Controlling Wasserstein distances by Kernel norms with application to Compressive Statistical Learning

Titouan Vayer

Post-doc ENS Lyon

CMAP

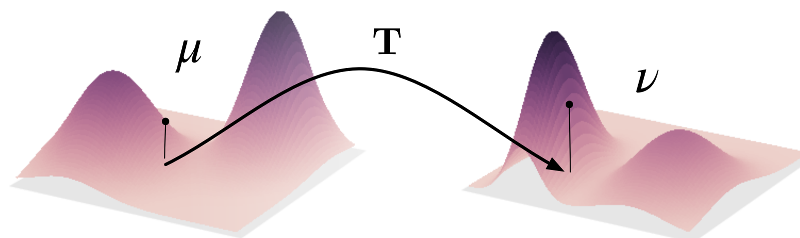
16/12/2021

Remi Gribonval

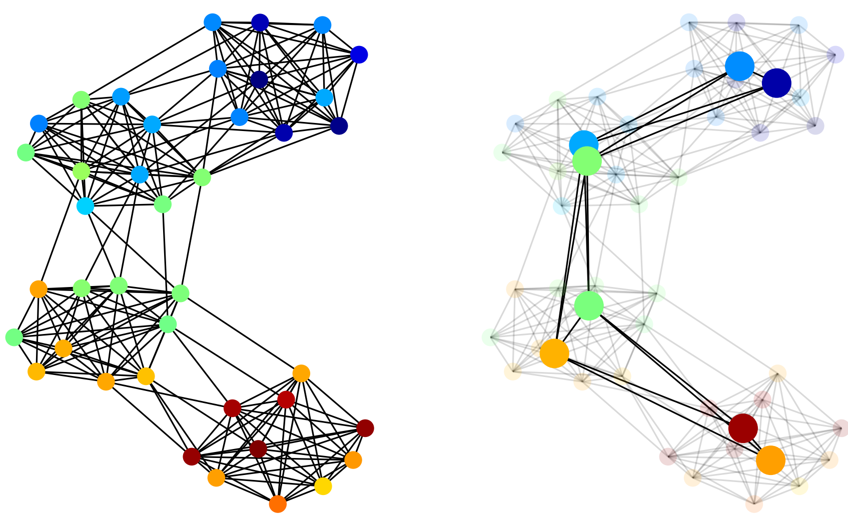
Other research topics

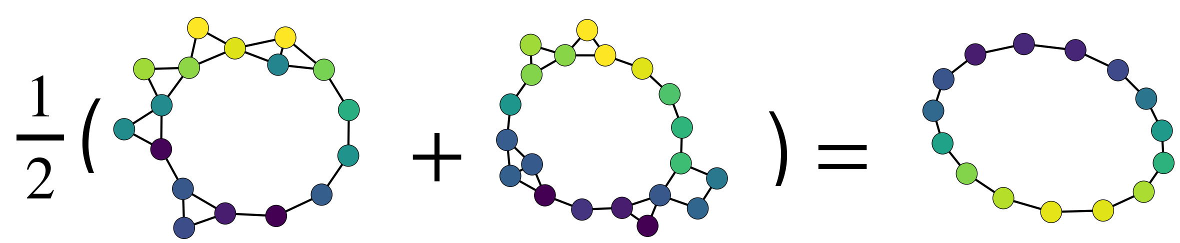
Learning from **structured and heterogeneous** data

Optimal Transport theory

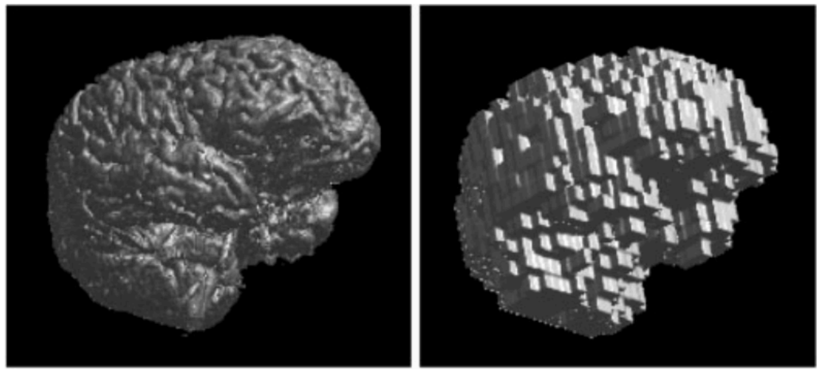


GraphS learning (reduction, classification, clustering, matching, barycenter)





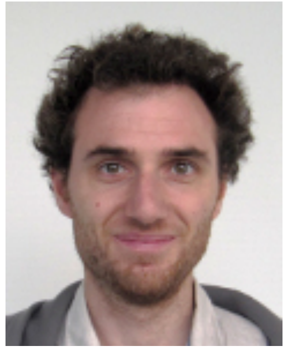
Heterogeneous data (domain adaptation)



Rémi Flamary

Nicolas Courty

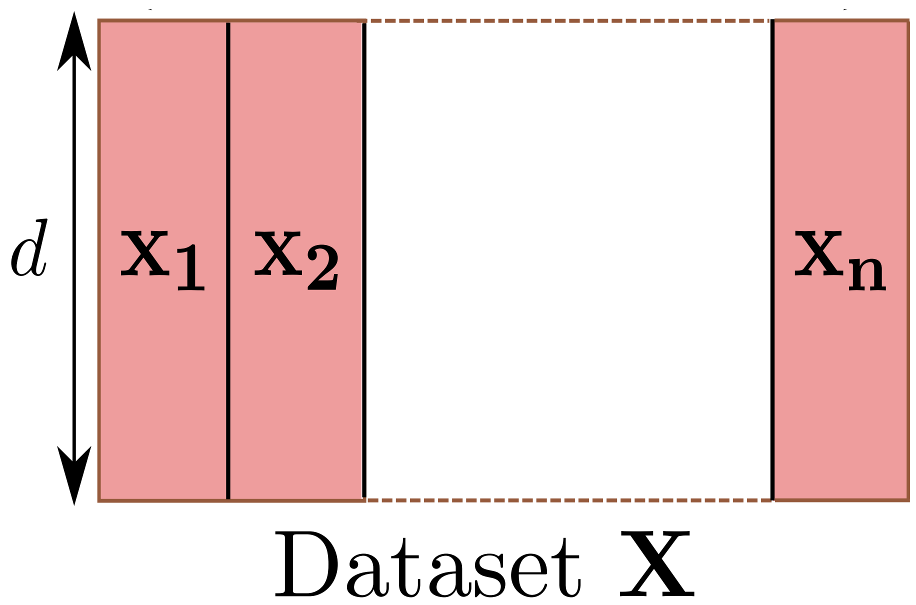
Laetitia Chapel



Romain Tavenard

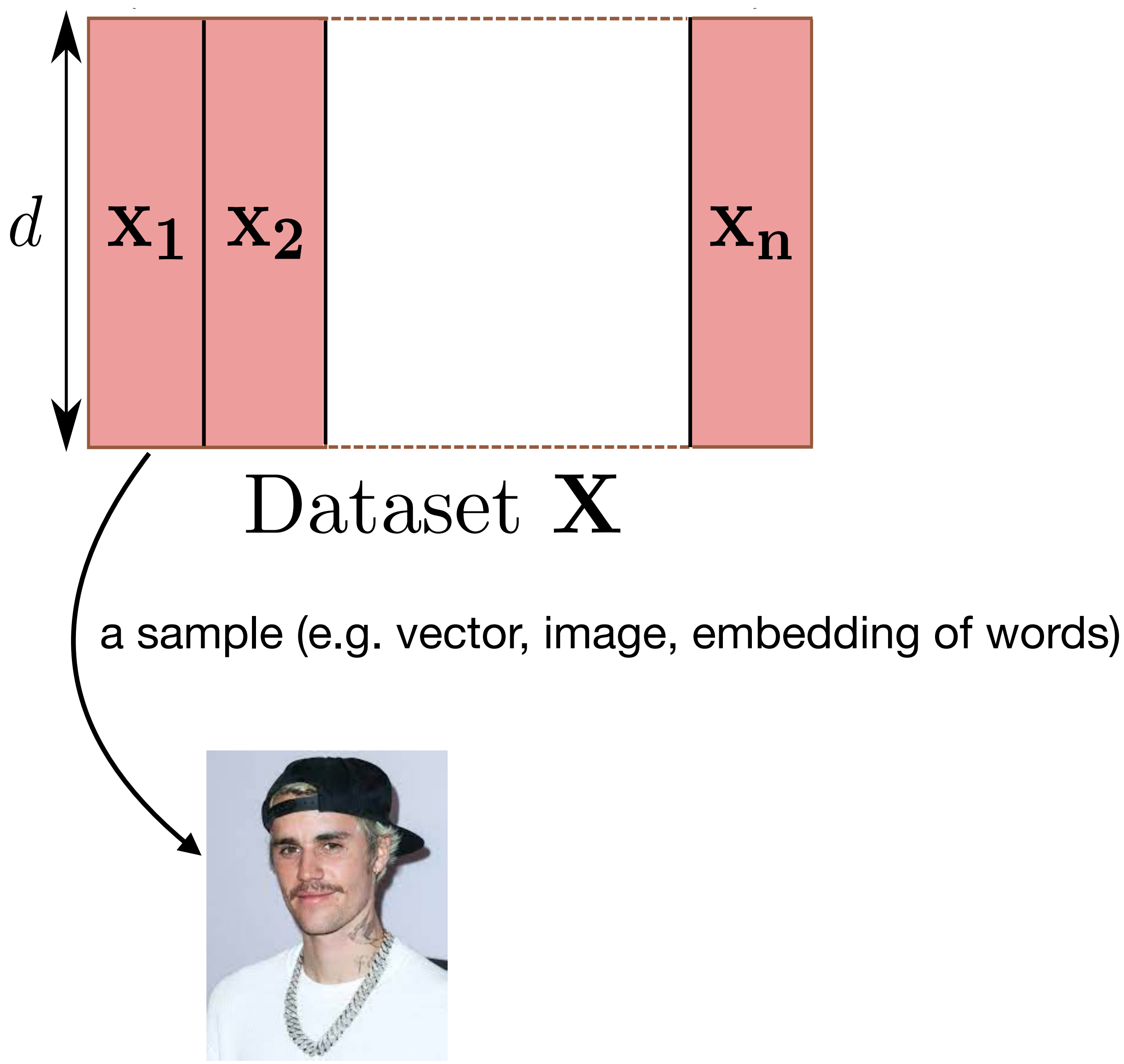
Motivations of this talk

Context: Machine learning



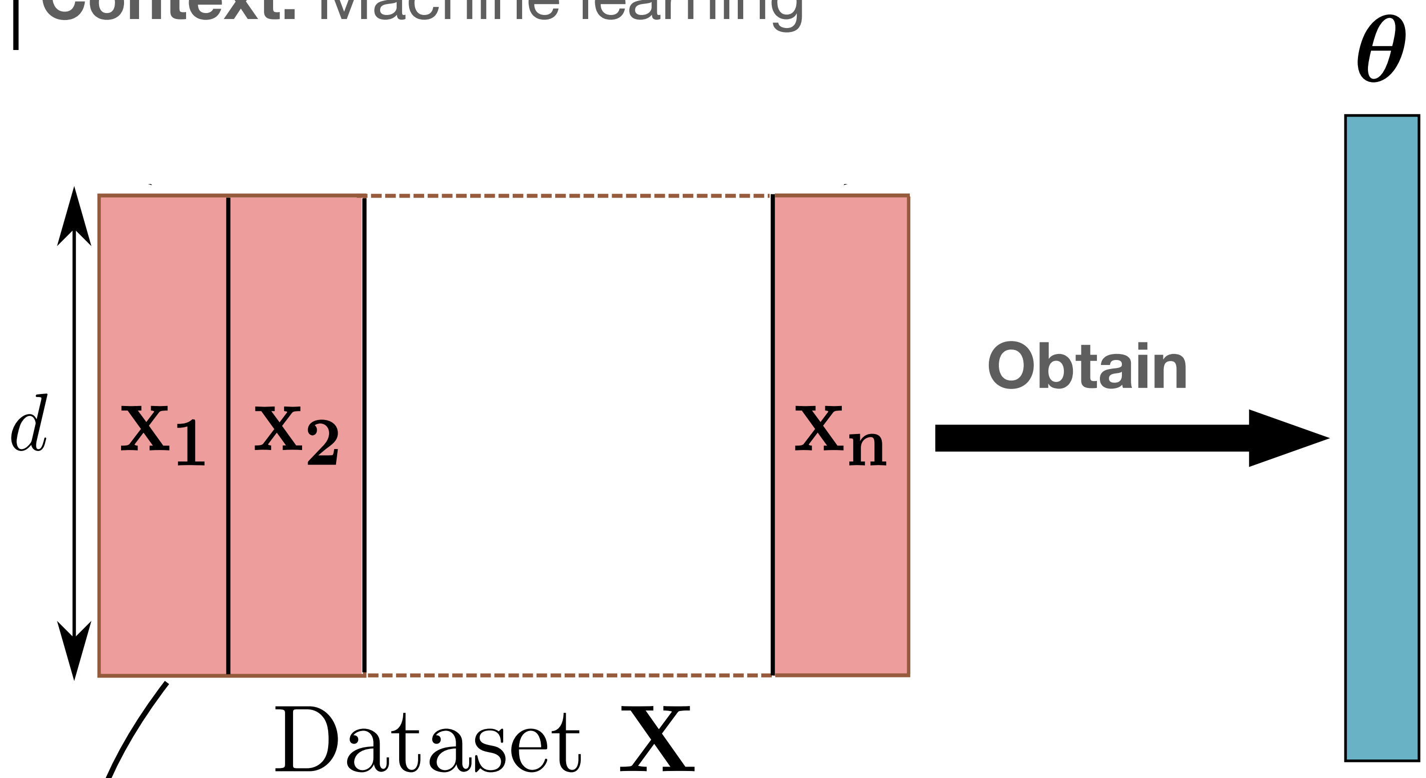
Motivations of this talk

Context: Machine learning



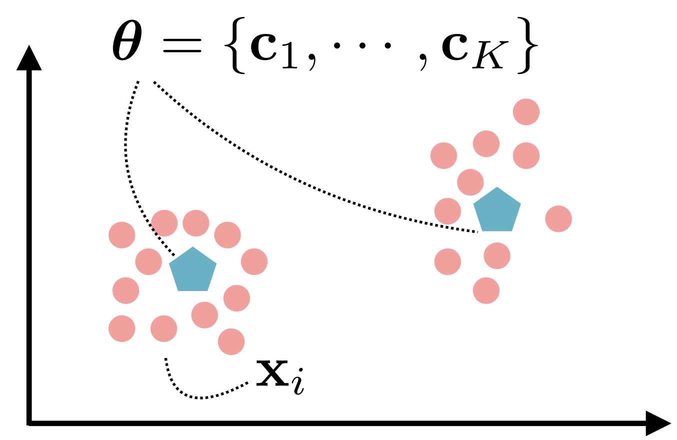
Motivations of this talk

Context: Machine learning



Parameters that solves a specific tasks

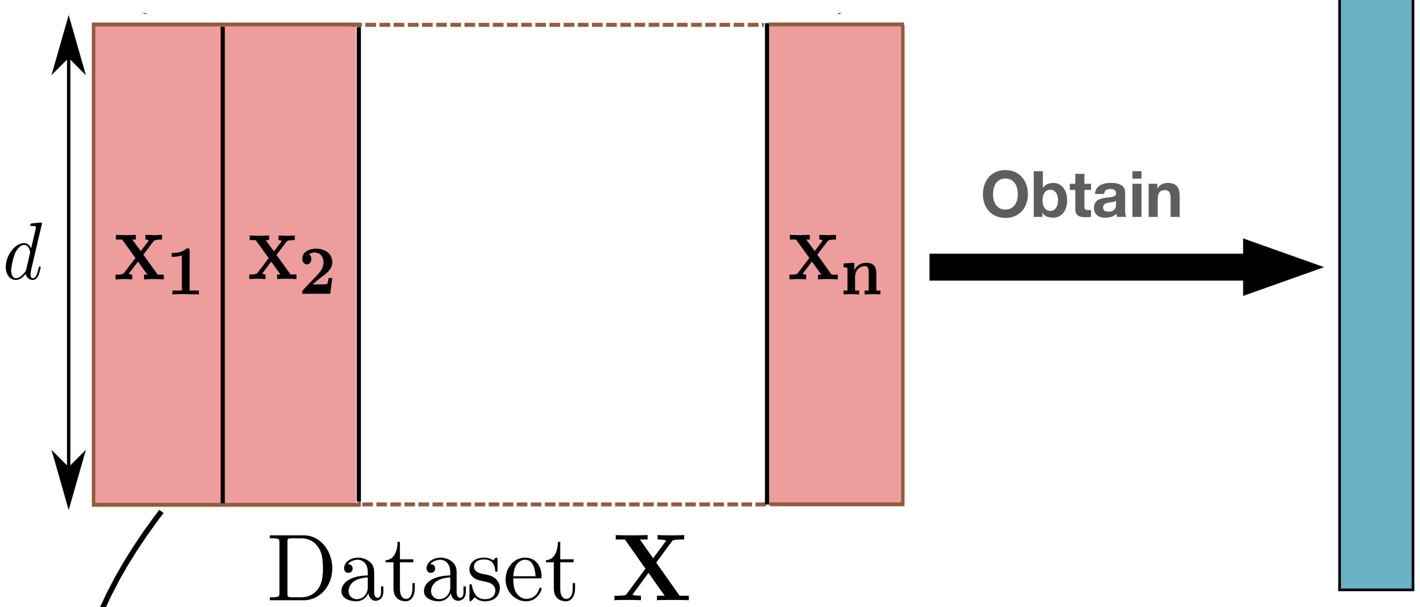
Example: K-means



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

Motivations of this talk

Context: Machine learning



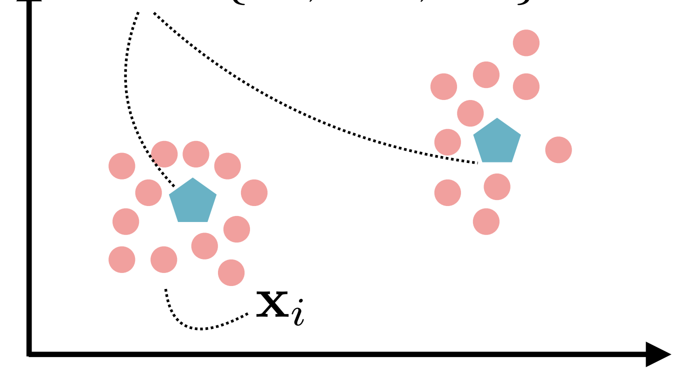
Dataset $\mathbf{X}$

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

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

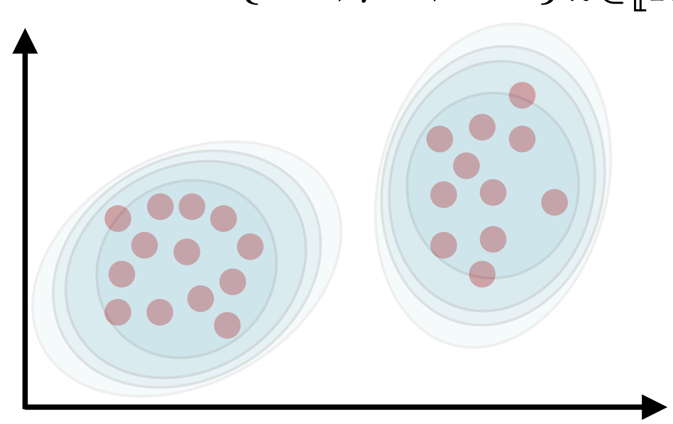
Example: K-means

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



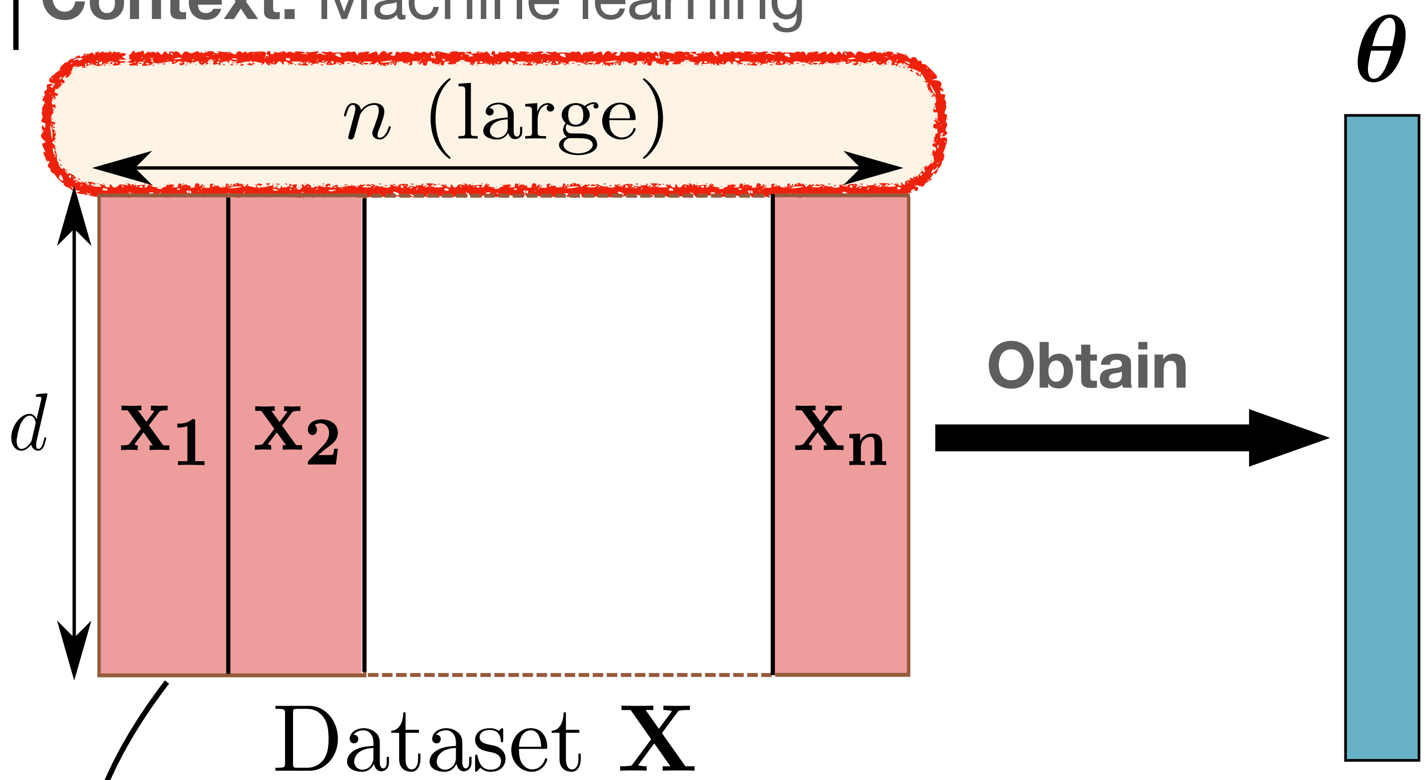
Example: GMM fitting

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



Motivations of this talk

Context: Machine learning

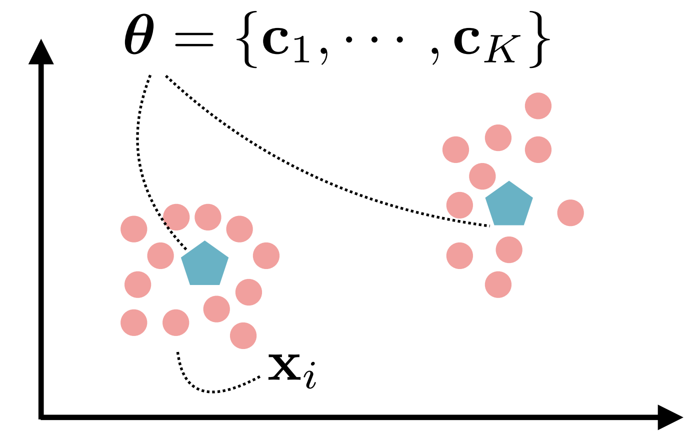


$\boldsymbol{\theta}$ **Parameters** that solves a specific tasks

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

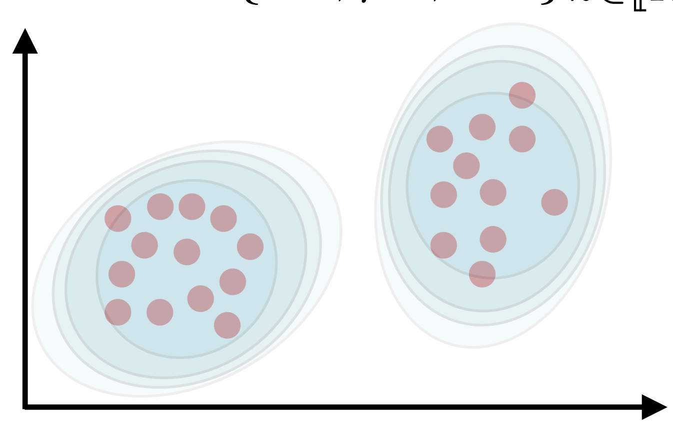
Large scale
Machine Learning

Example: K-means



Example: GMM fitting

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



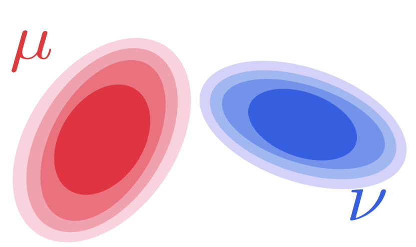
Overview of the talk

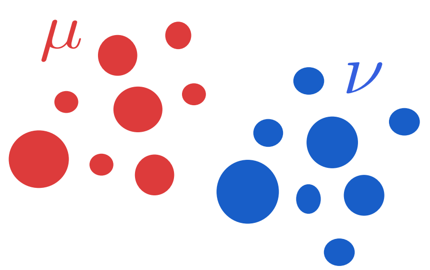
Part I: Optimal Transport and MMD

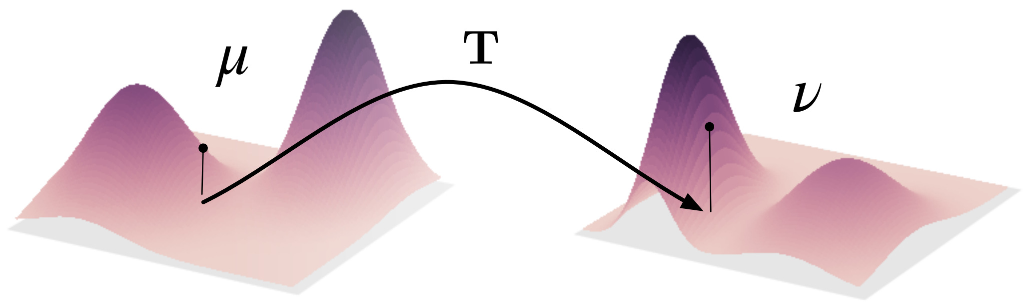
Part II: From Statistical Learning to Compressive Statistical Learning

