

# Comparer des distributions

Un florilège de présentations ETICS, et problématiques actuelles

**10ème édition d'ETICS**

**Sébastien Da Veiga — ENSAI / CREST**

# Outline

- Some divergences and distances between probability distributions
- How they are used in our community, and in ETICS talks
- Recent developments and selected topics

# **(Some) Divergences and distances between probability distributions**

# Common distances and divergences

Let  $P$  and  $Q$  denote two probability measures defined on the same measurable space  $(\mathcal{X}, \mathcal{F})$ , with respective densities  $p$  and  $q$  with respect to a common dominating measure (typically the Lebesgue measure)



# Common distances and divergences

Let  $P$  and  $Q$  denote two probability measures defined on the same measurable space  $(\mathcal{X}, \mathcal{F})$ , with respective densities  $p$  and  $q$  with respect to a common dominating measure (typically the Lebesgue measure)

The *total variation distance* measures the largest possible difference in probabilities assigned by  $P$  and  $Q$  to the same event:

$$d_{\text{TV}}(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|.$$

When  $P$  and  $Q$  admit densities  $p$  and  $q$ , this can be equivalently expressed as

$$d_{\text{TV}}(P, Q) = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| \, dx.$$

# Common distances and divergences

A large class of divergences, known as *f-divergences*, are defined for a convex function  $f : (0, +\infty) \rightarrow \mathbb{R}$  such that  $f(1) = 0$ :

$$D_f(P \parallel Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx.$$

Different choices of  $f$  yield well-known divergences. For example:

- **Kullback–Leibler divergence (KL divergence):**

$$\text{KL}(P \parallel Q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx.$$

- **Reverse KL divergence:**

$$\text{KL}(Q \parallel P) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} dx.$$

- **Total variation as an  $f$ -divergence:**

$$d_{\text{TV}}(P, Q) = \frac{1}{2} D_f(P \parallel Q) \quad \text{with} \quad f(t) = |t - 1|.$$

# Common distances and divergences

Mutual information between two random variables  $X$  and  $Y$  measures the divergence between the joint distribution  $P_{X,Y}$  and the product of the marginals  $P_X P_Y$ :

$$I(X; Y) = \text{KL}(P_{X,Y} \parallel P_X P_Y) = \int_{\mathcal{X} \times \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy.$$

It quantifies the amount of information shared between  $X$  and  $Y$ .

# Common distances and divergences

Unlike  $f$ -divergences, which compare probability densities pointwise, the *Wasserstein distance* (or *optimal transport distance*) takes the geometry of the space  $\mathcal{X}$  into account. For  $p \geq 1$ , the  $p$ -Wasserstein distance between  $P$  and  $Q$  is defined as

$$W_p(P, Q) = \left( \inf_{\pi \in \Pi(P, Q)} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|^p d\pi(x, y) \right)^{1/p},$$

where  $\Pi(P, Q)$  denotes the set of all *couplings* of  $P$  and  $Q$ , i.e., joint distributions on  $\mathcal{X} \times \mathcal{X}$  with marginals  $P$  and  $Q$  respectively.

# Sliced-Wasserstein distance

$$W_r(P, Q) = \left( \int_0^1 |F_P^{-1}(t) - F_Q^{-1}(t)|^r dt \right)^{\frac{1}{r}}$$

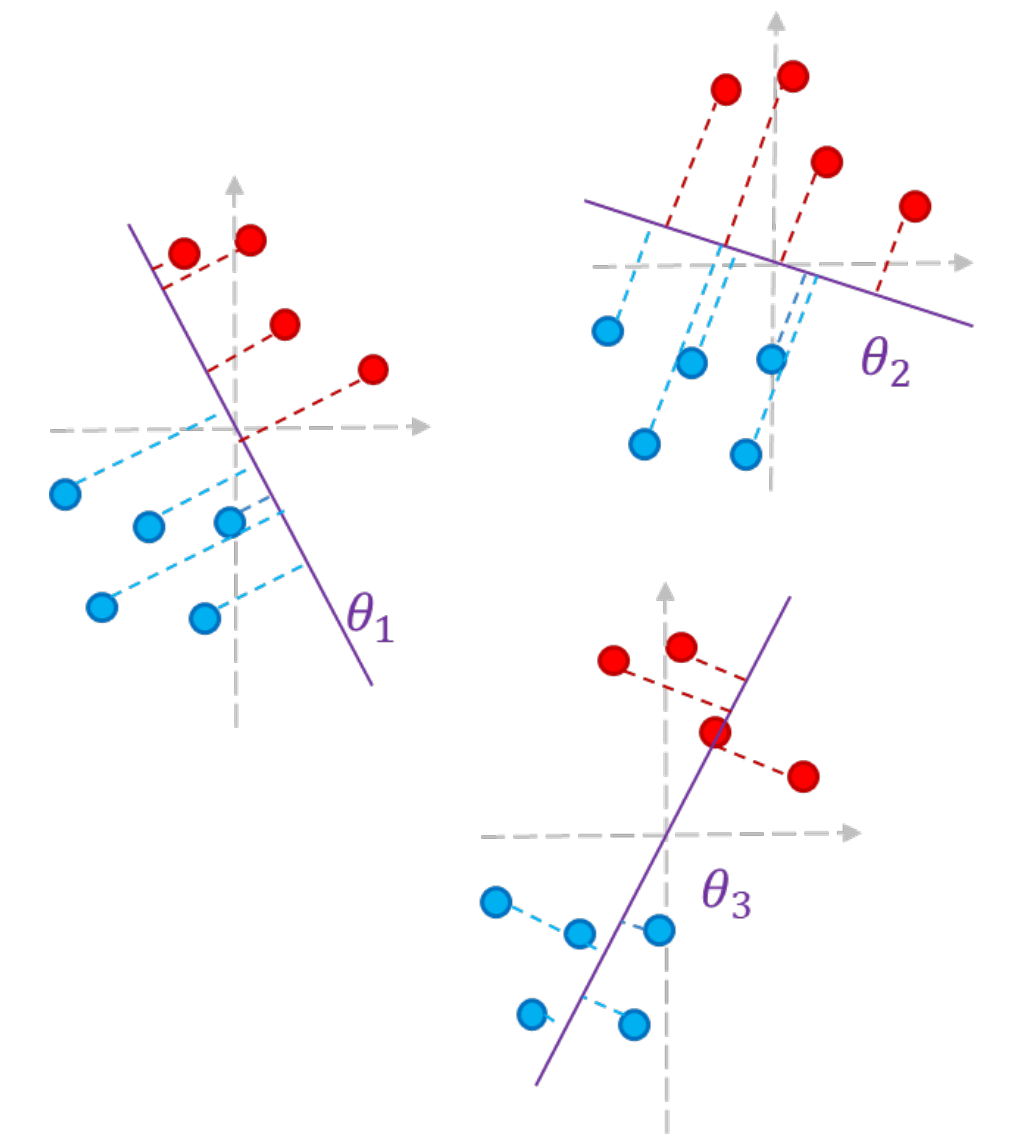
Wassertein in dimension 1 is easy

$$SW_r(P, Q) = \left( \int_{\mathbb{S}^{s-1}} W_r(\theta_{\#}^* P, \theta_{\#}^* Q)^r d\sigma(\theta) \right)^{\frac{1}{r}}$$

- $P, Q$  distribution on  $\mathbb{R}^s$
- $\mathbb{S}^{s-1}$  is the  $(s - 1)$ -dimensional unit sphere
- $\sigma$  uniform distribution on  $\mathbb{S}^{s-1}$
- $\theta^*$  projection function on direction  $\theta \in \mathbb{S}^{s-1}$
- $\theta_{\#}^* P$  push-forward measure of  $P$  by  $\theta^*$

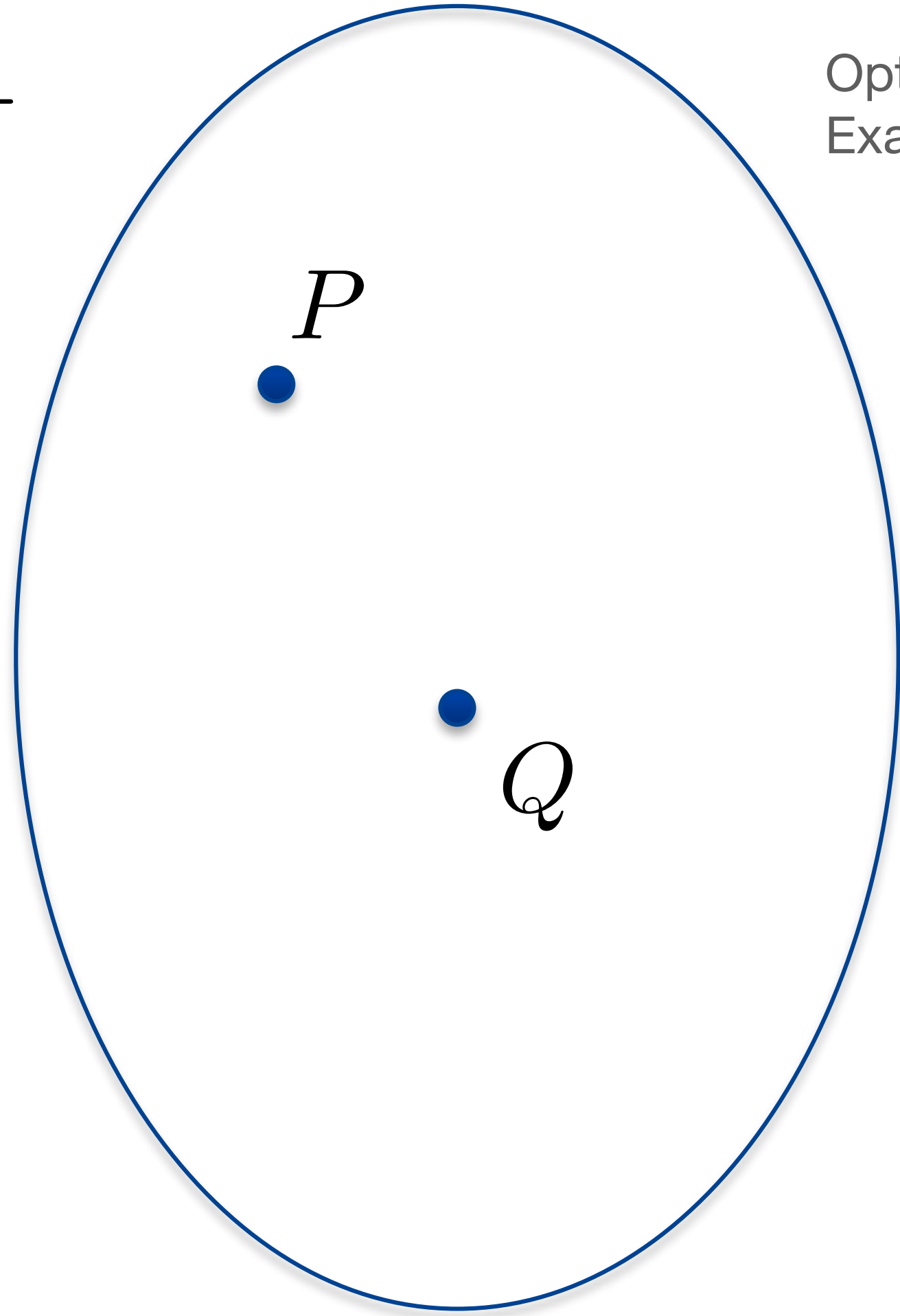
$$\widehat{SW}_2^2(P, Q) = \frac{1}{R} \sum_{r=1}^R \widehat{W}_2^2(\theta_{r,\#}^* P, \theta_{r,\#}^* Q)$$

$\theta_1, \dots, \theta_R$  projection directions  
uniformly drawn on  $\mathbb{S}^{s-1}$



