criteria than K-means

Sketching in practice

- **Sketching for large scale K-means:**
  - We aim at solving:

$$\min_{\widehat{\pi} \text{ s.t. } \widehat{\pi} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\mathbf{c}_k}} \|\mathbf{s} - \mathcal{A}\widehat{\pi}\|_2$$

  - Find a distribution of K diracs whose sketch is the closest to the sketch of the dataset
  - Different criteria than K-means
- **Reformulation:**

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \|\mathbf{s} - \frac{1}{K}\sum_{k=1}^{K} \Phi(\mathbf{c}_k)\|_2$$

Sketching in practice

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \|\mathbf{s} - \tfrac{1}{K}\textstyle\sum_{k=1}^K \Phi(\mathbf{c}_k)\|_2$$

- **Algorithm:**
  - Inspired from orthogonal matching pursuit (OMP)

Sketching in practice

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \|\mathbf{s} - \tfrac{1}{K}\textstyle\sum_{k=1}^{K} \Phi(\mathbf{c}_k)\|_2$$

- **Algorithm:**
  - Inspired from orthogonal matching pursuit (OMP)

$\Theta \leftarrow \emptyset \quad \mathbf{r} \leftarrow \mathbf{s}$ // Initialize

while $|\Theta| \leq K$:

$\widehat{\mathbf{c}} \in \arg\max_{\mathbf{c}\in\mathbb{R}^d} \left| \left\langle \frac{\Phi(\mathbf{c})}{\|\Phi(\mathbf{c})\|}, \mathbf{r} \right\rangle \right|$ // Find a new atom: minimizes the residuals

$\Theta \leftarrow \Theta \bigcup \{\widehat{\mathbf{c}}\}$ // add it to the support $\Theta = (\mathbf{c}_1, \cdots, \mathbf{c}_{|\Theta|})$

$\widehat{\boldsymbol{\alpha}} \in \arg\min_{\alpha_1,\cdots,\alpha_k} \|\mathbf{s} - \sum_{k=1}^{|\Theta|} \alpha_k \Phi(\mathbf{c}_k)\|^2$ // Adjust weights: least-squares

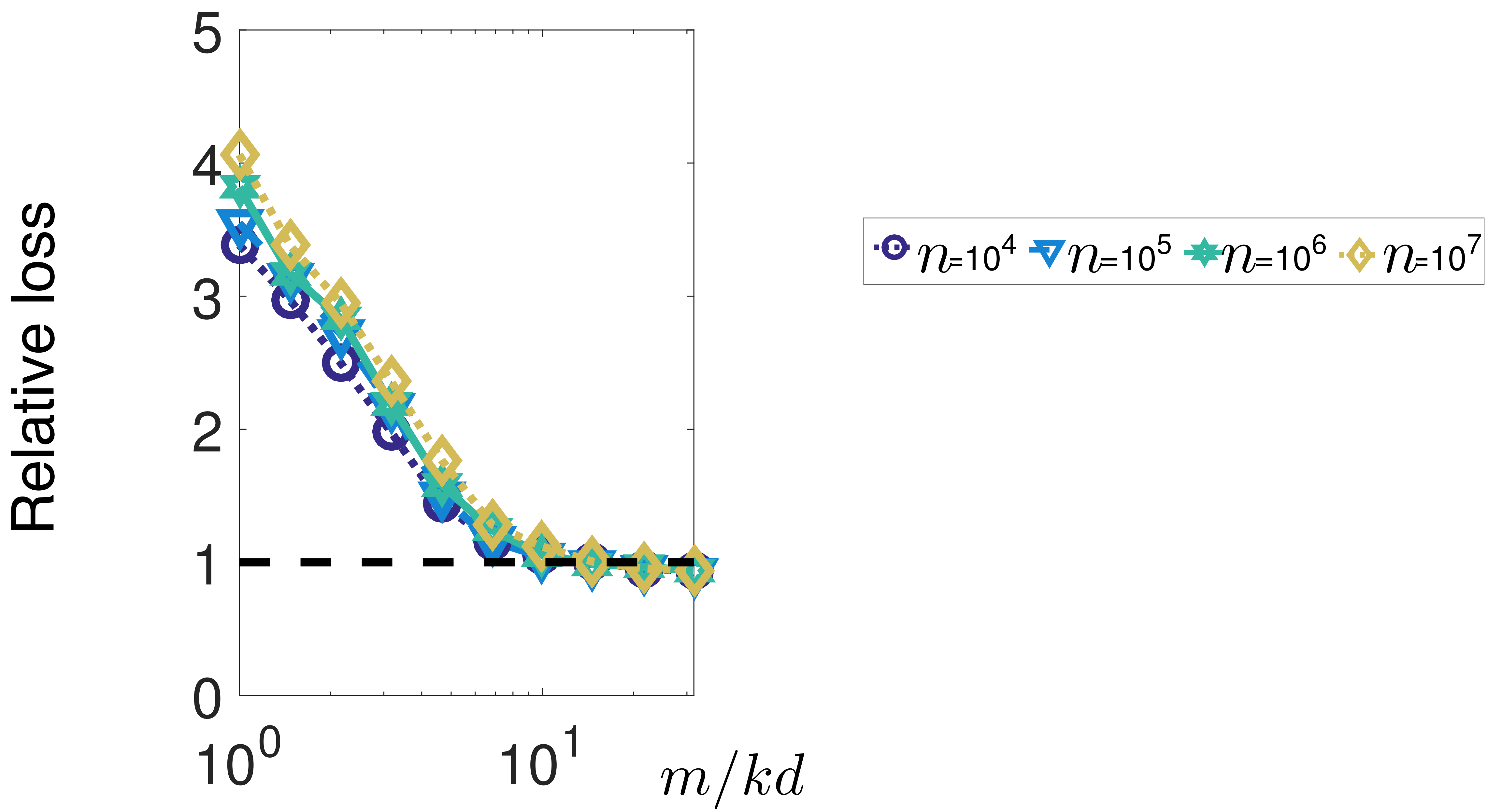
$\mathbf{r} \leftarrow \mathbf{s} - \sum_{k=1}^{|\Theta|} \widehat{\alpha_k} \Phi(\mathbf{c}_k)$ // update residuals

return: $(\mathbf{c}_1, \cdots, \mathbf{c}_{|\Theta|})$

Complexity: $\mathcal{O}(mdK^2)$

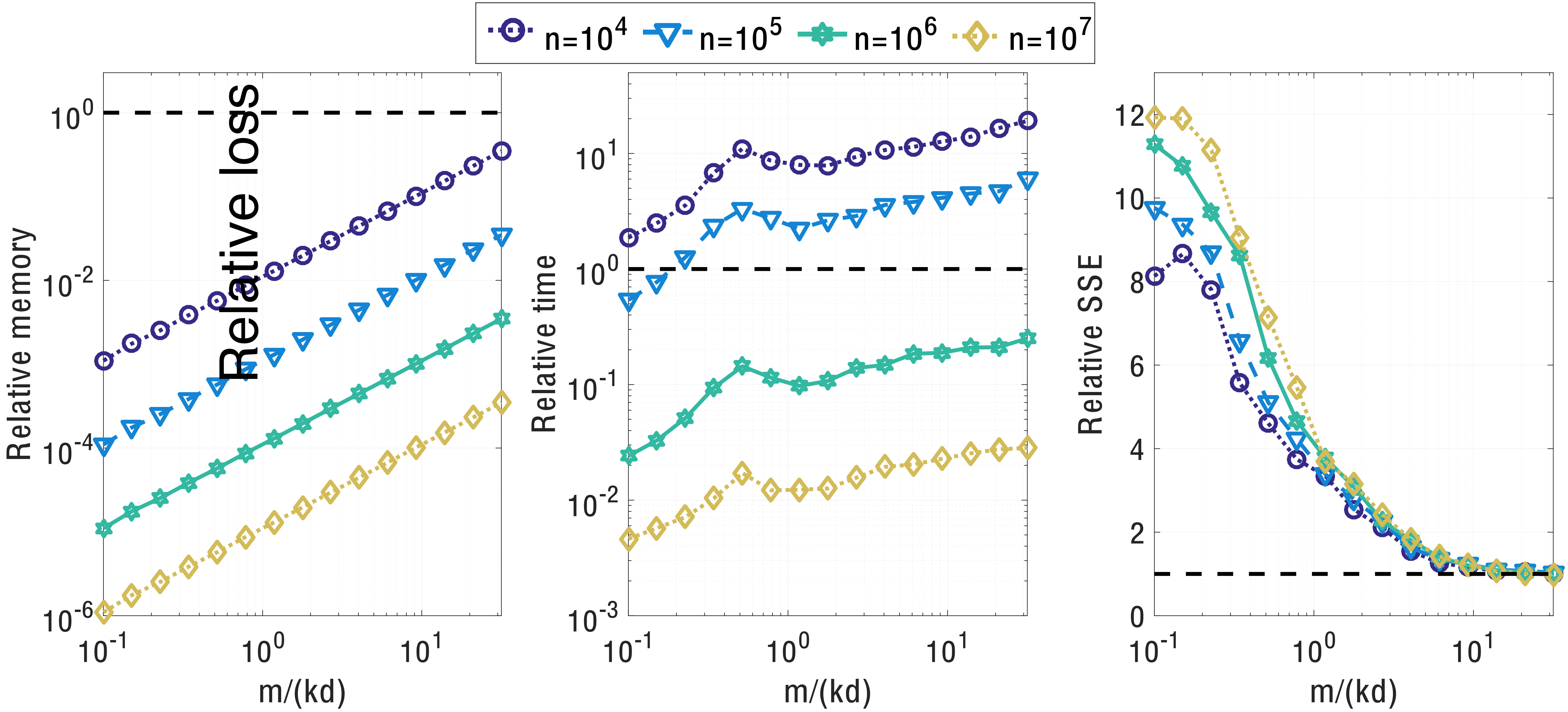
Sketching in practice

- **Results:** **How do we choose m ?** $\boxed{m \approx Kd}$ **number of params ?**
  - Synthetic dataset: $d = 10, K = 10$



Sketching in practice

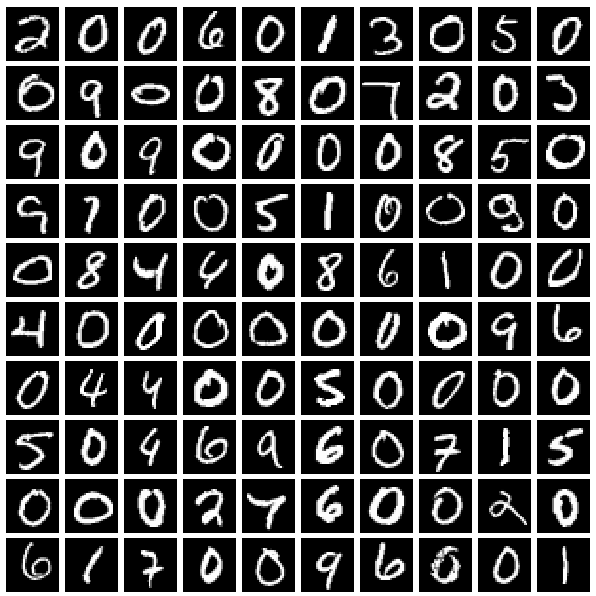
- **Results:** How do we choose m ? $\boxed{m \approx Kd}$ number of params ?
  - Synthetic dataset: $d = 10, K = 10$



Sketching in practice

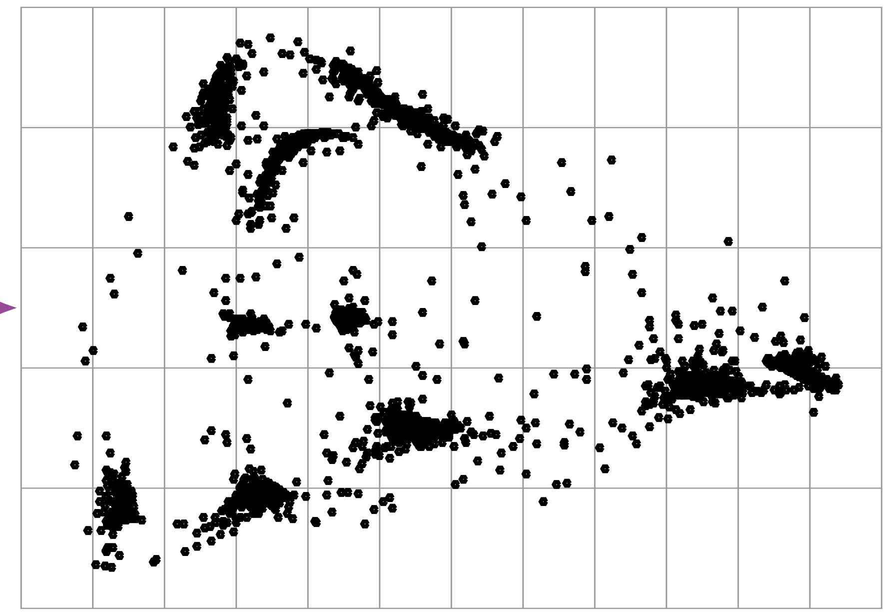
Results for K-means:

Handwritten digits



Pre-processing

Spectral embedding



$n = 70\ 000$

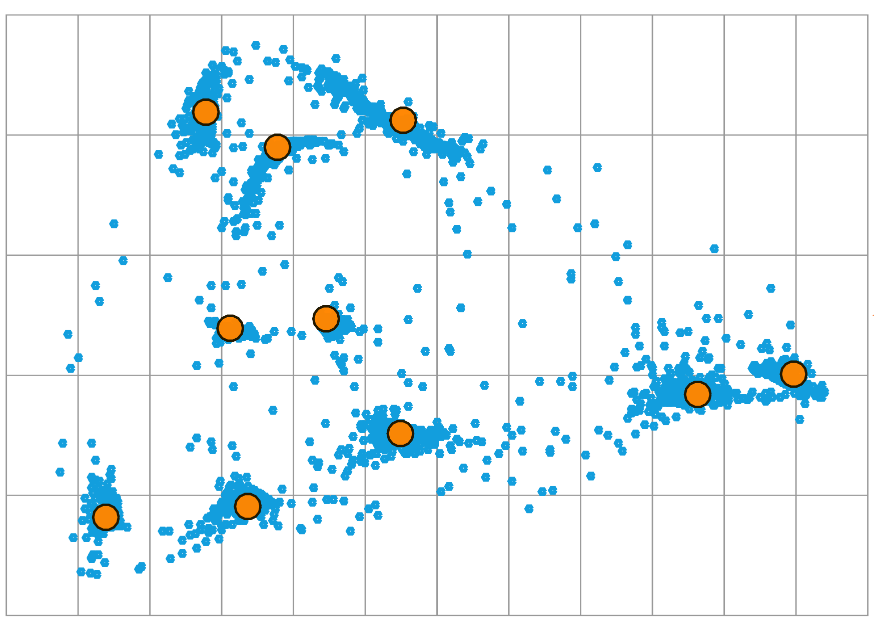
$K = 10$

$\tilde{z}$

Sketch

$\boxed{m \approx Kd}$

Sketched Clustering



Complexity:

$\mathcal{O}(mdK^2)$

- Remember K-means complexity:

  In time: $\mathcal{O}(nKd)$

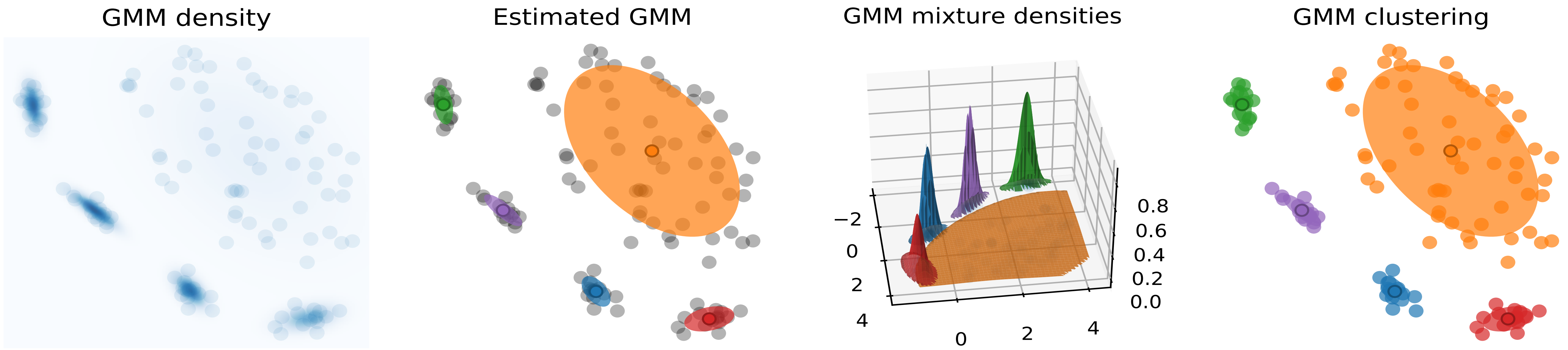
  Memory: $\mathcal{O}(nd)$

Sketching in practice

GMM: $\boxed{\pi_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k \pi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}(\mathbf{x})}$