Part III: Optimal Transport for Compressive Statistical Learning

Comparing probability distributions: Optimal Transport and MMD







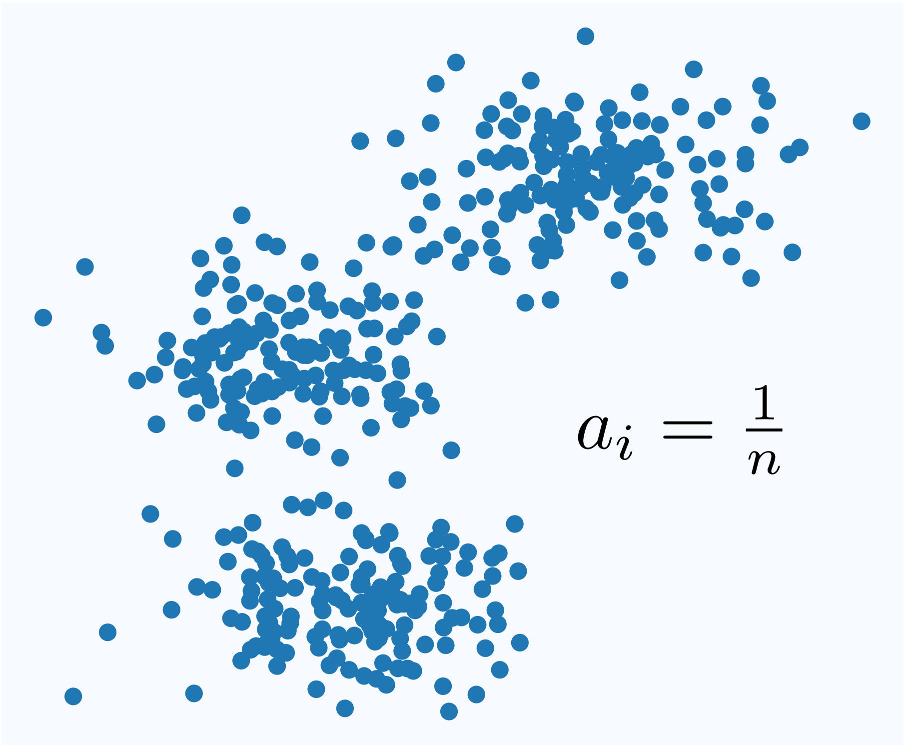
Probability distributions

Why do we care about probability distributions?

Measure and probability distributions are at the core of Machine learning

Data: $(\mathbf{x}_i)_{i \in [\![n]\!]}$; $\mathbf{x}_i \in \mathbb{R}^d$ ⟶ A probability distribution

Lagrangian: $\sum_{i=1}^n a_i \delta_{x_i}$



$a_i = \frac{1}{n}$

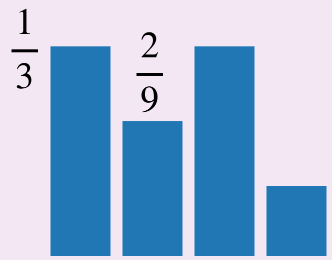
(point clouds)

$\delta_{\mathbf{x}_i}(\mathbf{x}) = 1 \text{ if } \mathbf{x} = \mathbf{x}_i \text{ else } 0$

Probability simplex

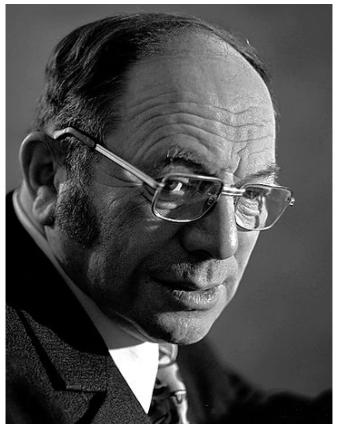
$\mathbf{a} = (a_i)_{i \in [\![n]\!]} \in \Sigma_n$

$a_i \geq 0, \sum_{i=1}^n a_i = 1$



Linear Optimal Transport

Formulation



Two probability distributions

$\pi \in \mathcal{P}(\mathcal{X}), \; \pi' \in \mathcal{P}(\mathcal{Y})$

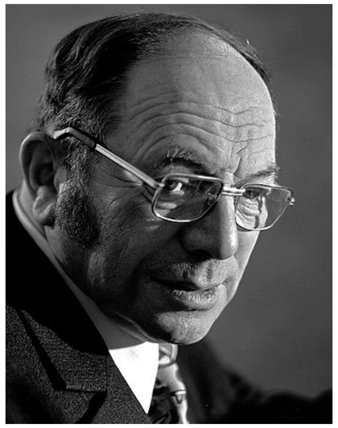
A cost function

$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

Optimal Transport

Linear Optimal Transport

Kantorovitch Formulation



Two probability distributions

$\pi \in \mathcal{P}(\mathcal{X}), \ \pi' \in \mathcal{P}(\mathcal{Y})$

A cost function

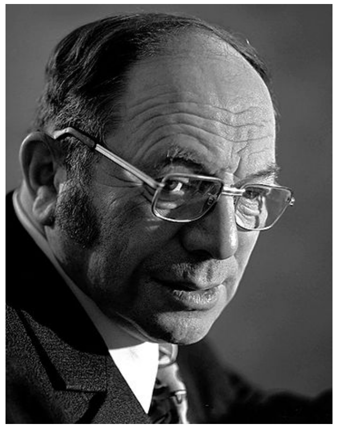
$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

Optimal Transport

All the mass of π is **transported** to π' by a **transport plan** $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$

Linear Optimal Transport

Kantorovitch Formulation



Two probability distributions

$\pi \in \mathcal{P}(\mathcal{X}), \ \pi' \in \mathcal{P}(\mathcal{Y})$

A cost function

$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

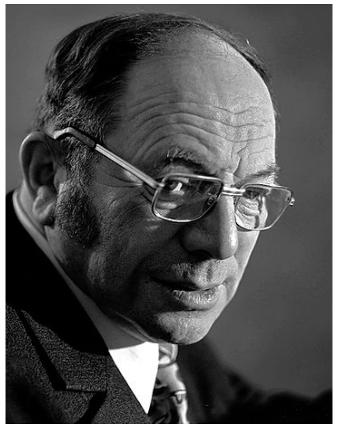
Optimal Transport

All the mass of π is **transported** to π' by a **transport plan** $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$

We want to find the plan that **minimizes the overall cost** of moving all the points

Linear Optimal Transport

Kantorovitch Formulation: an example



Two probability distributions

$\pi = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \pi' = \sum_{j=1}^{m} b_j \delta_{y_j}$

A cost function

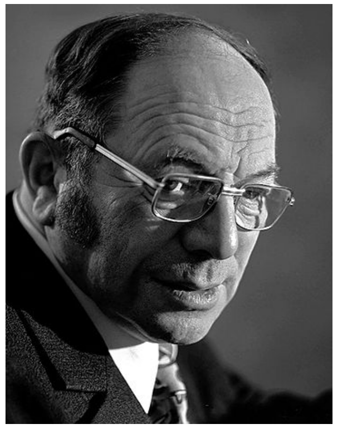
$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

Kantorovitch formulation

$$\min_{\mathbf{T} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i,j} c(x_i, y_j) T_{ij}$$

Linear Optimal Transport

Kantorovitch Formulation: an example



Two probability distributions

$\pi = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \pi' = \sum_{j=1}^{m} b_j \delta_{y_j}$

A cost function

$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

Kantorovitch formulation

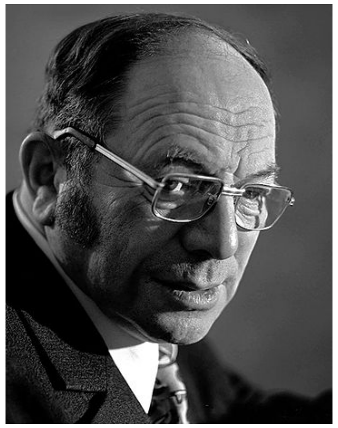
$$\min_{\mathbf{T} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i,j} c(x_i, y_j) T_{ij}$$

Set of couplings/ transport plans

$$\Pi(\mathbf{a}, \mathbf{b})$$

Linear Optimal Transport

Kantorovitch Formulation: an example



Two probability distributions

$\pi = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \pi' = \sum_{j=1}^{m} b_j \delta_{y_j}$

A cost function

$c(x, y) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$

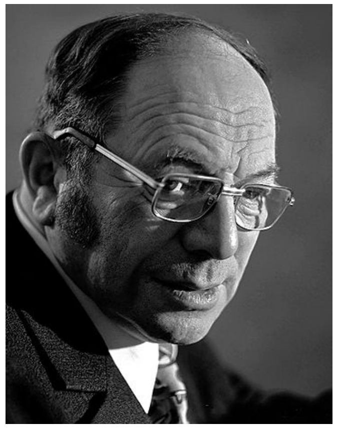
Kantorovitch formulation

$$\min_{\mathbf{T} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i,j} c(x_i, y_j) T_{ij}$$

How much is shifted from x_i to y_j

Linear Optimal Transport

Kantorovitch Formulation: an example



Two probability distributions

$\pi = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \pi' = \sum_{j=1}^{m} b_j \delta_{y_j}$

A cost function

$c(x, y) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$

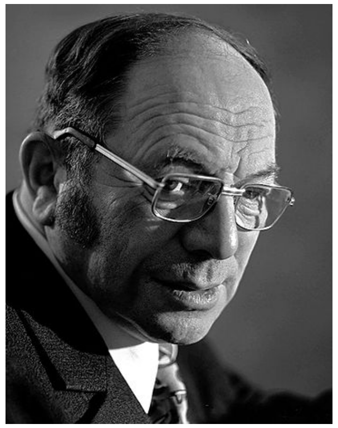
Kantorovitch formulation

$$\min_{\mathbf{T} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i,j} c(x_i, y_j) T_{ij}$$

Cost of moving masses from x_i to y_j

Linear Optimal Transport

Kantorovitch Formulation: an example



Two probability distributions

$\pi = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \pi' = \sum_{j=1}^{m} b_j \delta_{y_j}$

A cost function

$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

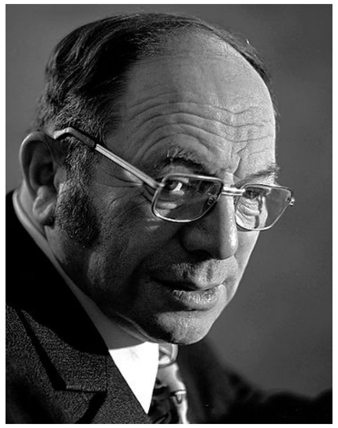
Kantorovitch formulation

$$\min_{\mathbf{T} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i,j} c(x_i, y_j) T_{ij}$$

Total cost

Linear Optimal Transport

Kantorovitch Formulation: an example



Two probability distributions

$\pi = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \pi' = \sum_{j=1}^{m} b_j \delta_{y_j}$

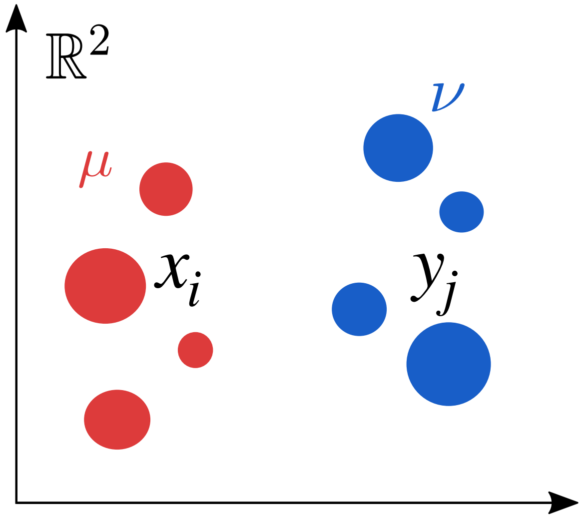
A cost function

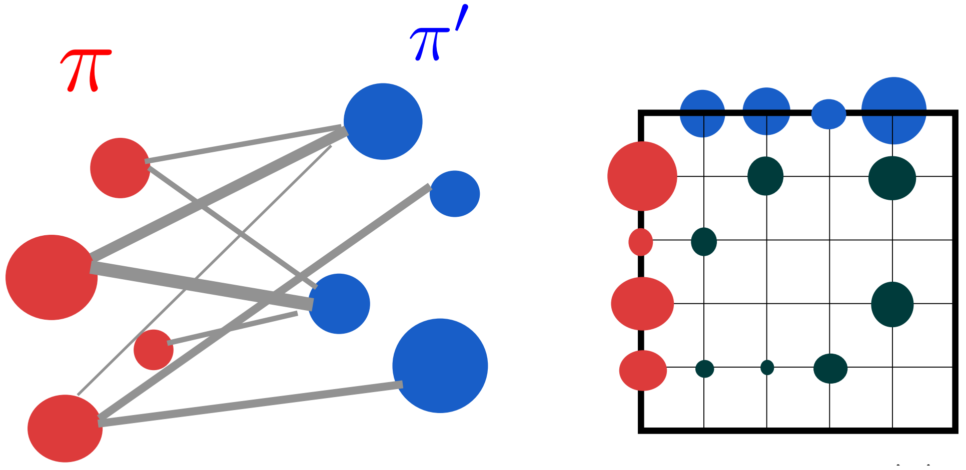
$c(x,y) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$

Kantorovitch formulation

$$\min_{\mathbf{T} \in \Pi(\mathbf{a},\mathbf{b})} \sum_{i,j} c(x_i, y_j) T_{ij}$$

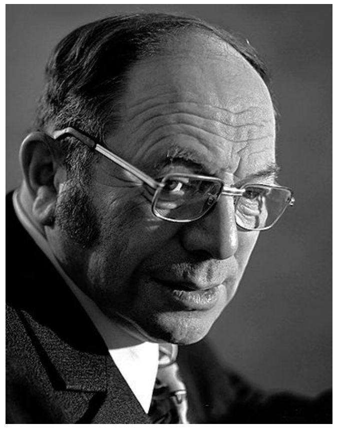
$$\Pi(\mathbf{a},\mathbf{b}) = \{\mathbf{T} \in \mathbb{R}_+^{n \times m};\ \forall (i,j), \sum_j T_{ij} = a_i, \sum_i T_{ij} = b_j\}$$





Linear Optimal Transport

Kantorovitch Formulation: general case



Two probability distributions

$\pi \in \mathcal{P}(\mathcal{X}), \ \pi' \in \mathcal{P}(\mathcal{Y})$

A cost function

$$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

Example: $\underline{\mathcal{X} = \mathcal{Y} = \mathbb{R}^d}$

Wasserstein distance

$$W_q^q(\pi, \pi') = \min_{T \in \Pi(\pi, \pi')} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|_2^q \mathrm{d}T(\mathbf{x}, \mathbf{y})$$

$(\mathcal{P}(\mathbb{R}^d), W_q)$ is a metric space

Maximum Mean Discrepancy

Kernel mean embedding and MMD

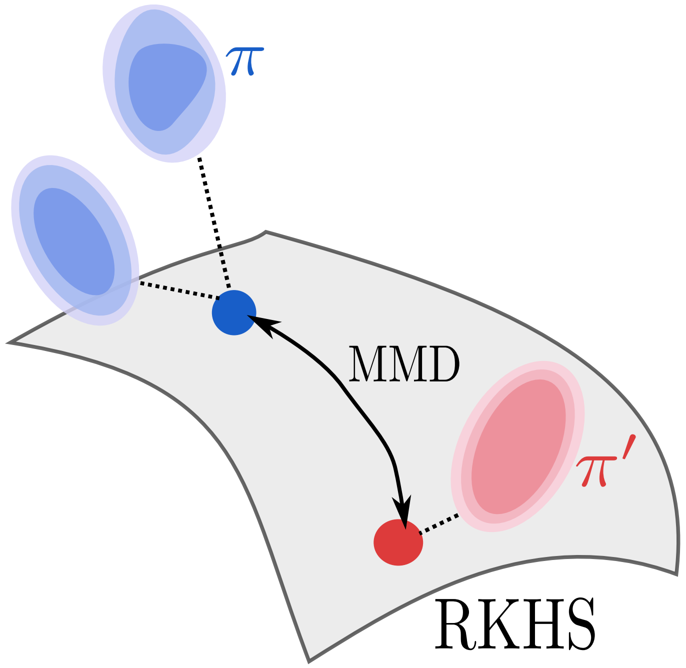
$\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ p.s.d kernel

$$\mathrm{MMD}^2_\kappa(\pi, \pi') := \int\int \kappa(\mathbf{x}, \mathbf{y}) \mathrm{d}(\pi - \pi')(\mathbf{x}) \mathrm{d}(\pi - \pi')(\mathbf{y})$$

Defines a (pseudo)metric

Distance in the RKHS after embedding of the distrib.

$$\| \int \kappa(\mathbf{x}, \cdot) \mathrm{d}\pi(\mathbf{x}) - \int \kappa(\mathbf{x}, \cdot) \mathrm{d}\pi'(\mathbf{x}) \|_{\mathcal{H}_\kappa}$$



Maximum Mean Discrepancy

Kernel mean embedding and MMD

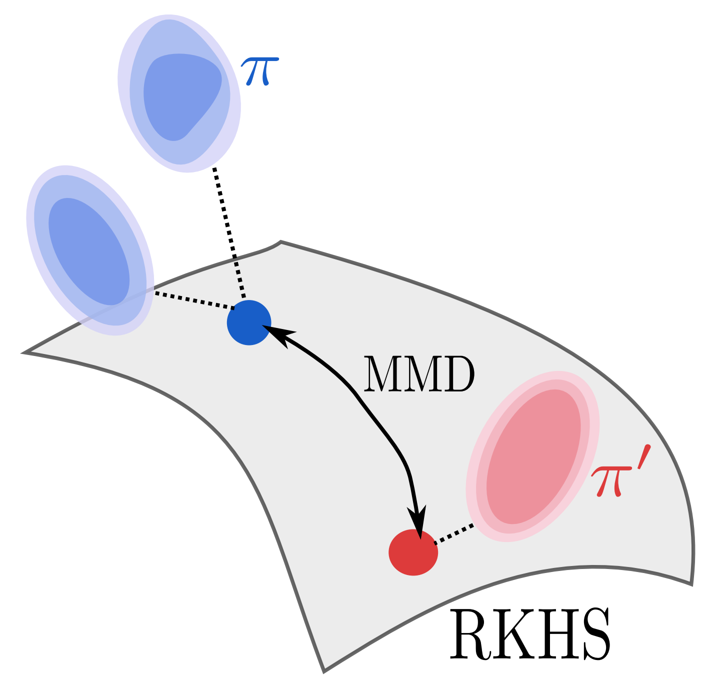
$\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ p.s.d kernel

$$\mathrm{MMD}_\kappa^2(\pi, \pi') := \int\int \kappa(\mathbf{x}, \mathbf{y}) \mathrm{d}(\pi - \pi')(\mathbf{x}) \mathrm{d}(\pi - \pi')(\mathbf{y})$$

Defines a (pseudo)metric

Distance in the RKHS after embedding of the distrib.

$$\| \int \kappa(\mathbf{x}, \cdot) \mathrm{d}\pi(\mathbf{x}) - \int \kappa(\mathbf{x}, \cdot) \mathrm{d}\pi'(\mathbf{x}) \|_{\mathcal{H}_\kappa}$$



Translation Invariant kernels (TI)

A p.s.d kernel is TI $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$

$\iff \kappa(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\boldsymbol{\omega} \sim \Lambda}[e^{-i\boldsymbol{\omega}^\top \mathbf{x}} e^{i\boldsymbol{\omega}^\top \mathbf{y}}]$ (Bochner)

Maximum Mean Discrepancy

Kernel mean embedding and MMD

$\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ p.s.d kernel

$$\mathrm{MMD}^2_\kappa(\pi, \pi') := \int\int \kappa(\mathbf{x}, \mathbf{y}) \mathrm{d}(\pi - \pi')(\mathbf{x}) \mathrm{d}(\pi - \pi')(\mathbf{y})$$

Defines a (pseudo)metric

Distance in the RKHS after embedding of the distrib.

$$\|\int \kappa(\mathbf{x}, \cdot)\mathrm{d}\pi(\mathbf{x}) - \int \kappa(\mathbf{x}, \cdot)\mathrm{d}\pi'(\mathbf{x})\|_{\mathcal{H}_\kappa}$$

Translation Invariant kernels (TI)

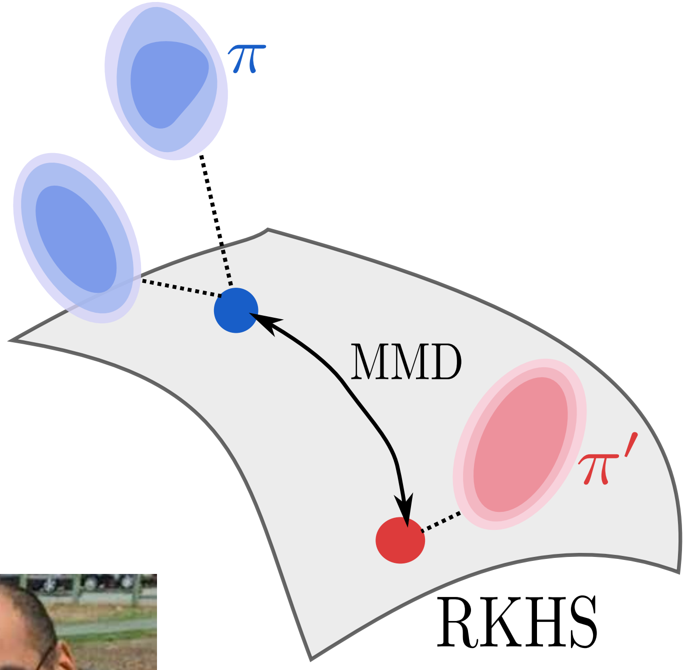
A p.s.d kernel is TI $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$

$$\iff \kappa(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\boldsymbol{\omega} \sim \Lambda}[e^{-i\boldsymbol{\omega}^\top \mathbf{x}} e^{i\boldsymbol{\omega}^\top \mathbf{y}}]$$ (Bochner)

Sample $\boldsymbol{\omega}_i \sim \Lambda, 1 \leq i \leq m$

$$\implies \kappa(\mathbf{x}, \mathbf{y}) \approx \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_{\mathbb{R}^m}$$

$\Phi(\mathbf{x}) = \frac{1}{\sqrt{m}} \left(\exp(-i\boldsymbol{\omega}_i^\top \mathbf{x})\right)_{i=1}^m$ Random Fourier Features



From Statistical Learning to Compressive Statistical Learning

From statistical learning…

Notations

Given data points $\mathbf{x}_i \sim \pi;\ 1 \leq i \leq n$

A hypothesis space $h \in \mathcal{H}$

Loss function $\ell : \mathcal{X} \times \mathcal{H} \to \mathbb{R}$

Find the best $h \in \mathcal{H}$ on the data

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{x} \sim \pi}[\ell(\mathbf{x}, h)]$$

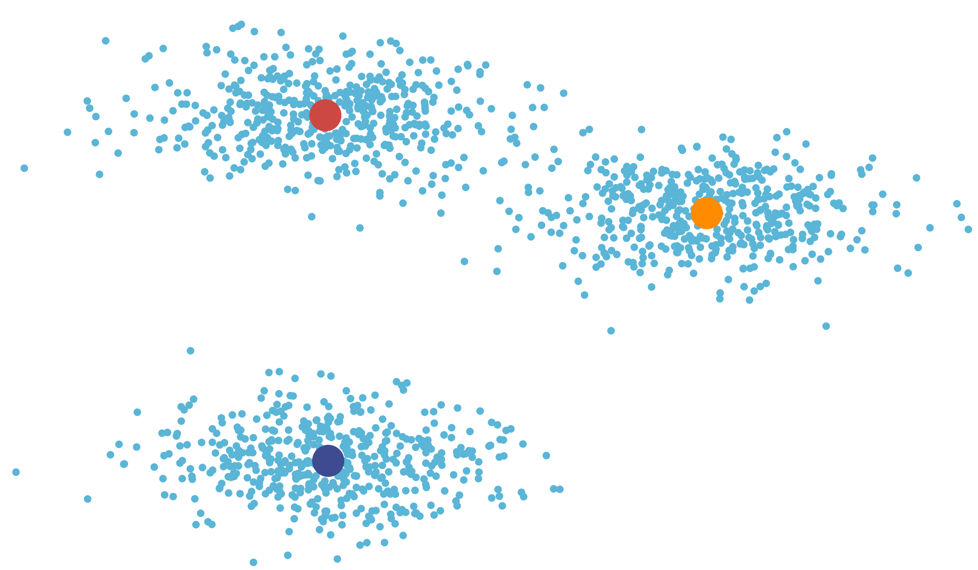
From statistical learning...