# Kernel-embedding of probability distributions

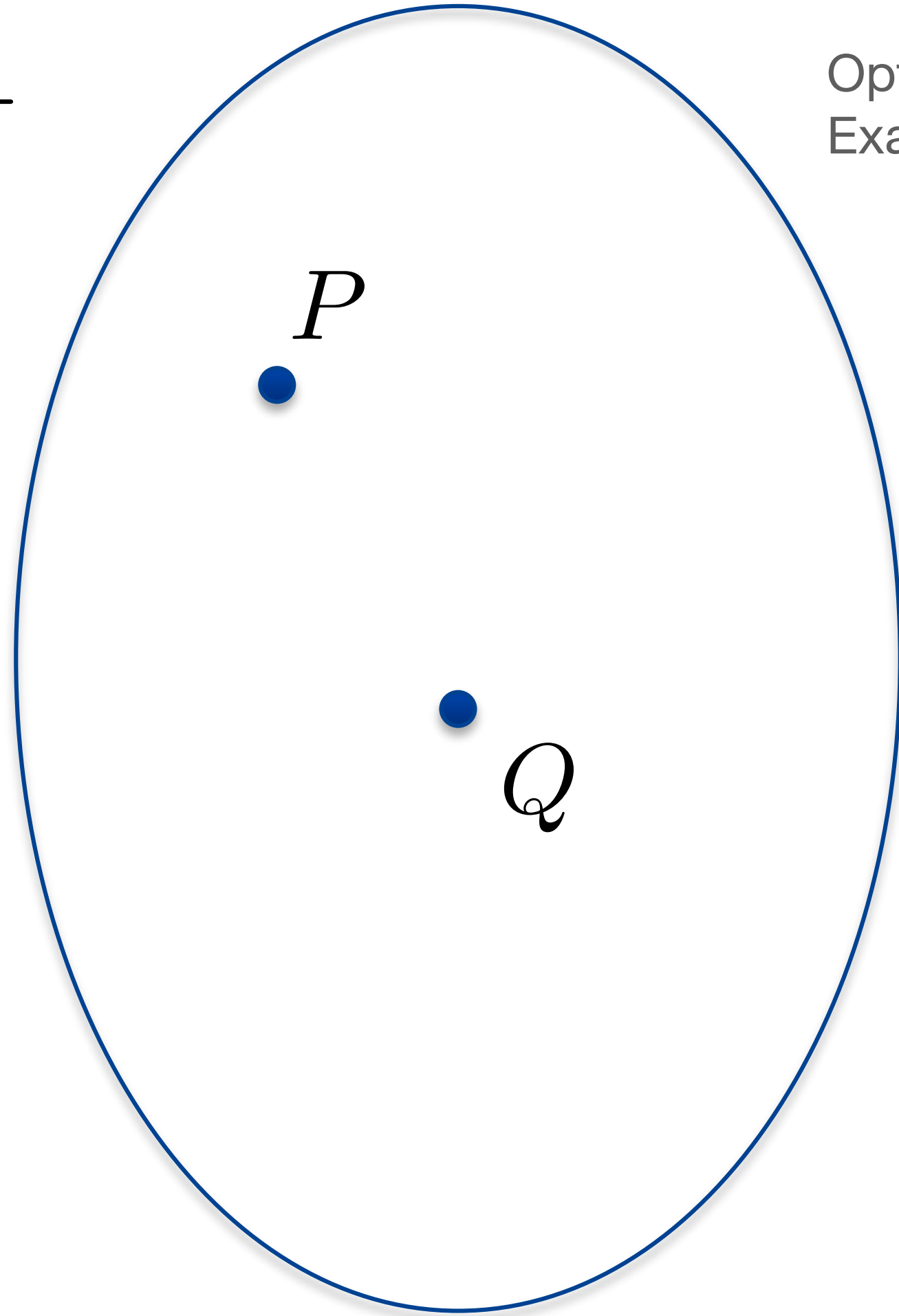
$\mathcal{M}_1^+$



Option 1: work directly in the space of probability measures  
Examples: KS, TV, KL, Hellinger, ...

# Kernel-embedding of probability distributions

$\mathcal{M}_1^+$

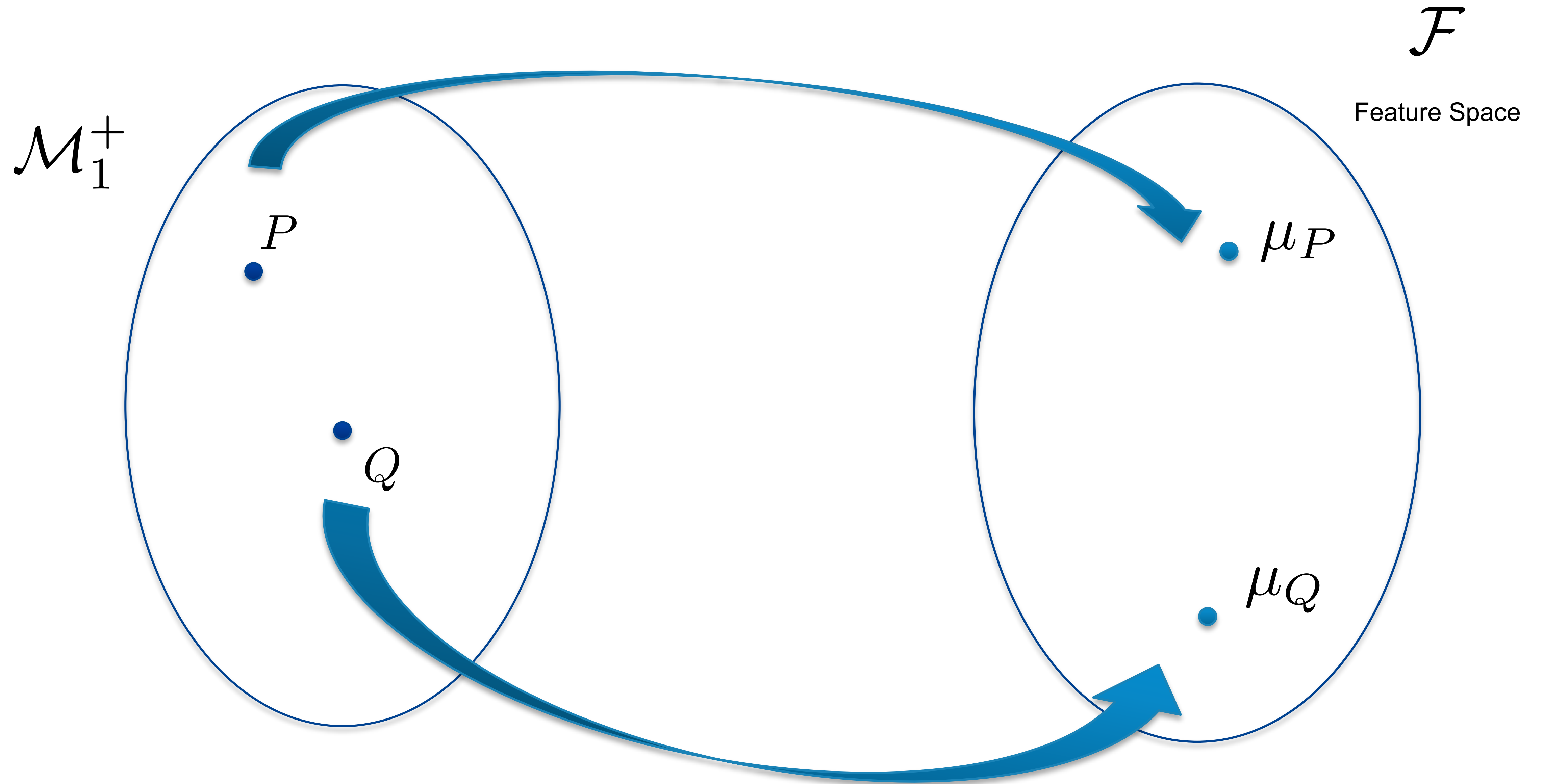


Option 1: work directly in the space of probability measures  
Examples: KS, TV, KL, Hellinger, ...