- **Come back to GMM:**

MLE estimate: $\boxed{\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} -\log(\pi_{\boldsymbol{\theta}}(\mathbf{x}_i))}$



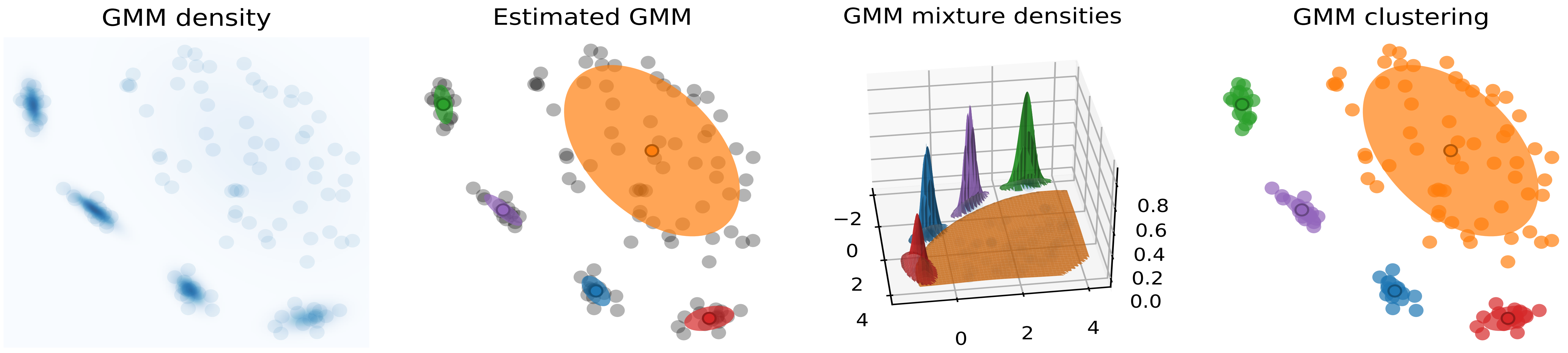
- Find a distribution $\pi_{\boldsymbol{\theta}}$ that bests approximate π

Sketching in practice

GMM: $$\pi_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k \pi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}(\mathbf{x})$$

- **Come back to GMM:**

MLE estimate: $$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} -\log(\pi_{\boldsymbol{\theta}}(\mathbf{x}_i))$$



  - Find a distribution $\pi_{\boldsymbol{\theta}}$ that bests approximate π

- **Same idea than before:**

$$\min_{\boldsymbol{\theta}=\{\alpha_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}} \|\mathbf{s} - \mathcal{A}\pi_{\boldsymbol{\theta}}\|_2$$

Find a GMM whose sketch is the closest to the sketch of the dataset

Sketching in practice

Sketching for large scale GMM:

- A little bit more delicate: need to evaluate $\mathcal{A}\pi_{\boldsymbol{\theta}}$
- Linearity: $\mathcal{A}\pi_{\boldsymbol{\theta}} = \sum_{k=1}^{K} \alpha_k \mathcal{A}(\pi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k})$ GMM
- When $\Phi = \mathrm{RFF}$:

$\mathcal{A}(\pi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k})$ is the Fourier transform of a Gaussian evaluated at m points

But Fourier transform of a Gaussian = Gaussian

-> Just need to sample m points from a Gaussian (easy)

Sketching in practice

- **Sketching for large scale GMM:**
  - A little bit more delicate: need to evaluate $\mathcal{A}\pi_{\boldsymbol{\theta}}$
  - Linearity: $\mathcal{A}\pi_{\boldsymbol{\theta}} = \sum_{k=1}^{K} \alpha_k \mathcal{A}(\pi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k})$ GMM
  - When $\Phi = \mathrm{RFF}$:

$\mathcal{A}(\pi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k})$ is the Fourier transform of a Gaussian evaluated at m points

But Fourier transform of a Gaussian = Gaussian

-> Just need to sample m points from a Gaussian (easy)

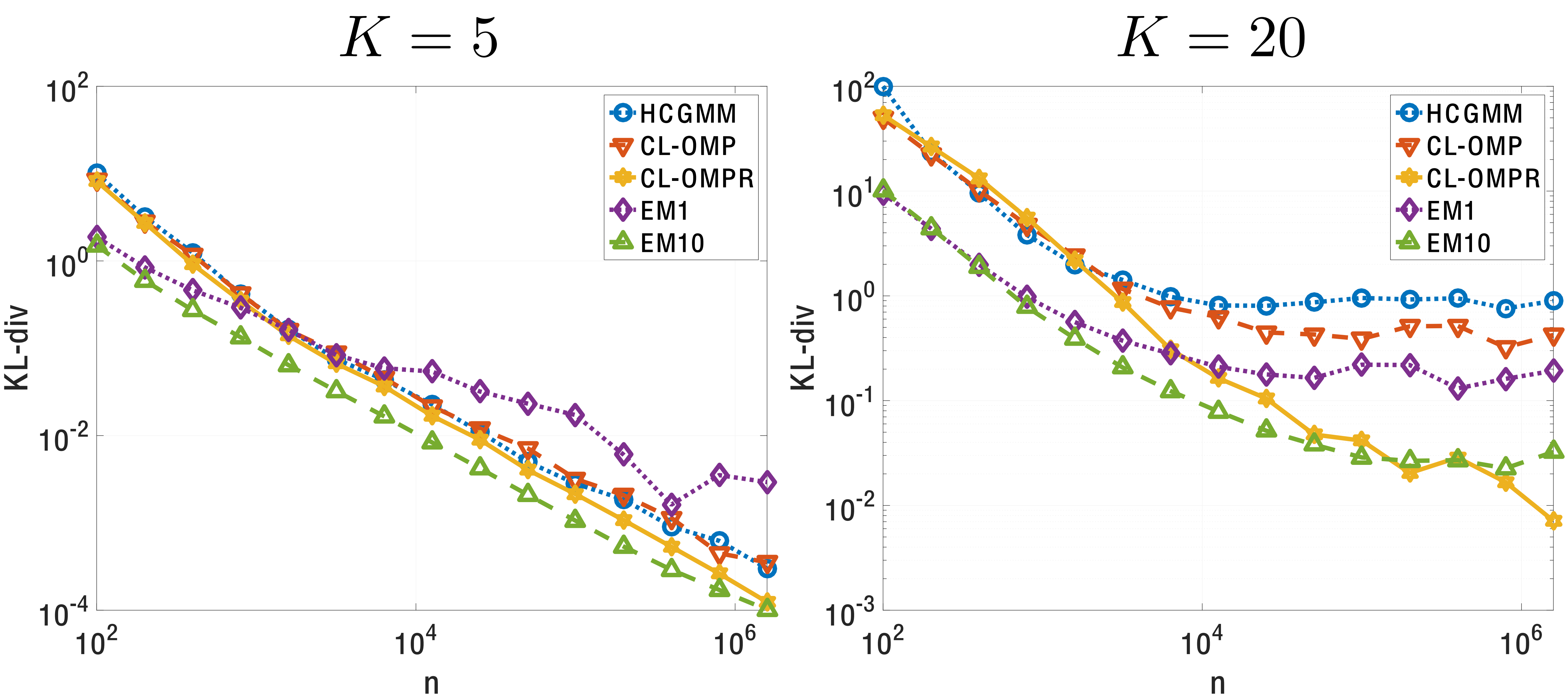
- **Algorithm:**
  - The algorithm of K-means adapts to this setting

$$\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}} \leftarrow \arg\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \left| \left\langle \frac{\mathcal{A}\pi_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}}{\|\mathcal{A}\pi_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}\|}, \mathbf{r} \right\rangle \right|$$

// Find a new atom: minimizes the residuals

Sketching in practice

Results: $d = 10$



Sketching in practice

Learn from sketch:

- Depending on the task, find the suitable set:

$$\boxed{\mathfrak{S}_{\Theta} = \{\pi_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}}$$

K-means: mixture of K diracs

GMM: mixture of K Gaussian

Generative model: distrib. parametrized by NN

- Solve the optimization problem:

Decoding problem:
- CL-OMP
- GD

$$\boxed{\min_{\boldsymbol{\theta} \in \Theta} \|\mathbf{s} - \mathcal{A}\pi_{\boldsymbol{\theta}}\|_2}$$

- Return the best parameter $\widehat{\boldsymbol{\theta}}$ and the distribution $\pi_{\widehat{\boldsymbol{\theta}}} \approx \pi$

Sketching in practice

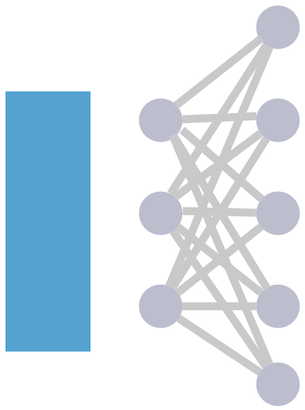
Another toy example:

Generative model:

$$\mathfrak{S}_{\Theta} = \{\frac{1}{p}\sum_{j=1}^{p} \delta_{\mathrm{NN}_{\boldsymbol{\theta}}(\mathbf{z}_j)}; \boldsymbol{\theta} \in \Theta\}$$

Latent space

$$\mathbf{z}_j \sim \Lambda$$



Sketching in practice

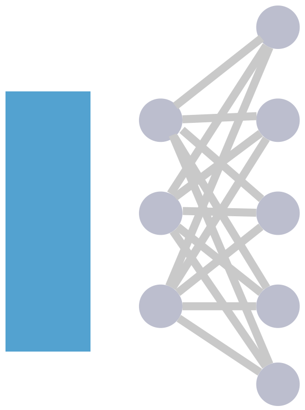
Another toy example:

Generative model:

$$\mathfrak{S}_\Theta = \{\frac{1}{p}\sum_{j=1}^{p}\delta_{\mathrm{NN}_{\boldsymbol{\theta}}(\mathbf{z}_j)}; \boldsymbol{\theta} \in \Theta\}$$

Latent space

$\mathbf{z}_j \sim \Lambda$



Sketching op:

$$\mathcal{A}\pi_{\boldsymbol{\theta}} = \frac{1}{p}\sum_{j=1}^{p}\Phi(\mathrm{NN}_{\boldsymbol{\theta}}(\mathbf{z}_j))$$

Optim.

SGD

Sketching in practice

Another toy example:

Generative model:

$$\mathfrak{S}_\Theta = \{\frac{1}{p}\sum_{j=1}^{p} \delta_{\mathrm{NN}_{\boldsymbol{\theta}}(\mathbf{z}_j)}; \boldsymbol{\theta} \in \Theta\}$$

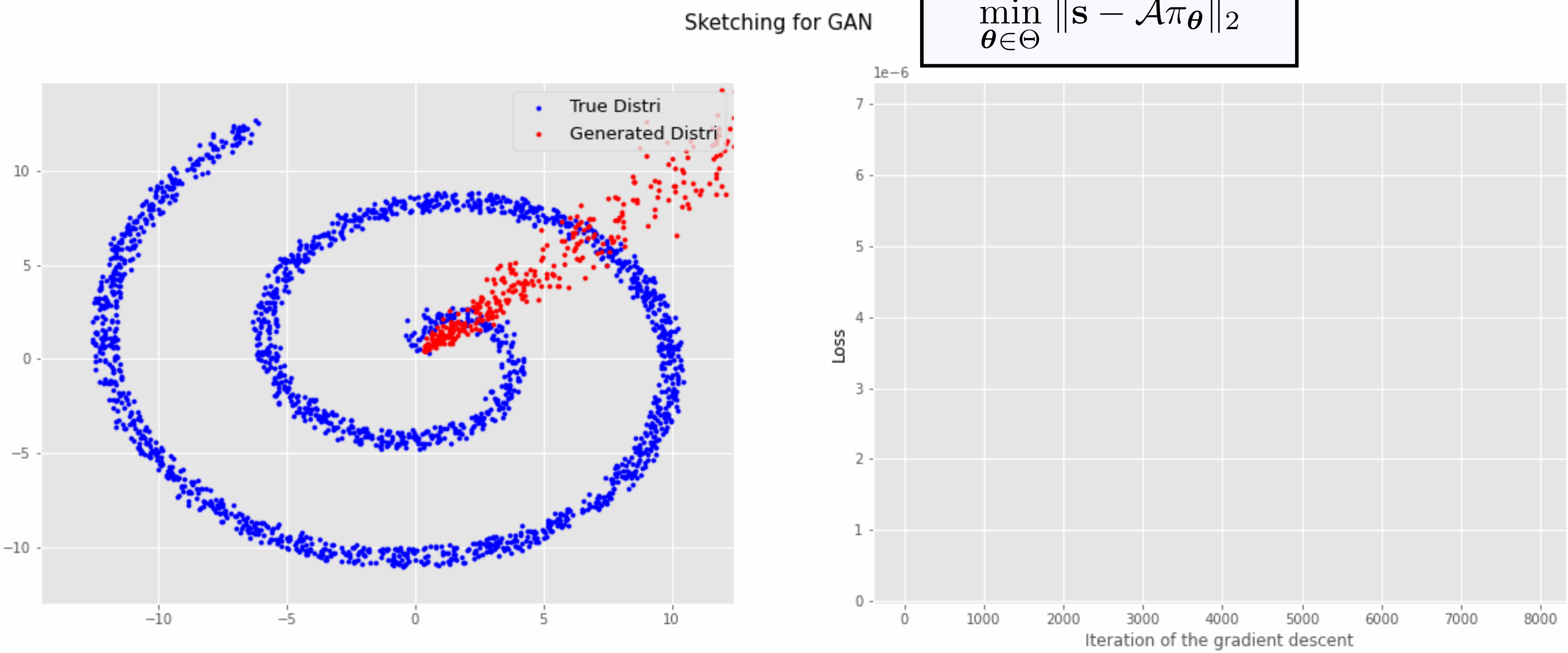
Sketching op:

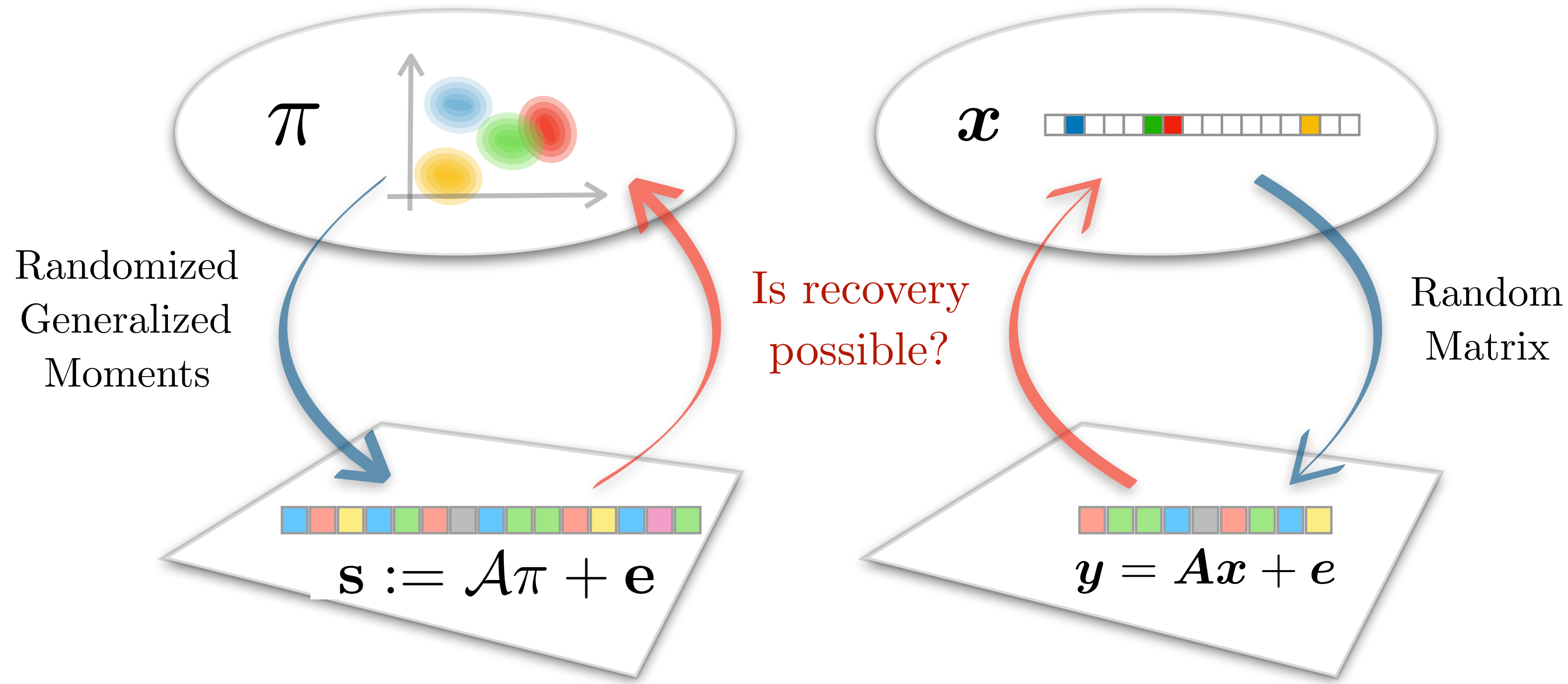
$$\mathcal{A}\pi_{\boldsymbol{\theta}} = \frac{1}{p}\sum_{j=1}^{p} \Phi(\mathrm{NN}_{\boldsymbol{\theta}}(\mathbf{z}_j))$$

Optim.