Notations

Given data points	$\mathbf{x}_i \sim \pi;\ 1 \leq i \leq n$
A hypothesis space	$h \in \mathcal{H}$
Loss function	$\ell : \mathcal{X} \times \mathcal{H} \rightarrow \mathbb{R}$

Find the best $h \in \mathcal{H}$ on the data

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{x} \sim \pi}[\ell(\mathbf{x}, h)]$$



$\mathbf{x}_i \in \mathbb{R}^d$

$h = (\mathbf{c}_1, \cdots, \mathbf{c}_K), \mathbf{c}_k \in \mathbb{R}^d$

$\ell(\mathbf{x}, h) = \min_{k \in [\![K]\!]} \|\mathbf{x} - \mathbf{c}_k\|_2^2$

From statistical learning…

Notations

Given data points	$\mathbf{x}_i \sim \pi;\ 1 \leq i \leq n$
A hypothesis space	$h \in \mathcal{H}$
Loss function	$\ell : \mathcal{X} \times \mathcal{H} \to \mathbb{R}$

Find the best $h \in \mathcal{H}$ on the data

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{x} \sim \pi}[\ell(\mathbf{x}, h)]$$

We do not have access to π

Empirical risk minimization

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, h)$$

From statistical learning…

Notations

Risk: $$\mathcal{R}(\pi, h) = \mathbb{E}_{\mathbf{x}\sim\pi}[\ell(\mathbf{x}, h)]$$

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

From statistical learning…

Notations

Risk: $\mathcal{R}(\pi, h) = \mathbb{E}_{\mathbf{x}\sim\pi}[\ell(\mathbf{x}, h)]$

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

Selected hypothesis

$$\hat{h} \in \arg\min_{h\in\mathcal{H}} \mathcal{R}(\pi_n, h)$$

Best hypothesis

$$h^* \in \arg\min_{h\in\mathcal{H}} \mathcal{R}(\pi, h)$$

From statistical learning…

Notations

Risk: $\mathcal{R}(\pi, h) = \mathbb{E}_{\mathbf{x} \sim \pi}[\ell(\mathbf{x}, h)]$

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

Selected hypothesis

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\pi_n, h)$$

Best hypothesis

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\pi, h)$$

Ultimate goal: small excess-risk

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \leq \eta_n \quad \text{w.h.p.}$$

Typically $\eta_n = O(\frac{1}{\sqrt{n}})$ or better [Shalev-Shwartz and Ben-David, 2014]

From statistical learning…

Ultimate goal: small excess-risk

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \leq \eta_n \quad \text{w.h.p.}$$

How to obtain these bounds -> control of the following

Central quantity:

$$\text{TaskMetric}(\pi, \pi') := \sup_{h \in \mathcal{H}} |\mathcal{R}(\pi, h) - \mathcal{R}(\pi', h)|$$

Defines a (pseudo) metric between probability distrib.

Depends on the learning task -> « task metric »

... to Compressive Statistical learning (CSL)

Problem

Finding $\hat{h}$ is often quite expensive in modern applications

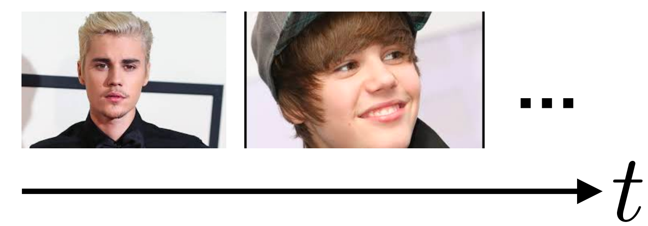
Very large dataset

$n >> 12$

Need to query the full training dataset many times (e.g. GD/SGD).

Distributed data

Streaming data



Algorithms need to adapt to these settings

… to Compressive Statistical learning (CSL)

Problem

Finding $\hat{h}$ is often quite expensive in modern applications

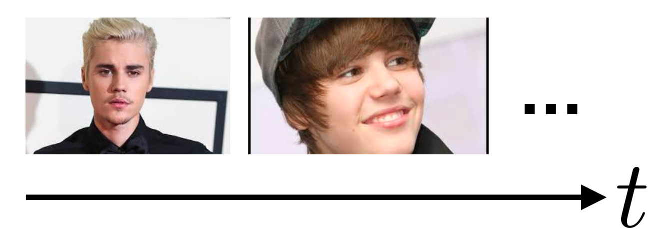
Very large dataset

$n >> 12$

Need to query the full training dataset many times (e.g. GD/SGD).

Distributed data

Streaming data

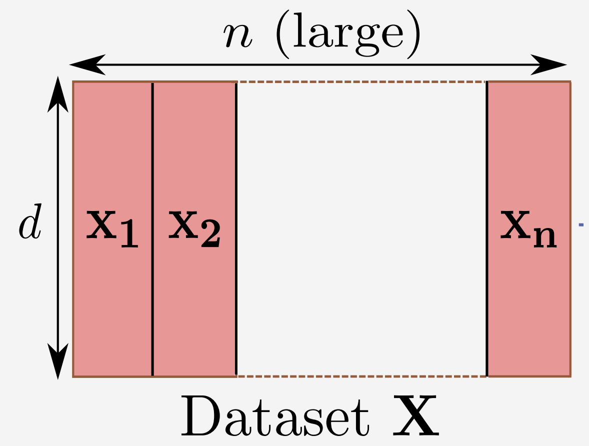


Algorithms need to adapt to these settings

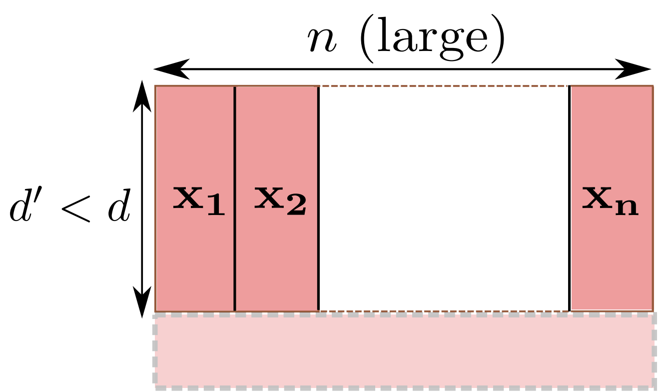
Compression ?

… to Compressive Statistical learning (CSL)

Find a small & faithfull representation of the data

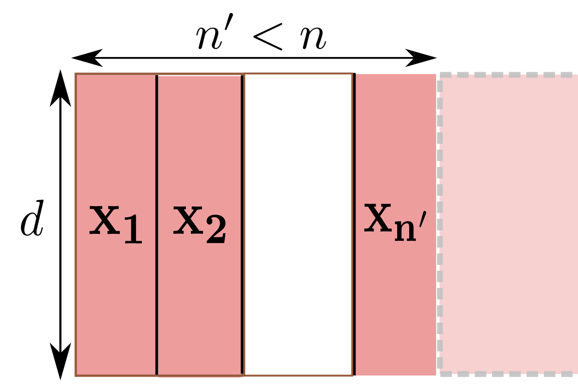


Dimension reduction



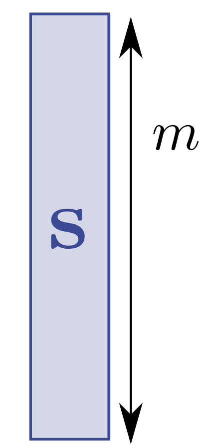
- Random projection (JL lemma)
- Feature selection
- Minimum distortion embedding, PCA

Subsampling



- Coresets
- Importance sampling

Here: linear « sketch »

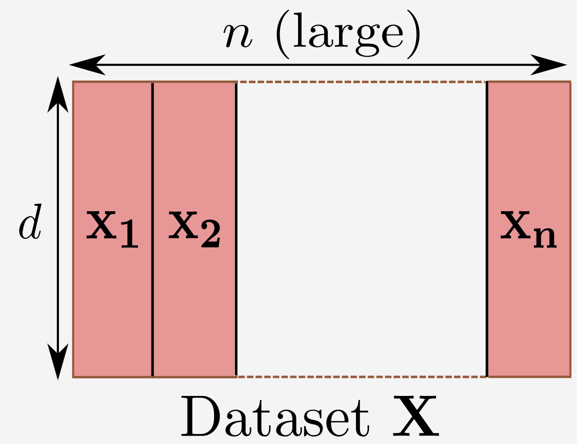


- Only one vector

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

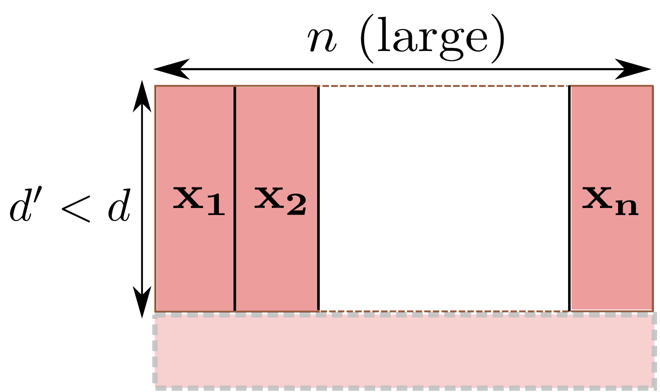
… to Compressive Statistical learning (CSL)

Find a small & faithfull representation of the data



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

Dimension reduction

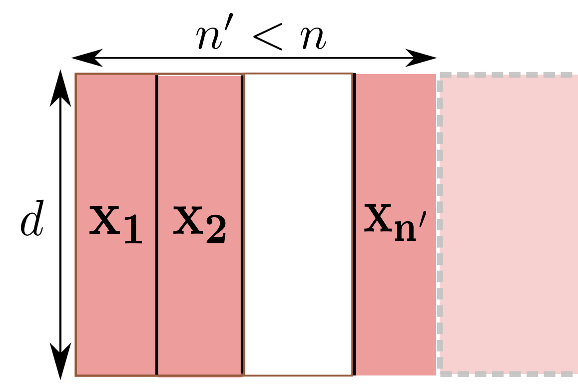


Random projection (JL lemma)

Feature selection

Minimum distortion embedding, PCA

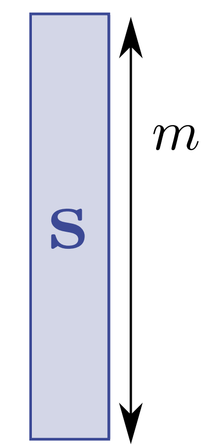
Subsampling



Coresets

Importance sampling

Here: linear « sketch »



Only one vector

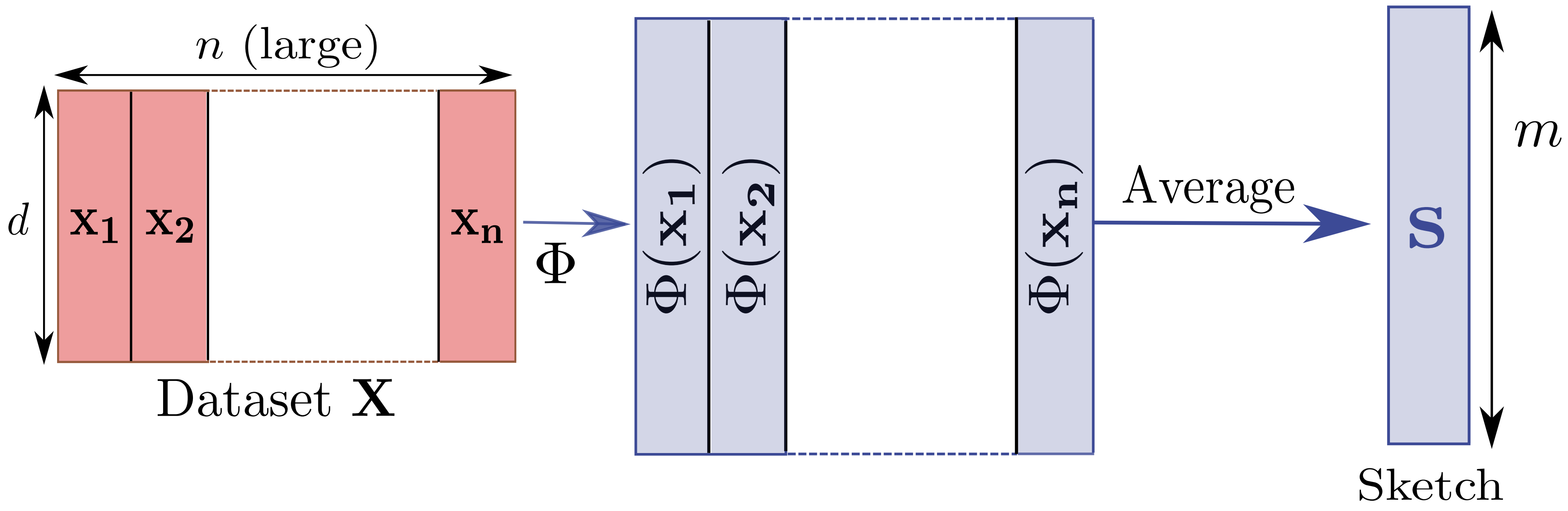
[Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, Yann Traonmilin, Antoine Chatalic, Vincent Schellekens, Laurent Jacques…]

… to Compressive Statistical learning (CSL)

Sketching

$\Phi : \mathcal{X} \to \mathbb{R}^m$ feature operator

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



… to Compressive Statistical learning (CSL)

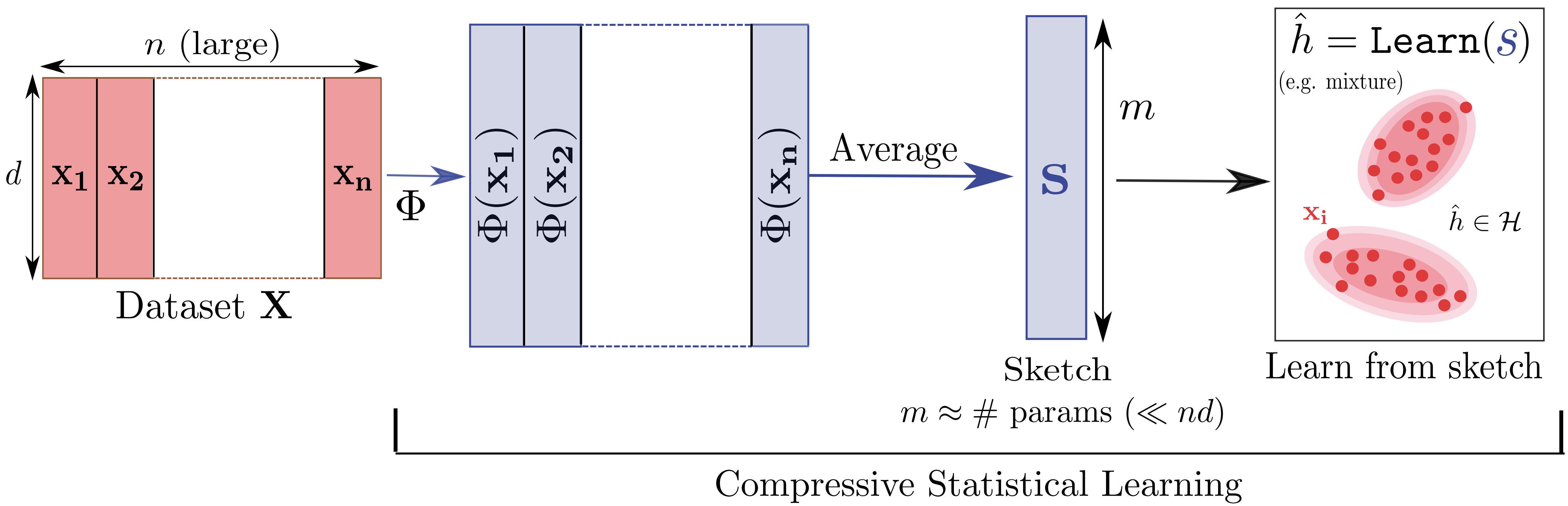
Sketching

$\Phi : \mathcal{X} \to \mathbb{R}^m$ feature operator

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

Pros

- Streaming + distributed scenario
- Storage



... to Compressive Statistical learning (CSL)

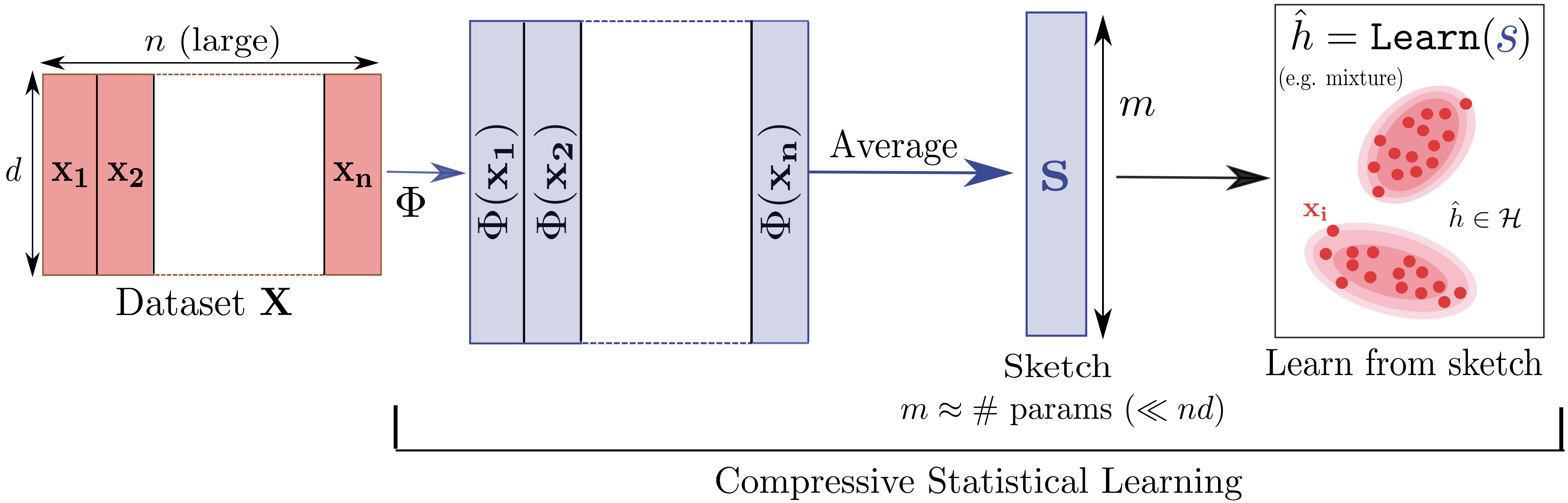
Sketching

$\Phi : \mathcal{X} \to \mathbb{R}^m$ feature operator

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

Pros

- Streaming + distributed scenario
- Storage



Random Fourier Features (RFF) [Rahimi and Recht, 2008]

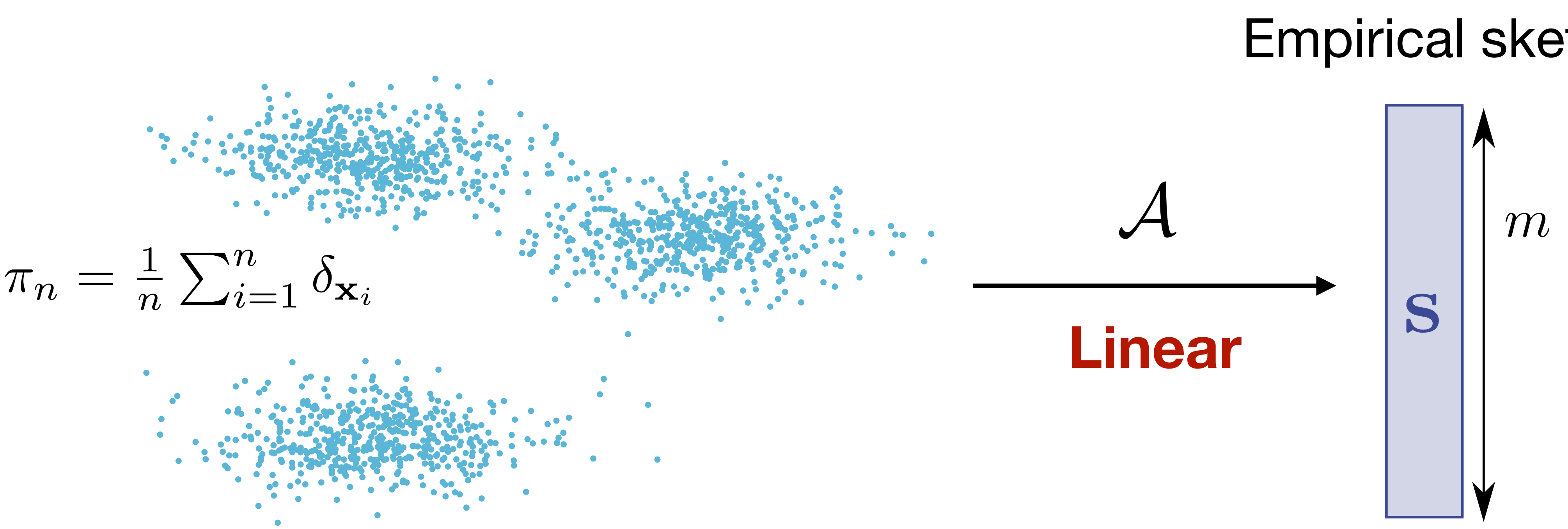
$$\Phi(\mathbf{x}) = \frac{1}{\sqrt{m}}(\exp(-i\boldsymbol{\omega}_1^\top \mathbf{x}), \cdots, \exp(-i\boldsymbol{\omega}_m^\top \mathbf{x}))^\top \quad \boldsymbol{\omega}_i \sim \Lambda \text{ i.i.d.}$$

Towards CSL guarantees: 1) Learn from sketch

Sketching operator

$$\mathcal{A} : \pi \to \mathcal{A}(\pi) := \int_{\mathcal{X}} \Phi(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) \in \mathbb{R}^m$$