**Option 2: represent probability measures with some features**

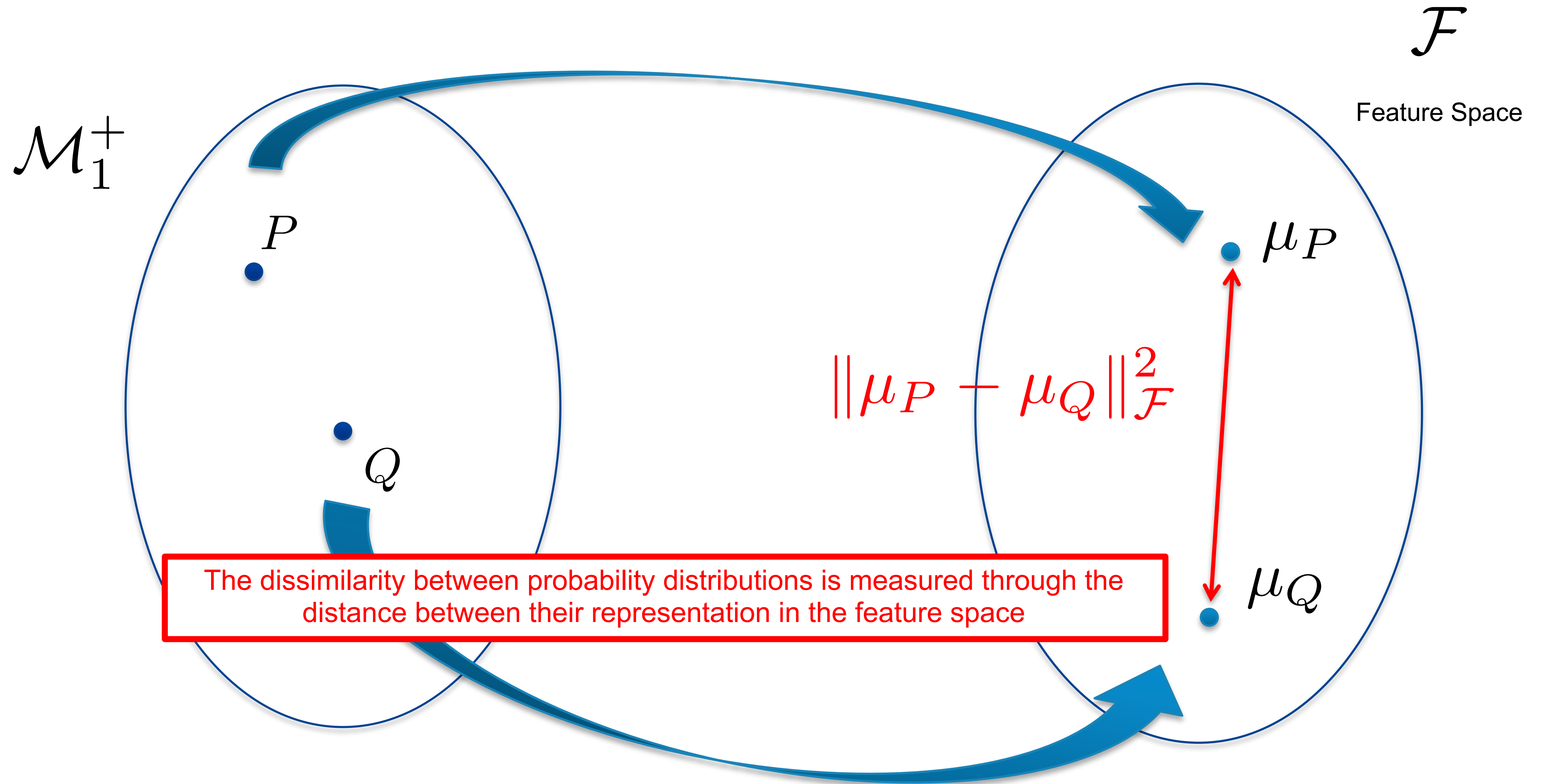


# Kernel-embedding of probability distributions



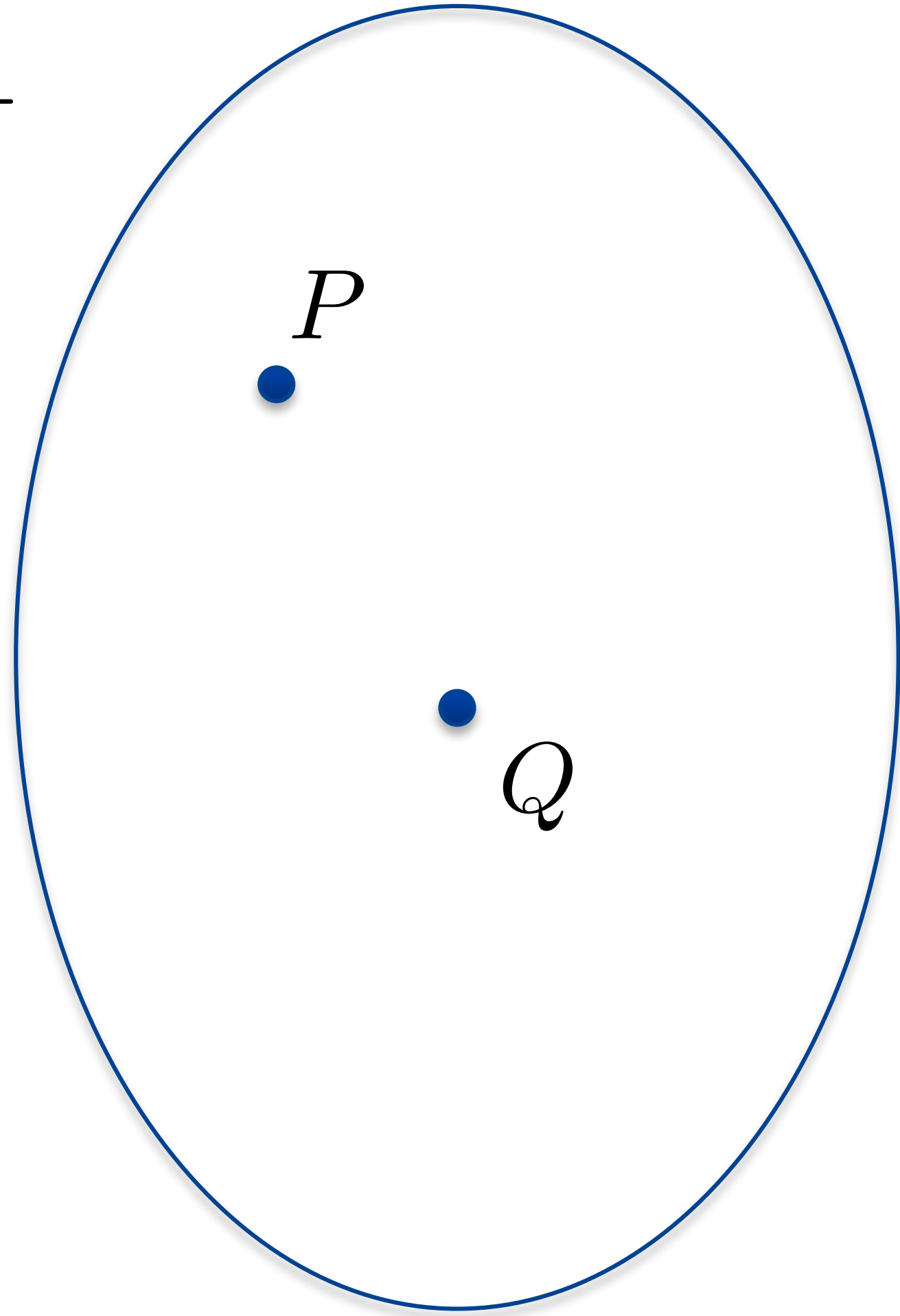


# Kernel-embedding of probability distributions



# Kernel-embedding of probability distributions

$\mathcal{M}_1^+$



$\mathcal{F}$

Feature Space

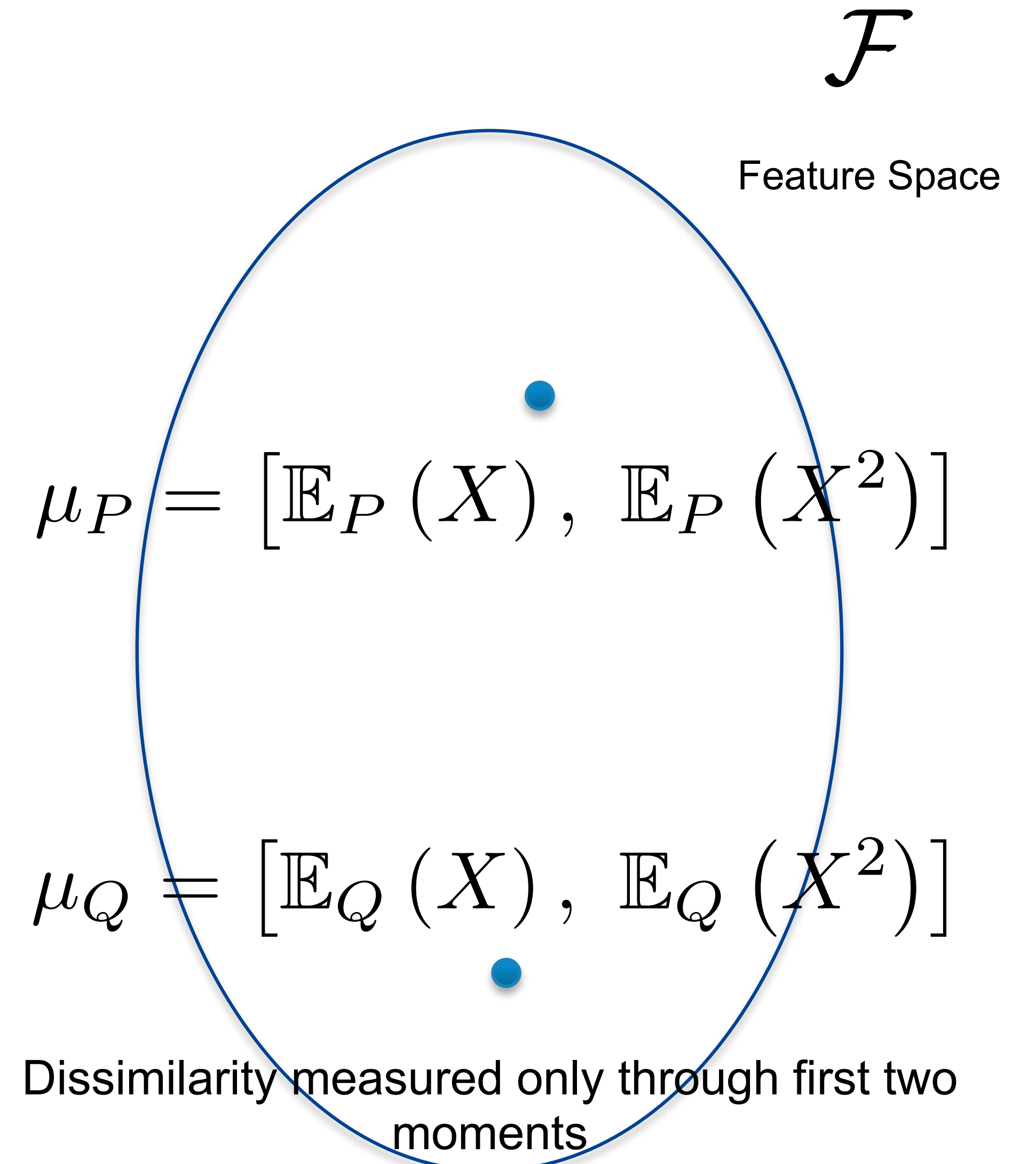
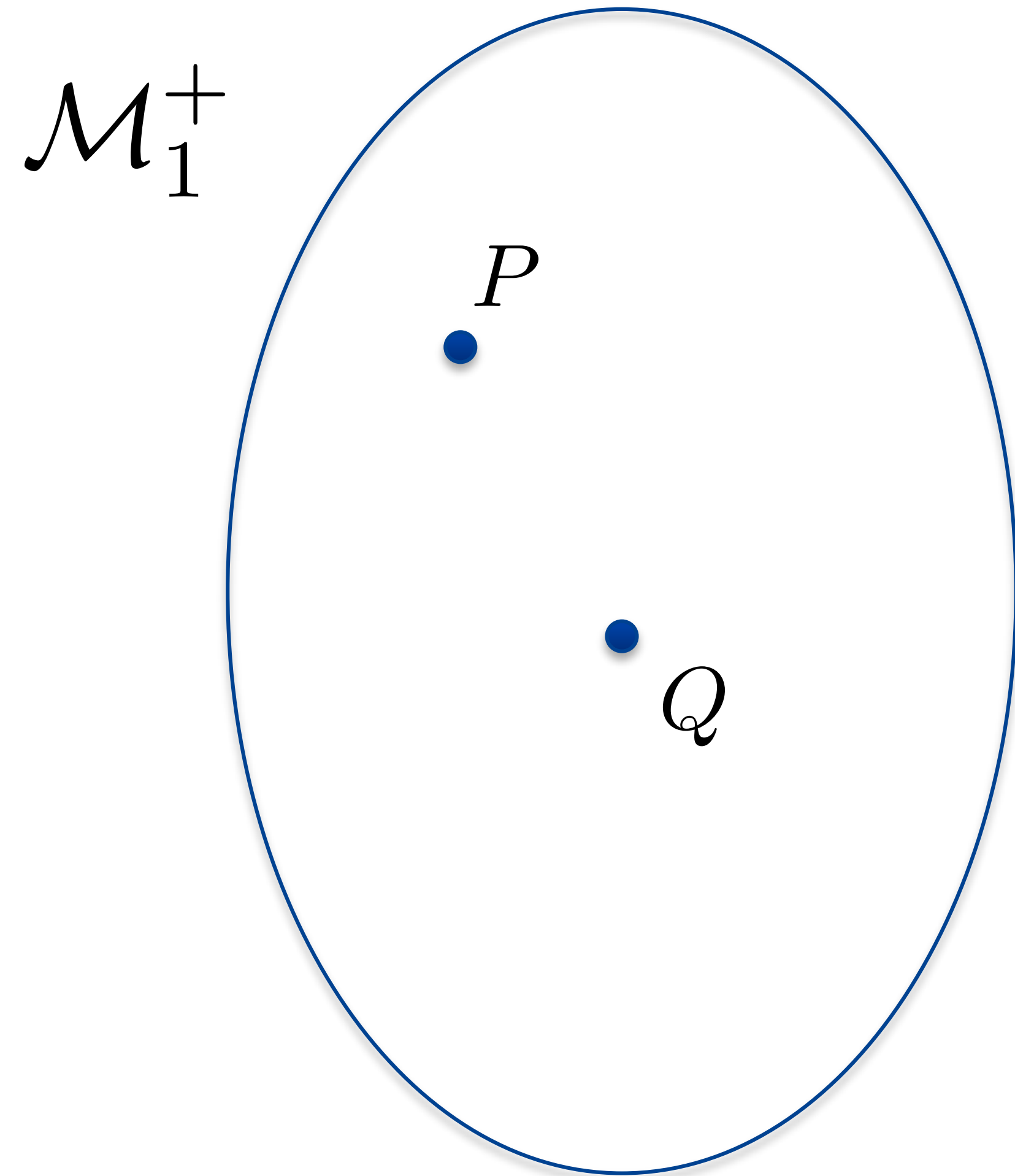
A large blue oval representing the Feature Space  $\mathcal{F}$ . Inside the oval, there are two blue dots. The upper dot is labeled  $\mu_P = \mathbb{E}_P(X)$  and the lower dot is labeled  $\mu_Q = \mathbb{E}_Q(X)$ .