SGD

$$\boxed{\min_{\boldsymbol{\theta}\in\Theta} \|\mathbf{s} - \mathcal{A}\pi_{\boldsymbol{\theta}}\|_2}$$





Compressive Learning

- Theory of sketching
- Sketching in practice
- Theoretical guarantees
- Limitations & perspectives

Theoretical guarantees

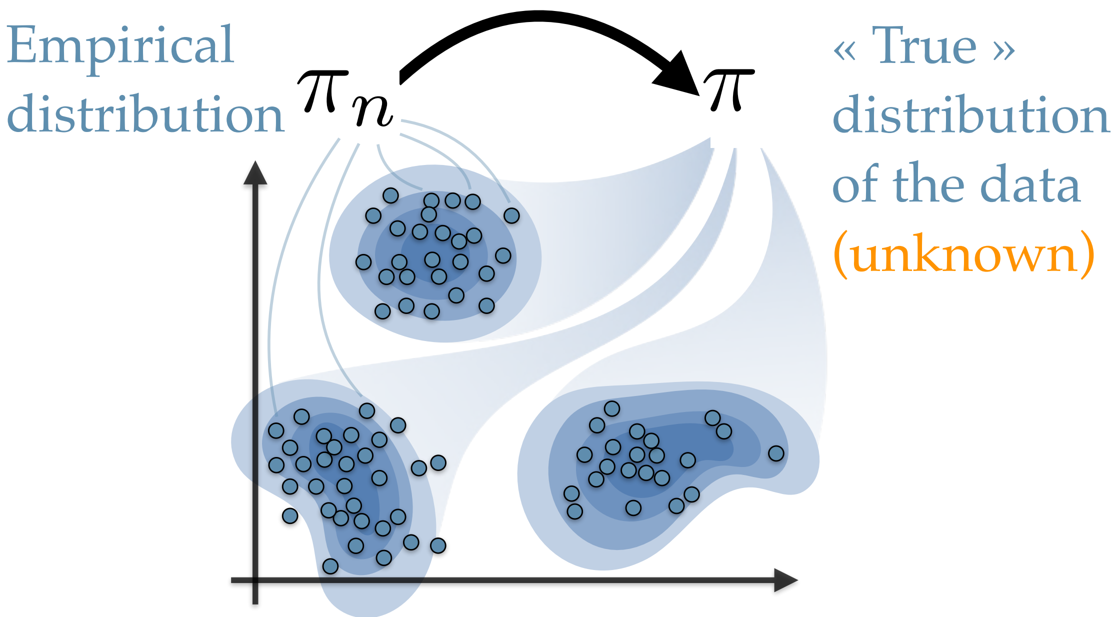
- **Analogy with compressed sensing**

Theoretical guarantees

Sketching operator

$$\mathcal{A}\pi = \mathop{\mathbb{E}}_{\mathbf{x}\sim\pi}[\Phi(\mathbf{x})]$$

Analogy with compressed sensing



We observe the sketch

$$\mathbf{s} = \mathcal{A}\pi_n \in \mathbb{R}^m$$

Theoretical guarantees

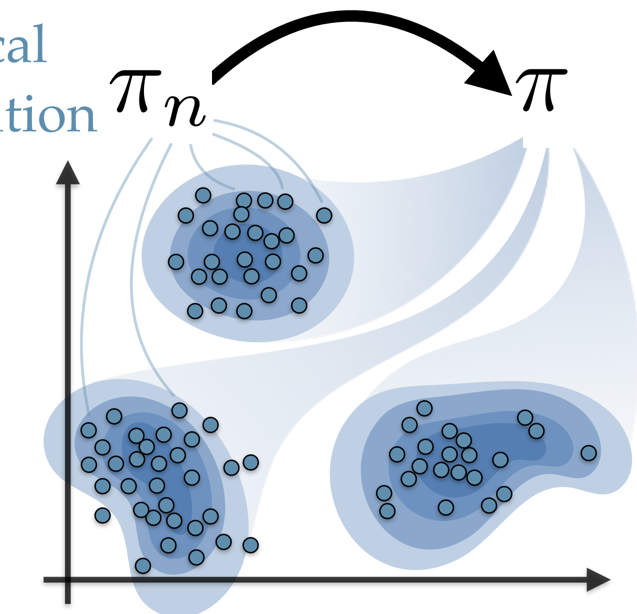
Sketching operator

$$\boxed{\mathcal{A}\pi = \mathbb{E}_{\mathbf{x}\sim\pi}[\Phi(\mathbf{x})]}$$

Analogy with compressed sensing

Empirical distribution π_n

« True » distribution of the data (unknown) π



We observe the sketch

$$\mathbf{s} = \mathcal{A}\pi_n \in \mathbb{R}^m$$

$$= \mathcal{A}\pi + \mathbf{e}$$

where $\mathbf{e} := \mathcal{A}(\pi_n - \pi)$

noise

- Noisy, linear and finite-dimensional measurements of a distribution

IN CS

Underdetermined $m < d$

$$\mathbf{y} = \mathbf{A}\,\mathbf{x} + \mathbf{e}$$

IN Sketching

Very underdetermined $m < +\infty$

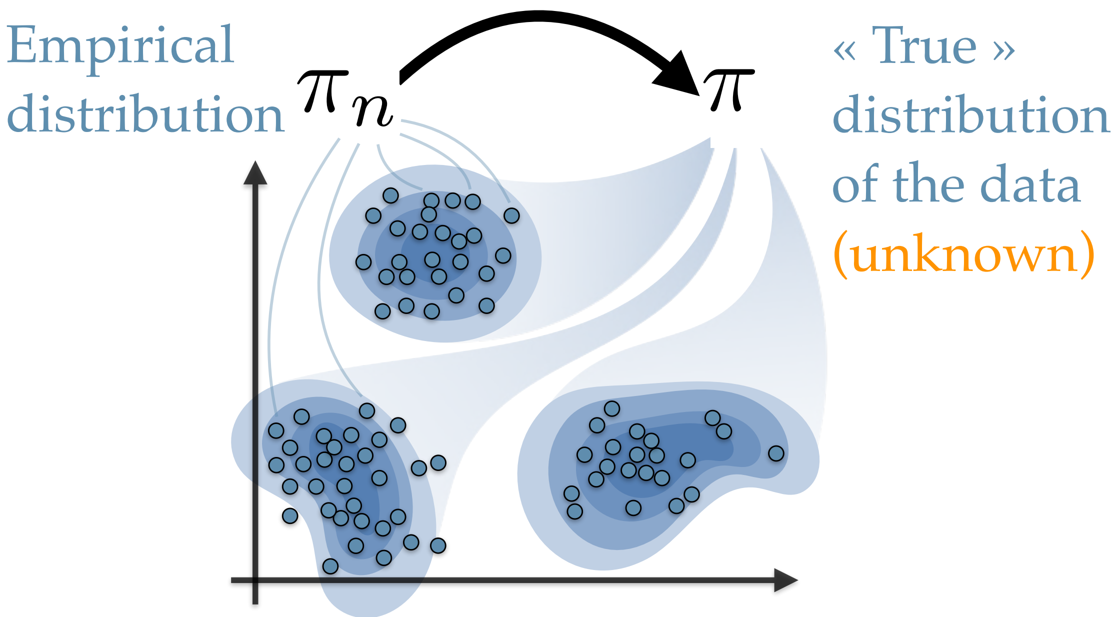
$$\mathbf{s} = \mathcal{A}\,\pi + \mathbf{e}$$

Theoretical guarantees

Sketching operator

$$\boxed{\mathcal{A}\pi = \mathop{\mathbb{E}}_{\mathbf{x}\sim\pi}[\Phi(\mathbf{x})]}$$

Analogy with compressed sensing



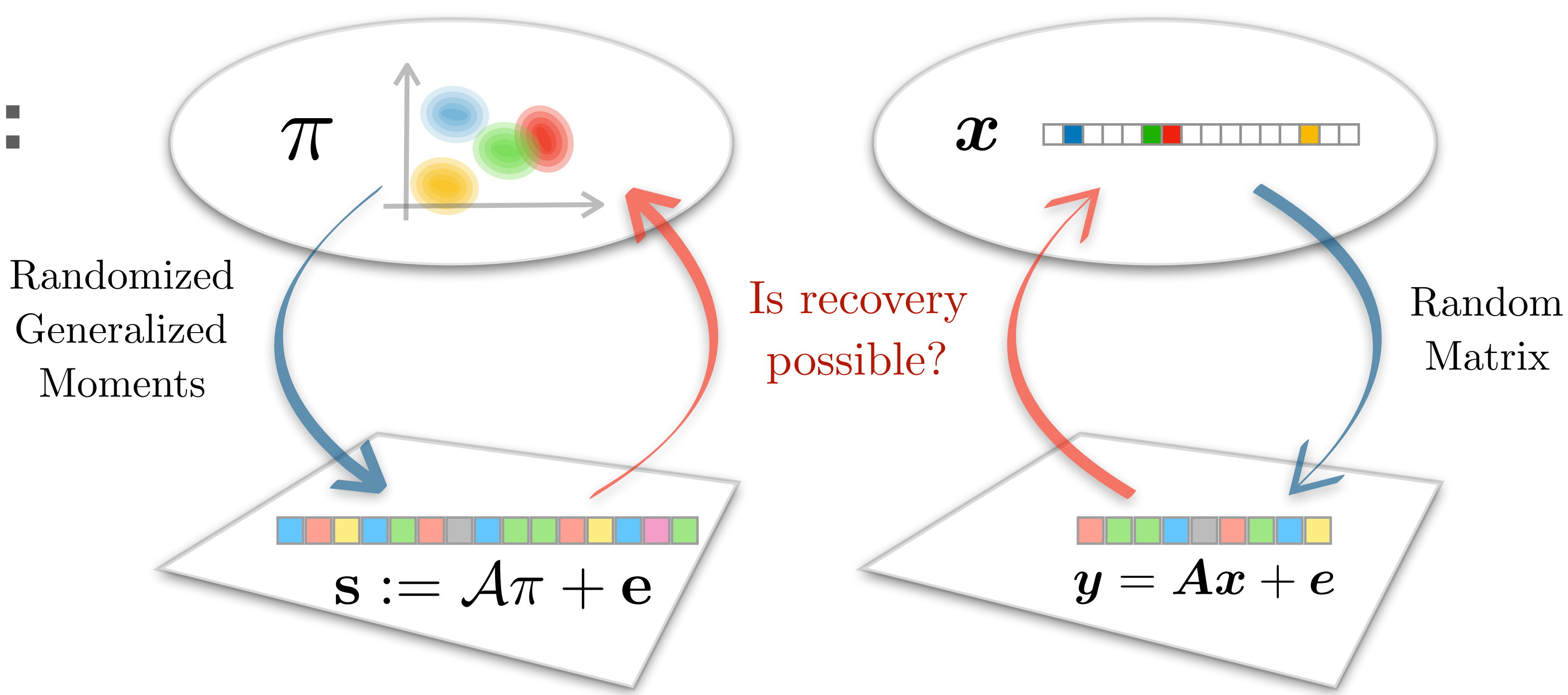
We observe the sketch

$$\boxed{\begin{aligned} \mathbf{s} &= \mathcal{A}\pi_n \in \mathbb{R}^m \\ &= \mathcal{A}\pi + \mathbf{e} \\ \text{where } \mathbf{e} &:= \mathcal{A}(\pi_n - \pi) \end{aligned}}$$

noise

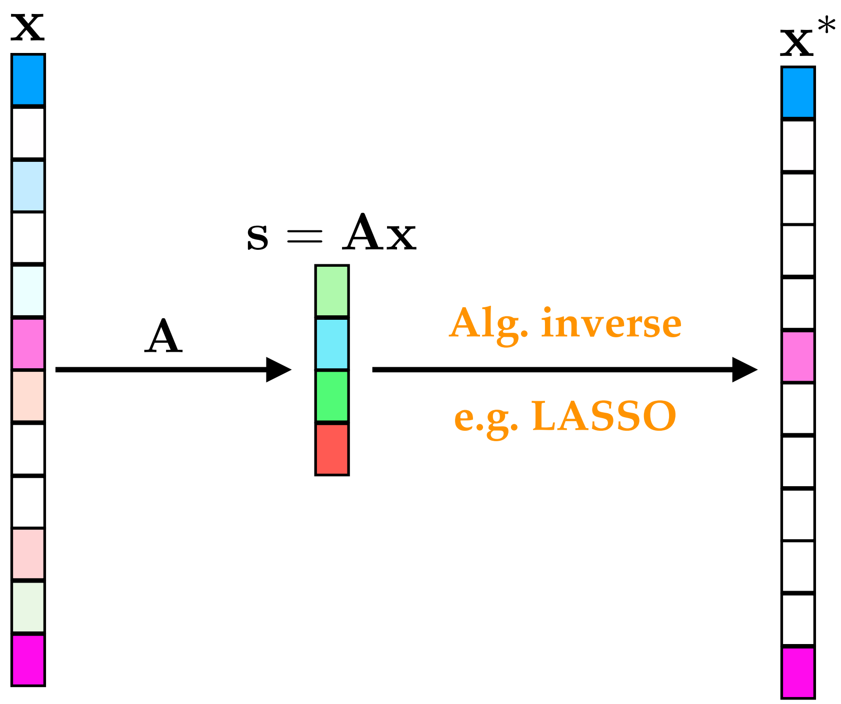
- Noisy, linear and finite-dimensional measurements of a distribution

Question:

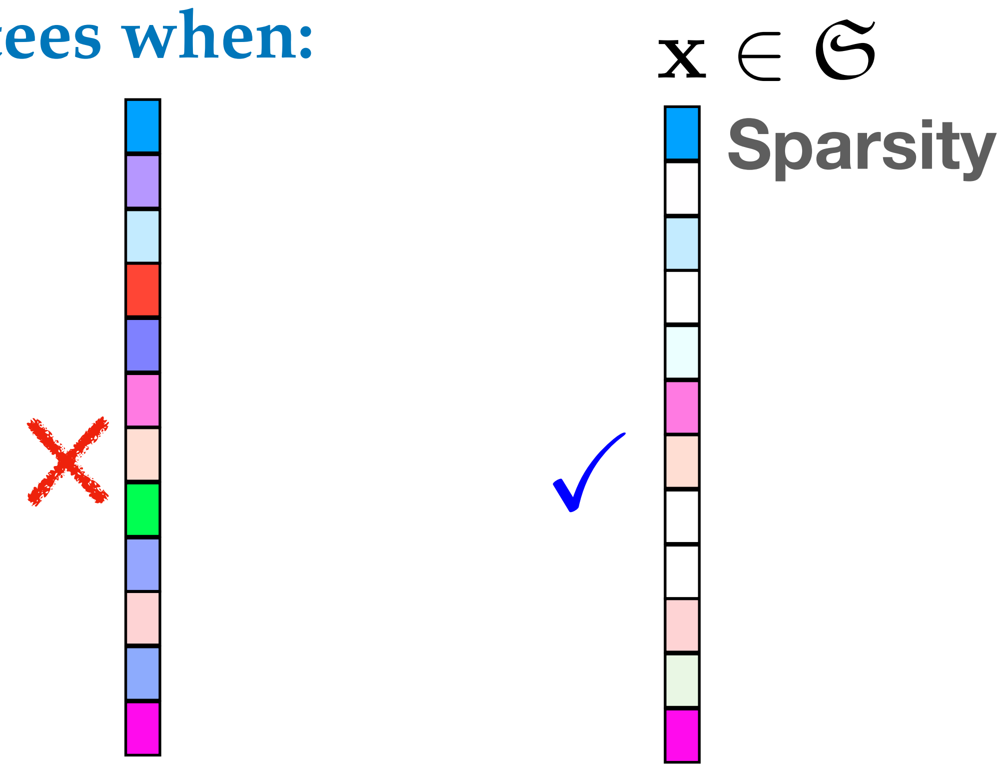


Theoretical guarantees

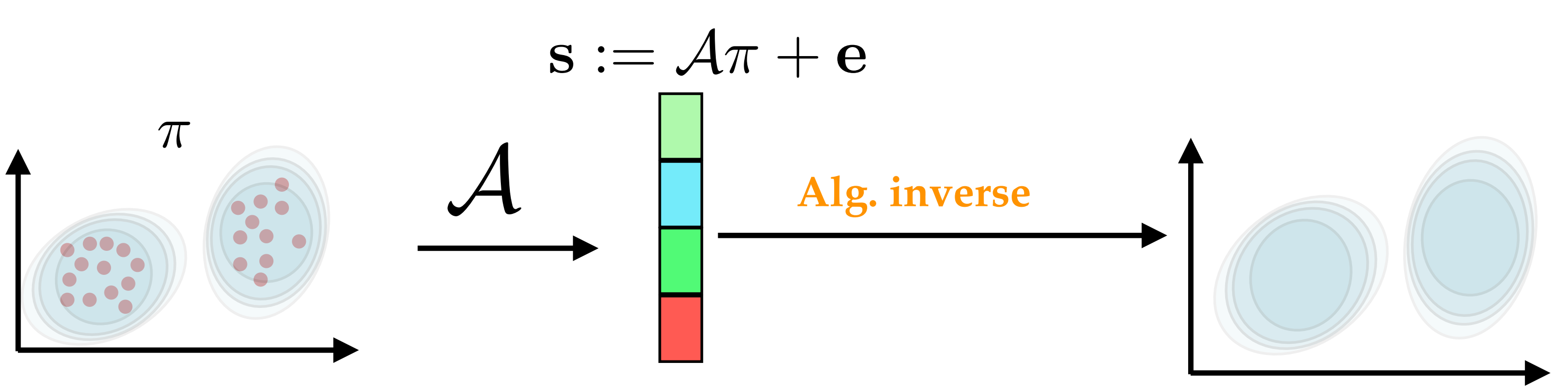
Analogy with compressed sensing

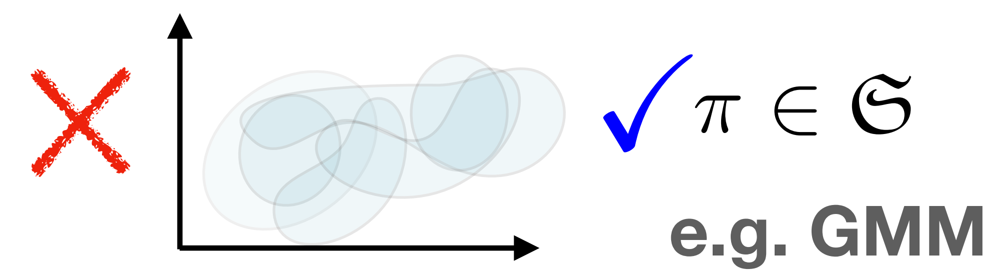


Guarantees when:



…Need a « low-dimensional » distribution

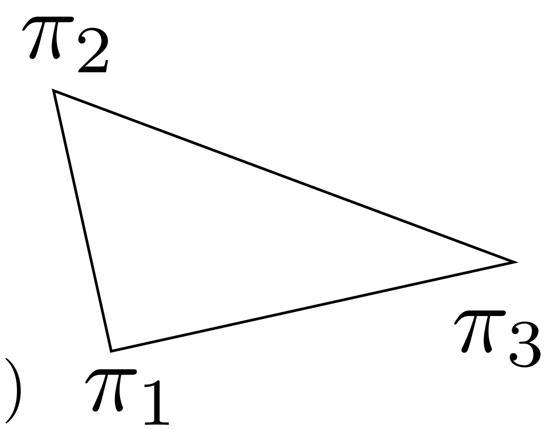




Theoretical guarantees

- **Metric between distributions:**

  - Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric



1) $D(\pi, \pi') = 0 \iff \pi = \pi'$

2) $D(\pi, \pi') = D(\pi', \pi)$

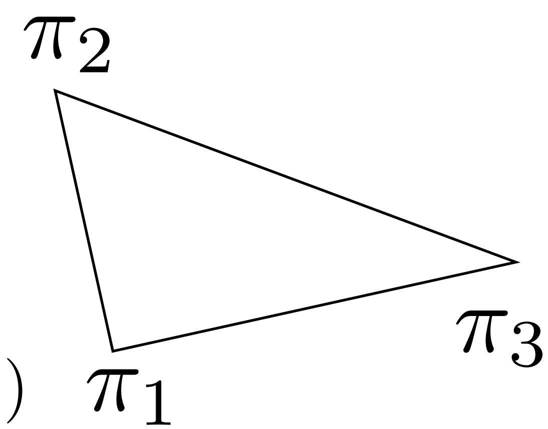
3) $D(\pi_1, \pi_2) \leq D(\pi_1, \pi_3) + D(\pi_3, \pi_2)$

Quantifies the distance between distrib.

Theoretical guarantees

Metric between distributions:

- Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric



1) $D(\pi, \pi') = 0 \iff \pi = \pi'$ **3)** $D(\pi_1, \pi_2) \leq D(\pi_1, \pi_3) + D(\pi_3, \pi_2)$

2) $D(\pi, \pi') = D(\pi', \pi)$

Quantifies the distance between distrib.

- Numerous examples in the literature…

Total Variation

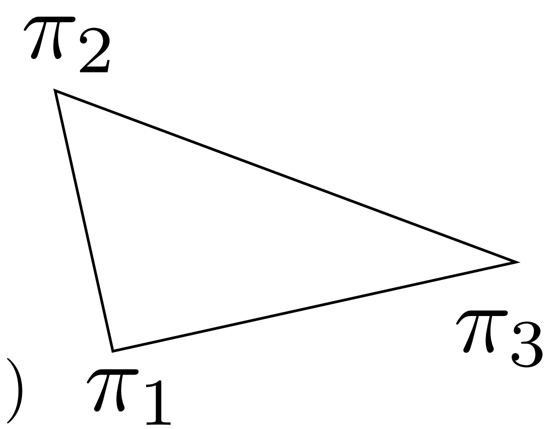
$$D(\pi, \pi') = \|\pi - \pi'\|_{\mathrm{TV}}$$

if densities f,g: $= \frac{1}{2}\|f - g\|_{L_1}$

Theoretical guarantees

Metric between distributions:

- Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric



1) $D(\pi, \pi') = 0 \iff \pi = \pi'$ **3)** $D(\pi_1, \pi_2) \leq D(\pi_1, \pi_3) + D(\pi_3, \pi_2)$

2) $D(\pi, \pi') = D(\pi', \pi)$

Quantifies the distance between distrib.

- Numerous examples in the literature…

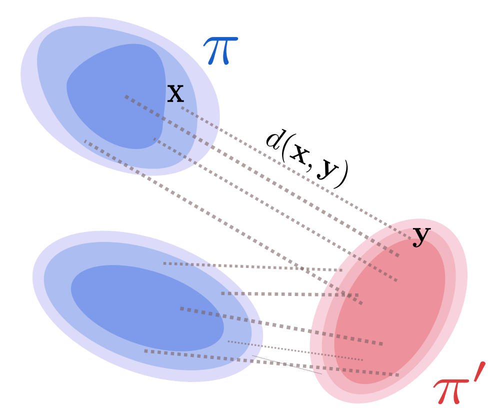
Total Variation

$D(\pi, \pi') = \|\pi - \pi'\|_{\mathrm{TV}}$

if densities f,g: $= \frac{1}{2}\|f - g\|_{L_1}$

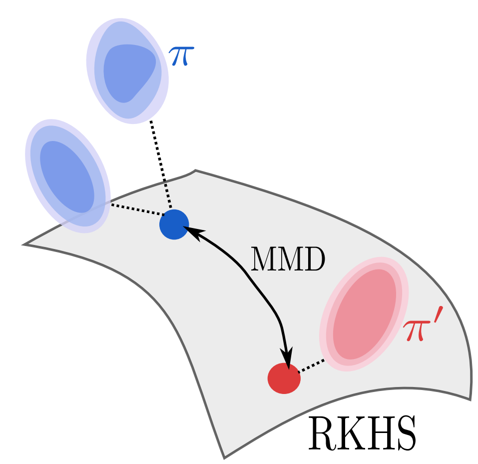
Optimal Transport

$D(\pi, \pi') = W_p(\pi, \pi')$



MMD

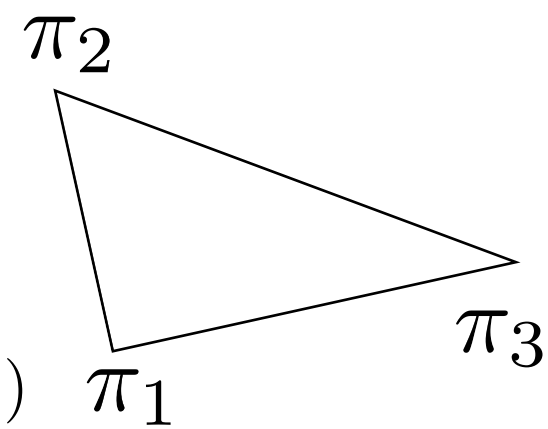
$D(\pi, \pi') = \|\pi - \pi'\|_\kappa$



Theoretical guarantees

Metric between distributions:

- Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric



1) $D(\pi, \pi') = 0 \iff \pi = \pi'$ **3)** $D(\pi_1, \pi_2) \leq D(\pi_1, \pi_3) + D(\pi_3, \pi_2)$

2) $D(\pi, \pi') = D(\pi', \pi)$

Quantifies the distance between distrib.

- Numerous examples in the literature…

Total Variation

$$D(\pi, \pi') = \|\pi - \pi'\|_{\mathrm{TV}}$$

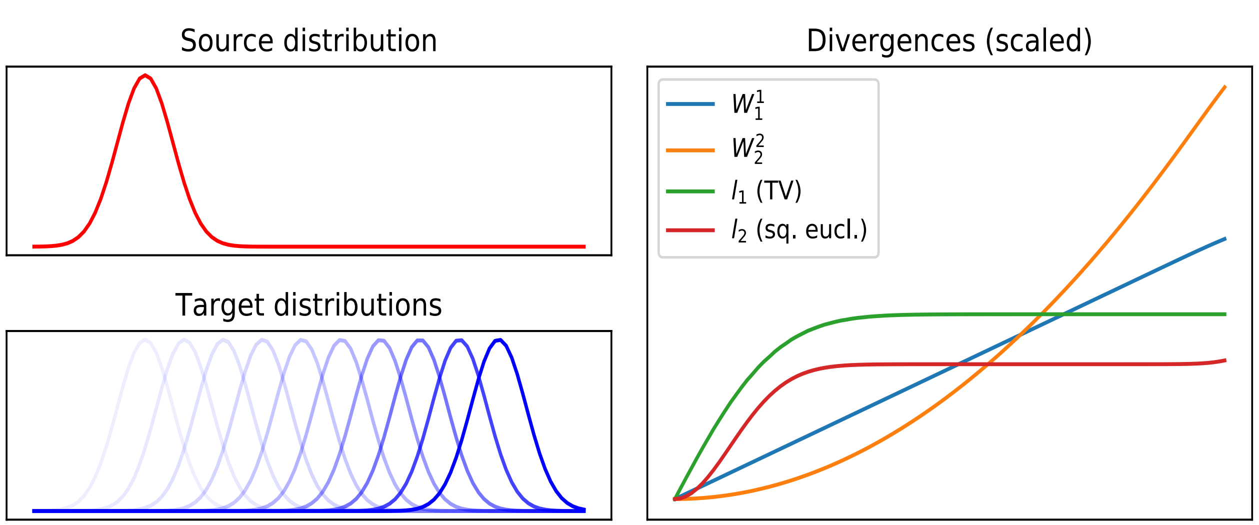
if densities f,g: $= \frac{1}{2}\|f - g\|_{L_1}$

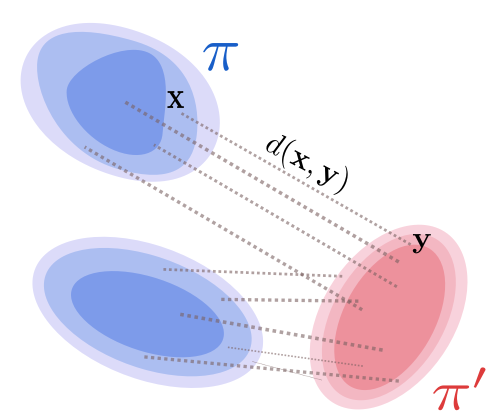
Optimal Transport

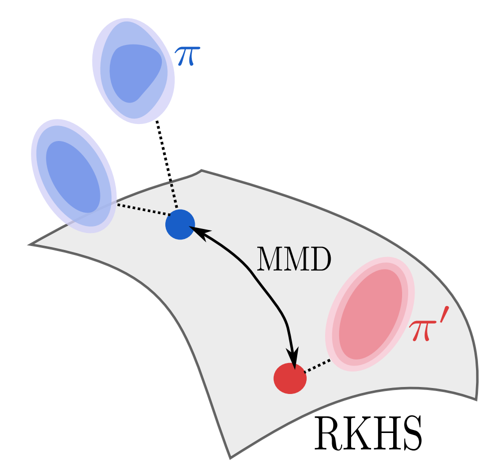
$$D(\pi, \pi') = W_p(\pi, \pi')$$

MMD

$$D(\pi, \pi') = \|\pi - \pi'\|_{\kappa}$$







Theoretical guarantees

The lower RIP:

- Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric
- The model set related to the task $\mathfrak{S}_\Theta = \{\pi_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$

Lower RIP

$$\boxed{\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta,\ D(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq C\|\mathcal{A}\pi_{\boldsymbol{\theta}} - \mathcal{A}\pi_{\boldsymbol{\theta}'}\|_2}$$

Theoretical guarantees

The lower RIP:

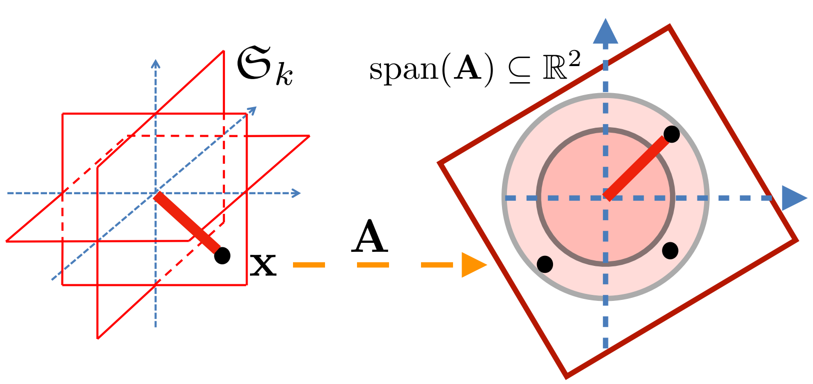
- Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric
- The model set related to the task $\mathfrak{S}_\Theta = \{\pi_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$

Lower RIP

$$\boxed{\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta, \; D(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq C \| \mathcal{A}\pi_{\boldsymbol{\theta}} - \mathcal{A}\pi_{\boldsymbol{\theta}'} \|_2}$$

Connection with the RIP:

$$\boxed{\begin{array}{l} \exists \delta_k \in [0,1[\;\; \forall \mathbf{x} \; k\text{-sparse} \\ (1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2 \end{array}}$$



E.g. $D(\pi, \pi') = \|\pi - \pi'\|_{\mathrm{TV}}$

$$\boxed{\begin{array}{l} \exists C > 0 \;\; \forall x = \pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'} \\ \qquad \|x\|_{\mathrm{TV}} \leq C \|\mathcal{A}x\|_2 \end{array}}$$

same kind of idea

Theoretical guarantees

The lower RIP:

- Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric
- The model set related to the task $\mathfrak{S}_\Theta = \{\pi_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$

Lower RIP

$$\boxed{\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta, \; D(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq C \|\mathcal{A}\pi_{\boldsymbol{\theta}} - \mathcal{A}\pi_{\boldsymbol{\theta}'}\|_2}$$

Suppose Lower LRIP holds:

Take $\pi \in \mathcal{P}(\mathbb{R}^d)$ emp. distrib. π_n

Sketch $\mathbf{s} = \mathcal{A}\pi_n \in \mathbb{R}^m$

Solve:

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \|\mathbf{s} - \mathcal{A}\pi_{\boldsymbol{\theta}}\|_2$$

Theoretical guarantees

The lower RIP:

- Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric
- The model set related to the task $\mathfrak{S}_\Theta = \{\pi_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$

Lower RIP

$$\boxed{\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta,\ D(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq C\|\mathcal{A}\pi_{\boldsymbol{\theta}} - \mathcal{A}\pi_{\boldsymbol{\theta}'}\|_2}$$

Suppose Lower LRIP holds:

Take $\pi \in \mathcal{P}(\mathbb{R}^d)$ emp. distrib. π_n

Sketch $\mathbf{s} = \mathcal{A}\pi_n \in \mathbb{R}^m$

Solve:

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \|\mathbf{s} - \mathcal{A}\pi_{\boldsymbol{\theta}}\|_2$$

$$\Longrightarrow \boxed{D(\pi_{\widehat{\boldsymbol{\theta}}}, \pi) \leq d^\circ(\pi, \mathfrak{S}_\Theta) + 2C\|\mathcal{A}\pi - \mathcal{A}\pi_n\|_2}$$

Theoretical guarantees

The lower RIP:

- Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric
- The model set related to the task $\mathfrak{S}_\Theta = \{\pi_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$

Lower RIP

$$\boxed{\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta,\ D(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq C\|\mathcal{A}\pi_{\boldsymbol{\theta}} - \mathcal{A}\pi_{\boldsymbol{\theta}'}\|_2}$$

Suppose Lower LRIP holds:

Take $\pi \in \mathcal{P}(\mathbb{R}^d)$ emp. distrib. π_n

Sketch $\mathbf{s} = \mathcal{A}\pi_n \in \mathbb{R}^m$

Solve:

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \|\mathbf{s} - \mathcal{A}\pi_{\boldsymbol{\theta}}\|_2$$

$$\Longrightarrow \boxed{D(\pi_{\widehat{\boldsymbol{\theta}}}, \pi) \leq d^\circ(\pi, \mathfrak{S}_\Theta) + 2C\|\mathcal{A}\pi - \mathcal{A}\pi_n\|_2}$$

distance between the true distrib. and the estimated one

Theoretical guarantees

- **The lower RIP:**
  - Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric
  - The model set related to the task $\mathfrak{S}_\Theta = \{\pi_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$

Lower RIP

$$\boxed{\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta,\ D(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq C \|\mathcal{A}\pi_{\boldsymbol{\theta}} - \mathcal{A}\pi_{\boldsymbol{\theta}'}\|_2}$$

Suppose Lower LRIP holds:

Take $\pi \in \mathcal{P}(\mathbb{R}^d)$ emp. distrib. π_n

Sketch $\mathbf{s} = \mathcal{A}\pi_n \in \mathbb{R}^m$

Solve:

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \|\mathbf{s} - \mathcal{A}\pi_{\boldsymbol{\theta}}\|_2$$

$$\Longrightarrow \boxed{D(\pi_{\widehat{\boldsymbol{\theta}}}, \pi) \leq d^\circ(\pi, \mathfrak{S}_\Theta) \ldots}$$

How far is the true distrib. from the model = approx. error

Theoretical guarantees

The lower RIP:

- Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric
- The model set related to the task $\mathfrak{S}_\Theta = \{\pi_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$

Lower RIP

$$\boxed{\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta,\ D(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq C \|\mathcal{A}\pi_{\boldsymbol{\theta}} - \mathcal{A}\pi_{\boldsymbol{\theta}'}\|_2}$$

Suppose Lower LRIP holds:

Take $\pi \in \mathcal{P}(\mathbb{R}^d)$ emp. distrib. π_n

Sketch $\mathbf{s} = \mathcal{A}\pi_n \in \mathbb{R}^m$

Solve:

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \|\mathbf{s} - \mathcal{A}\pi_{\boldsymbol{\theta}}\|_2$$

$$\Longrightarrow \boxed{D(\pi_{\widehat{\boldsymbol{\theta}}}, \pi) \leq d^\circ(\pi, \mathfrak{S}_\Theta) + 2C\|\mathcal{A}\pi - \mathcal{A}\pi_n\|_2}$$

Error term $\mathcal{O}(n^{-1/2})$

Theoretical guarantees

The lower RIP:

- Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric
- The model set related to the task $\mathfrak{S}_\Theta = \{\pi_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$

Lower RIP

$$\boxed{\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta,\ D(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq C \|\mathcal{A}\pi_{\boldsymbol{\theta}} - \mathcal{A}\pi_{\boldsymbol{\theta}'}\|_2}$$

Suppose Lower LRIP holds:

Take $\pi \in \mathcal{P}(\mathbb{R}^d)$ emp. distrib. π_n

Sketch $\mathbf{s} = \mathcal{A}\pi_n \in \mathbb{R}^m$

Solve:

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \|\mathbf{s} - \mathcal{A}\pi_{\boldsymbol{\theta}}\|_2$$

$$\Longrightarrow \boxed{D(\pi_{\widehat{\boldsymbol{\theta}}}, \pi) \leq d^\circ(\pi, \mathfrak{S}_\Theta) + 2C\|\mathcal{A}\pi - \mathcal{A}\pi_n\|_2}$$

$\Longrightarrow$ If $\pi \in \mathfrak{S}_\Theta$ then $D(\pi_{\widehat{\boldsymbol{\theta}}}, \pi) \underset{n \to +\infty}{\to} 0$

Theoretical guarantees

- **The lower RIP:**
  - Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric
  - The model set related to the task $\mathfrak{S}_\Theta = \{\pi_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$

Lower RIP

$$\boxed{\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta,\ D(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq C \|\mathcal{A}\pi_{\boldsymbol{\theta}} - \mathcal{A}\pi_{\boldsymbol{\theta}'}\|_2}$$

- **We want the LRIP:**
  - The Lower RIP implies interesting theoretical guarantees
  - However difficult to prove in general ….

Theoretical guarantees

The lower RIP:

- Let $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ be a metric
- The model set related to the task $\mathfrak{S}_\Theta = \{\pi_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$

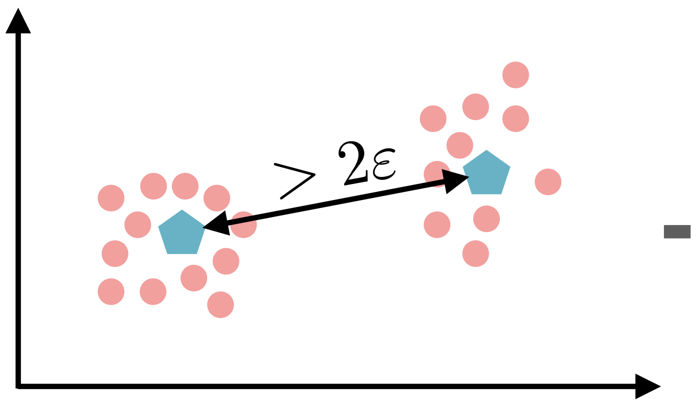
Lower RIP

$$\boxed{\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta,\ D(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \le C \|\mathcal{A}\pi_{\boldsymbol{\theta}} - \mathcal{A}\pi_{\boldsymbol{\theta}'}\|_2}$$

We want the LRIP:

- The Lower RIP implies interesting theoretical guarantees
- However difficult to prove in general …. **with separability of the clusters**

K-means

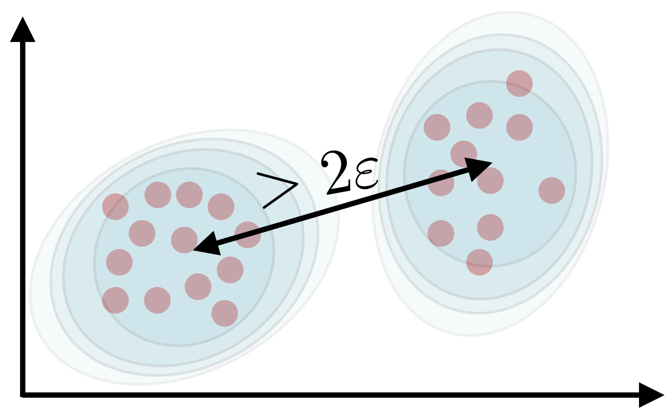


Lower RIP with:

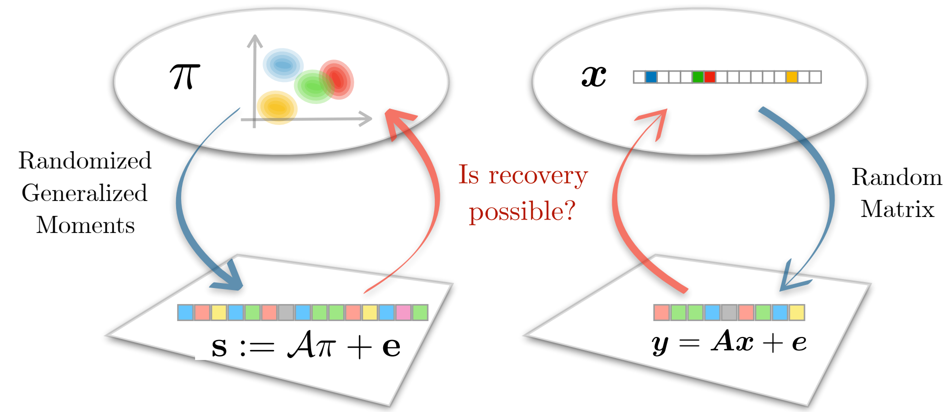
$$\boxed{m = \mathcal{O}(K^2 d)}$$

(and $\Phi = \text{RFF}$)

GMM

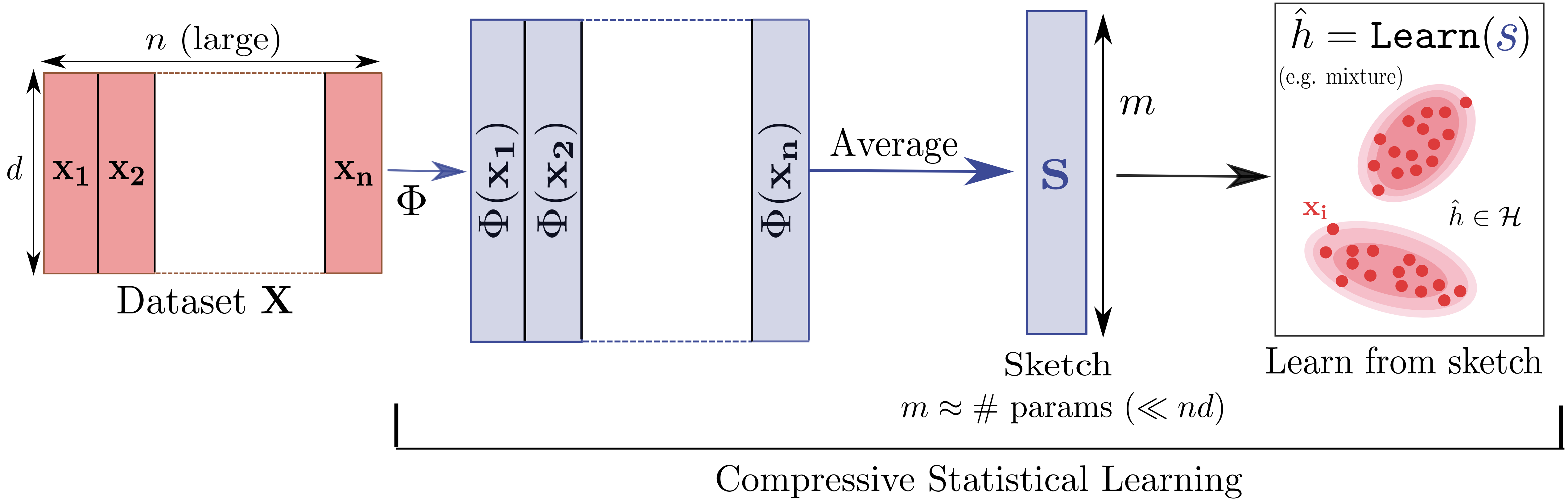


Conclusion

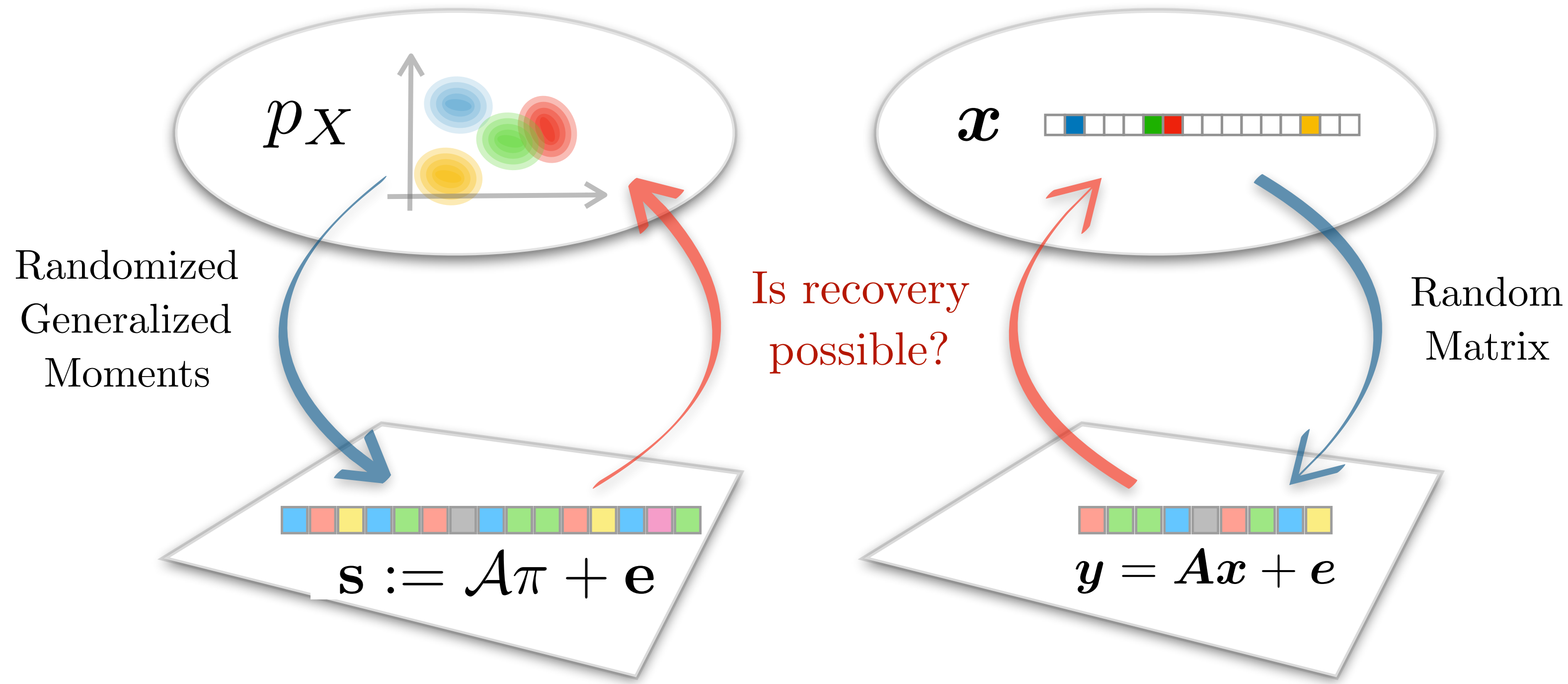


Sketching theory

- Method for ressource efficient large-scale machine learning



- Comes with statistical guarantees inspired by compressed sensing
- Useful for online, distributed and private learning



Compressive Learning

- Theory of sketching
- Sketching in practice
- Theoretical guarantees
- Limitations & perspectives

Limitations & perspectives

Complexity in space

- Sketching reduces drastically the dimension but….

Need to store somewhere:

$$\mathbf{W} \in \mathbb{R}^{m \times d}$$

dense random matrix …

Need to calculate the sketch: $\mathcal{O}(nmd)$

$$\mathbf{s} = \frac{1}{n} \sum_{i=1}^{n} \exp(-i\mathbf{W}\mathbf{x}_i)$$

Limitations & perspectives

Complexity in space

- Sketching reduces drastically the dimension but….

Need to store somewhere:

$$\mathbf{W} \in \mathbb{R}^{m \times d}$$

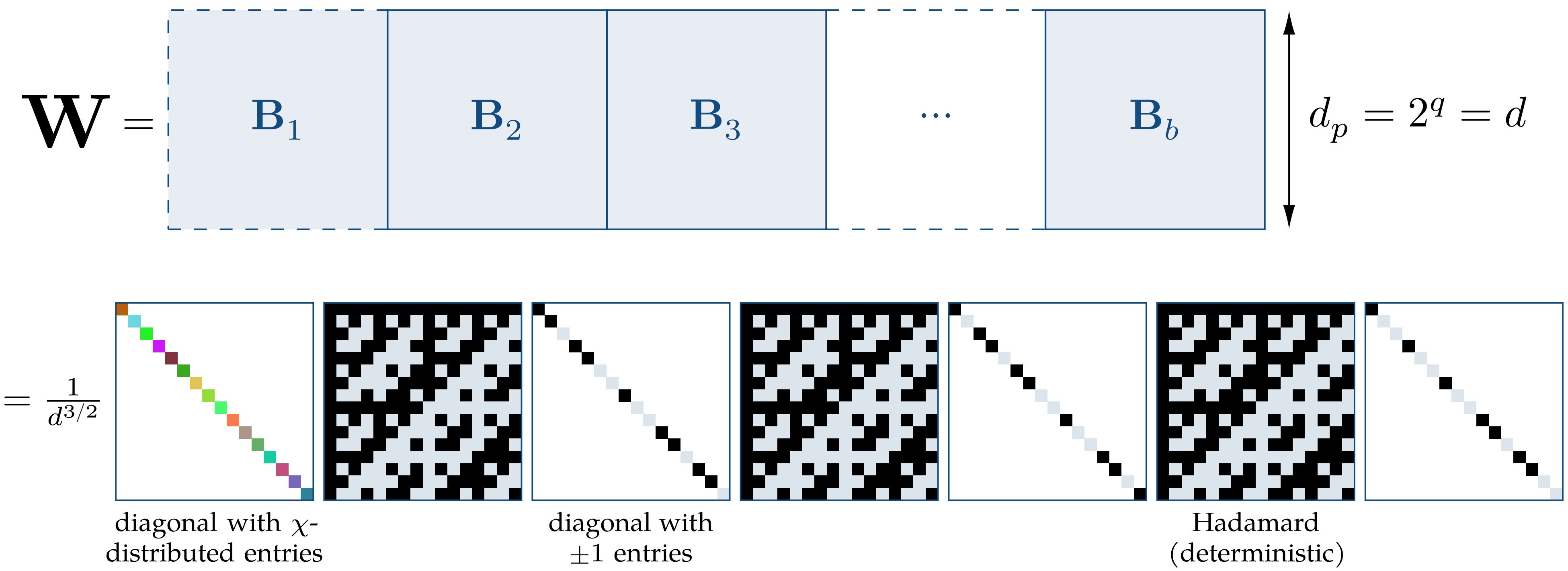
dense random matrix …

Need to calculate the sketch: $\mathcal{O}(nmd)$

$$\mathbf{s} = \frac{1}{n} \sum_{i=1}^{n} \exp(-i\mathbf{W}\mathbf{x}_i)$$

Perspectives

- Use random structured matrices:



Limitations & perspectives

Complexity in space

- Sketching reduces drastically the dimension but….