Finite dimensional Mean Map embedding

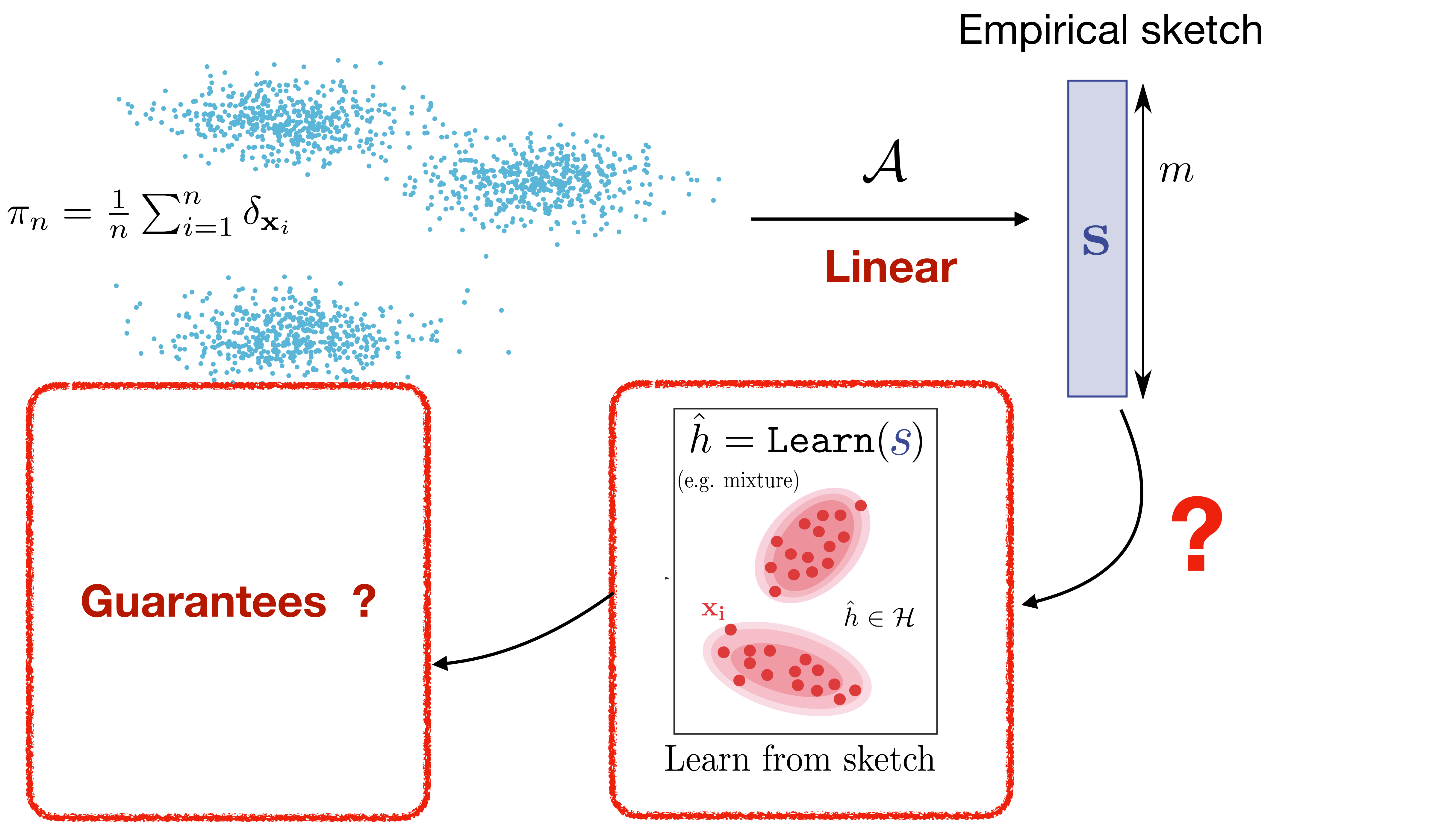


Towards CSL guarantees: 1) Learn from sketch

Sketching operator

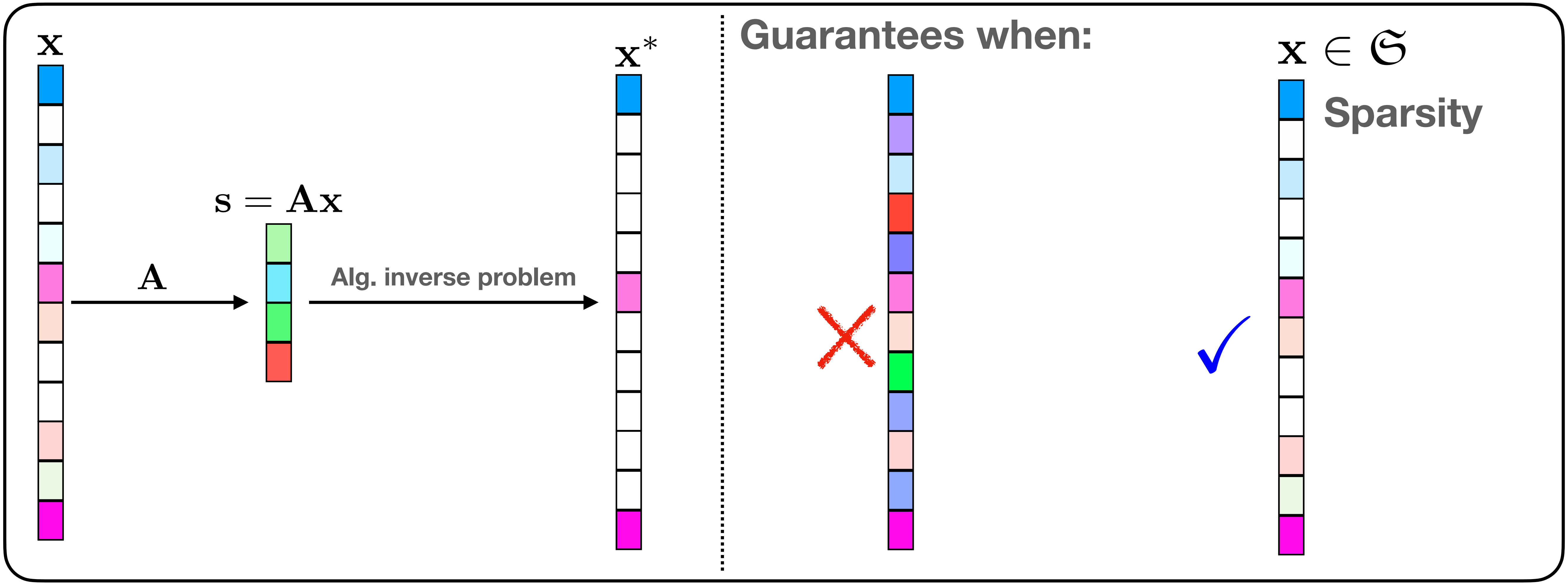
$$\mathcal{A} : \pi \to \mathcal{A}(\pi) := \int_{\mathcal{X}} \Phi(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) \in \mathbb{R}^m$$

Finite dimensional Mean Map embedding



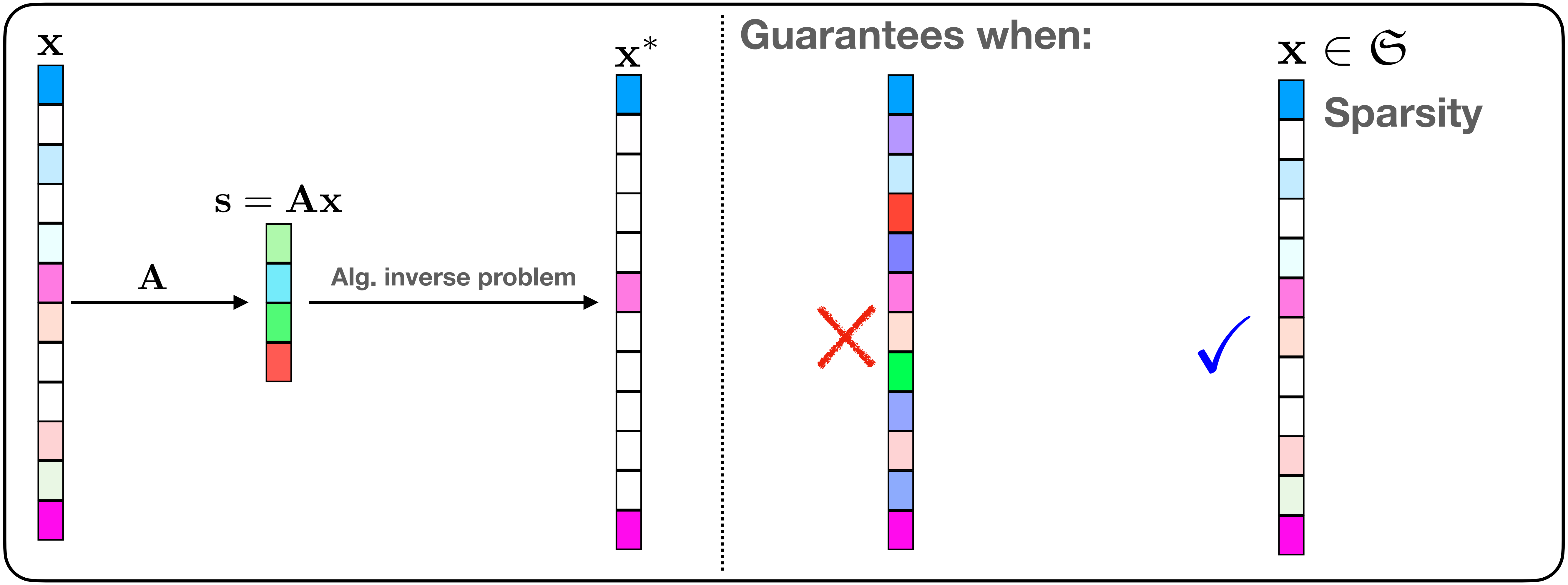
Towards CSL guarantees: 1) Learn from sketch

Analogy with compressed sensing:



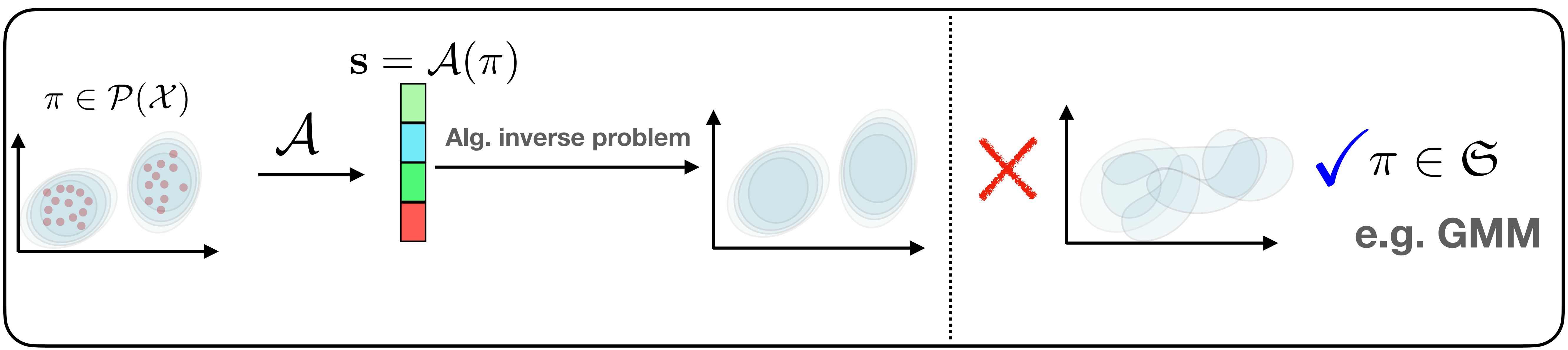
Towards CSL guarantees: 1) Learn from sketch

Analogy with compressed sensing:



…Need a « low-dimensional » model

Idea here

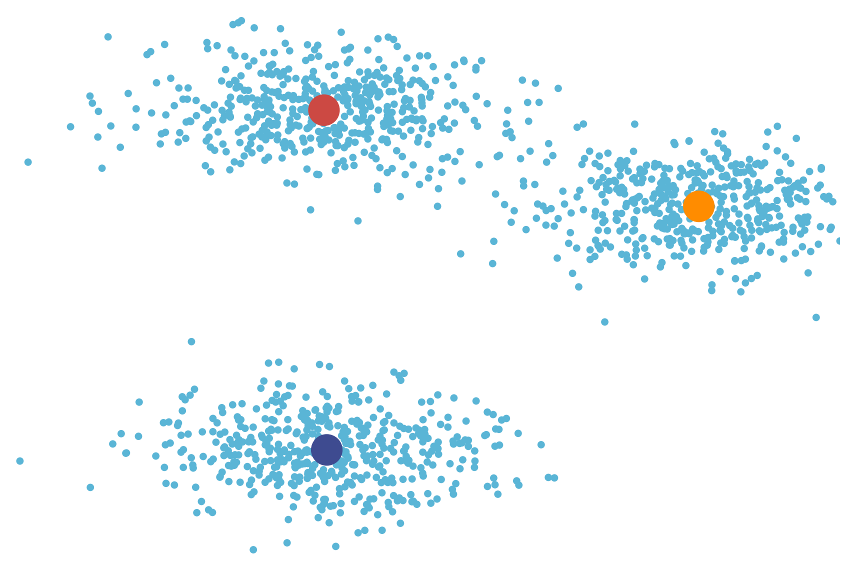


Towards CSL guarantees: 1) Learn from sketch

Learn from sketch $\hat{h} = \texttt{Learn}(s)$

K-means

Model set of distributions. $\pi_{\boldsymbol{\theta}} \in \left\{ \frac{1}{K} \sum_{k=1}^{K} \delta_{\mathbf{c}_k} \right\}$ $\boldsymbol{\theta} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$



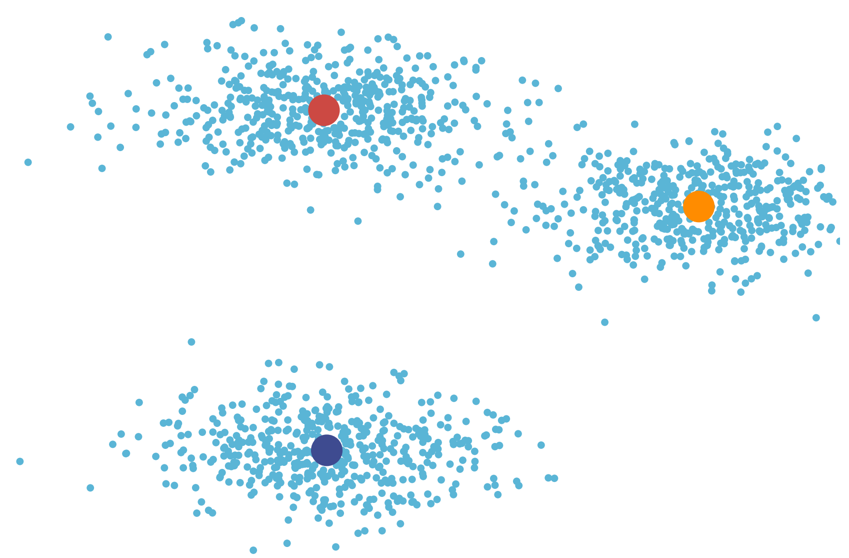
Towards CSL guarantees: 1) Learn from sketch

Learn from sketch $\hat{h} = \texttt{Learn}(s)$

K-means

Model set of distributions. $\pi_{\boldsymbol{\theta}} \in \left\{ \frac{1}{K} \sum_{k=1}^{K} \delta_{\mathbf{c}_k} \right\}$ $\boldsymbol{\theta} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$

Solve: $$\min_{\boldsymbol{\theta}} \|\mathbf{s} - \mathcal{A}(\pi_{\boldsymbol{\theta}})\|_2$$



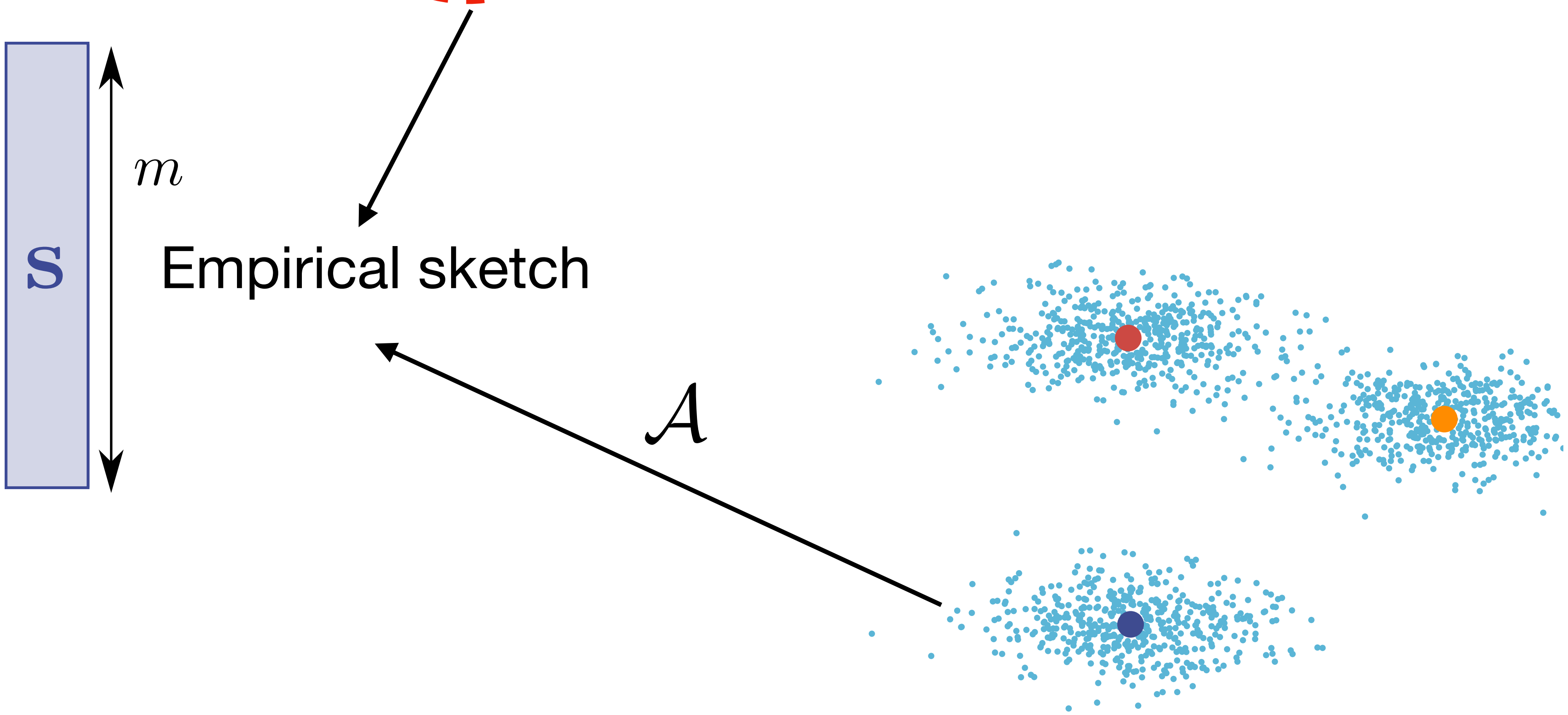
Towards CSL guarantees: 1) Learn from sketch

Learn from sketch $\hat{h} = \texttt{Learn}(s)$

K-means

Model set of distributions. $\pi_{\boldsymbol{\theta}} \in \left\{ \frac{1}{K} \sum_{k=1}^{K} \delta_{\mathbf{c}_k} \right\}$ $\boldsymbol{\theta} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$

Solve: $\min_{\boldsymbol{\theta}} \|\mathbf{s} - \mathcal{A}(\pi_{\boldsymbol{\theta}})\|_2$



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

Towards CSL guarantees: 1) Learn from sketch

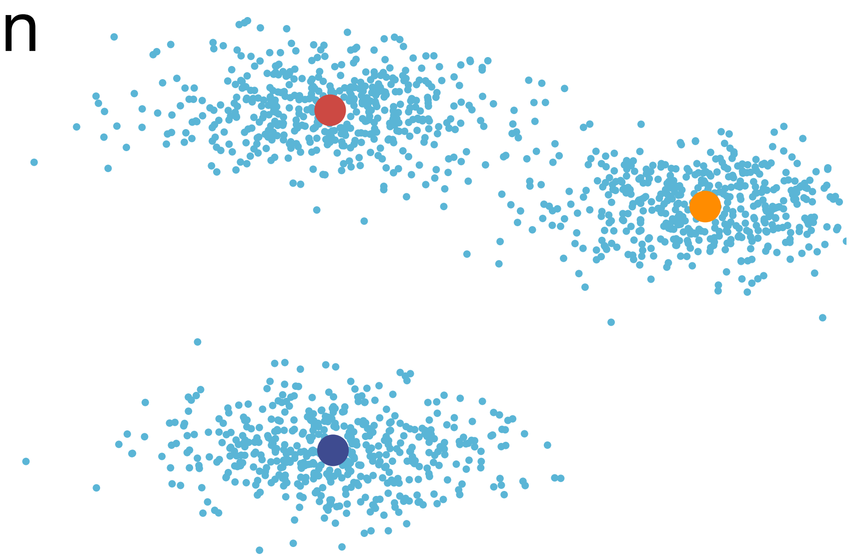
Learn from sketch $\hat{h} = \texttt{Learn}(s)$

K-means

Model set of distributions. $\pi_{\boldsymbol{\theta}} \in \left\{ \frac{1}{K} \sum_{k=1}^{K} \delta_{\mathbf{c}_k} \right\}$ $\boldsymbol{\theta} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$

Solve: $\min_{\boldsymbol{\theta}} \|\mathbf{s} - \mathcal{A}(\pi_{\boldsymbol{\theta}})\|_2$

Sketch of the parametrized distribution



Towards CSL guarantees: 1) Learn from sketch

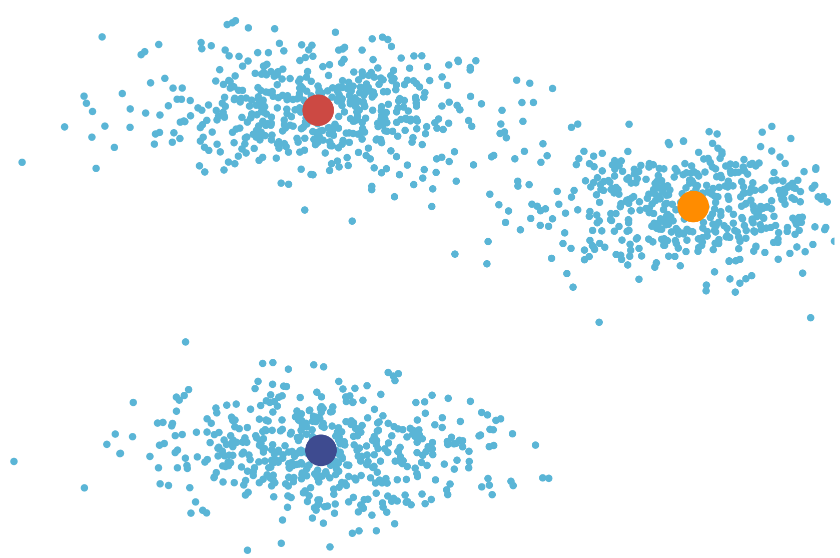
Learn from sketch $\hat{h} = \mathtt{Learn}(s)$

K-means

Model set of distributions. $\pi_{\boldsymbol{\theta}} \in \left\{ \frac{1}{K} \sum_{k=1}^{K} \delta_{\mathbf{c}_k} \right\}$ $\boldsymbol{\theta} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$

Solve: $\min_{\boldsymbol{\theta}} \|\mathbf{s} - \mathcal{A}(\pi_{\boldsymbol{\theta}})\|_2$

Find the distribution whose sketch
is the closest to the empirical sketch



Towards CSL guarantees: 1) Learn from sketch

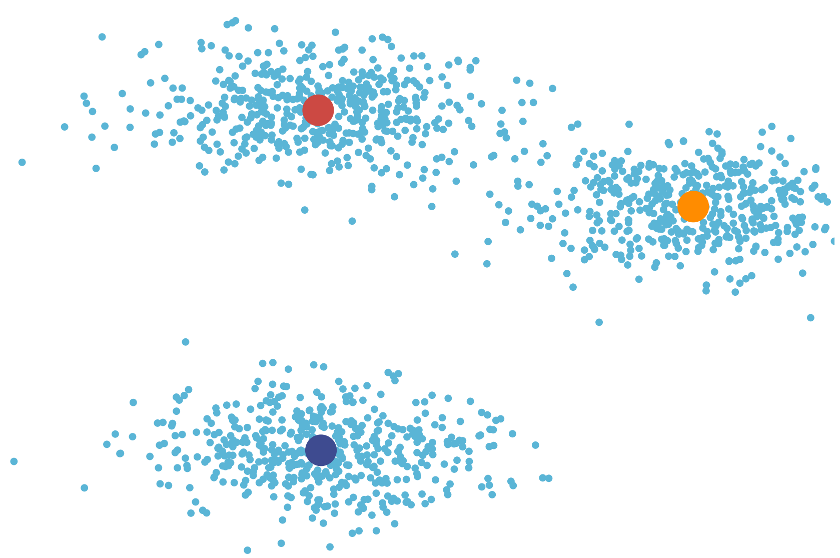
Learn from sketch $\hat{h} = \texttt{Learn}(s)$

K-means

Model set of distributions. $\pi_{\boldsymbol{\theta}} \in \left\{ \frac{1}{K} \sum_{k=1}^{K} \delta_{\mathbf{c}_k} \right\}$ $\boldsymbol{\theta} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$

Solve: $\min_{\boldsymbol{\theta}} \|\mathbf{s} - \mathcal{A}(\pi_{\boldsymbol{\theta}})\|_2$

Find the distribution whose sketch
is the closest to the empirical sketch



Return $\hat{h} = \boldsymbol{\theta}^* = (\mathbf{c}_1, \cdots, \mathbf{c}_K)$

Towards CSL guarantees: 1) Learn from sketch

Learn from sketch: $\hat{h} = \texttt{Learn}(s)$

Design a model set of distrib. $\pi \in \mathfrak{S} \subseteq \mathcal{P}(\mathcal{X})$ **Step 1**

Solve a moment matching prob. (inverse, preimage prob.)