$$\mu_P = \mathbb{E}_P(X)$$

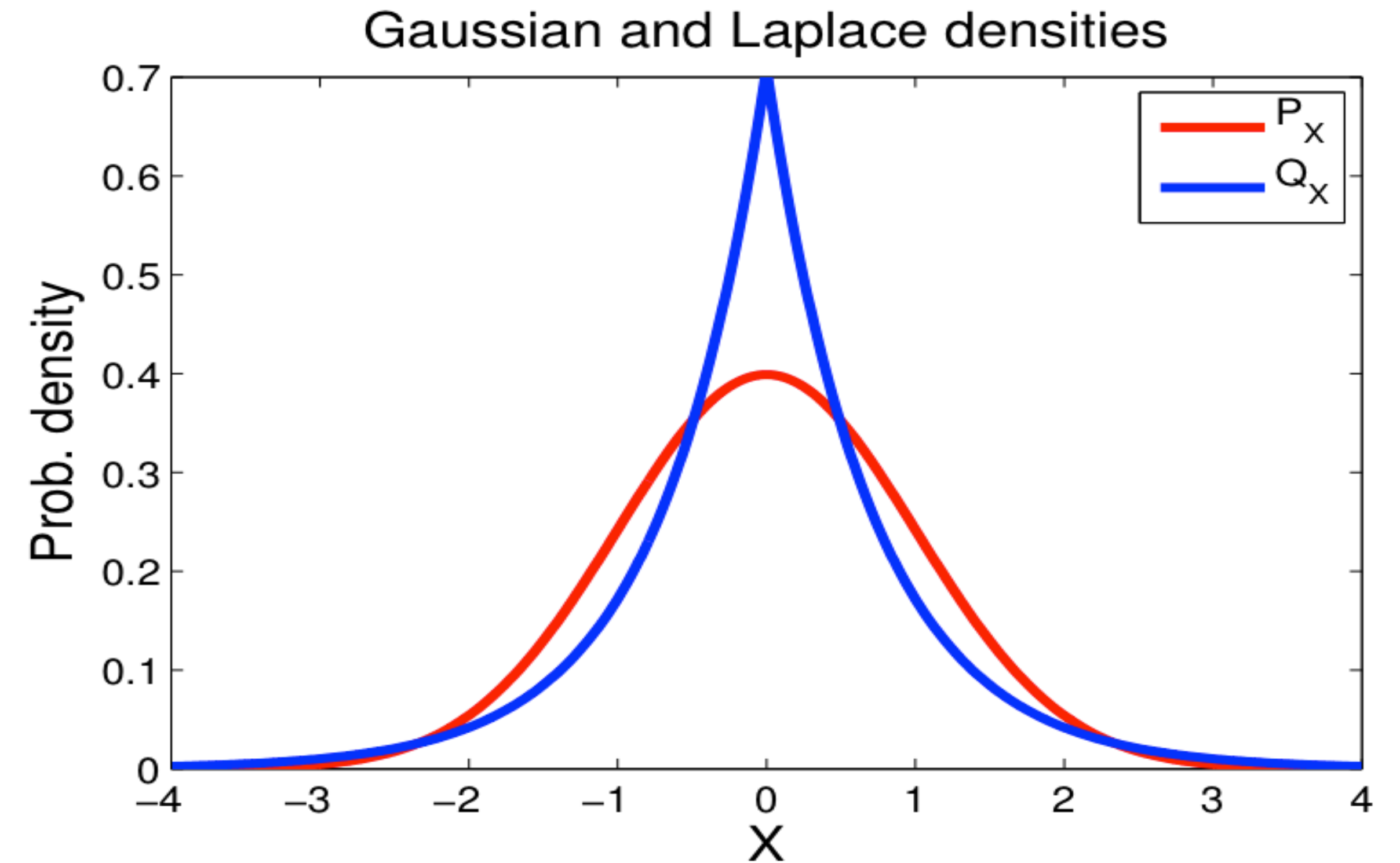
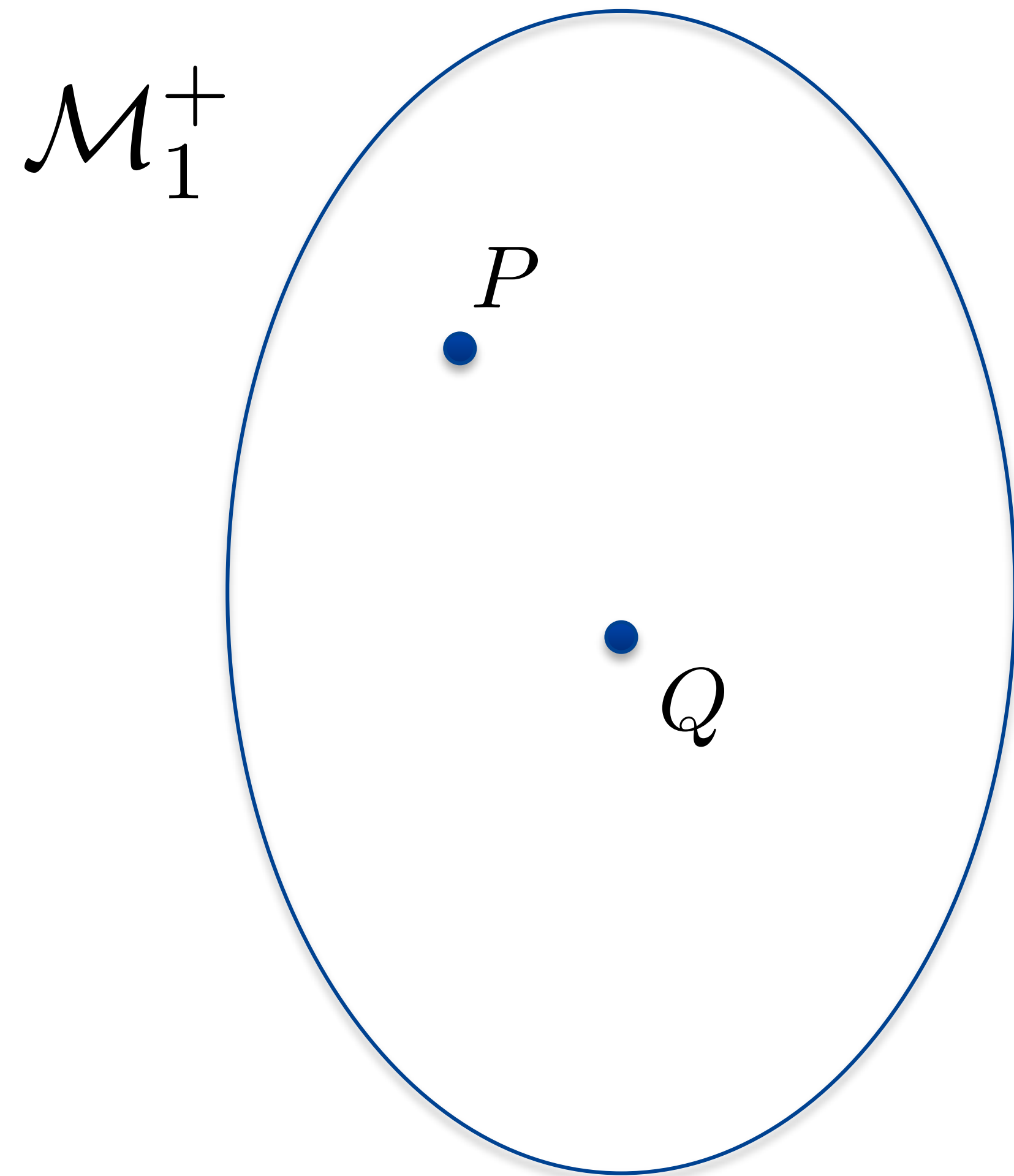
$$\mu_Q = \mathbb{E}_Q(X)$$

Dissimilarity measured only through the means

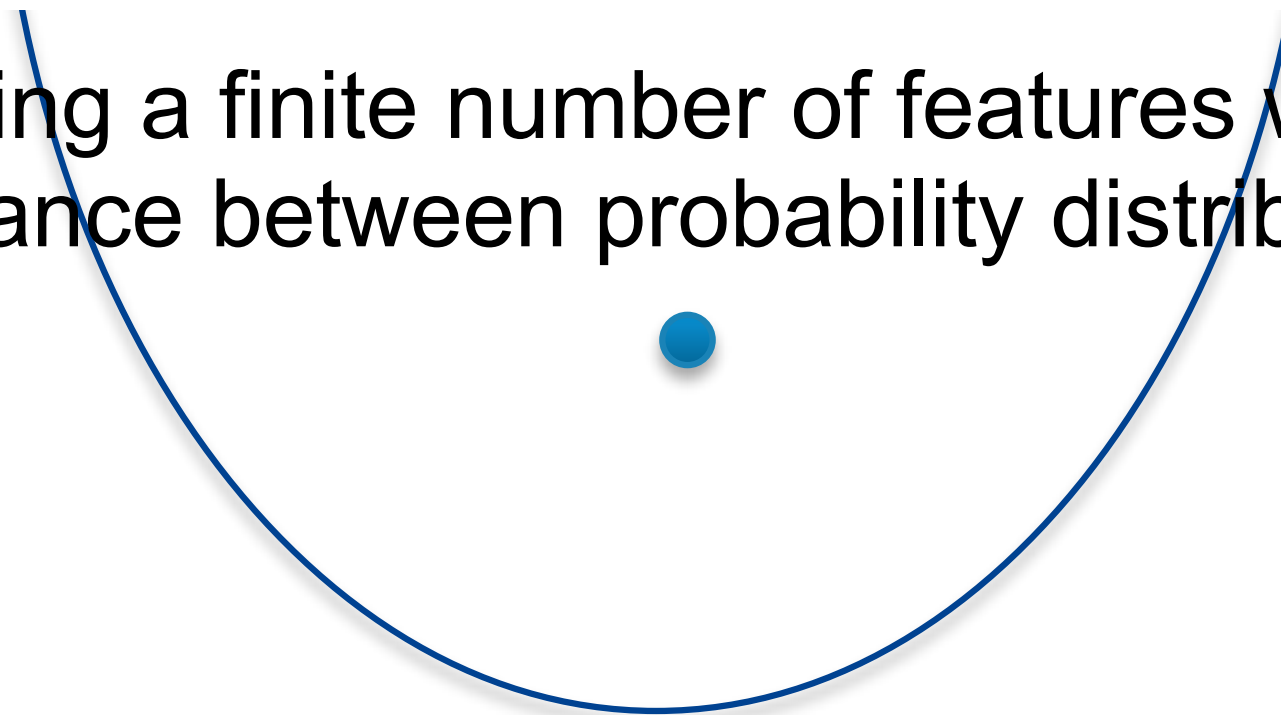
# Kernel-embedding of probability distributions