Need to store somewhere:

$$\mathbf{W} \in \mathbb{R}^{m\times d}$$

dense random matrix …

Need to calculate the sketch: $\mathcal{O}(nmd)$

$$\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n} \exp(-i\mathbf{W}\mathbf{x}_i)$$

Fast transform

$\mathcal{O}(nm\ln(d))$

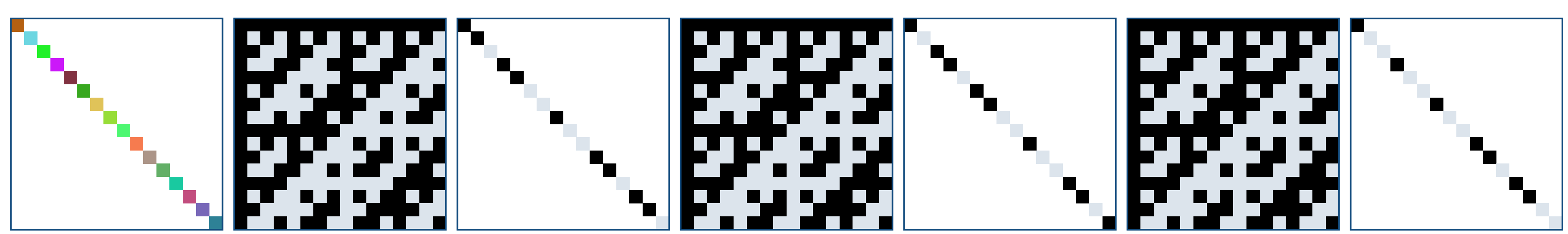
Perspectives

- Use random structured matrices:

sparse

$$\mathbf{W} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_2 & \mathbf{B}_3 & \cdots & \mathbf{B}_b \end{bmatrix} \quad d_p = 2^q = d$$

$$\mathbf{B}^T = \frac{1}{d^{3/2}}$$



diagonal with χ-distributed entries

diagonal with ± 1 entries

Hadamard (deterministic)

Limitations & perspectives

Complexity in space

- Sketching reduces drastically the dimension but….

Need to store somewhere:

$$\mathbf{W} \in \mathbb{R}^{m\times d}$$

dense random matrix …

sparse

Need to calculate the sketch: $\mathcal{O}(nmd)$

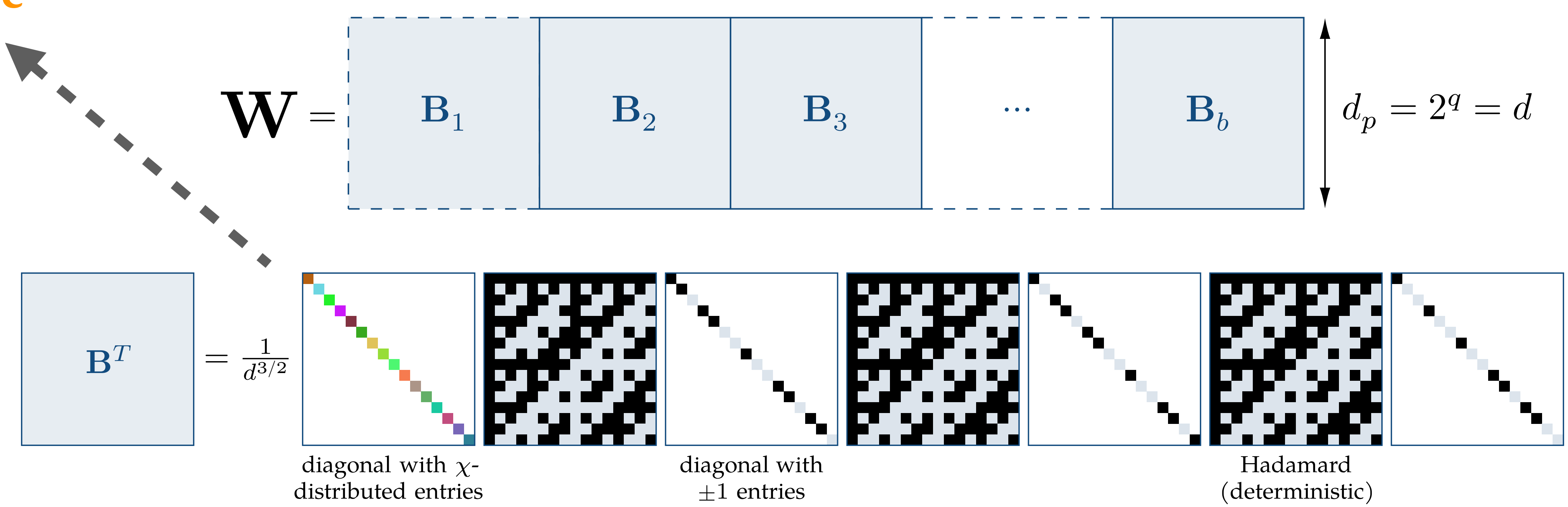
$$\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n} \exp(-i\mathbf{W}\mathbf{x}_i)$$

Fast transform

$\mathcal{O}(nm\ln(d))$

Perspectives

- Use random structured matrices **+ parallelize**



Limitations & perspectives

Complexity in space

- Sketching reduces drastically the dimension but….

Need to store somewhere:

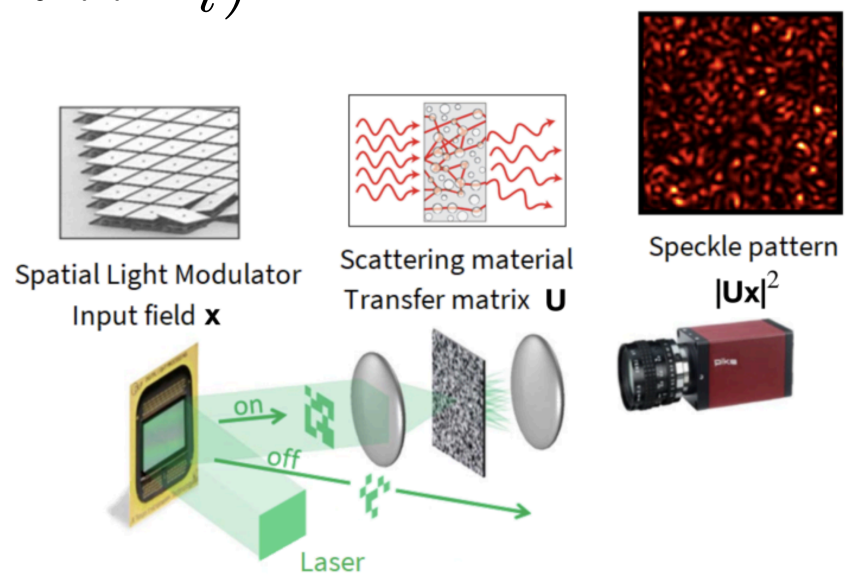
$$\mathbf{W} \in \mathbb{R}^{m \times d}$$

Need to calculate the sketch: $\mathcal{O}(nmd)$

$$\mathbf{s} = \frac{1}{n} \sum_{i=1}^{n} \exp(-i\mathbf{W}\mathbf{x}_i)$$

Perspectives

- Use random structured matrices
- Optical Processing Unit (OPU)

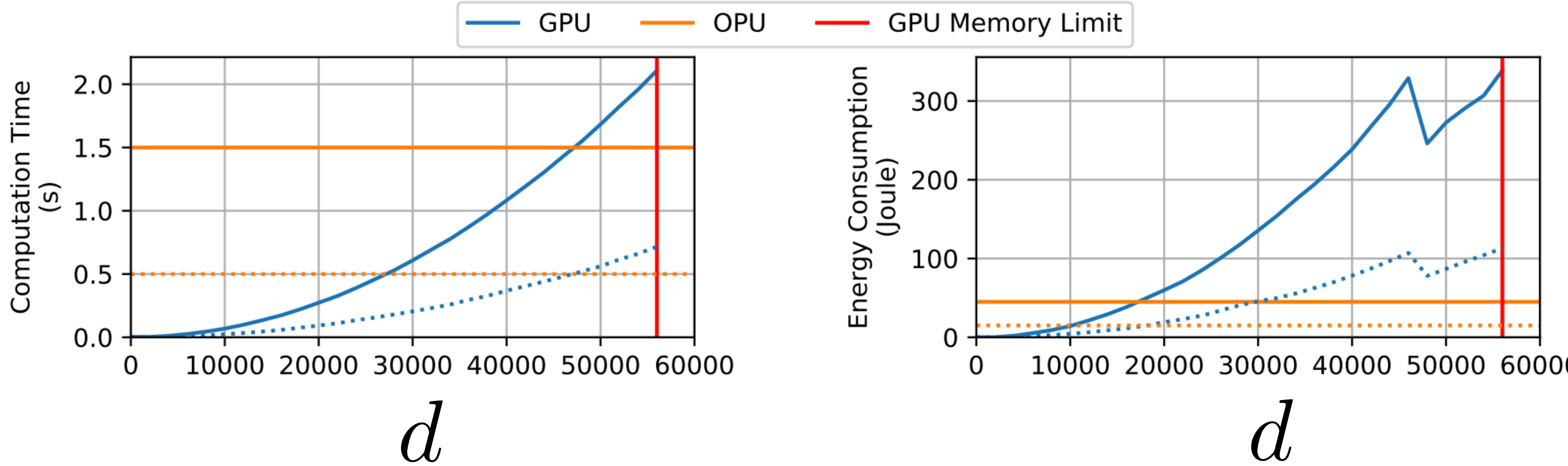


[Giffon & Gribonval, 2022]

- Constant time for matrix multiplication (with random matrix)