Step 2

$$\Delta[\mathrm{s}] \in \arg\min_{\pi \in \mathfrak{S}} \|\mathrm{s} - \mathcal{A}(\pi)\|_2$$

Find the hypothesis

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathrm{s}], h)$$

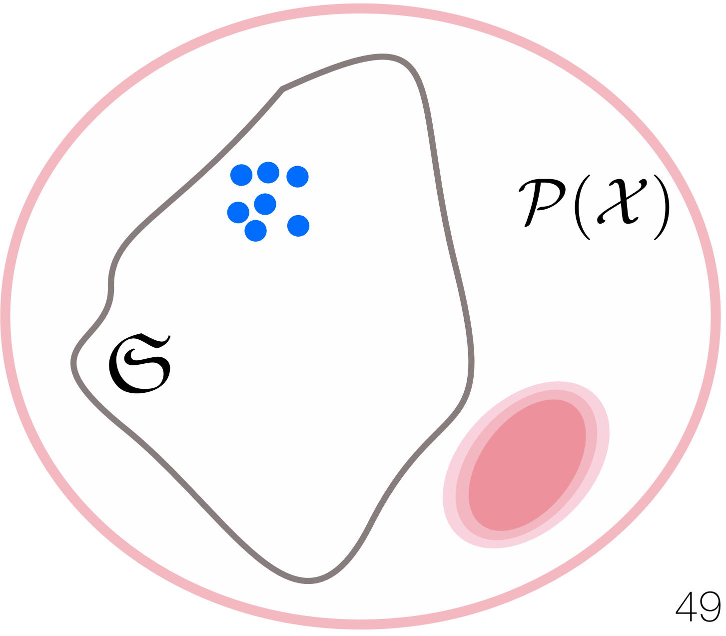
Step 3

Prior on the « true » distribution

K-means = K-sparse, GMM = mixture of Gaussian…

Related to the learning problem

Small « complexity » -> learnable with sketch



Towards CSL guarantees: 1) Learn from sketch

Learn from sketch: $\hat{h} = \texttt{Learn}(s)$

Design a model set of distrib. $\pi \in \mathfrak{S} \subseteq \mathcal{P}(\mathcal{X})$

Step 1

Solve a moment matching prob. (inverse prob. $\approx$ **compressed sensing**)

Step 2

$$\Delta[\mathbf{s}] \in \arg\min_{\pi \in \mathfrak{S}} \|\mathbf{s} - \mathcal{A}(\pi)\|_2$$

Find the hypothesis

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathrm{s}], h)$$

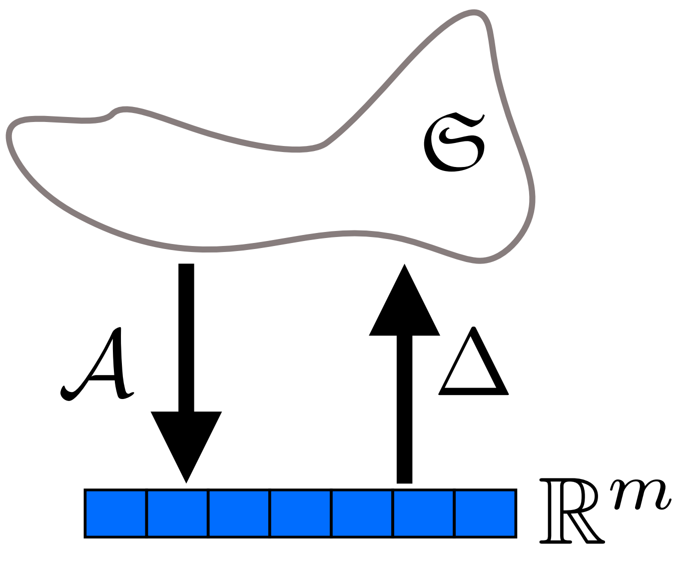
Step 3

Inverse problem in the space of measure

Beurling LASSO, CLOMP, super-resolution, flows …

[Candès, Keriven, De Castro, Poon, Peyré, Denoyelle, Duval, Chizat, Boyd …]

Δ **is the decoder** $\mathbb{R}^m \to \mathfrak{S}$



Towards CSL guarantees: 1) Learn from sketch

Learn from sketch: $\hat{h} = \texttt{Learn}(s)$

Design a model set of distrib. $\pi \in \mathfrak{S} \subseteq \mathcal{P}(\mathcal{X})$ **Step 1**

Solve a moment matching prob. (inverse prob. $\approx$ **compressed sensing**)

Step 2

$$\Delta[\mathbf{s}] \in \arg\min_{\pi \in \mathfrak{S}} \|\mathbf{s} - \mathcal{A}(\pi)\|_2$$

Find the hypothesis

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$$

Step 3

Usually easier than ERM !

K-means, GMM: comes for free

Uses that $\Delta[\mathbf{s}]$ is in a low complexity model set

Towards CSL guarantees: 2) Theoretical guarantees

Why should it work ? Goal:

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \leq \eta_n \quad \text{w.h.p.}$$

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\pi_n, h)$$ (struck through)

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$$

Towards CSL guarantees: 2) Theoretical guarantees

Why should it work ? Goal:

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \leq \eta_n \quad \text{w.h.p.}$$

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\pi_n, h) \qquad \hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$$

Lower Restricted Isometric Property (LRIP)

$$\forall \pi, \pi' \in \mathfrak{S}, \text{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$$

Towards CSL guarantees: 2) Theoretical guarantees

Why should it work ? Goal:

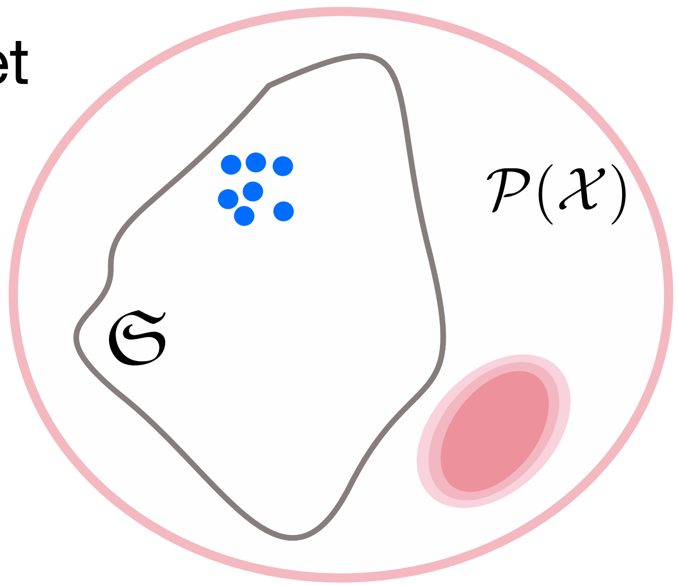
$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \leq \eta_n \quad \text{w.h.p.}$$

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\pi_n, h) \qquad \hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$$

Lower Restricted Isometric Property (LRIP)

$$\forall \pi, \pi' \in \mathfrak{S}, \ \text{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$$

Model set



Towards CSL guarantees: 2) Theoretical guarantees

Why should it work ? Goal:

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \leq \eta_n \quad \text{w.h.p.}$$

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\pi_n, h) \qquad \hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$$

Lower Restricted Isometric Property (LRIP)

$$\forall \pi, \pi' \in \mathfrak{S}, \text{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$$

Task-metric (we want to control it)

Towards CSL guarantees: 2) Theoretical guarantees

Why should it work ? Goal:

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \leq \eta_n \quad \text{w.h.p.}$$

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\pi_n, h)$$ (crossed out)

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$$

Lower Restricted Isometric Property (LRIP)

$$\forall \pi, \pi' \in \mathfrak{S}, \text{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$$

Distance between the sketches of the distrib.

Towards CSL guarantees: 2) Theoretical guarantees

Lower Restricted Isometric Property (LRIP)

$$\Downarrow$$

Statistical Guarantees $\underline{\forall \pi \in \mathcal{P}(\mathcal{X})}$

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim d^\circ(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2$$

Towards CSL guarantees: 2) Theoretical guarantees

Lower Restricted Isometric Property (LRIP)

$\Downarrow$

Statistical Guarantees $\forall \pi \in \mathcal{P}(\mathcal{X})$

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim d^\circ(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2$$

Excess-risk

$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$

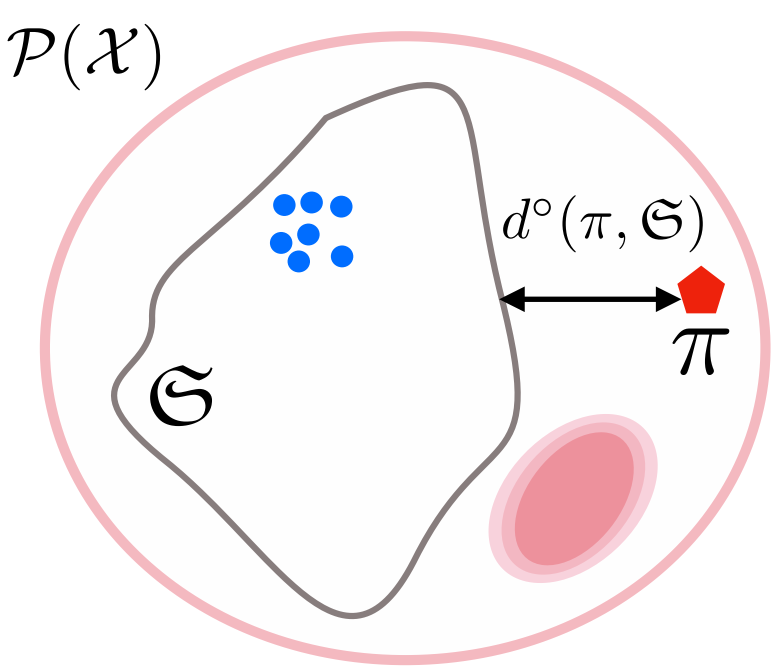
Towards CSL guarantees: 2) Theoretical guarantees

Lower Restricted Isometric Property (LRIP)

$$\Downarrow$$

Statistical Guarantees $\forall \pi \in \mathcal{P}(\mathcal{X})$

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim d^\circ(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2$$



Notion of **distance** to the model set

Bias term: vanishes when the true distrib. in the model

$$\pi \in \mathfrak{S} \implies d^\circ(\pi, \mathfrak{S}) = 0$$

$\approx$ **Approximation error is SL**

Towards CSL guarantees: 2) Theoretical guarantees

Lower Restricted Isometric Property (LRIP)

$$\Downarrow$$

Statistical Guarantees $\underline{\forall \pi \in \mathcal{P}(\mathcal{X})}$

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim d^\circ(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2$$

$\mathbf{s} = \mathcal{A}(\pi_n)$

Distance between the empirical and true sketch

Basically converges to zero in $O(\frac{1}{\sqrt{n}})$

$\approx$ **Estimation error is SL**

Towards CSL guarantees: 2) Theoretical guarantees

Lower Restricted Isometric Property (LRIP)

$\Downarrow$

Statistical Guarantees

How to obtain the LRIP = DIFFICULT

Towards CSL guarantees: 3) The LRIP

Setting $\underline{\mathcal{X} = \mathbb{R}^d}$ $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ $\Phi = \mathrm{RFF}$

How to prove the LRIP

Step 1 $\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \mathrm{MMD}_\kappa(\pi, \pi')$ **Kernel LRIP**

Step 2 $\forall \pi, \pi' \in \mathfrak{S}, \mathrm{MMD}_\kappa(\pi, \pi') \approx \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$ m large enough

Towards CSL guarantees: 3) The LRIP

Setting $\underline{\mathcal{X} = \mathbb{R}^d}$ $\quad \kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y}) \quad \Phi = \mathrm{RFF}$

How to prove the LRIP

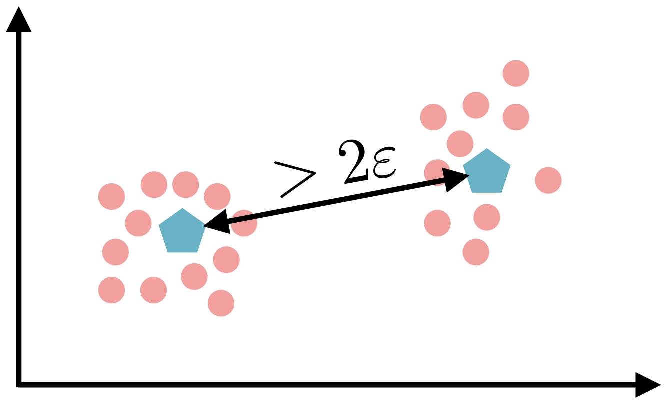
Step 1 $\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi,\pi') \lesssim \mathrm{MMD}_\kappa(\pi,\pi')$ **Kernel LRIP**

Step 2 $\forall \pi, \pi' \in \mathfrak{S}, \mathrm{MMD}_\kappa(\pi,\pi') \approx \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$ $\quad m$ large enough

Examples

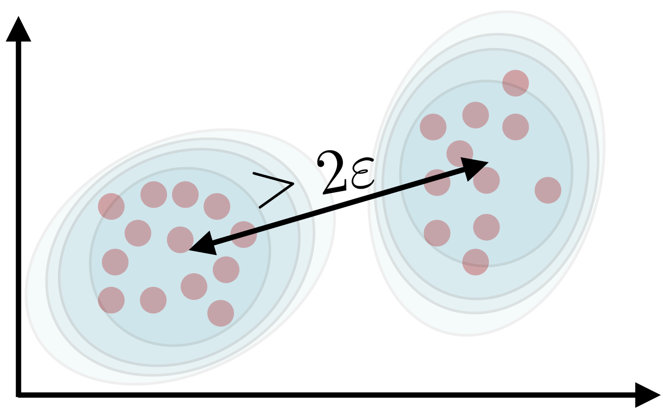
K-means $\mathfrak{S} = \{\frac{1}{K}\sum_{k=1}^K \delta_{\mathbf{c}_k}\}$

+ separability of the clusters



GMM $\mathfrak{S} = \{\sum_{k=1}^K \alpha_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})\}$

+ separability of the means



LRIP and statistical guarantees with:

$$m = O(k^2 d)$$

Optimal Transport for CSL

Optimal Transport for CSL

We will look for:

Hölder LRIP

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^{\delta}, 0 < \delta \leq 1$$

We will show

Similar statistical guarantees

Easier to obtain via optimal transport

Optimal Transport for CSL

We will look for:

Hölder LRIP

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^{\delta}, 0 < \delta \leq 1$$

We will show

Similar statistical guarantees | Easier to obtain via optimal transport

$\Downarrow$ Not so difficult

Statistical Guarantees

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim d^{\circ}(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2^{\delta}$$

Optimal Transport for CSL

We will look for:

Hölder LRIP

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^{\delta}, 0 < \delta \leq 1$$

We will show

Similar statistical guarantees | Easier to obtain via optimal transport

$\Downarrow$ Not so difficult

Statistical Guarantees

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim d^{\circ}(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2^{\delta}$$

Slow rates $O(n^{-\delta/2})$

Optimal Transport for CSL

We will look for:

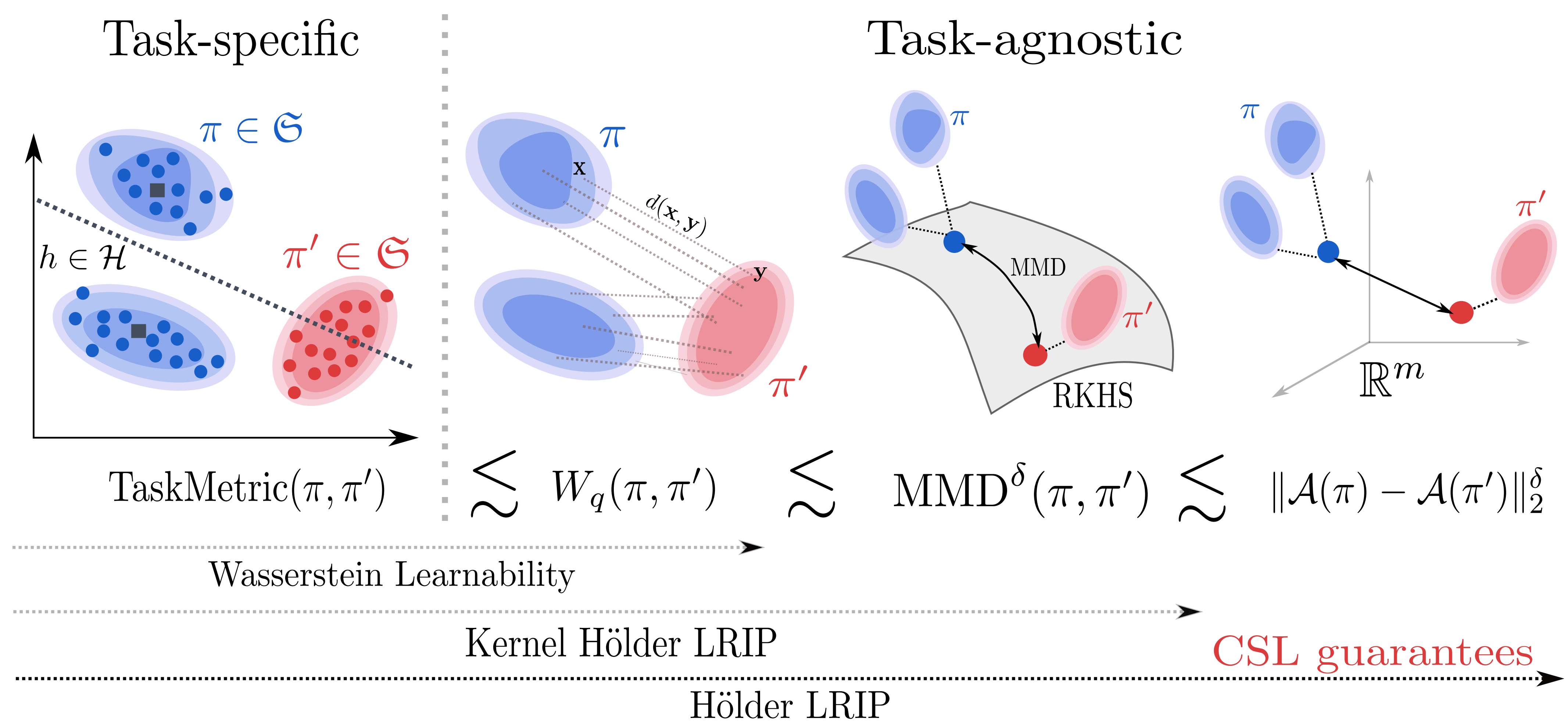
Hölder LRIP

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^{\delta}, 0 < \delta \leq 1$$

We will show

Similar statistical Guarantees | Easier to obtain via optimal transport

Roadmap



Bounding the task metric

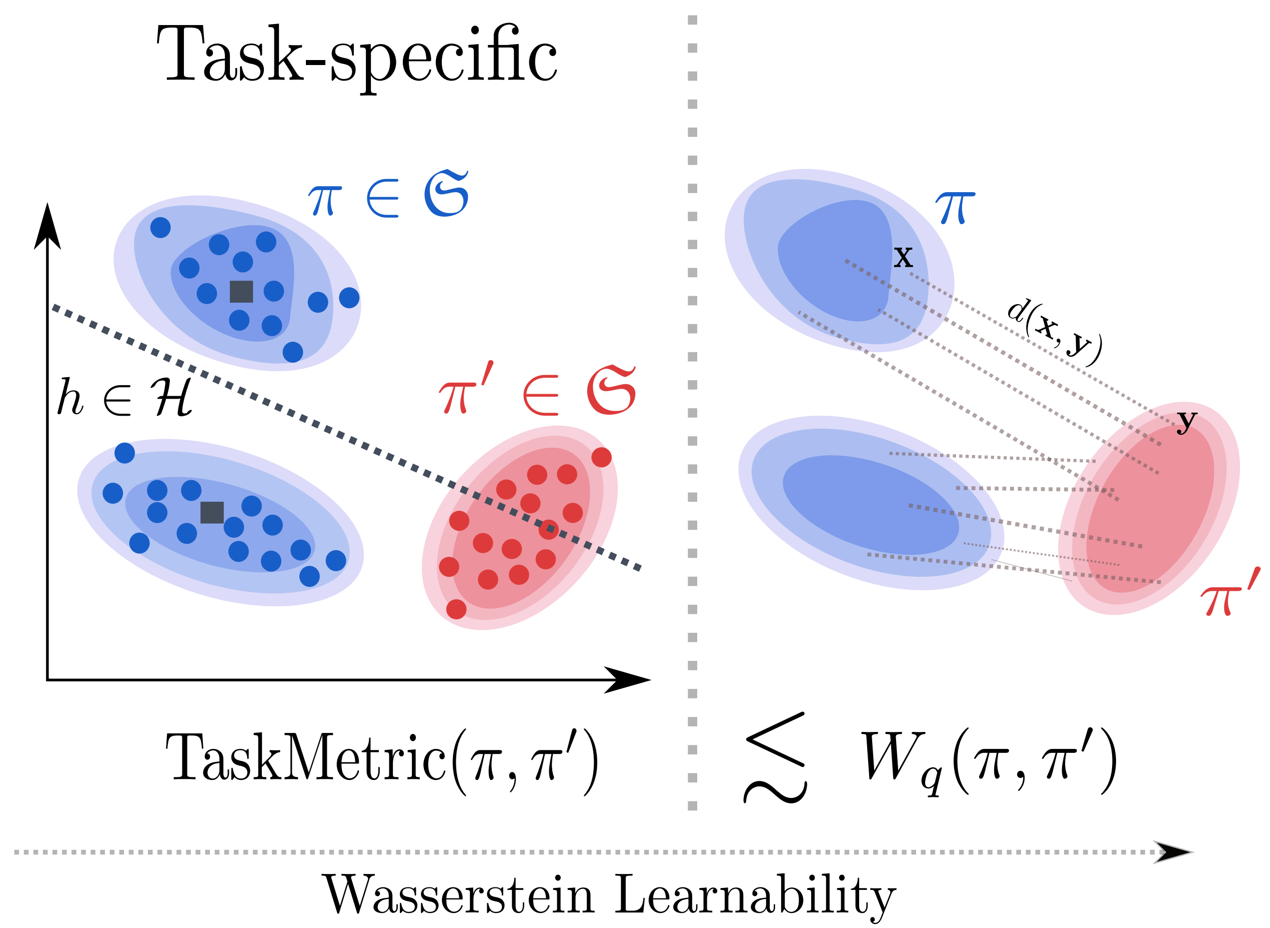
Optimal Transport for CSL: 1) Wasserstein Learnability

Goal

$$\forall \pi, \pi' \in \mathcal{P}(\mathbb{R}^d), \mathrm{TaskMetric}(\pi, \pi') \lesssim W_q(\pi, \pi')$$

Remarks

- Seems unexpected
- Depends only on the learning task



Optimal Transport for CSL: 1) Wasserstein Learnability

Wasserstein learnability

$$\forall \pi, \pi' \in \mathcal{P}(\mathbb{R}^d), \mathrm{TaskMetric}(\pi, \pi') \lesssim W_q(\pi, \pi')$$

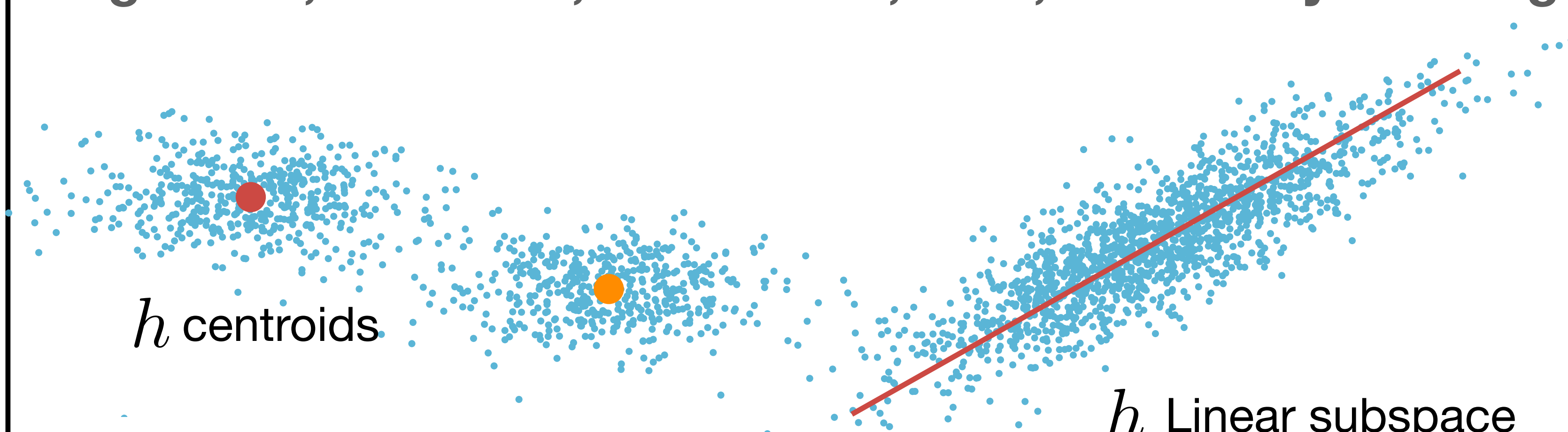
True for many unsupervised learning tasks

Compression type-tasks

- $\ell(\mathbf{x}, h) = \|\mathbf{x} - P_h(\mathbf{x})\|_2^q$
- P_h projection func.

$$\implies \mathcal{R}(\pi, h) = W_q^q(\pi, P_h \# \pi)$$

E.g.: PCA, K-means, K-medians, NMF, Dictionary learning…



Optimal Transport for CSL: 1) Wasserstein Learnability

Wasserstein learnability

$$\forall \pi, \pi' \in \mathcal{P}(\mathbb{R}^d), \mathrm{TaskMetric}(\pi, \pi') \lesssim W_q(\pi, \pi')$$