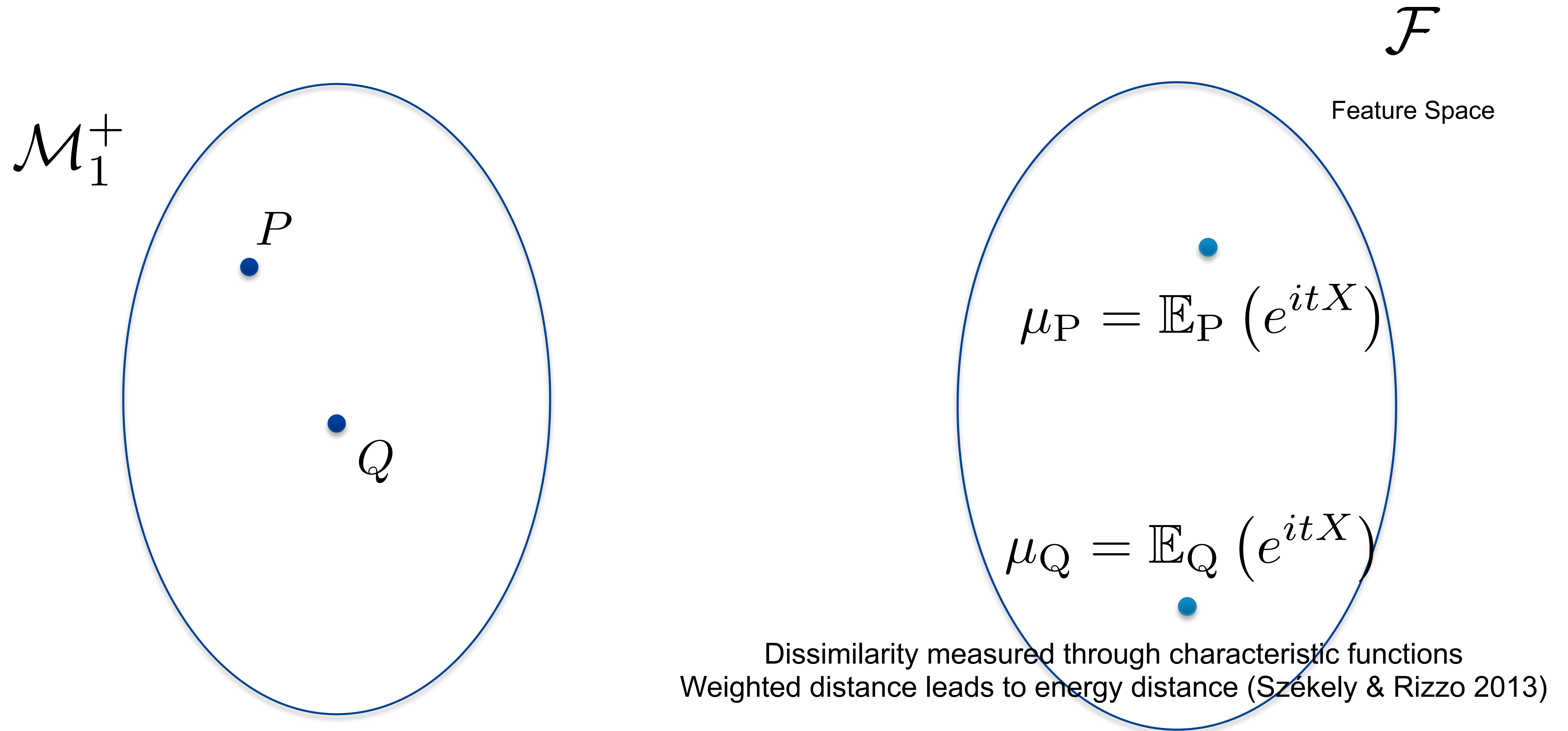
# Kernel-embedding of probability distributions



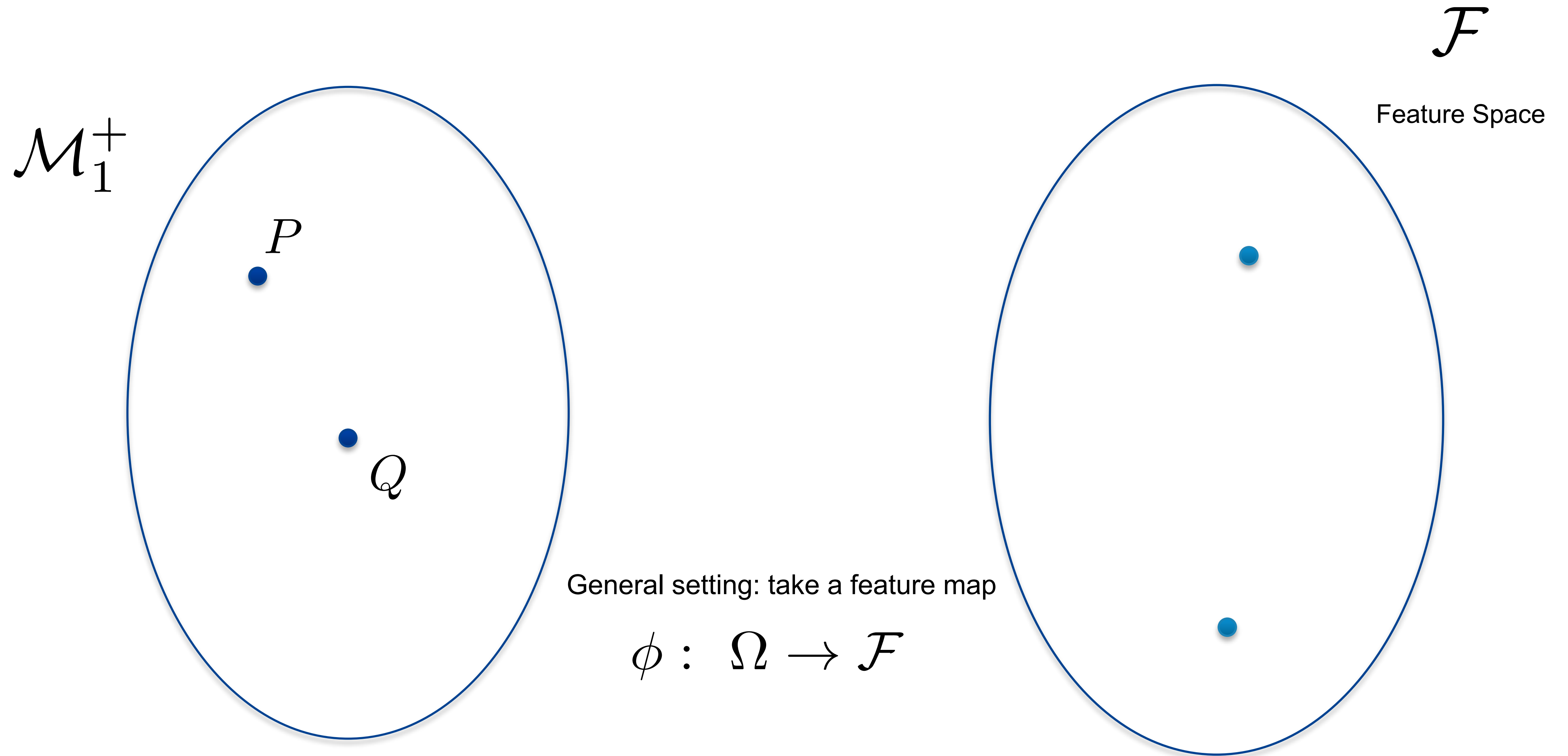
Obviously using a finite number of features will not lead to a distance between probability distributions