$$\mathbf{W}\mathbf{x}_i \text{ in } \mathcal{O}(1)$$

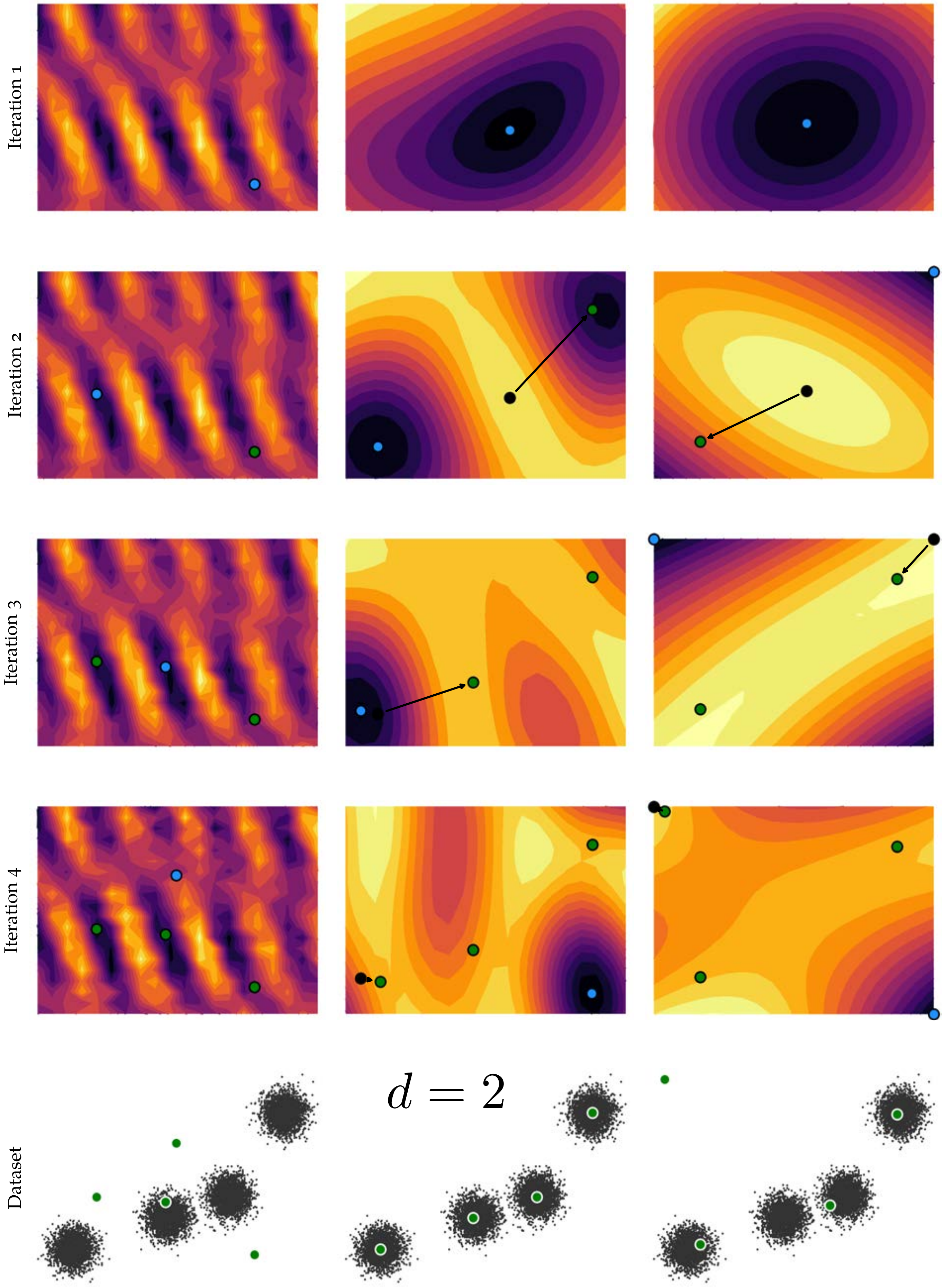
- Constant power consumption



Limitations & perspectives

Hyperparameter

$$W_{ij} \sim \mathcal{N}(0, \sigma^2)$$ **Very sensitive**



Limitations & perspectives

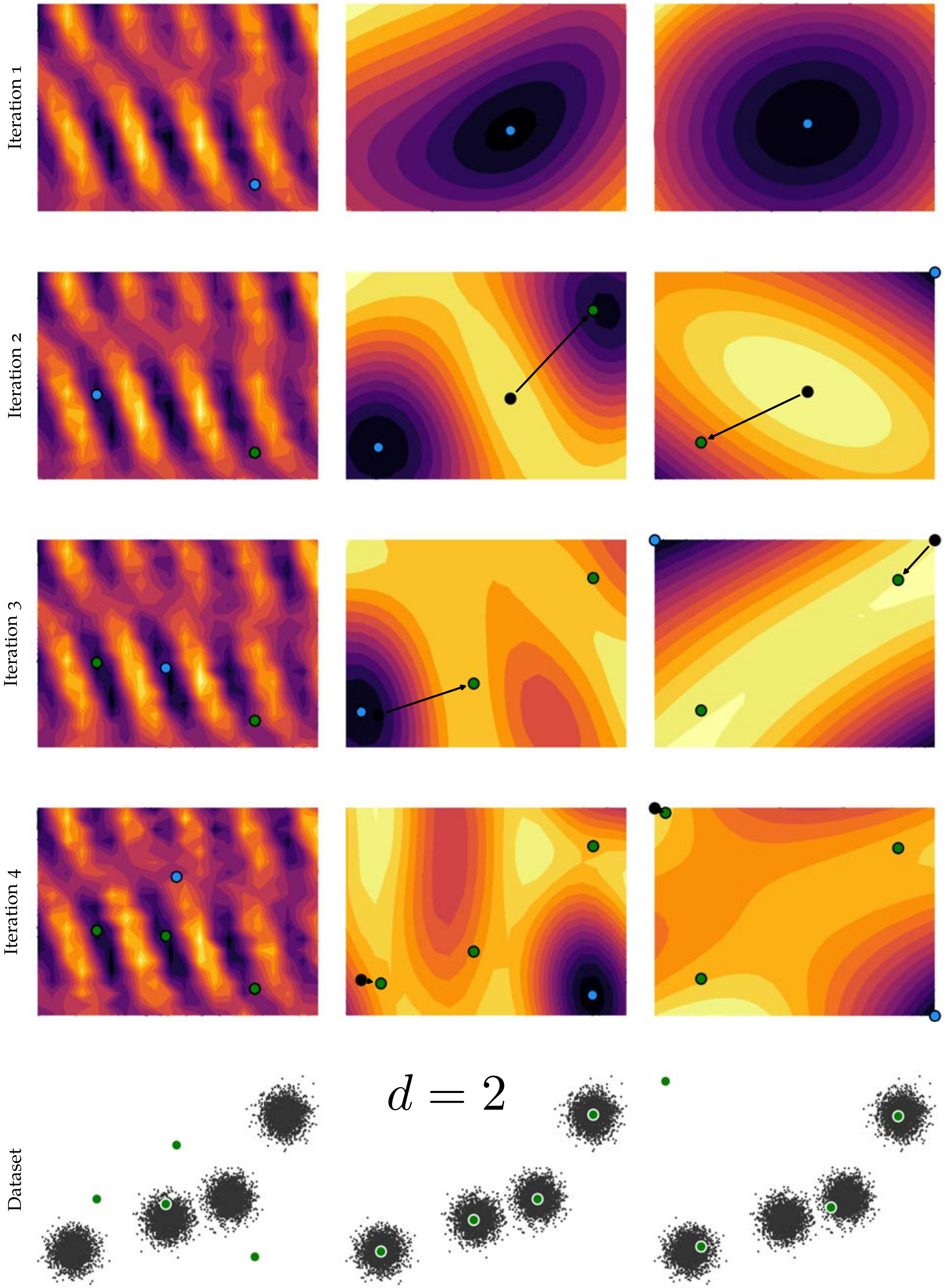
Hyperparameter

$W_{ij} \sim \mathcal{N}(0, \sigma^2)$ **Very sensitive**

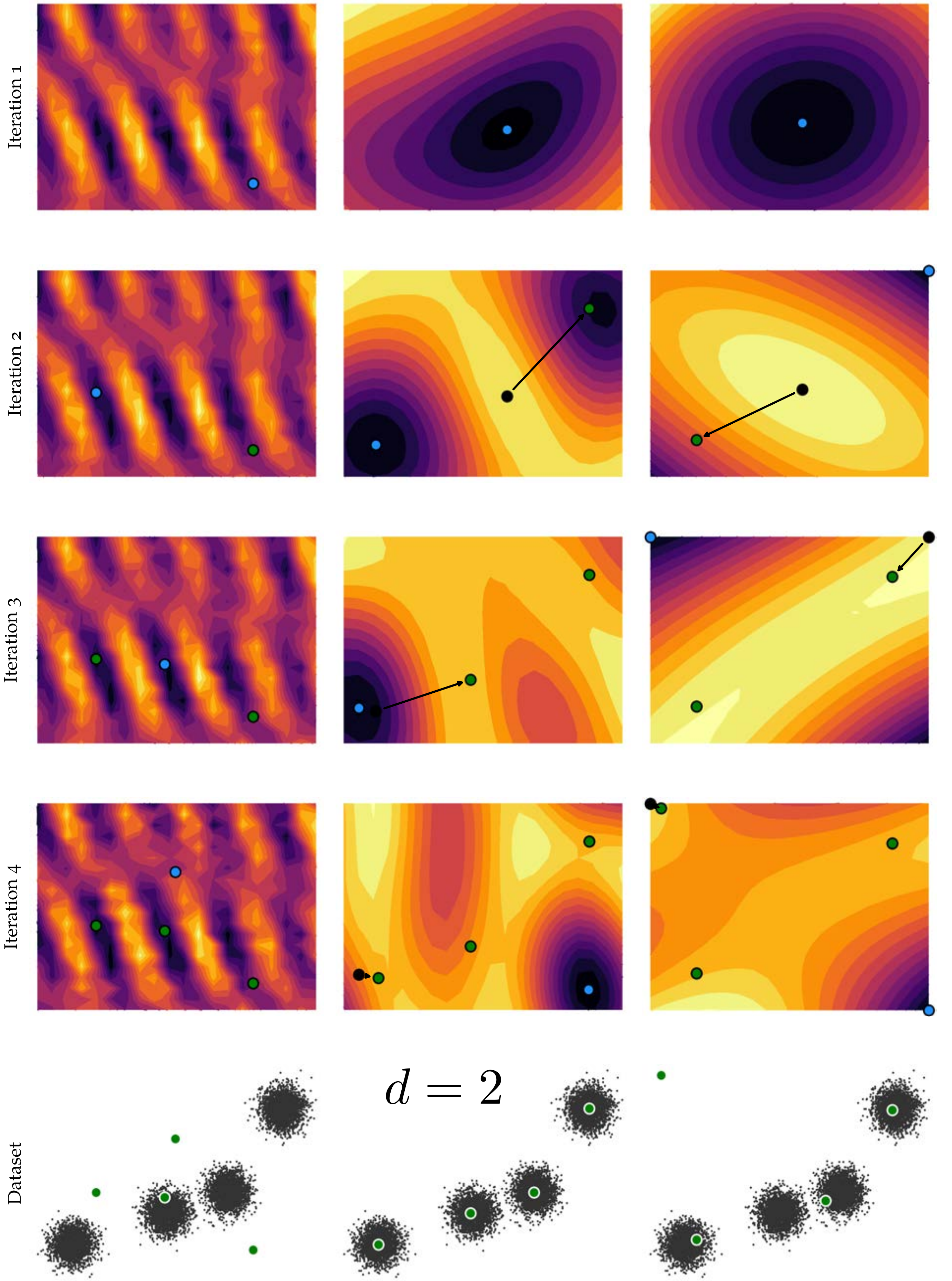
Theoretical guarantees

Lower RIP difficult to prove

Decoding = non-convex problem



Limitations & perspectives



Hyperparameter

$W_{ij} \sim \mathcal{N}(0, \sigma^2)$ **Very sensitive**

Theoretical guarantees

Lower RIP difficult to prove

Decoding = non-convex problem

More tasks ?

Supervised machine learning ?

Limitations & perspectives

- **Hyperparameter**

$W_{ij} \sim \mathcal{N}(0, \sigma^2)$ **Very sensitive**

- **Theoretical guarantees**

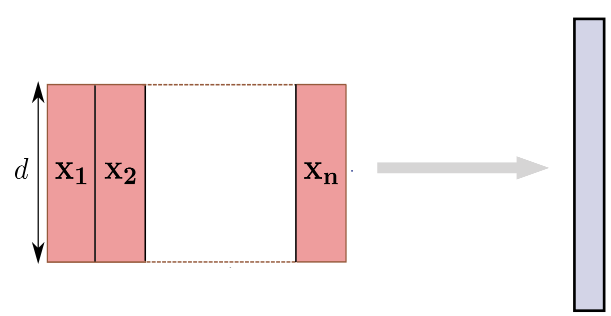
Lower RIP difficult to prove

Decoding = non-convex problem

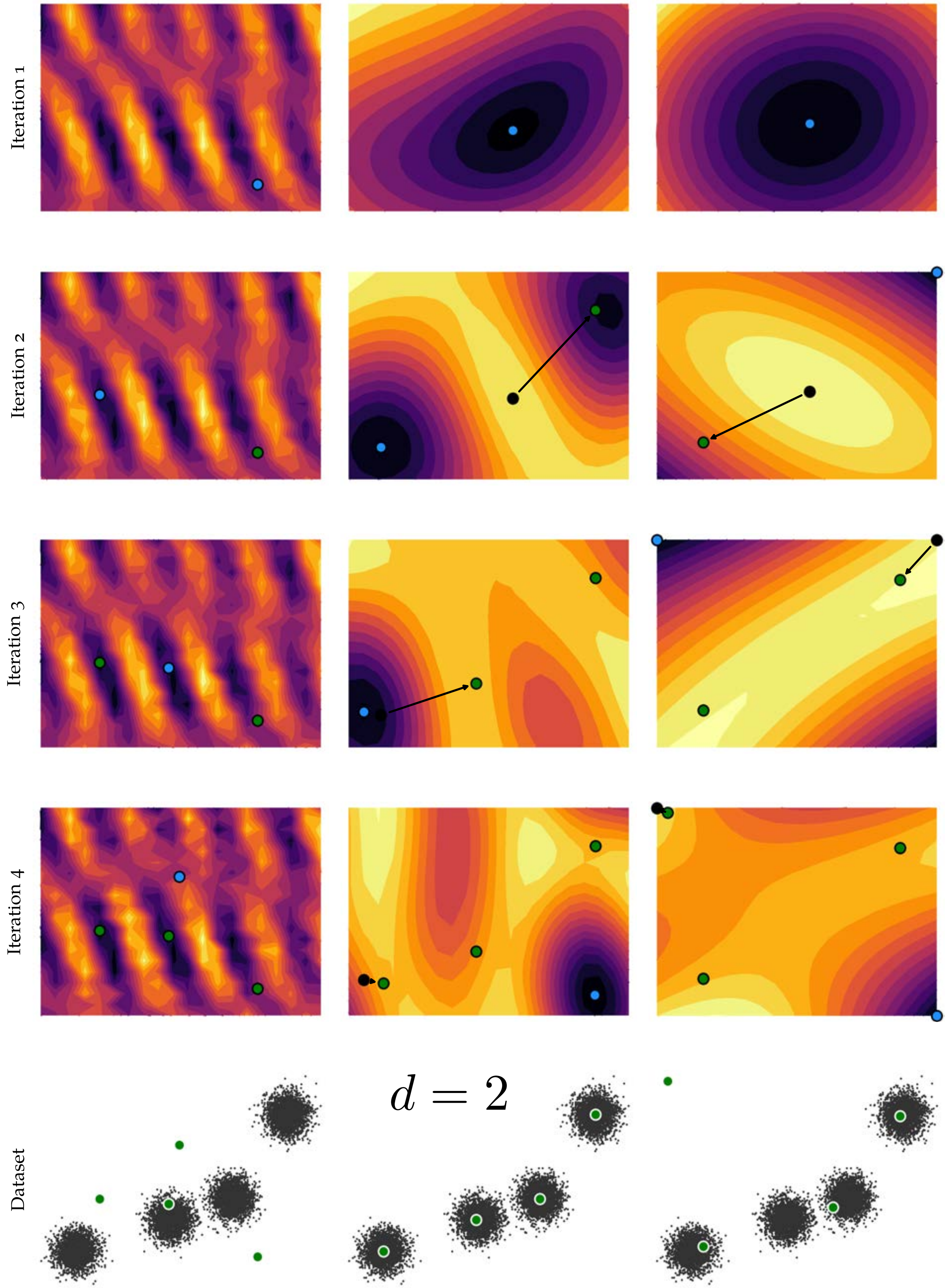
- **More tasks ?**

Supervised machine learning ?

- **Privacy**



[Chatalic, 2020]

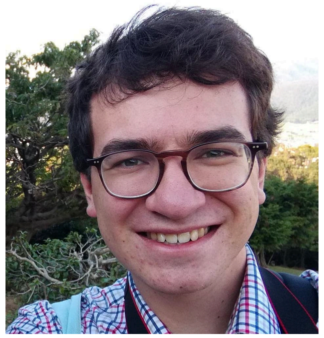


The sketching theory

- Series of past work with **Rémi Gribonval**
- And also Anthony Bourrier, **Nicolas Keriven**, **Antoine Chatalic**, Ayoub Belhadji, Luc Giffon, Gilles Puy, Nicolas Tremblay, Yann Traonmilin, Clément Elvira, Patrick Perez, Mike Davies, Gilles Blanchard, Pierre Vandergheynst, Laurent Jacques, **Vincent Schellekens**, Florimond Houssiau, Phil Schniter, Evan Byrne, …

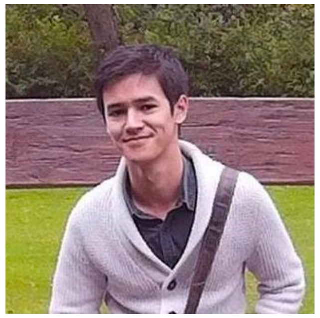
Extending the Compressive Statistical Learning Framework: Quantization, Privacy, and Beyond

Vincent Schellekens



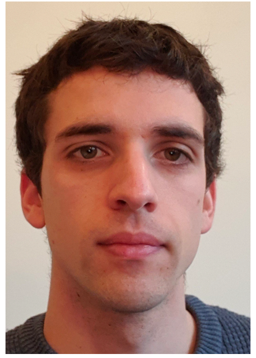
Sketching for large-scale learning of mixture models

Nicolas Keriven

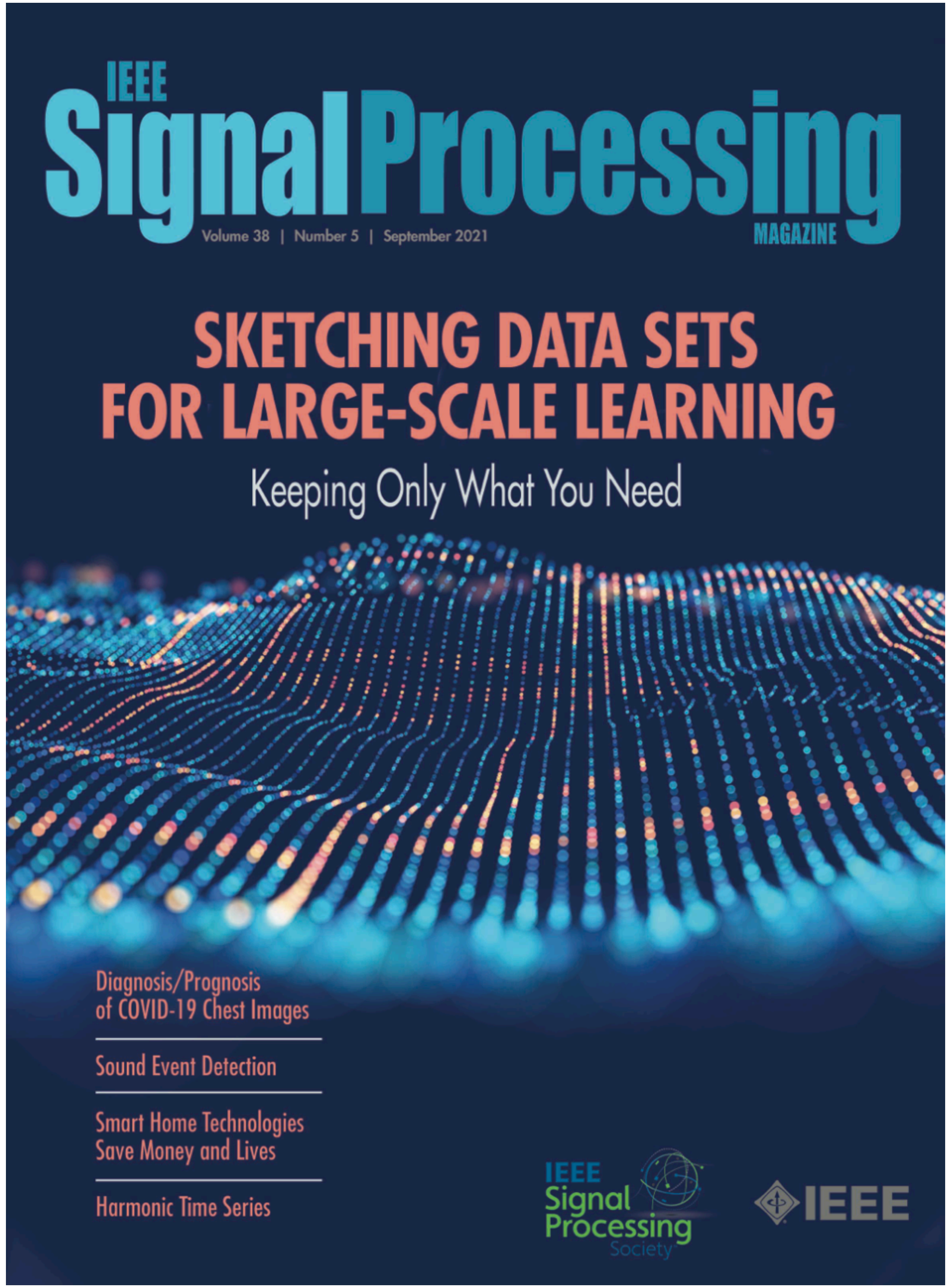


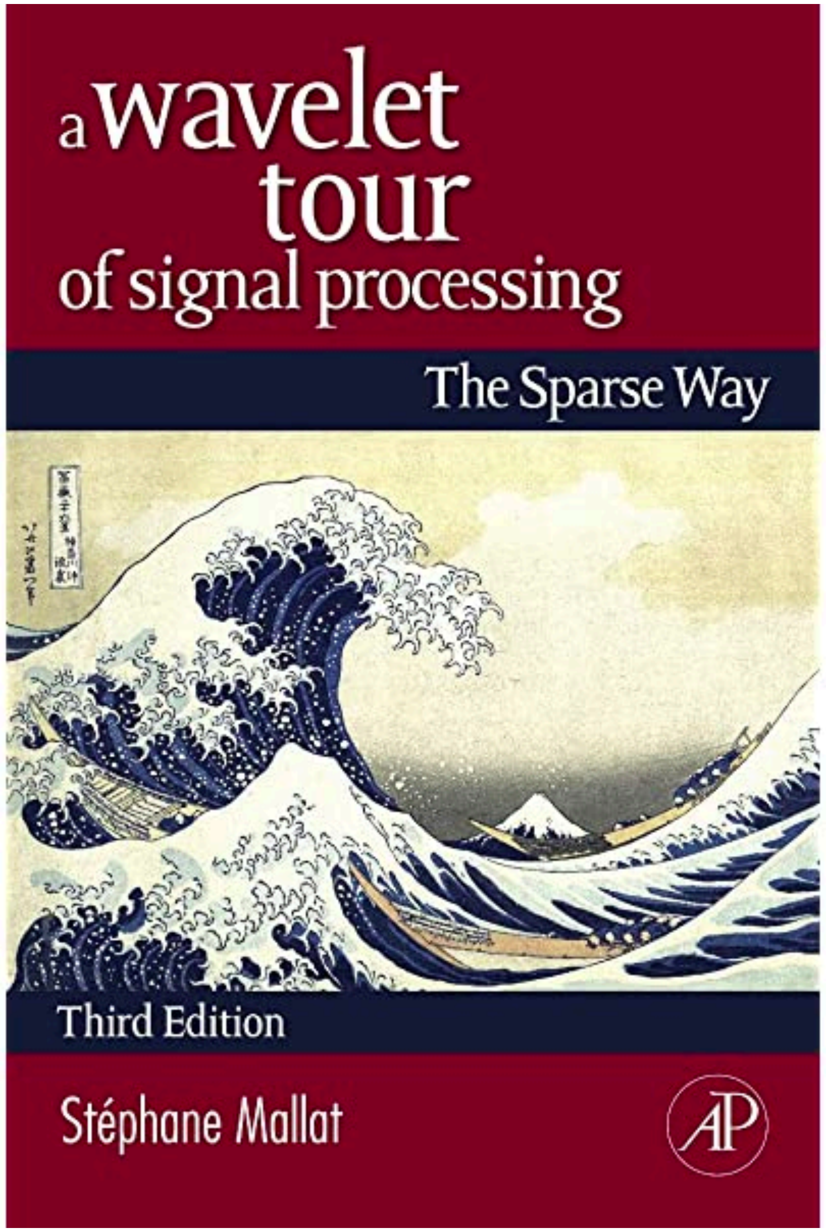
Efficient and privacy-preserving compressive learning