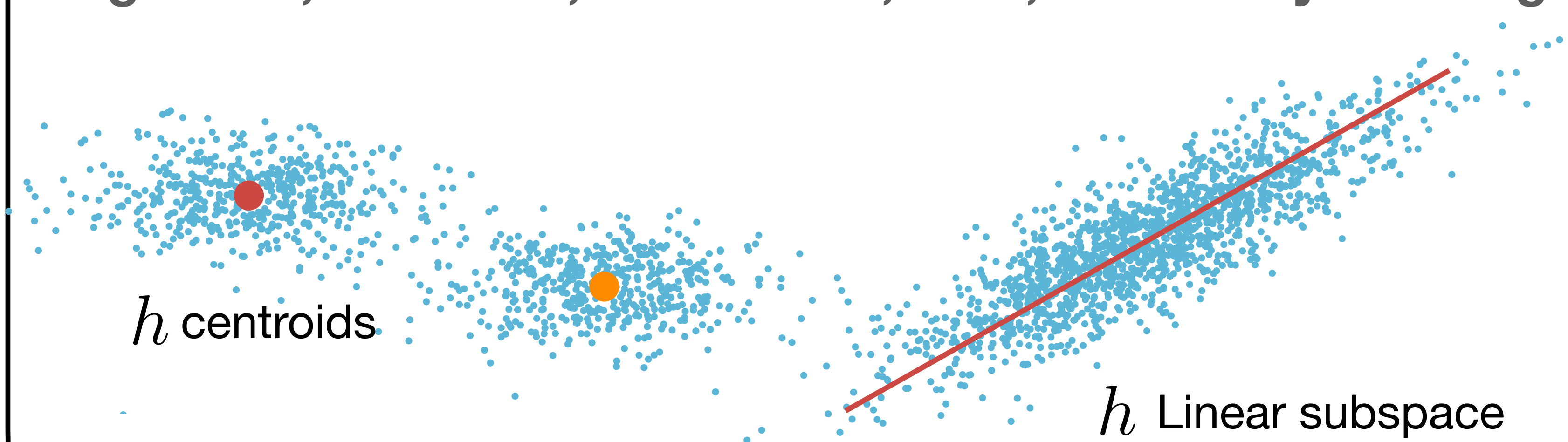
True for many unsupervised learning tasks

Compression type-tasks

- $\ell(\mathbf{x}, h) = \|\mathbf{x} - P_h(\mathbf{x})\|_2^q$
- P_h projection func.

$$\implies \mathcal{R}(\pi, h) = W_q^q(\pi, P_h \# \pi)$$

E.g.: PCA, K-means, K-medians, NMF, Dictionary learning…

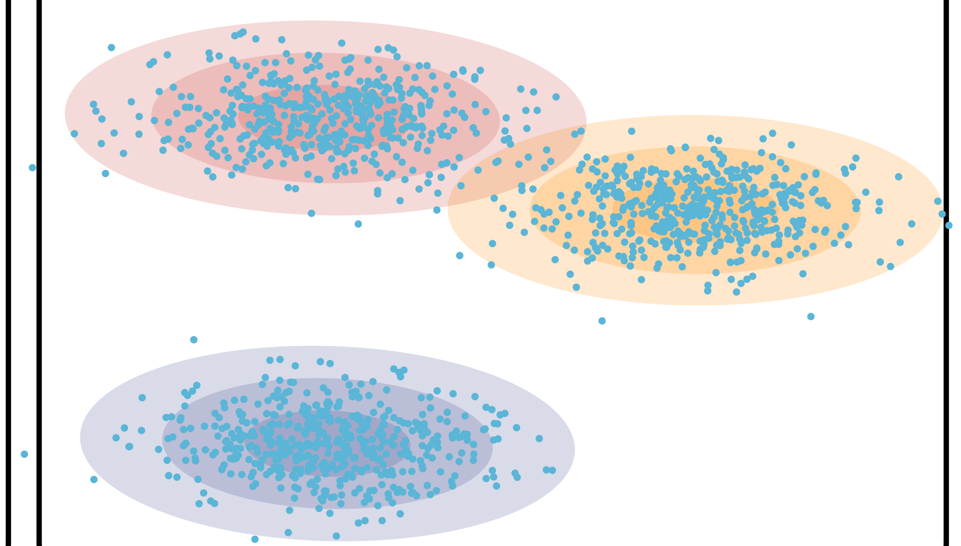


Parametrized density estimation

h: parameters

$$\mathcal{R}(\pi, h) = W_1(\pi, \pi_h)$$

E.g.: GAN, GMM



Optimal Transport for CSL: 1) Wasserstein Learnability

Wasserstein learnability

$$\forall \pi, \pi' \in \mathcal{P}(\mathbb{R}^d), \mathrm{TaskMetric}(\pi, \pi') \lesssim W_q(\pi, \pi')$$

Supervised Learning

Condition on the task $\mathcal{L}(\mathcal{H})$	Condition on q	Examples
Regression tasks Hypothesis : h Lipschitz function, Loss : square-loss	$q = 2$	Linear regression, regression using MLP with bounded params
Binary classification Hypothesis : h Lipschitz function, Loss : convex surrogate $\ell(\mathbf{x} = (\mathbf{z}, y), h) = \varphi(yh(\mathbf{z}))$	$q = 1$	MLP classifier (bounded params) + Lipschitz ouput layer

Remarks

Encompasses all the known tasks in CSL + other

Wasserstein vs MMD

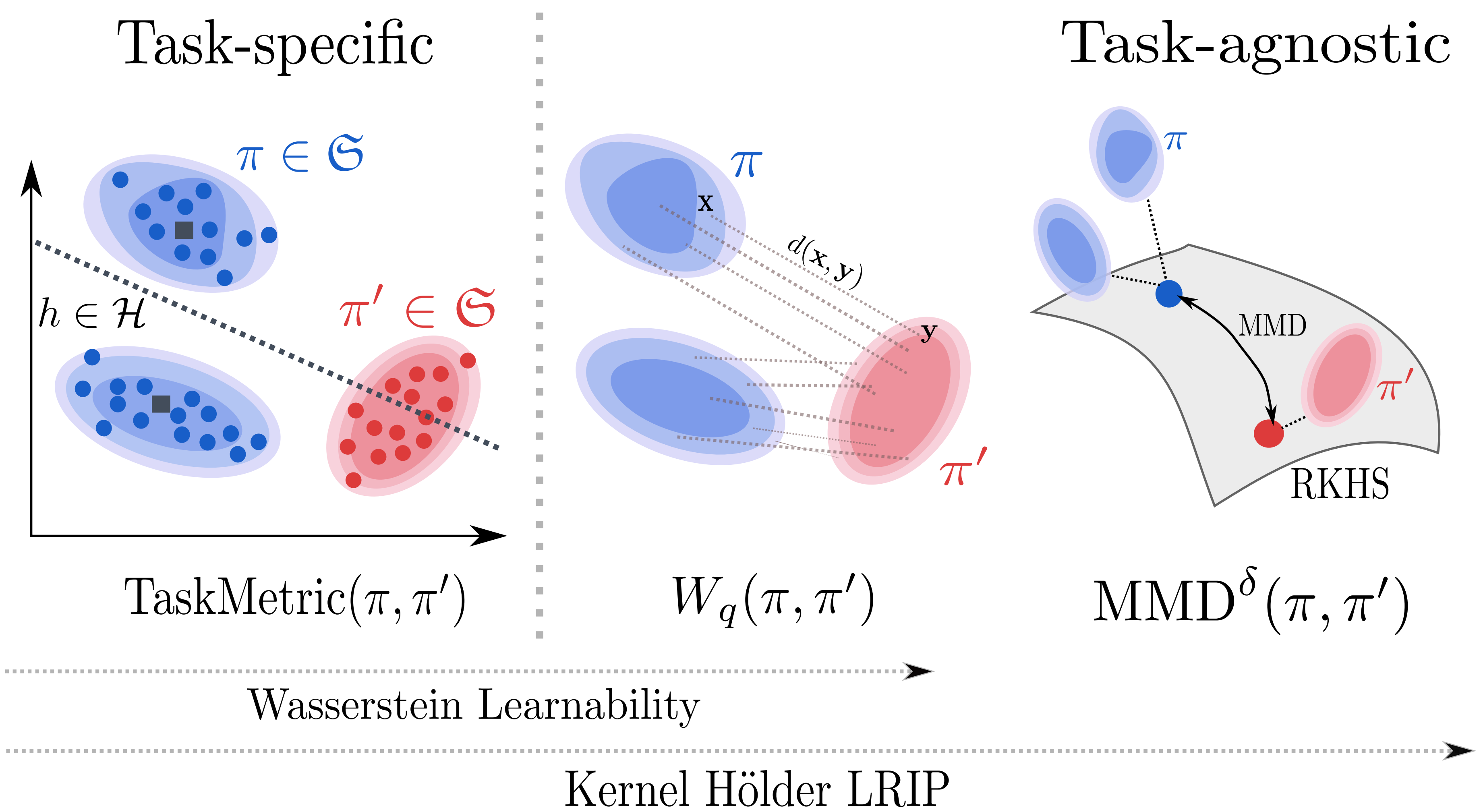
Optimal Transport for CSL: 2) Wass vs MMD

Goal

$$\forall \pi, \pi' \in \mathfrak{S}, W_q(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{\delta}(\pi, \pi'), 0 < \delta \leq 1$$

Remarks

- Focus on TI kernels
- Uniform control
- Do we need model set ?



Optimal Transport for CSL: 2) Wass vs MMD

Goal

(1) $\forall \pi, \pi' \in \mathfrak{S}, W_q(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{\delta}(\pi, \pi'), 0 < \delta \leq 1$

A bunch of negative results

- κ bounded
- Any $\mathfrak{S}$

If (1) then:

$$\sup_{\pi,\pi' \in \mathfrak{S}} \|\,\mathrm{mean}(\pi) - \mathrm{mean}(\pi')\|_2 < +\infty$$

Optimal Transport for CSL: 2) Wass vs MMD

Goal

(1) $\forall \pi, \pi' \in \mathfrak{S}, W_q(\pi, \pi') \lesssim \text{MMD}_\kappa^{\delta}(\pi, \pi'), 0 < \delta \leq 1$

A bunch of negative results

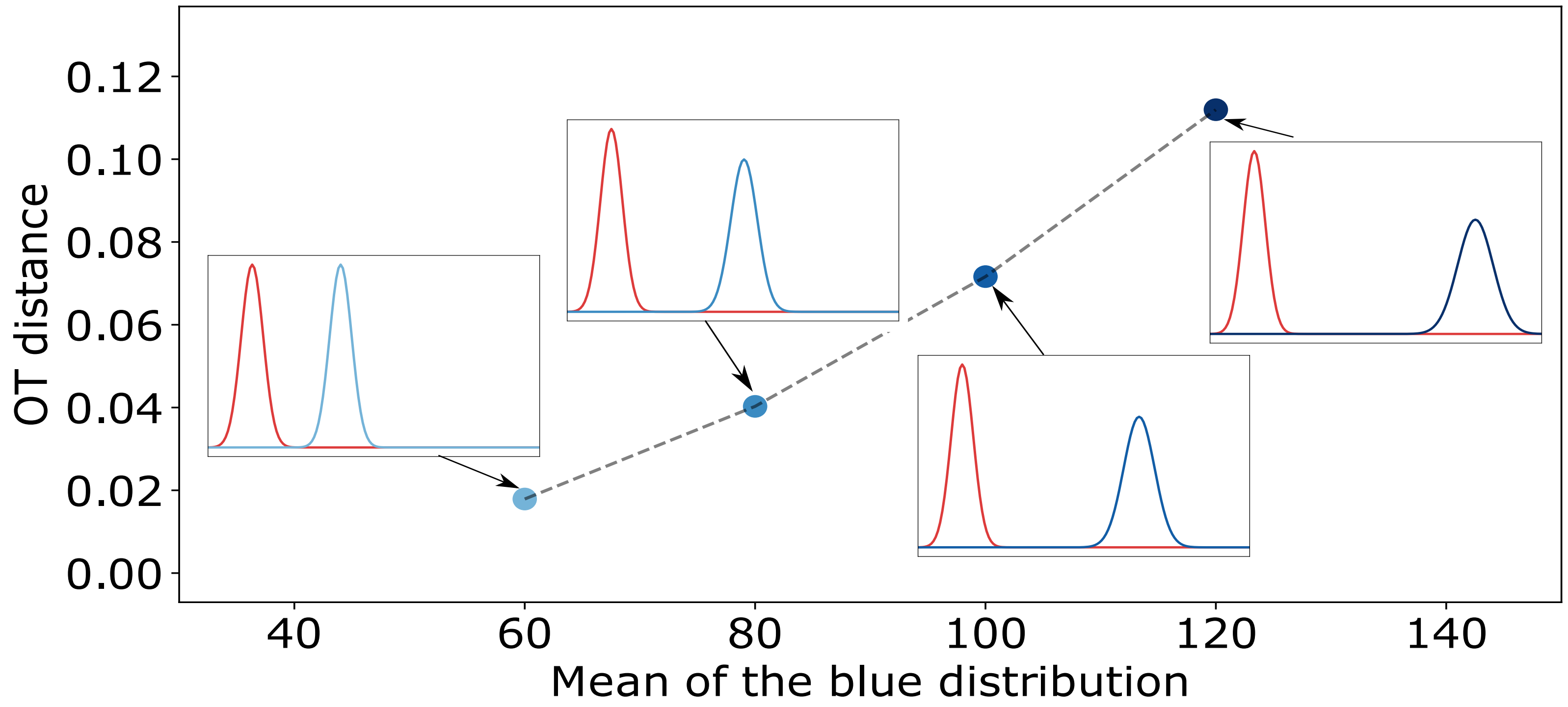
- κ bounded
- Any $\mathfrak{S}$

If (1) then:

$$\sup_{\pi, \pi' \in \mathfrak{S}} \|\text{mean}(\pi) - \text{mean}(\pi')\|_2 < +\infty$$

Since:

$$\text{MMD}_\kappa \leq \text{cte} < +\infty$$



Optimal Transport for CSL: 2) Wass vs MMD

Goal

(1) $\forall \pi, \pi' \in \mathfrak{S}, W_q(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{\delta}(\pi, \pi'), 0 < \delta \leq 1$

A bunch of negative results

- κ bounded
- $\mathfrak{S}$ contains all distrib. with compact support

If (1) then:

$$\delta \leq 2/d$$

Since:

Convergence of finite samples

Wass = curse of dim.
MMD not

$$\mathbb{E}[W_1(\pi, \pi_n)] \gtrsim n^{-1/d}$$

$$\mathbb{E}[\mathrm{MMD}_\kappa^{\delta}(\pi, \pi_n)] \lesssim n^{-\delta/2}$$

Optimal Transport for CSL: 2) Wass vs MMD

Goal

(1) $\forall \pi, \pi' \in \mathfrak{S}, W_q(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{\delta}(\pi, \pi'), 0 < \delta \leq 1$

A bunch of negative results

- κ bounded
- $\mathfrak{S}$ contains all distrib. with compact support

If (1) then:

$\delta \leq 2/d$

Very slow rate for CSL

Since:

Convergence of finite samples

Wass = curse of dim.
MMD not

$\mathbb{E}[W_1(\pi, \pi_n)] \gtrsim n^{-1/d}$

$\mathbb{E}[\mathrm{MMD}_\kappa^{\delta}(\pi, \pi_n)] \lesssim n^{-\delta/2}$

Optimal Transport for CSL: 2) Wass vs MMD

Goal

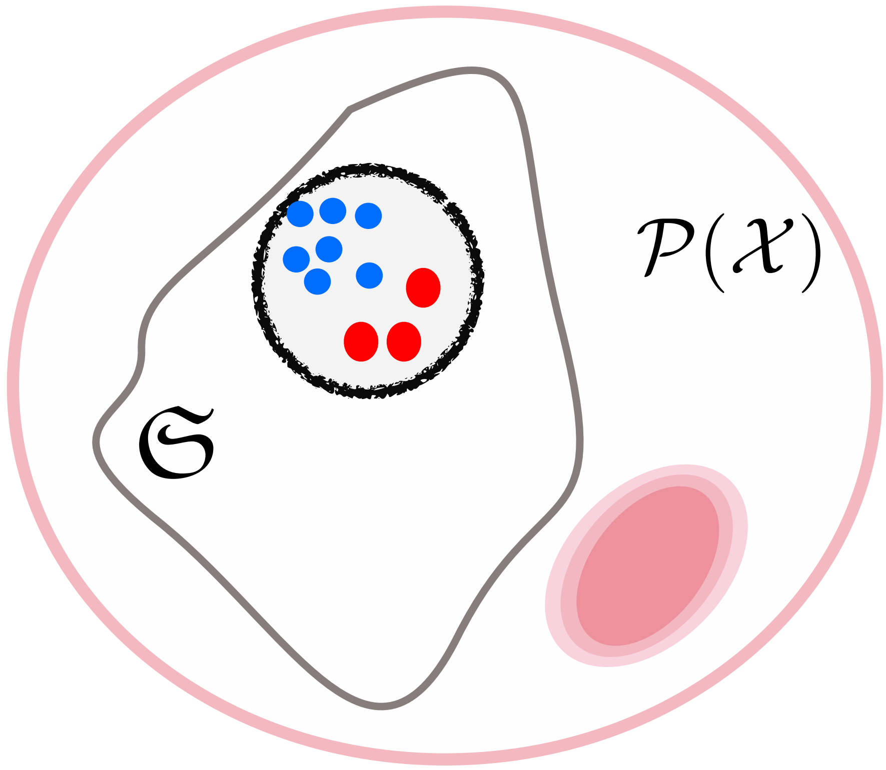
(1) $\forall \pi, \pi' \in \mathfrak{S}, W_q(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{\delta}(\pi, \pi'), 0 < \delta \leq 1$

A bunch of negative results

- $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ with $\kappa_0 \in C^k$
- $\mathfrak{S}$ contains mixtures of $\lfloor \frac{k}{2} \rfloor + 1$ diracs on a ball

If (1) then:

$$\delta \leq 2/k$$



Optimal Transport for CSL: 2) Wass vs MMD

Goal

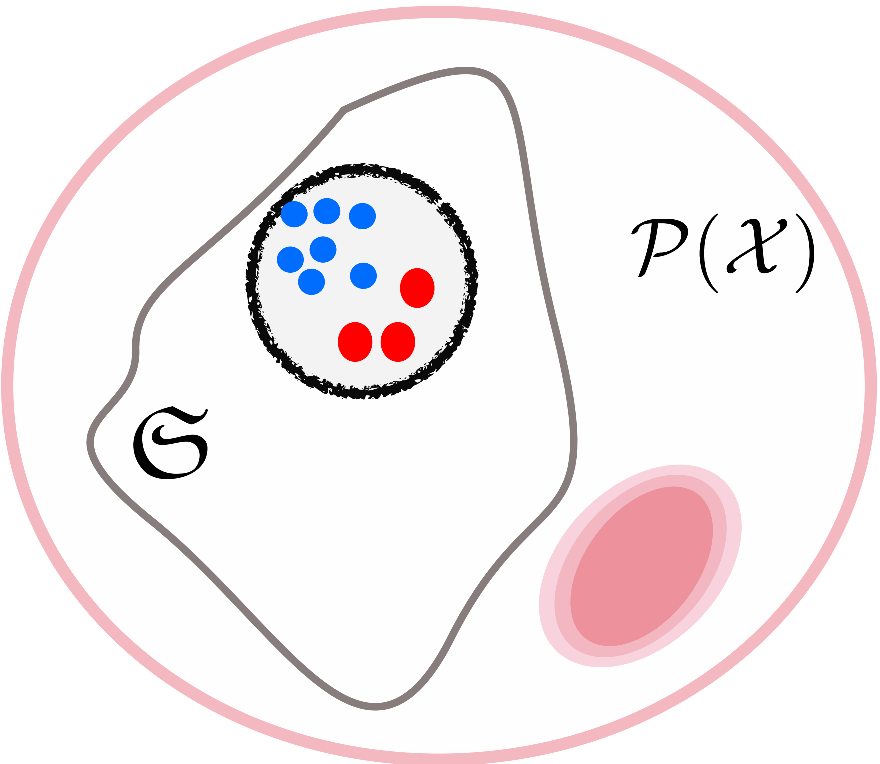
(1) $\forall \pi, \pi' \in \mathfrak{S}, W_q(\pi, \pi') \lesssim \mathrm{MMD}^{\delta}_{\kappa}(\pi, \pi'), 0 < \delta \leq 1$

A bunch of negative results

- $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ with $\kappa_0 \in C^\infty$ smooth
- $\mathfrak{S}$ contains mixtures of K diracs on a ball

If (1) then:

$$\delta \leq 2/K$$



Optimal Transport for CSL: 2) Wass vs MMD

Goal

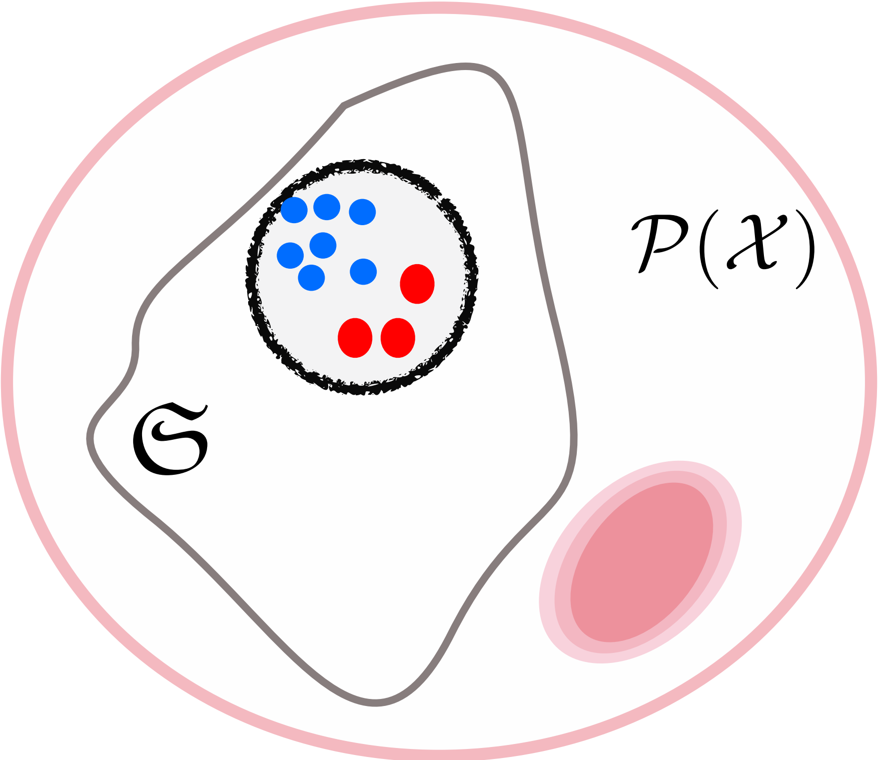
(1) $\forall \pi, \pi' \in \mathfrak{S}, W_q(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{\delta}(\pi, \pi'), 0 < \delta \leq 1$

A bunch of negative results

- $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ with $\kappa_0 \in C^\infty$ **smooth**
- $\mathfrak{S}$ contains mixtures of K diracs on a ball

If (1) then:

$$\delta \leq 2/K$$



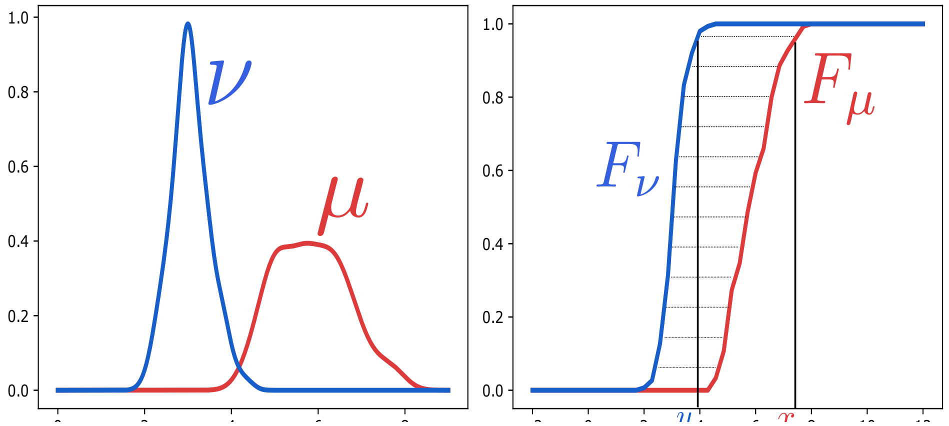
Trade off between **regularity of the kernel** and δ

We <u>**can not**</u> control Wass by MMD **uniformly over all discrete distrib.** (even compact) **for a smooth TI kernel**

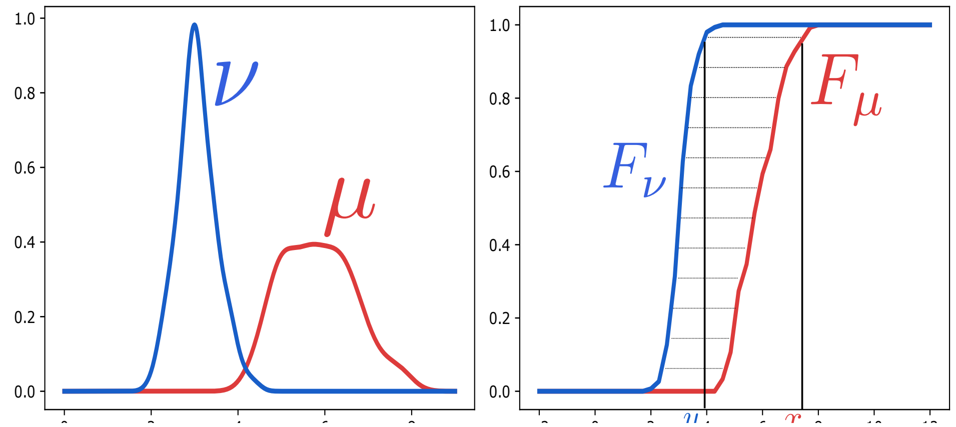
Optimal Transport for CSL: 2) Wass vs MMD