# Kernel-embedding of probability distributions

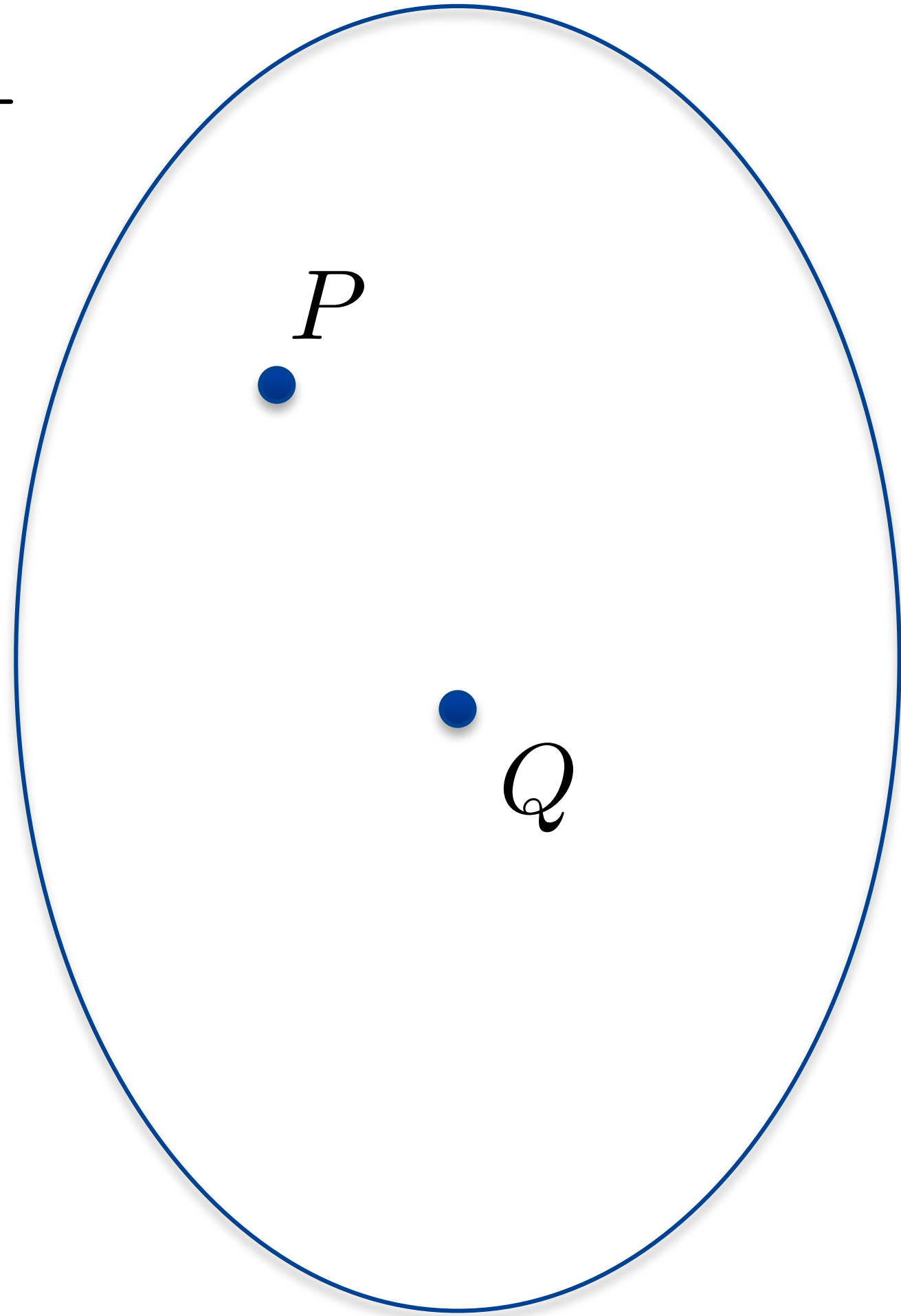


# Kernel-embedding of probability distributions



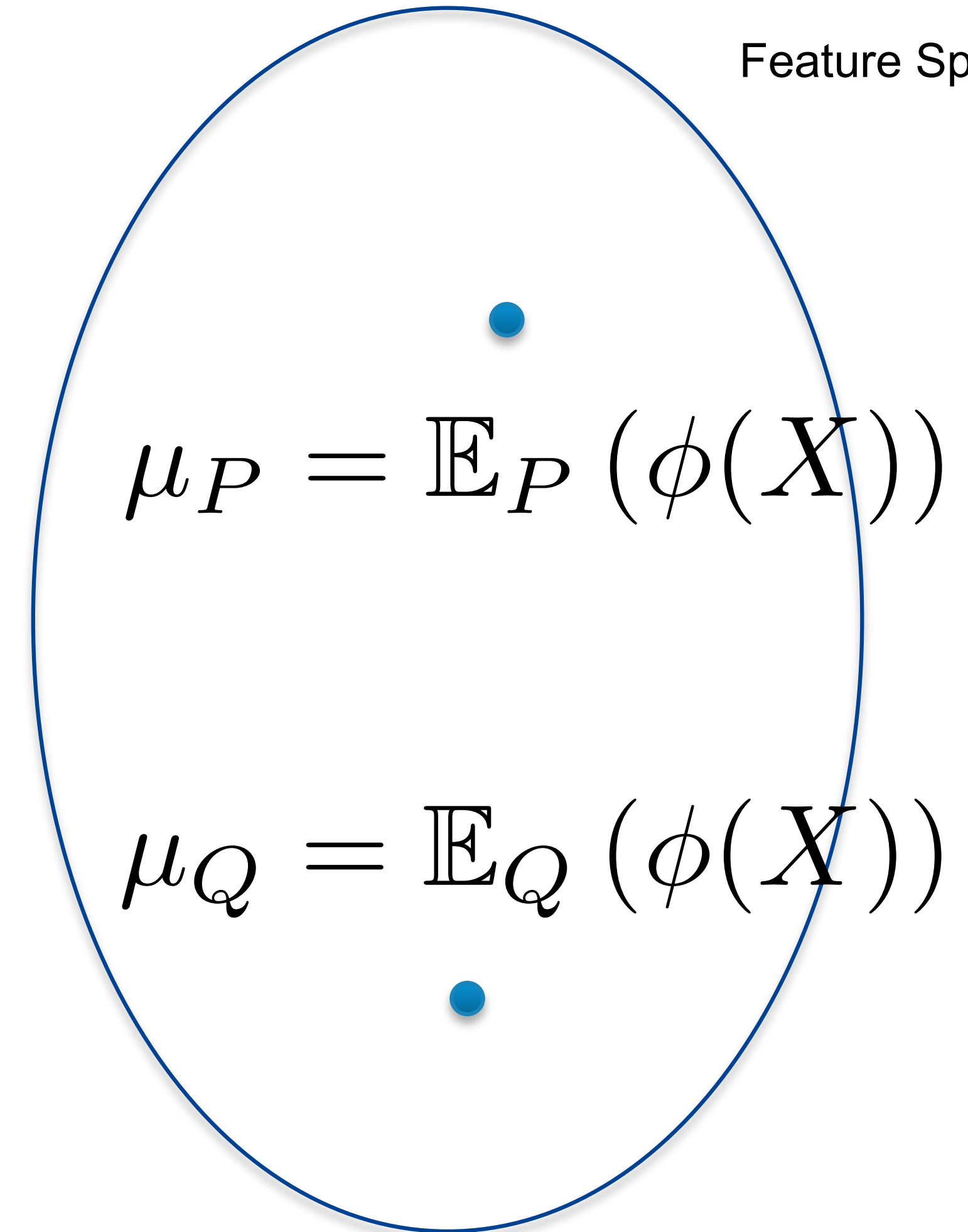
# Kernel-embedding of probability distributions

$\mathcal{M}_1^+$



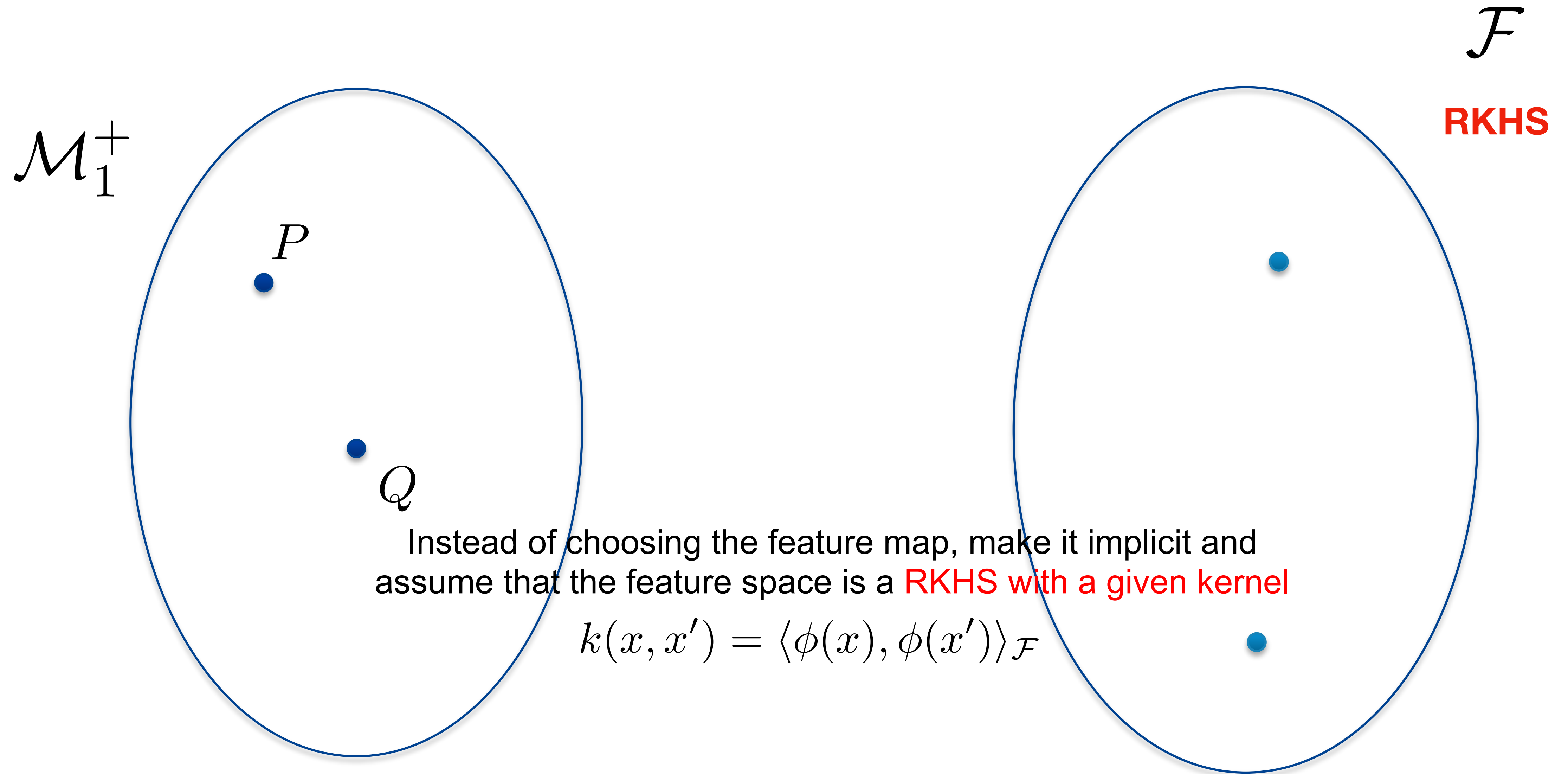
$\mathcal{F}$

Feature Space





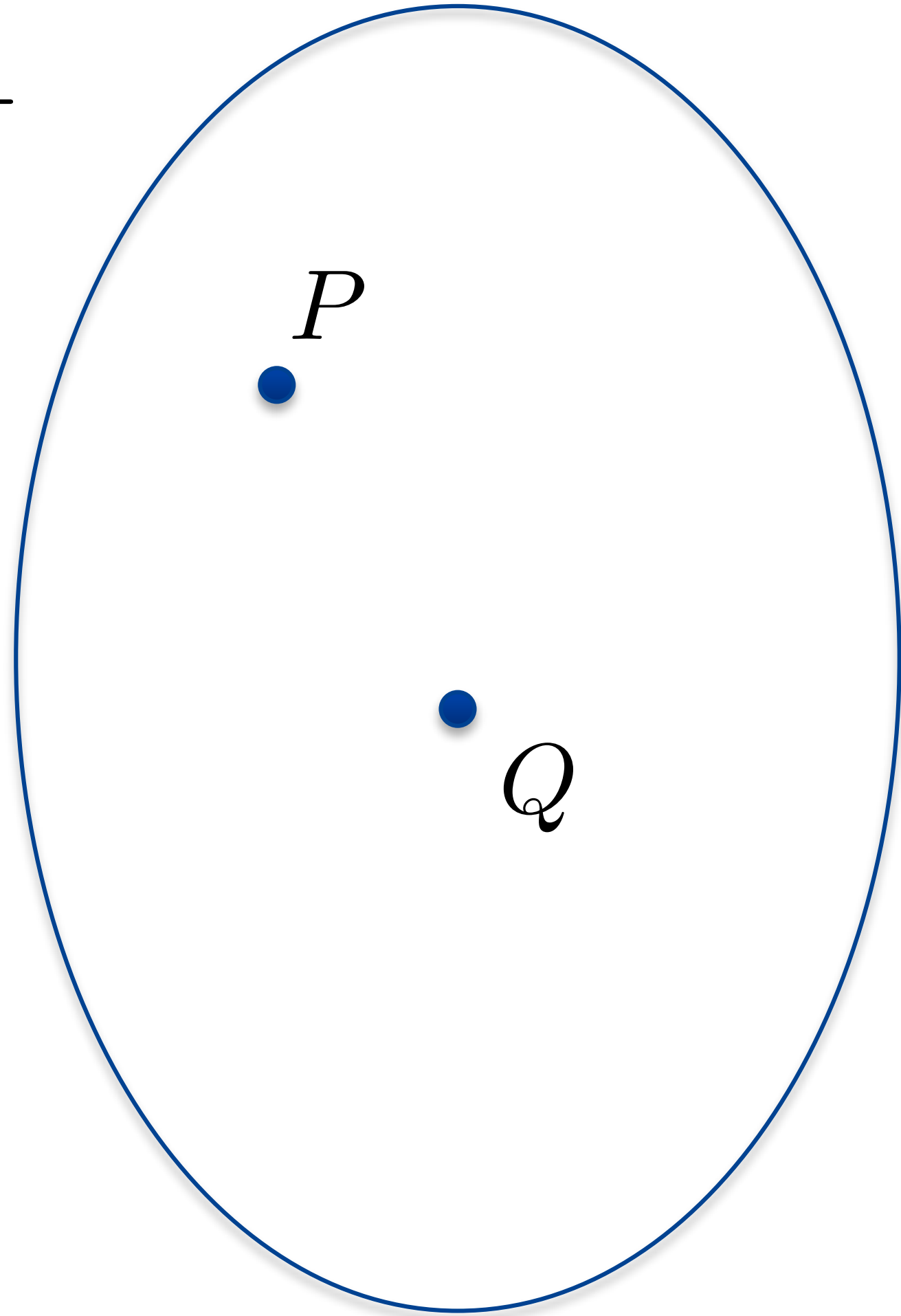
# Kernel-embedding of probability distributions