Antoine Chatalic

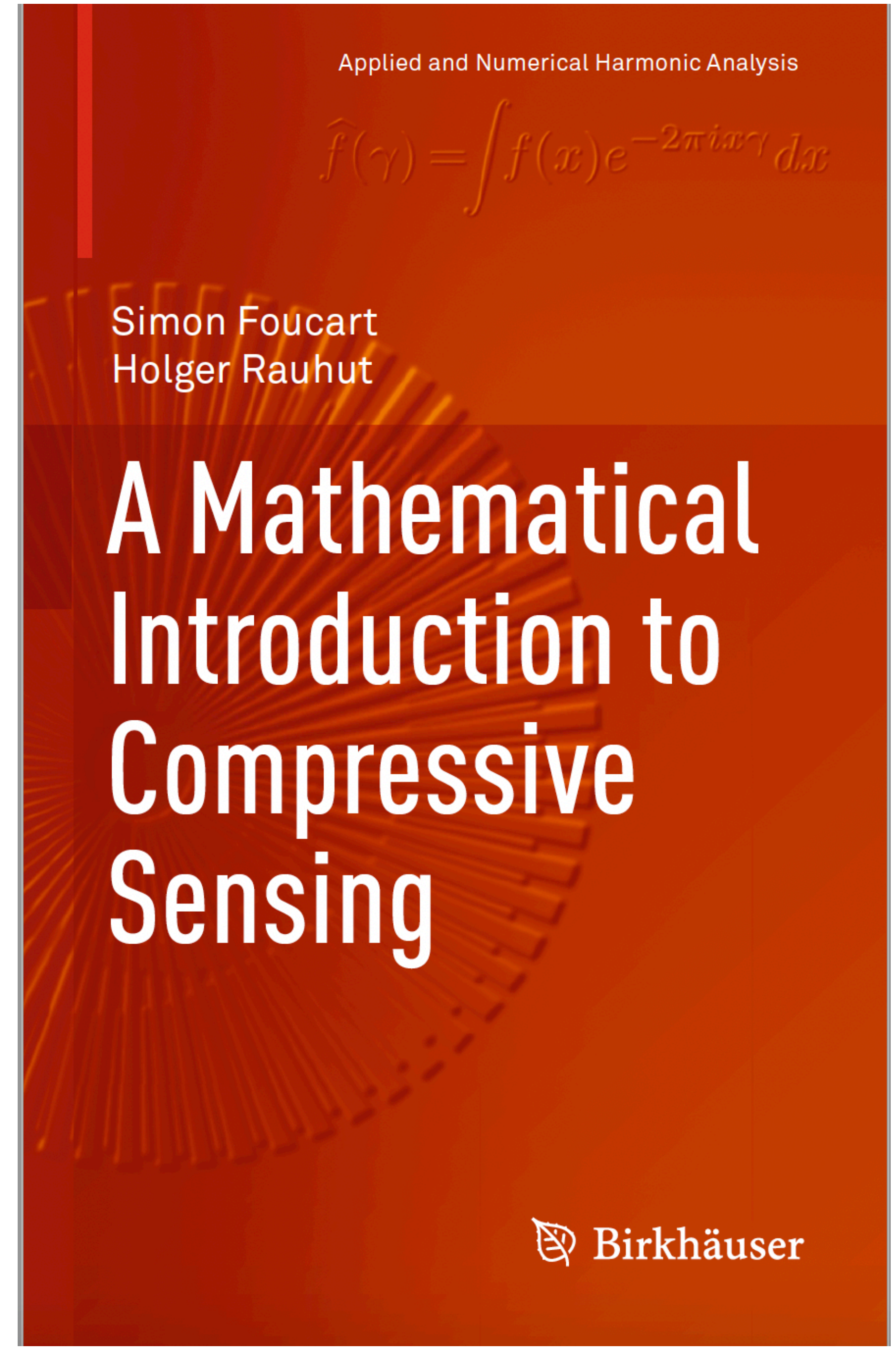


also thanks to Rémi Flamary for some figures…


IEEE
Signal Processing
MAGAZINE
Volume 38 | Number 5 | September 2021
SKETCHING DATA SETS
FOR LARGE-SCALE LEARNING
Keeping Only What You Need
Diagnosis/Prognosis
of COVID-19 Chest Images
Sound Event Detection
Smart Home Technologies
Save Money and Lives
Harmonic Time Series
IEEE
Signal
Processing
Society
IEEE


a wavelet tour of signal processing
The Sparse Way
Third Edition
Stéphane Mallat


Numerical Tours
of Data Sciences


Applied and Numerical Harmonic Analysis
Simon Foucart
Holger Rauhut
A Mathematical Introduction to Compressive Sensing
Birkhäuser

Some other references

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Convex optimization with sparsity-inducing norms. Foundations and Trends in Machine Learning, 4(1):1–106, 2012.

Beck, A. and Teboulle, M. A fast iterative shrinkage thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci., 2(1):183–202, 2009.

Friedman, J., Hastie, T. J., Hofling, H., and Tibshirani, R. Pathwise coordinate optimization. Ann. Appl. Stat., 1(2): 302–332, 2007.

E. J. Candes and T. Tao, Decoding by Linear Programming, IEEE Trans. Inf. Th., 51(12): 4203–4215 (2005).

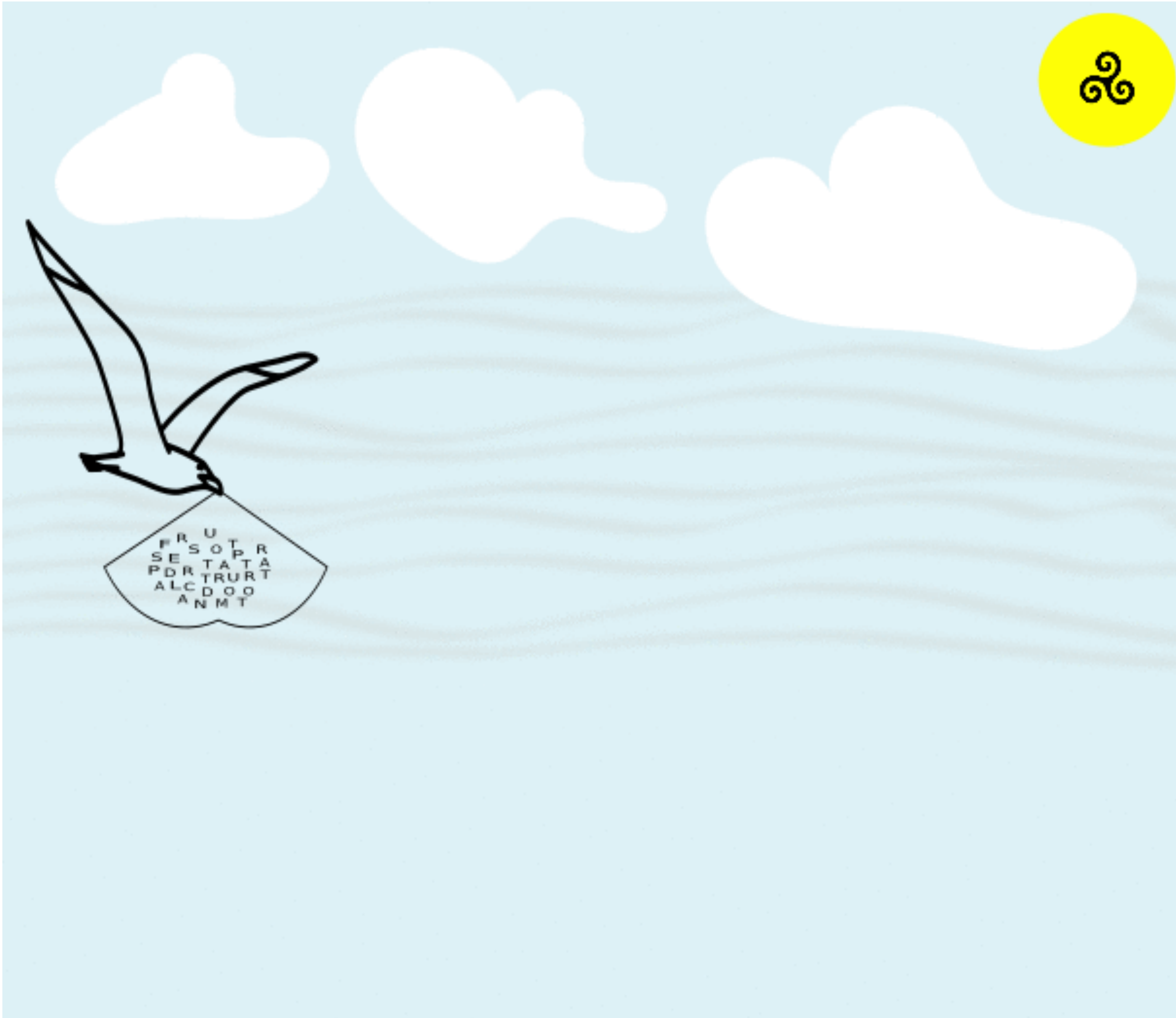
Arthur, D., Vassilvitskii, S. *k*-means++: the advantages of careful seeding . *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035.

Luc Giffon, Rémi Gribonval. Compressive Clustering with an Optical Processing Unit. GRETSI. 2022

Chen, Scott Shaobing and Donoho, David L. and Saunders, Michael A. Atomic Decomposition by Basis Pursuit. SIAM Journal on Scientific Computing.1995

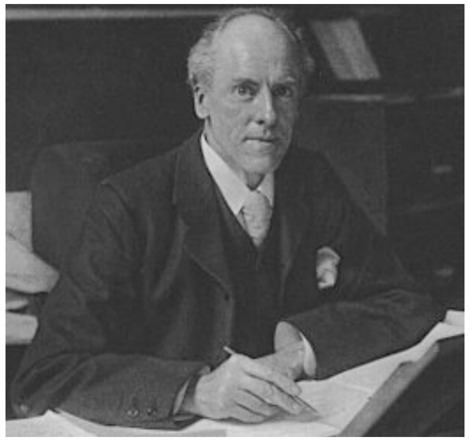
Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. 1996

Thank you!



Machine learning theory

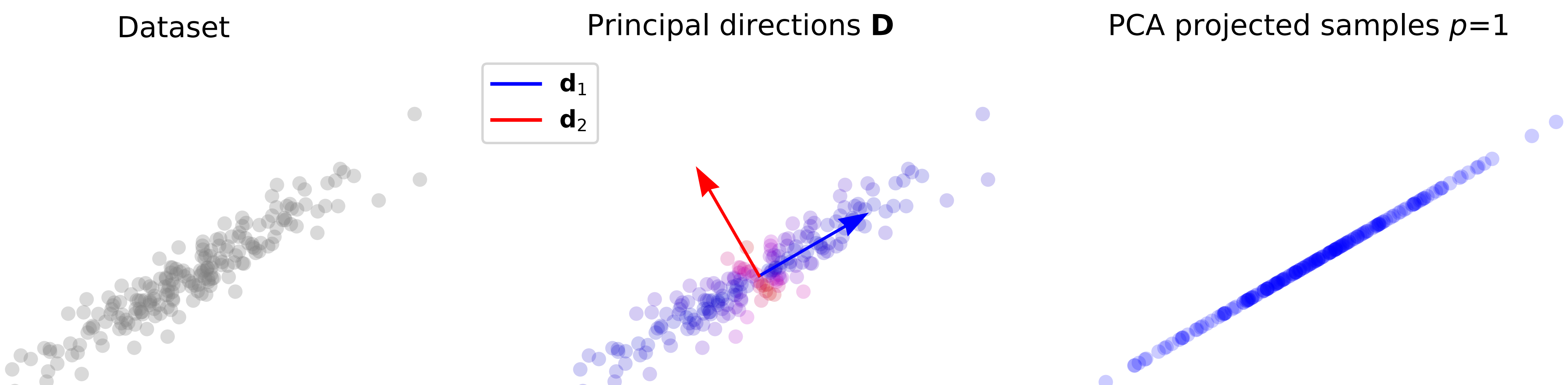
Dimension reduction: PCA



- We aim at solving: [Pearson, 1901]

$$\boxed{\min_{\mathbf{D}\in\mathbb{R}^{d\times k},\ \mathbf{D}^\top\mathbf{D}=\mathbf{I}_k} \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{D}\mathbf{D}^\top\mathbf{x}_i\|_2^2}$$

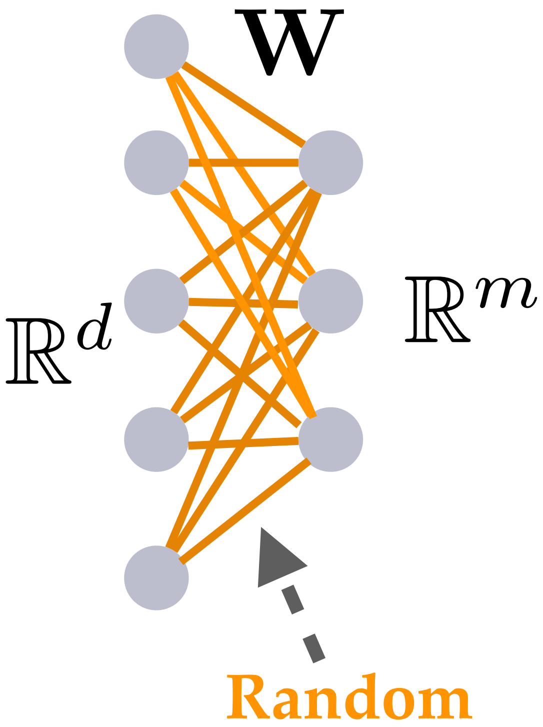
- Find a linear subspace that minimizes the reconstruction error
- $h = \mathbf{D}$, $\mathcal{H} = \text{Stiefel}$ and $\ell(\mathbf{x}_i, h) = \|\mathbf{x}_i - \mathbf{D}\mathbf{D}^\top\mathbf{x}_i\|_2^2$
- Eigen. decomposition of covariance matrix: $\mathcal{O}(nd^2 + d^3)$



Theory of sketching

Randomization: the core of sketching

- A function called feature operator $\boxed{\Phi : \mathbb{R}^d \to \mathbb{R}^m}$
- In practice $\Phi(\mathbf{x}) = \rho(\mathbf{W}\mathbf{x})$
- Where $\mathbf{W} \in \mathbb{R}^{m \times d}$ is a random matrix
- Where $\rho : \mathbb{R}^m \to \mathbb{R}^m$ is a non-linear activation



Random Fourier Features (RFF)

- Where $\mathbf{W} \in \mathbb{R}^{m \times d}$ is **Gaussian** $W_{ij} \sim \mathcal{N}(0, \sigma^2)$
- Where $\rho(\mathbf{y}) = \frac{1}{\sqrt{m}} (\exp(-iy_1), \cdots, \exp(-iy_m))$

$$\boxed{\Phi(\mathbf{x}) = \frac{1}{\sqrt{m}}(\exp(-i\boldsymbol{\omega}_1^\top \mathbf{x}), \cdots, \exp(-i\boldsymbol{\omega}_m^\top \mathbf{x}))}$$

[Rahimi & Recht, 2008]

$$\mathbf{W} = [\boldsymbol{\omega}_1^\top, \cdots, \boldsymbol{\omega}_m^\top]$$

- For PCA:
  $\Phi(\mathbf{x}) = \frac{1}{\sqrt{m}} \left(|\mathbf{x}^\top \boldsymbol{\omega}_1|^2, \cdots, |\mathbf{x}^\top \boldsymbol{\omega}_m|^2\right)$