Let us be positive now: the real line

On $\mathbb{R}$ Wasserstein admits a closed-form



Optimal Transport for CSL: 2) Wass vs MMD



Let us be positive now: the real line

On $\mathbb{R}$ Wasserstein admits a closed-form

Hypothesis

For any $\kappa(x, y) = \kappa_0(x - y)$

$(\widehat{\kappa_0})^{-1}$ continuous $(\widehat{\kappa_0})^{-1}(\omega) = O_{+\infty}(\omega^k)$

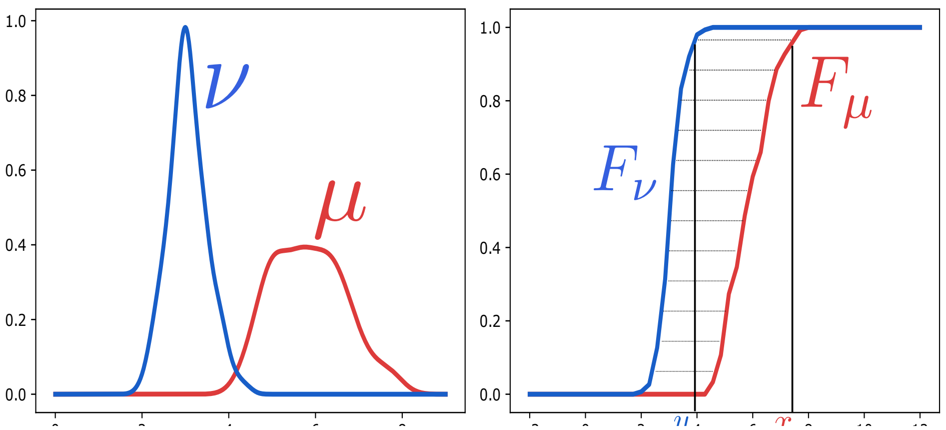
Remarks

True for any T.I. with some regularities

Optimal Transport for CSL: 2) Wass vs MMD

Let us be positive now: the real line

On $\mathbb{R}$ Wasserstein admits a closed-form



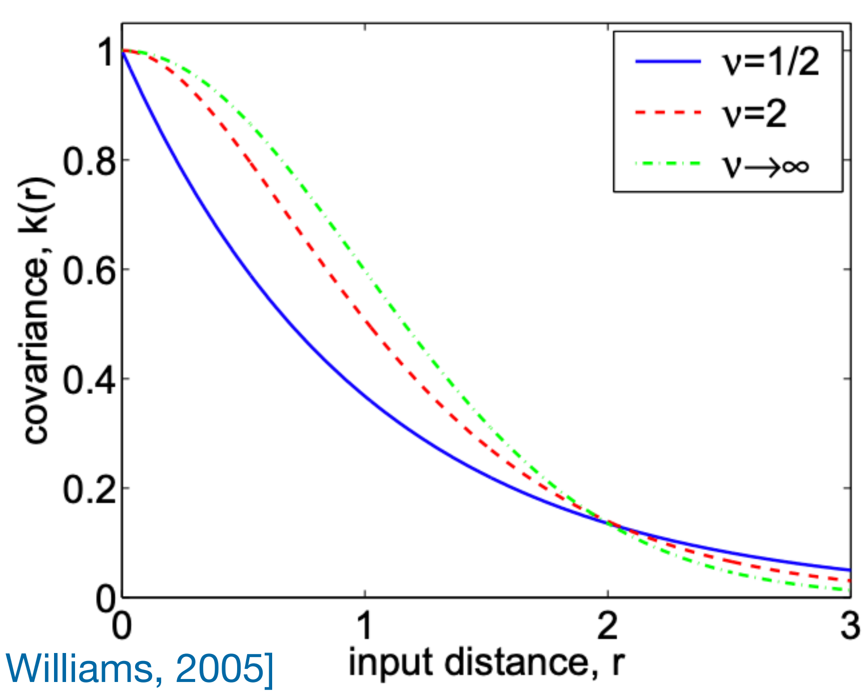
Hypothesis

<u>For any</u> $\kappa(x,y) = \kappa_0(x-y)$

$(\widehat{\kappa_0})^{-1}$ continuous $(\widehat{\kappa_0})^{-1}(\omega) = O_{+\infty}(\omega^k)$

~~Gaussian~~

Matérn class, splines, polyharmonic curves



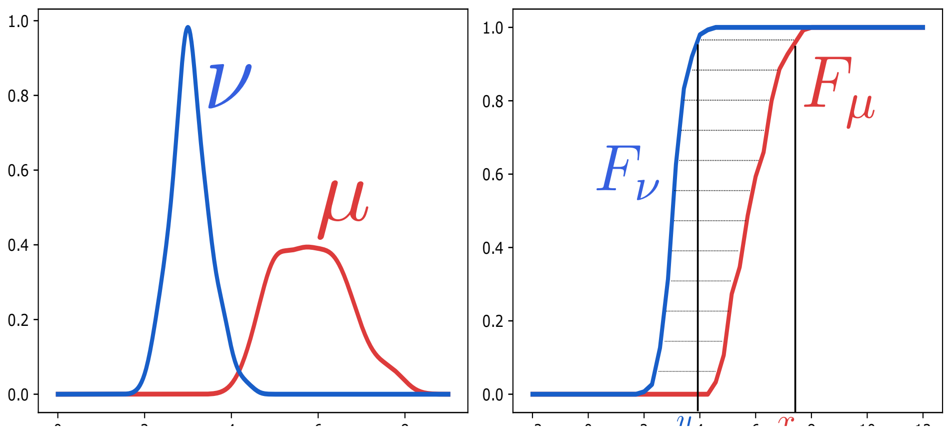
[Rsmussen and Williams, 2005]

Remarks

True for any T.I. with some regularities

Optimal Transport for CSL: 2) Wass vs MMD

Let us be positive now: the real line



On $\mathbb{R}$ Wasserstein admits a closed-form

Hypothesis

For any $\kappa(x,y) = \kappa_0(x-y)$

$(\widehat{\kappa_0})^{-1}$ continuous $(\widehat{\kappa_0})^{-1}(\omega) = O_{+\infty}(\omega^k)$

$\mathfrak{S} \subseteq \{\pi \ll f\mathrm{d}x, \mathrm{mean}(\pi) = m, \|f\|_{W^{s,1}} \leq M\}$ $s \geq k/2+1$

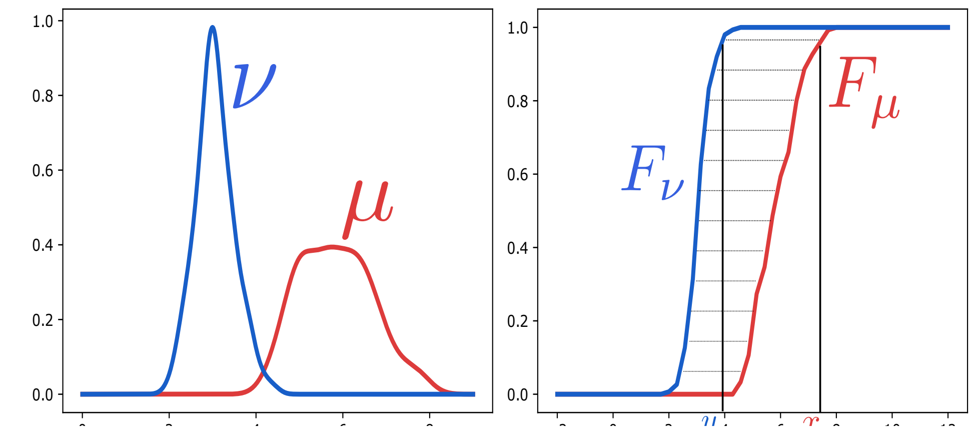
Remarks

True with some regularities on the distrib

Optimal Transport for CSL: 2) Wass vs MMD

Let us be positive now: the real line

On $\mathbb{R}$ Wasserstein admits a closed-form



Hypothesis

For any $\kappa(x,y) = \kappa_0(x-y)$

$(\widehat{\kappa_0})^{-1}$ continuous $(\widehat{\kappa_0})^{-1}(\omega) = O_{+\infty}(\omega^k)$

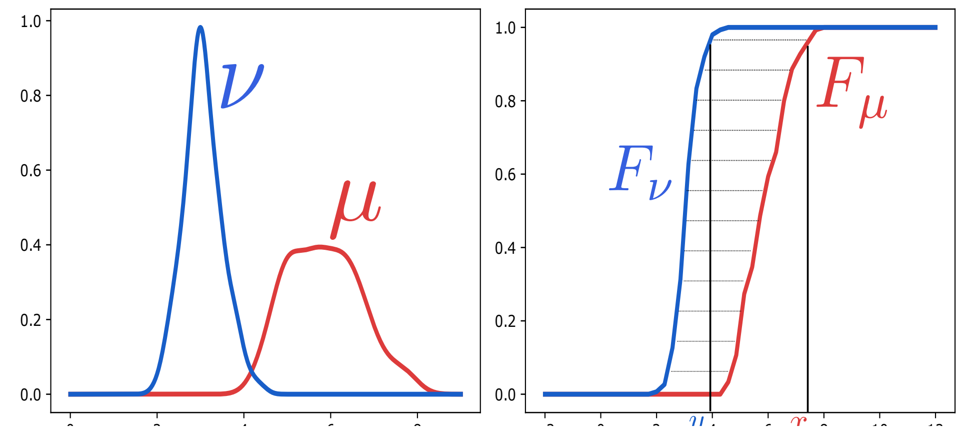
$\mathfrak{S} \subseteq \{\pi \ll f\mathrm{d}x, \mathrm{mean}(\pi) = m, \|f\|_{W^{s,1}} \leq M\}$ $s \geq k/2+1$

Density

Remarks

True with some regularities on the distrib

Optimal Transport for CSL: 2) Wass vs MMD



Let us be positive now: the real line

On $\mathbb{R}$ Wasserstein admits a closed-form

Hypothesis

For any $\kappa(x,y) = \kappa_0(x-y)$

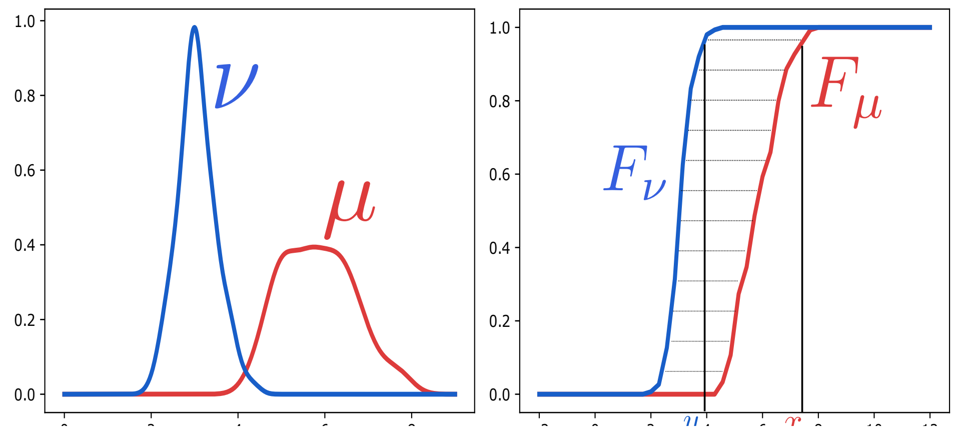
- $(\widehat{\kappa_0})^{-1}$ continuous $(\widehat{\kappa_0})^{-1}(\omega) = O_{+\infty}(\omega^k)$
- $\mathfrak{S} \subseteq \{\pi \ll f\mathrm{d}x, \mathrm{mean}(\pi) = m, \|f\|_{W^{s,1}} \leq M\}$ $s \geq k/2+1$

Same mean (centered)

Remarks

True with some regularities on the distrib

Optimal Transport for CSL: 2) Wass vs MMD



Let us be positive now: the real line

- On $\mathbb{R}$ Wasserstein admits a closed-form

Hypothesis

For any $\kappa(x,y) = \kappa_0(x-y)$

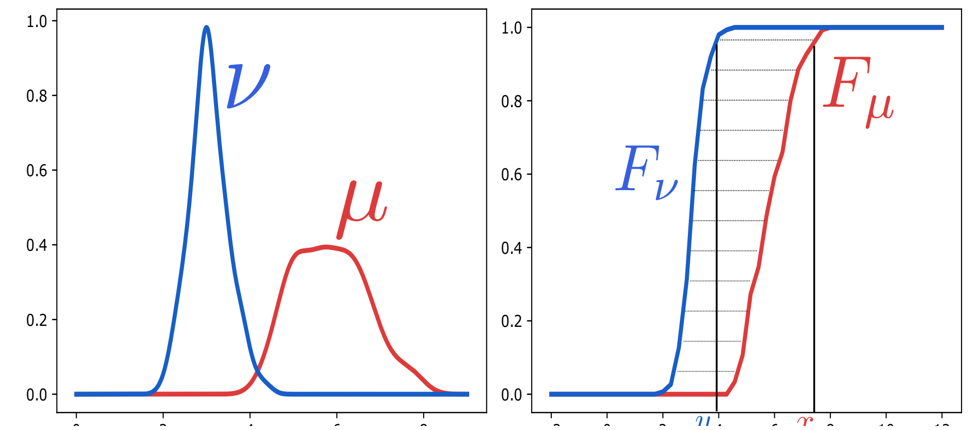
- $(\widehat{\kappa_0})^{-1}$ continuous $(\widehat{\kappa_0})^{-1}(\omega) = O_{+\infty}(\omega^k)$
- $\mathfrak{S} \subseteq \{\pi \ll f\mathrm{d}x, \mathrm{mean}(\pi) = m, \|f\|_{W^{s,1}} \leq M\} \quad s \geq k/2+1$

Sobolev Ball

Remarks

- True with some regularities on the distrib

Optimal Transport for CSL: 2) Wass vs MMD



Let us be positive now: the real line

On $\mathbb{R}$ Wasserstein admits a closed-form

Hypothesis

<u>For any</u> $\kappa(x, y) = \kappa_0(x - y)$

$(\widehat{\kappa_0})^{-1}$ continuous $(\widehat{\kappa_0})^{-1}(\omega) = O_{+\infty}(\omega^k)$

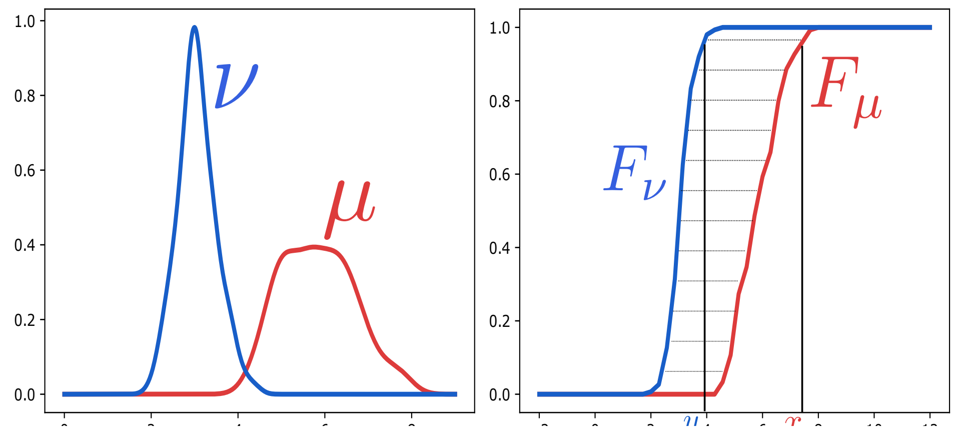
$\mathfrak{S} \subseteq \{\pi \ll f\mathrm{d}x, \mathrm{mean}(\pi) = m, \|f\|_{W^{s,1}} \leq M\} \quad s \geq k/2 + 1$

$$\Longrightarrow \forall \pi, \pi' \in \mathfrak{S}, W_2(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{1/2}(\pi, \pi')$$

Optimal Transport for CSL: 2) Wass vs MMD

Let us be positive now: the real line

On $\mathbb{R}$ Wasserstein admits a closed-form



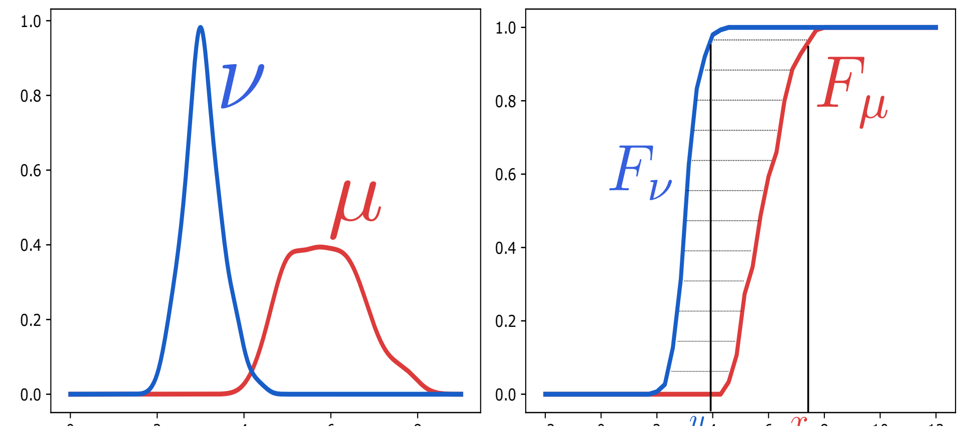
$$\forall \pi, \pi' \in \mathfrak{S}, W_2(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{1/2}(\pi, \pi')$$

Sketch:

CDF

On $\mathbb{R}$ $W_2(\pi, \pi') = \|F - G\|_{L_2} = \frac{1}{2\pi}\|\hat{F} - \hat{G}\|_{L_2}$

Optimal Transport for CSL: 2) Wass vs MMD



Let us be positive now: the real line

On $\mathbb{R}$ Wasserstein admits a closed-form

$$\forall \pi, \pi' \in \mathfrak{S}, W_2(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{1/2}(\pi, \pi')$$

Sketch:

On $\mathbb{R}$ $W_2(\pi, \pi') = \|F - G\|_{L_2} = \frac{1}{2\pi}\|\hat{F} - \hat{G}\|_{L_2}$

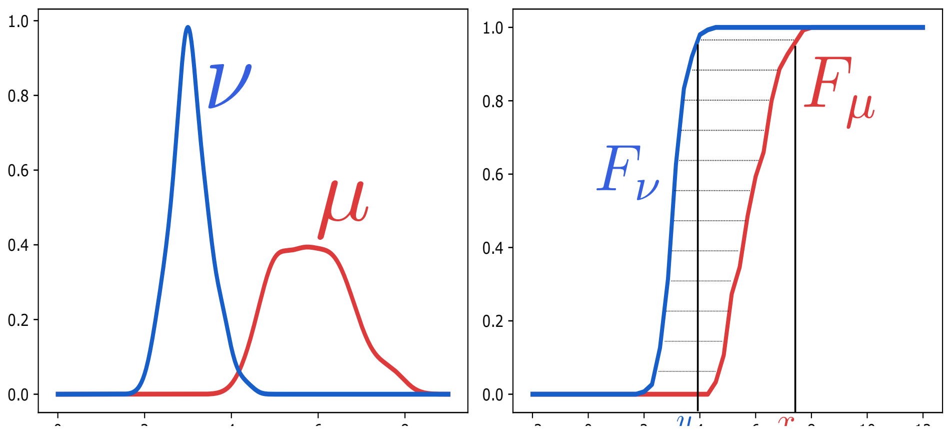
So $W_2^2(\pi, \pi') = \frac{1}{2\pi}\int |\omega|^{-2}|\hat{f}(\omega) - \hat{g}(\omega)|^2 \mathrm{d}\omega$

$$\leq \frac{1}{2\pi}\left(\int \frac{|\hat{f}(\omega) - \hat{g}(\omega)|^2}{|\omega|^4 \widehat{\kappa_0}(\omega)}\mathrm{d}\omega\right)^{\frac{1}{2}}\left(\int \widehat{\kappa_0}(\omega)|\hat{f}(\omega) - \hat{g}(\omega)|^2 \mathrm{d}\omega\right)^{\frac{1}{2}}$$

Optimal Transport for CSL: 2) Wass vs MMD

Let us be positive now: the real line

On $\mathbb{R}$ Wasserstein admits a closed-form



$$\forall \pi, \pi' \in \mathfrak{S}, W_2(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{1/2}(\pi, \pi')$$

Sketch:

On $\mathbb{R}$ $\quad W_2(\pi, \pi') = \|F - G\|_{L_2} = \frac{1}{2\pi} \|\hat{F} - \hat{G}\|_{L_2}$

So $\quad W_2^2(\pi, \pi') = \frac{1}{2\pi} \int |\omega|^{-2} |\hat{f}(\omega) - \hat{g}(\omega)|^2 \mathrm{d}\omega$

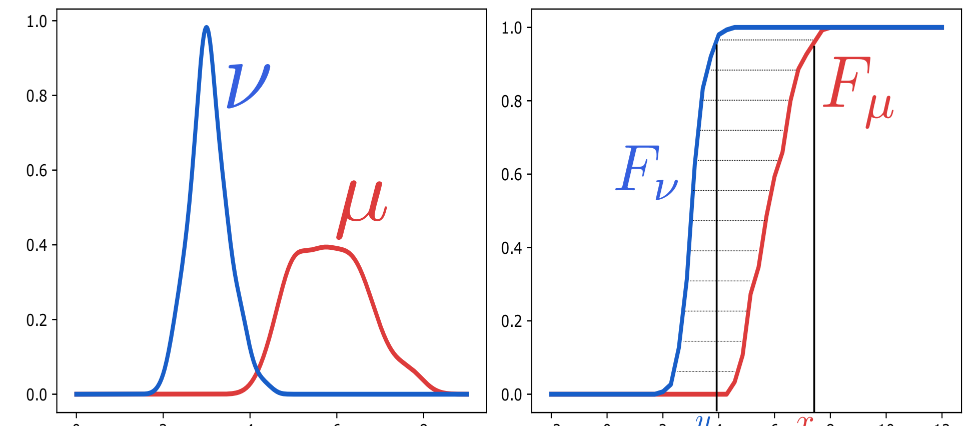
$$\leq \frac{1}{2\pi} \left(\int \frac{|\hat{f}(\omega) - \hat{g}(\omega)|^2}{|\omega|^4 \widehat{\kappa_0}(\omega)} \mathrm{d}\omega \right)^{\frac{1}{2}} \left(\int \widehat{\kappa_0}(\omega) |\hat{f}(\omega) - \hat{g}(\omega)|^2 \mathrm{d}\omega \right)^{\frac{1}{2}}$$

MMD

Optimal Transport for CSL: 2) Wass vs MMD

Let us be positive now: the real line

On $\mathbb{R}$ Wasserstein admits a closed-form



$$\forall \pi, \pi' \in \mathfrak{S}, W_2(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{1/2}(\pi, \pi')$$

Sketch:

On $\mathbb{R}$ $\quad W_2(\pi, \pi') = \|F - G\|_{L_2} = \frac{1}{2\pi}\|\hat{F} - \hat{G}\|_{L_2}$

So $\quad W_2^2(\pi, \pi') = \frac{1}{2\pi}\int |\omega|^{-2}|\hat{f}(\omega) - \hat{g}(\omega)|^2 \mathrm{d}\omega$

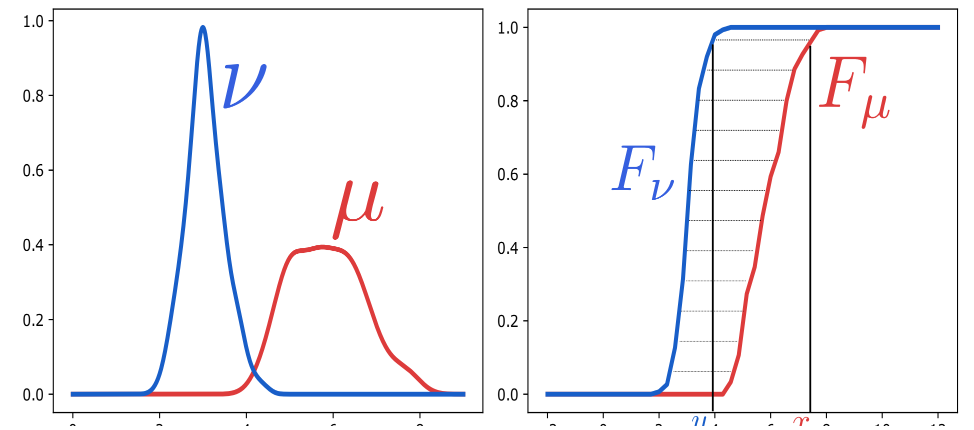
$$\leq \frac{1}{2\pi}\left(\int \frac{|\hat{f}(\omega) - \hat{g}(\omega)|^2}{|\omega|^4 \widehat{\kappa_0}(\omega)}\mathrm{d}\omega\right)^{\frac{1}{2}}\left(\int \widehat{\kappa_0}(\omega)|\hat{f}(\omega) - \hat{g}(\omega)|^2 \mathrm{d}\omega\right)^{\frac{1}{2}}$$

$\lesssim \mathrm{cte}$

Optimal Transport for CSL: 2) Wass vs MMD

Let us be positive now: the real line

On $\mathbb{R}$ Wasserstein admits a closed-form



Hypothesis: without the mean

For any $\kappa(x,y) = \kappa_0(x-y) + xy$ → Not TI

$(\widehat{\kappa_0})^{-1}$ continuous $\quad (\widehat{\kappa_0})^{-1}(\omega) = O_{+\infty}(\omega^k) \quad \kappa_0$ Lipschitz

$\mathfrak{S} \subseteq \{\pi \ll f\mathrm{d}x, \|f\|_{W^{s,1}(\mathbb{R})} \leq M\} \qquad s \geq k/2+1$

$$\Longrightarrow \forall \pi, \pi' \in \mathfrak{S}, W_2(\pi,\pi') \lesssim \mathrm{MMD}_\kappa^{1/2}(\pi,\pi')$$

Optimal Transport for CSL: 2) Wass vs MMD

Let us be positive now:

From $\mathbb{R}$ to $\mathbb{R}^d$ -> Sliced Wasserstein distance !

$$\implies \forall \pi, \pi' \in \mathfrak{S}, W_2(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{\frac{1}{2q(d+1)}}(\pi, \pi')$$

Compactness + regularity assumptions

$$\mathfrak{S} \subseteq \{\pi \ll f\mathrm{d}\mathbf{x}, \|f\|_{W^{s,1}(\mathbb{R}^d)} \leq M, \mathrm{supp}(f) \subseteq B(0, R)\}$$

Sliced kernel

$$\kappa(\mathbf{x}, \mathbf{y}) = \mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim \mathbb{S}^{d-1}} [\kappa_0(\boldsymbol{\theta}^\top \mathbf{x} - \boldsymbol{\theta}^\top \mathbf{y})] + \frac{1}{d}\mathbf{x}^\top \mathbf{y}$$

Non-compactly supported distributions ? No density ?