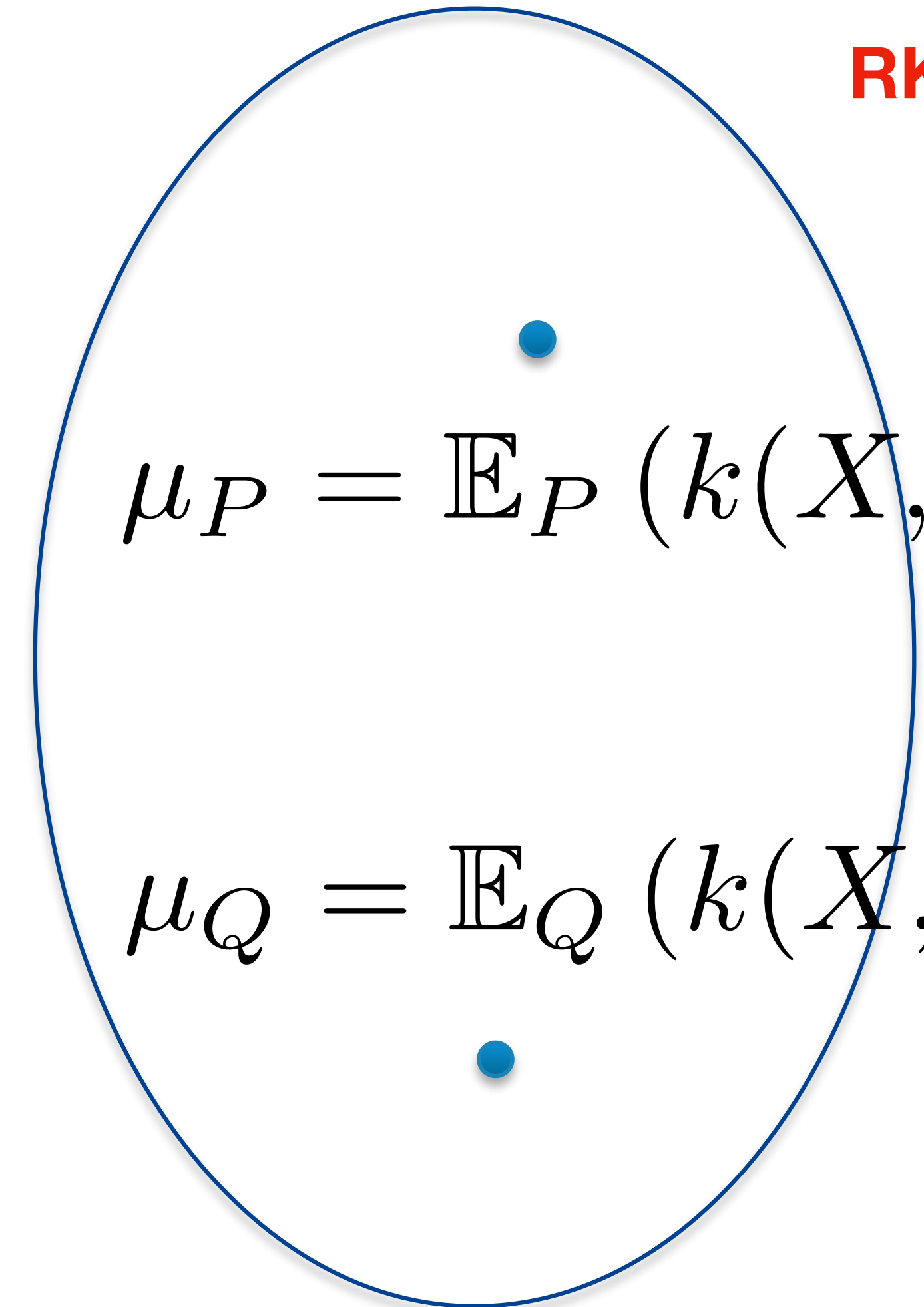
# Kernel-embedding of probability distributions

$\mathcal{M}_1^+$



$\mathcal{F}$

**RKHS**



# Kernel-embedding of probability distributions

The kernel mean embedding of a probability measure is defined as

$$\mu_P = \mathbb{E}_{\xi \sim P} k_{\mathcal{X}}(\xi, \cdot) = \int_{\mathcal{X}} k_{\mathcal{X}}(\xi, \cdot) dP(\xi)$$

A distance between probability measures is then given by the **Maximum Mean Discrepancy**

$$\text{MMD}(P_1, P_2) = \|\mu_{P_1} - \mu_{P_2}\|_{\mathcal{H}}$$

The reproducing property in the RKHS gives the central result

$$\text{MMD}^2(P_1, P_2) = \mathbb{E}_{\xi, \xi'} k_{\mathcal{X}}(\xi, \xi') - 2\mathbb{E}_{\xi, \zeta} k_{\mathcal{X}}(\xi, \zeta) + \mathbb{E}_{\zeta, \zeta'} k_{\mathcal{X}}(\zeta, \zeta')$$

# Kernel-embedding of probability distributions

## Advantages of this distance vs others

- Thanks to the RKHS, only involves **expectations of kernels**
- Less prone to the curse of dimensionality
- **Can easily handle structured objects** (curves, images, graphs, probability measures, sets) by using specific kernels
- (This is a distance only if a *characteristic kernel* is used)

# Kernel-embedding of probability distributions

Other major use: testing independence of random vectors

$$\text{MMD}^2(P_{\mathbf{UV}}, P_{\mathbf{U}} \otimes P_{\mathbf{V}}) = \|\mu_{P_{\mathbf{UV}}} - \mu_{P_{\mathbf{U}}} \otimes \mu_{P_{\mathbf{V}}}\|_{\mathcal{H}}^2$$

$$\begin{aligned} \text{HSIC}(\mathbf{U}, \mathbf{V}) &= \text{MMD}^2(P_{\mathbf{UV}}, P_{\mathbf{U}} \otimes P_{\mathbf{V}}) \\ &= \mathbb{E}_{\mathbf{U}, \mathbf{U}', \mathbf{V}, \mathbf{V}'} k_{\mathcal{X}}(\mathbf{U}, \mathbf{U}') k_{\mathcal{Y}}(\mathbf{V}, \mathbf{V}') \\ &\quad + \mathbb{E}_{\mathbf{U}, \mathbf{U}'} k_{\mathcal{X}}(\mathbf{U}, \mathbf{U}') \mathbb{E}_{\mathbf{V}, \mathbf{V}'} k_{\mathcal{Y}}(\mathbf{V}, \mathbf{V}') \\ &\quad - 2\mathbb{E}_{\mathbf{U}, \mathbf{V}} [\mathbb{E}_{\mathbf{U}'} k_{\mathcal{X}}(\mathbf{U}, \mathbf{U}') \mathbb{E}_{\mathbf{V}'} k_{\mathcal{Y}}(\mathbf{V}, \mathbf{V}')] \end{aligned}$$

Gretton et al. 2005a,b

Many applications: goodness-of-fit, independence tests, feature selection, ...



# Kernel-embedding of probability distributions



**ETICS 2022**

Other major use: testing independence

$$\text{MMD}^2(P_{UV}, P_U \otimes P_V)$$

$$\begin{aligned} \text{HSIC}(U, V) &= \mathbb{E}[\text{K}(U, U)] \\ &= \mathbb{E}_U[\text{K}(U, U)] \\ &+ \mathbb{E}_V[\text{K}(V, V)] \\ &- 2\mathbb{E}[\text{K}(U, V)] \end{aligned}$$

Many applications: goodness-of-fit,

**École Thématique sur les Incertitudes en Calcul Scientifique**  
**Research School on Uncertainty in Scientific Computing**

<https://www.gdr-mascotnum.fr/etics.html>

October, 2-7, Belhambra, Belgodère Golfe de Lozari , France -  
<https://www.belambra.com/club-belgodere-golfe-de-lozari/summer>



Talk of D. J. Sutherland

**Several appearances in ETICS community**

# **Several appearances in ETICS community**

## **1 - Sensitivity analysis**

# Sensitivity analysis

- **What is sensitivity analysis?**

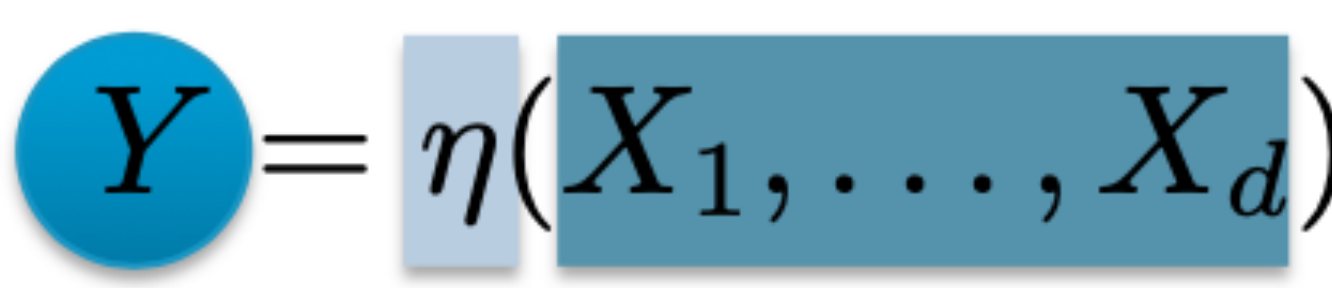
- Originates from the field of **computer experiments**
- Main goal: **identify and rank** the input parameters according to their impact on the output of a computer code
- Why?
  - Simplify the model
  - Improve the knowledge of the physical phenomenon
  - For uncertainty quantification, we can reduce the output uncertainty by focusing on the main input contributors

- **Notation**

Computer code

Output  $Y = \eta(X_1, \dots, X_d)$

Input parameters



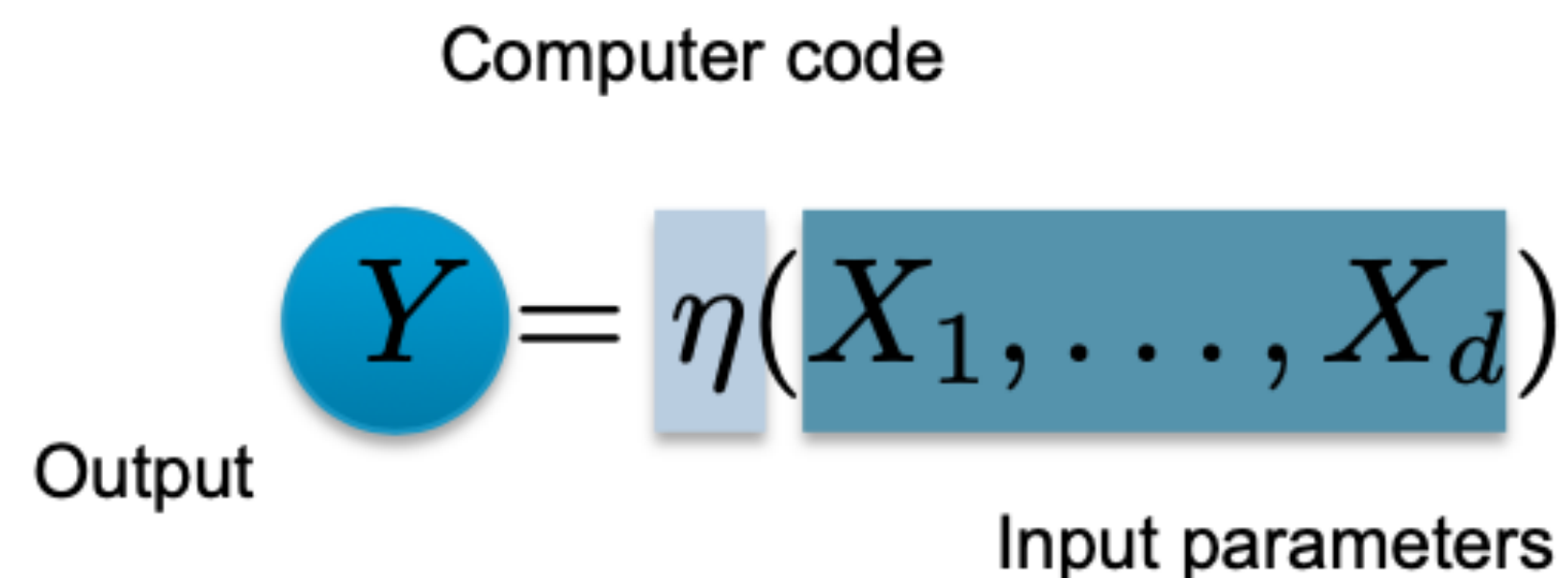


# Sensitivity analysis

- **What is sensitivity analysis?**

- Originates from the field of **computer experiments**
- Main goal: **identify and rank** the input parameters according to their impact on the output of a computer code
- Why?
  - Simplify the model
  - Improve the knowledge of the physical phenomenon
  - For uncertainty quantification, we can reduce the output uncertainty by focusing on the main input contributors

- **Notation**



**As a side note**

- Replace « computer code » by « ML model » trained on a data set
- The goal of SA actually corresponds to assessing the **feature importance** in a given ML model
- Consequently, SA has many strong links with the field of **explainability and interpretability** in modern ML

# Sensitivity analysis: beyond Sobol' indices

- **General framework for moment independent indices**

$$\mathcal{S}_l = \mathbb{E}_{X_l} \left( d(P_Y, P_{Y|X_l}) \right)$$

Baucells & Borgonovo 2013  
D. 2015

- If the output probability distribution and the conditional one are « close », the input parameter has little influence