Optimal Transport for CSL: 2) Wass vs MMD

The general case on $\mathbb{R}^d$

For any $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ where $\kappa_0 = \alpha * \alpha$

Hypothesis on the kernel

$\alpha \geq 0, \int \alpha(\mathbf{x})\mathrm{d}\mathbf{x} = 1$

Decomposition true for « classical » T.I. kernels (Gaussian, Matérn, Laplace)

e.g. Gaussian kernel obtained with $\alpha(\mathbf{x}) = (2\pi)^{-d/2}\sigma^{-d}\exp(-\|\mathbf{x}\|_2^2/\sigma^2)$

« Convolution » or « Boas–Kac » root of the kernel

Optimal Transport for CSL: 2) Wass vs MMD

The general case on $\mathbb{R}^d$

For any $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ where $\kappa_0 = \alpha * \alpha$

$$\mathfrak{S} \subseteq \{\pi \in \mathcal{P}(\mathbb{R}^d), \mathbb{E}_{\mathbf{x}\sim\pi}[||\mathbf{x}||^s] \leq M\} \quad s \geq 1$$

Hypothesis on the distrib.

- All the distributions have uniformly s-bounded moments
- Obtained e.g. parametric densities with bounded params or discrete distrib.

Optimal Transport for CSL: 2) Wass vs MMD

The general case on $\mathbb{R}^d$

For any $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ where $\kappa_0 = \alpha * \alpha$

$$\mathfrak{S} \subseteq \{\pi \in \mathcal{P}(\mathbb{R}^d), \mathbb{E}_{\mathbf{x}\sim\pi}[||\mathbf{x}||^s] \leq M\} \quad s \geq 1$$

For $1 \leq q < s$

$$\forall \pi, \pi' \in \mathfrak{S}, W_q(\pi,\pi') \lesssim \mathrm{MMD}_{\kappa}^{\frac{2(s-q)}{(d+2s)q}}(\pi,\pi') + \eta$$

Conclusion

- $\delta = \frac{2(s-q)}{(d+2s)q}$ The regularity of the distrib. mitigates the curse of dim

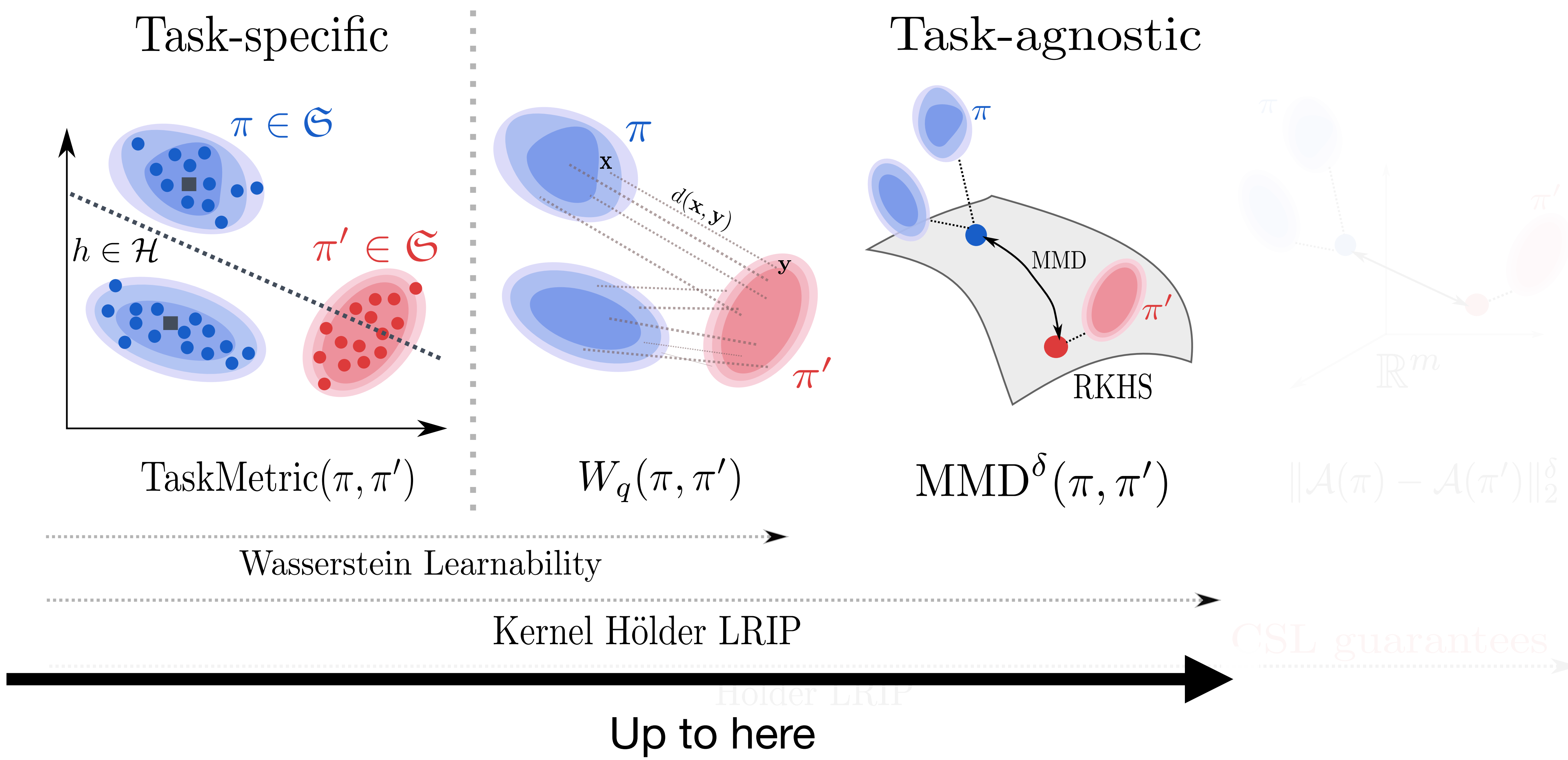
  s big $\delta \approx \frac{1}{q}$

- $\eta > 0$ will add an error term for the Holder LRIP
- Can be chosen arbitrary small -> sharper kernel

Obtaining sketching operator

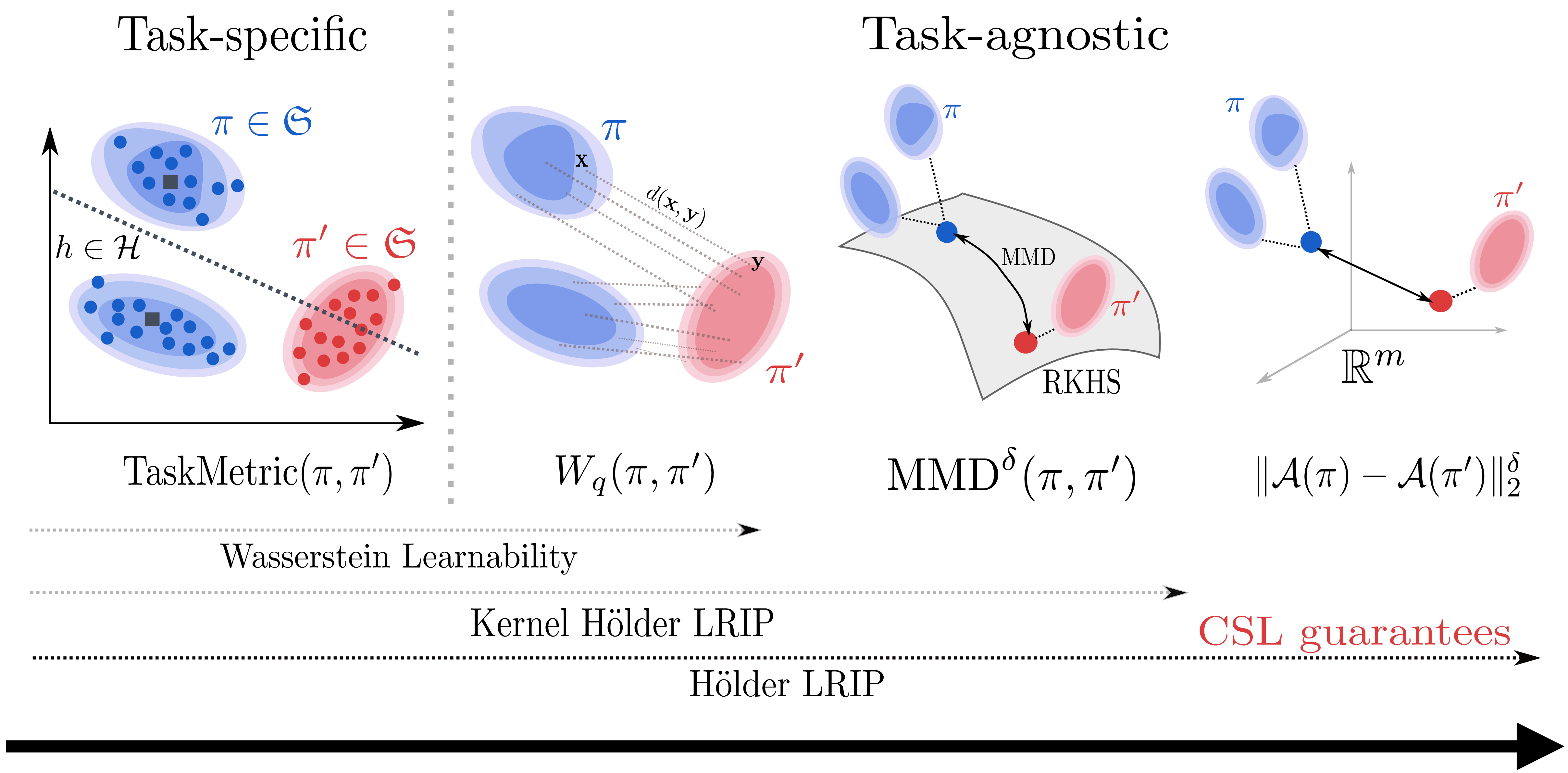
Optimal Transport for CSL

Roadmap



Optimal Transport for CSL

Roadmap



To the end

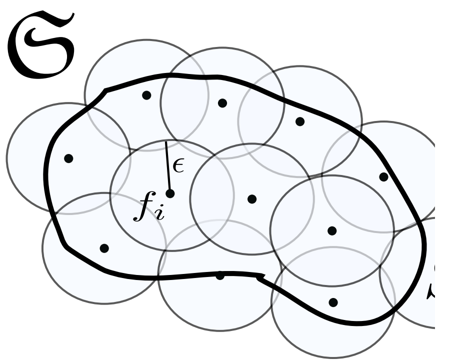
Convergence empirical MMD. Need to control the « size » of $\mathfrak{S}$ (covering numbers)

Optimal Transport for CSL

Existence of sketching operator with CSL guarantees

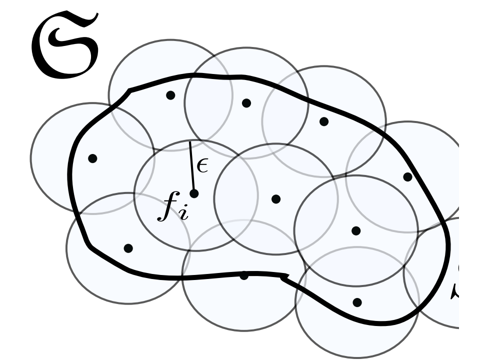
Suppose

- $\forall \pi, \pi' \in \mathfrak{S},\ \mathrm{TaskMetric}(\pi, \pi') \lesssim \mathrm{MMD}_{\kappa}^{\delta}(\pi, \pi')$
- Box-counting dimension: $d(\mathfrak{S}) < +\infty$ (covering TV norm)



Optimal Transport for CSL

Existence of sketching operator with CSL guarantees



Suppose

- $\forall \pi, \pi' \in \mathfrak{S},\ \text{TaskMetric}(\pi, \pi') \lesssim \text{MMD}_\kappa^{\delta}(\pi, \pi')$
- Box-counting dimension: $d(\mathfrak{S}) < +\infty$ (covering TV norm)

Then with: $m > 2d(\mathfrak{S})$

$$\exists \mathcal{A} : \mathcal{P}(\mathcal{X}) \to \mathbb{R}^m \text{ Hölder LRIP with } \beta < \delta$$

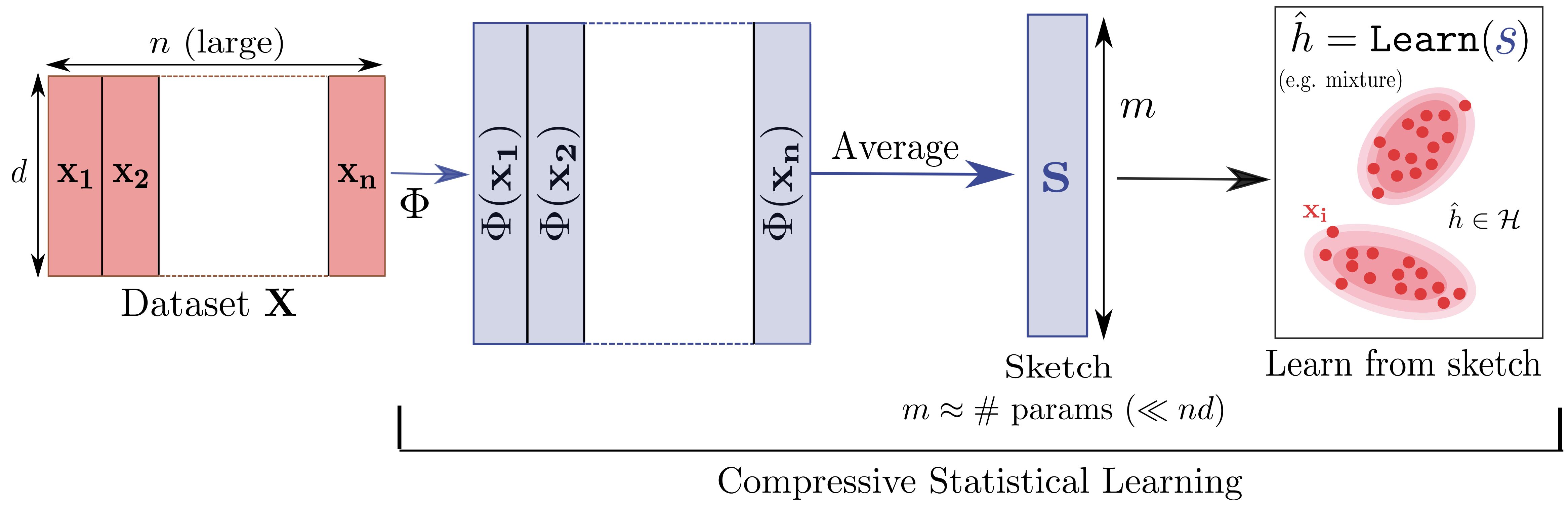
CSL guarantees

$$\underline{\forall \pi \in \mathcal{P}(\mathcal{X})}, \text{excess-risk}(\pi) \lesssim d^{\circ}(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2^{\beta}$$

A gentle recap

Compressive Statistical Learning

\+ Ressource efficient ML framework
\+ suitable distributed/streaming learning
Privacy



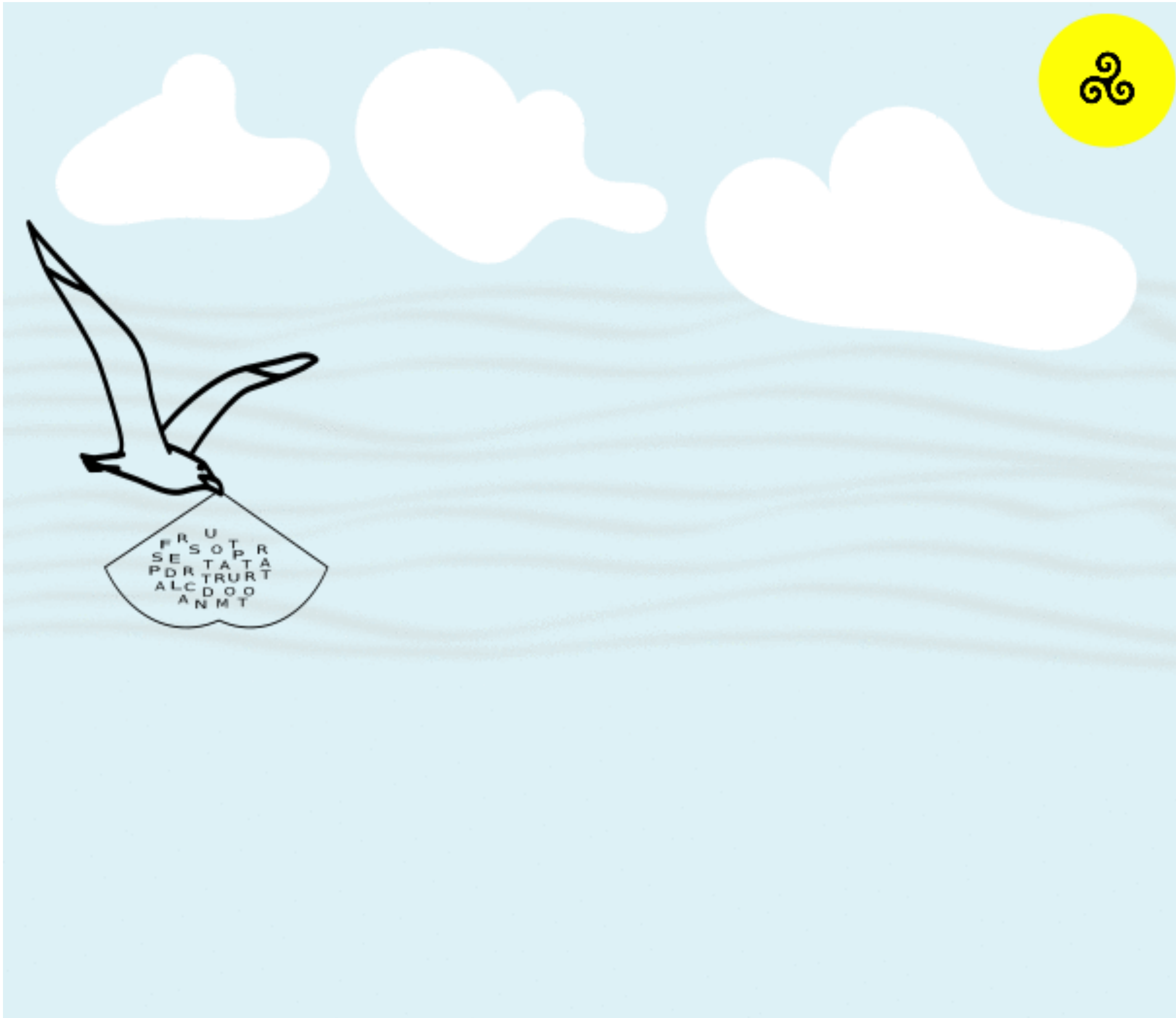
Statistical learning guarantees

| LRIP -> difficult to prove

| Hölder LRIP: easier + control of Wass by kernel norms

-> Need to design a model set of distrib.

Thank you!



Optimal Transport for CSL: 2) Wass vs MMD

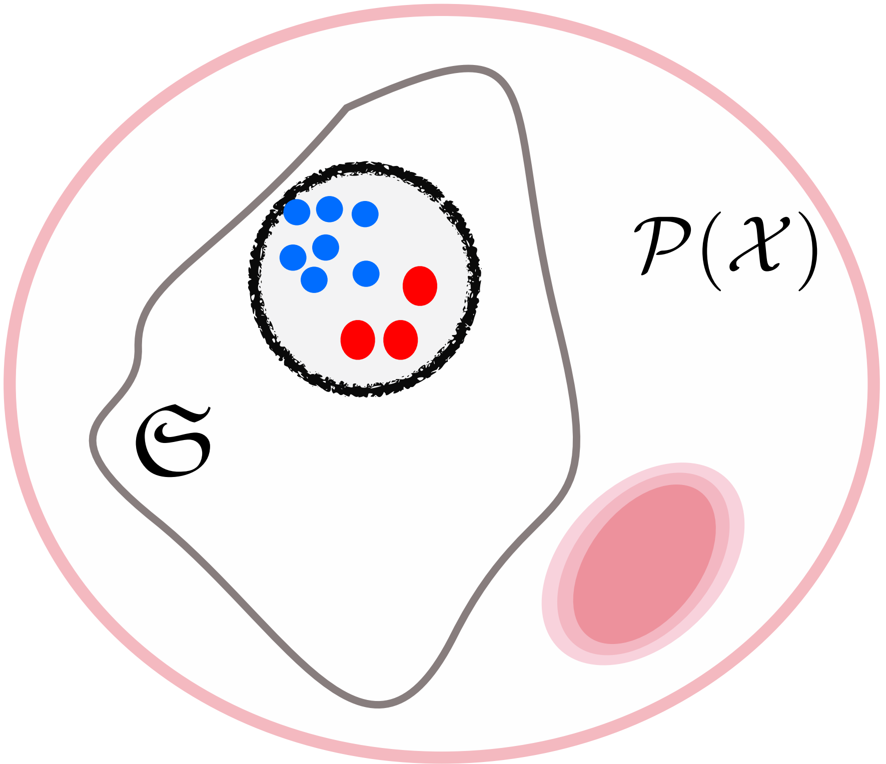
Goal

(1) $\forall \pi, \pi' \in \mathfrak{S}, W_q(\pi, \pi') \lesssim \mathrm{MMD}_\kappa^{\delta}(\pi, \pi'), 0 < \delta \leq 1$

A bunch of negative results

- κ bounded
- $\mathfrak{S}$ contains a segment $[\pi_0, \pi_1]$

$$\mathrm{supp}(\pi_0) \cap \mathrm{supp}(\pi_1) = \emptyset$$



If (1) then:

$$\delta \leq 1/p$$

Towards CSL guarantees: 1) Learn from sketch

Feature operator $\underline{\mathcal{X} = \mathbb{R}^d}$

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

$\mathbf{W} = (\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_m) \in \mathbb{R}^{d \times m}$ random matrix (e.g. i.i.d. normal entries)

ρ non-linear function applied pointwise

Example:

$$\rho(t) = \exp(-it)$$

Random Fourier Features (RFF) [Rahimi and Recht, 2008]

$$\Phi(\mathbf{x}) = \frac{1}{\sqrt{m}}(\exp(-i\boldsymbol{\omega}_1^\top \mathbf{x}), \cdots, \exp(-i\boldsymbol{\omega}_m^\top \mathbf{x}))^\top \quad \boldsymbol{\omega}_i \sim \Lambda \text{ i.i.d.}$$

Towards CSL guarantees: 3) The LRIP

Setting $\underline{\mathcal{X} = \mathbb{R}^d}$ $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ $\Phi = \mathrm{RFF}$

How to prove the LRIP

Step 1 $\forall \pi, \pi' \in \mathfrak{S}, \mathrm{TaskMetric}(\pi, \pi') \lesssim \mathrm{MMD}_\kappa(\pi, \pi')$ **Kernel LRIP**

Step 2 $\forall \pi, \pi' \in \mathfrak{S}, \mathrm{MMD}_\kappa(\pi, \pi') \approx \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$ m large enough

Problems:

Step 1: not trivial at all !

Few tasks (K-means, GMM) + need separability assumptions

How to prove it for more tasks ?

Step 2: a little bit easier

Convergence of empirical MMD to the true MMD

+

Need to control the « size » of $\mathfrak{S}$