# Sensitivity analysis: beyond Sobol' indices

- **General framework for moment independent indices**

$$\mathcal{S}_l = \mathbb{E}_{X_l} \left( d(P_Y, P_{Y|X_l}) \right)$$

Baucells & Borgonovo 2013  
D. 2015

- If the output probability distribution and the conditional one are « close », the input parameter has little influence
- Example: f-divergence (D. 2015, Rahman 2016), with particular cases TV & KL

# Sensitivity analysis: beyond Sobol' indices

- **General framework for moment independent indices**

$$\mathcal{S}_l = \mathbb{E}_{X_l} \left( d(P_Y, P_{Y|X_l}) \right)$$

Baucells & Borgonovo 2013  
D. 2015

- If the output probability distribution and the conditional one are « close », the input parameter has little influence
- Example: f-divergence (D. 2015, Rahman 2016), with particular cases TV & KL



# Sensitivity analysis: beyond Sobol' indices

- **General framework for moment independent indices**

$$\mathcal{S}_l = \mathbb{E}_{X_l} \left( d(P_Y, P_{Y|X_l}) \right)$$

Baucells & Borgonovo 2013  
D. 2015

- If the output probability distribution and the conditional one are « close », the input parameter has little influence

- Example: f-divergence (D. 2015, Rahman 2016), with particular cases TV & KL

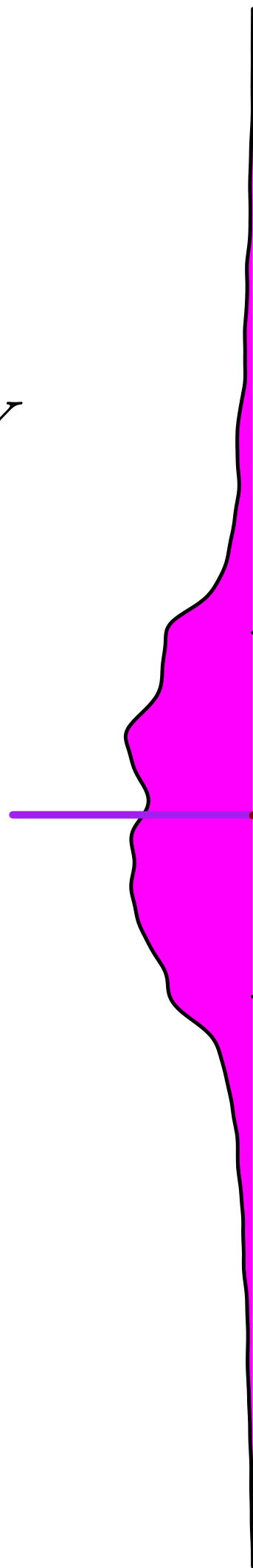
- Toy example

$$Y = \sin(X_1) + 7 \sin(X_2)^2 + X_3^4 \sin(X_1)$$

$$X_l \sim \mathcal{U}(-\pi, \pi) \text{ for } l = 1, \dots, 4$$

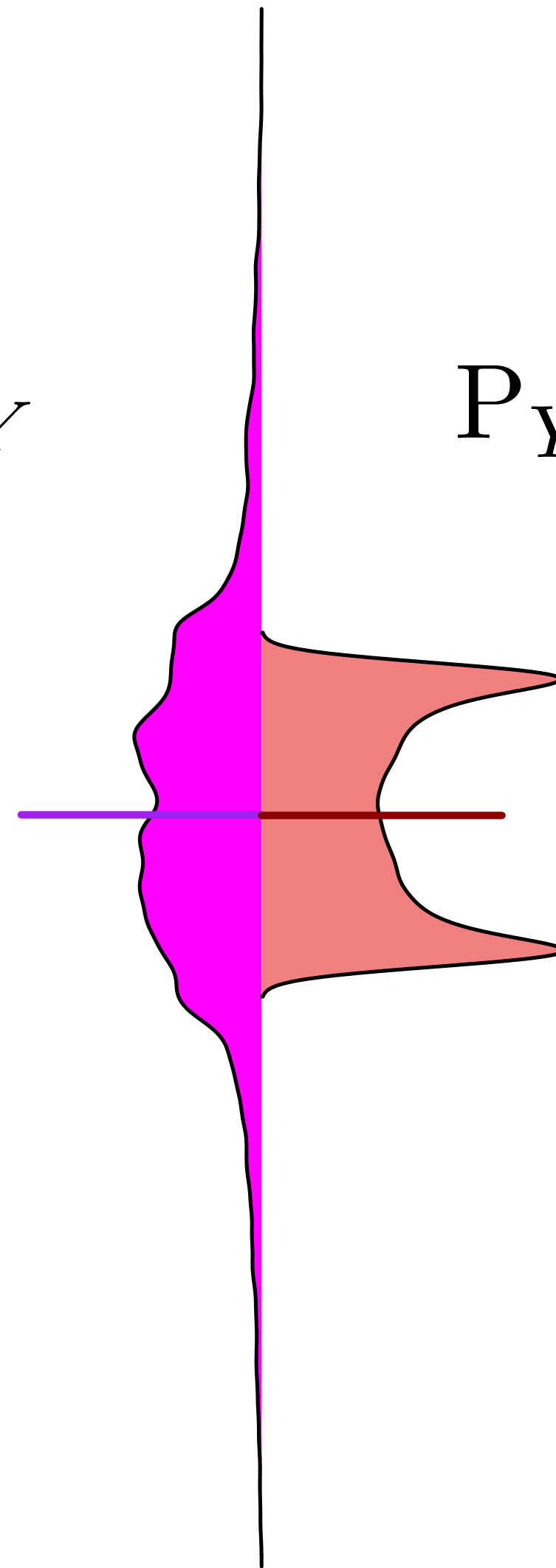


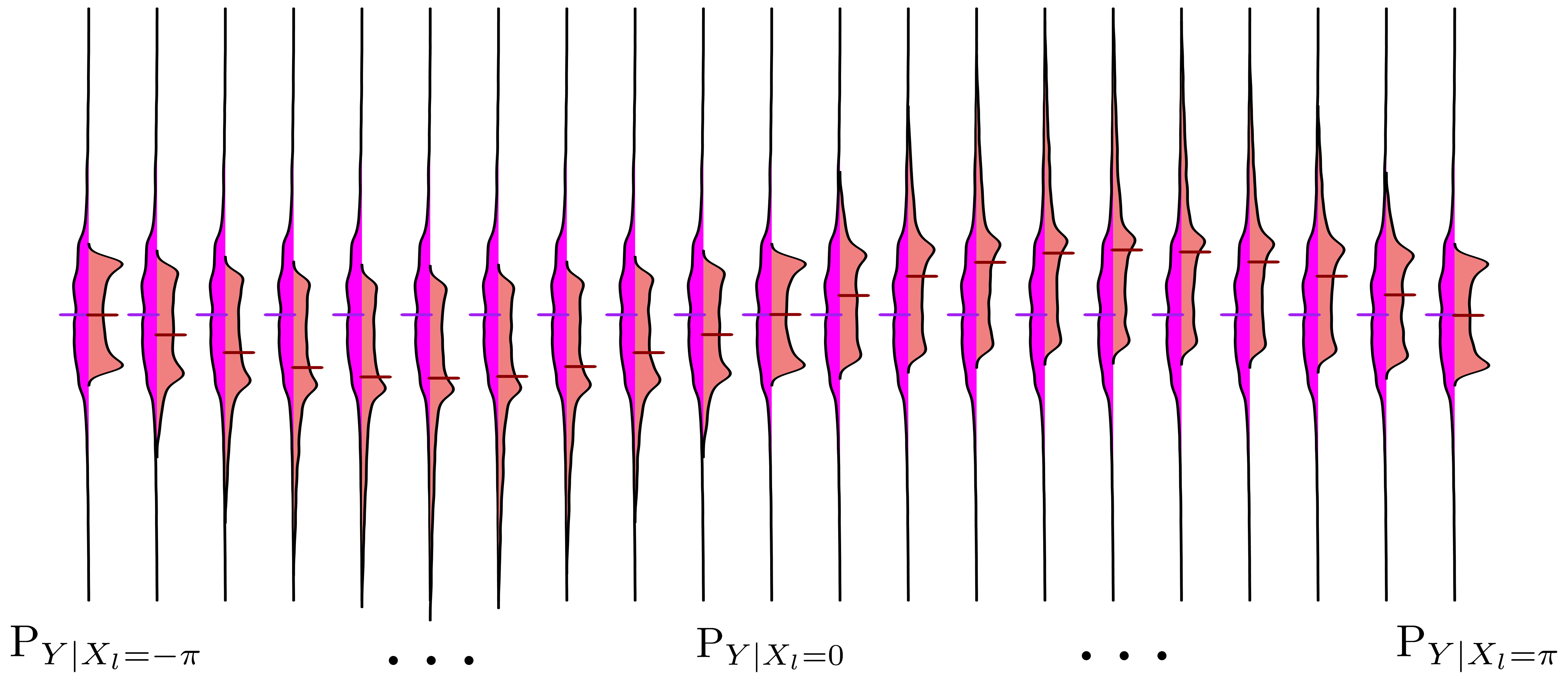
$P_Y$



$P_Y$

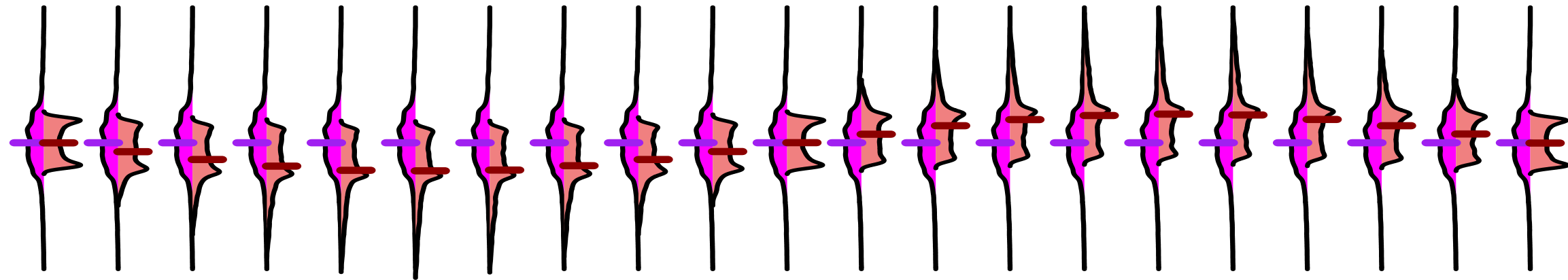
$P_{Y|X_l=0}$



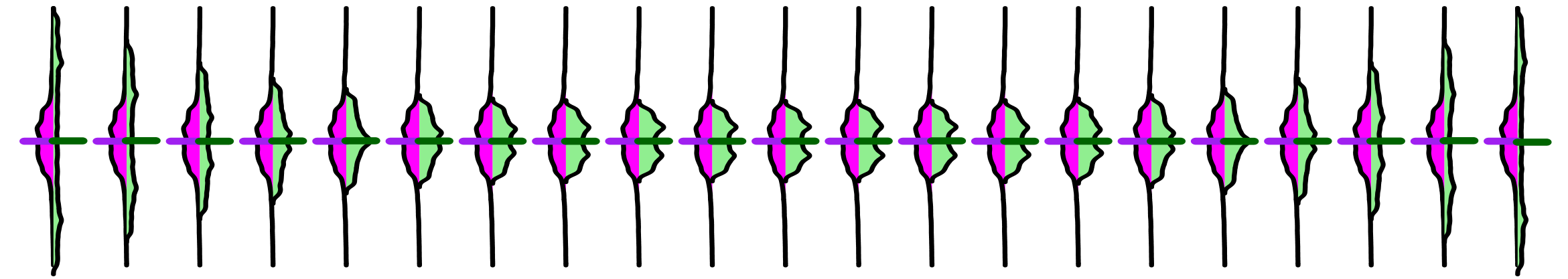




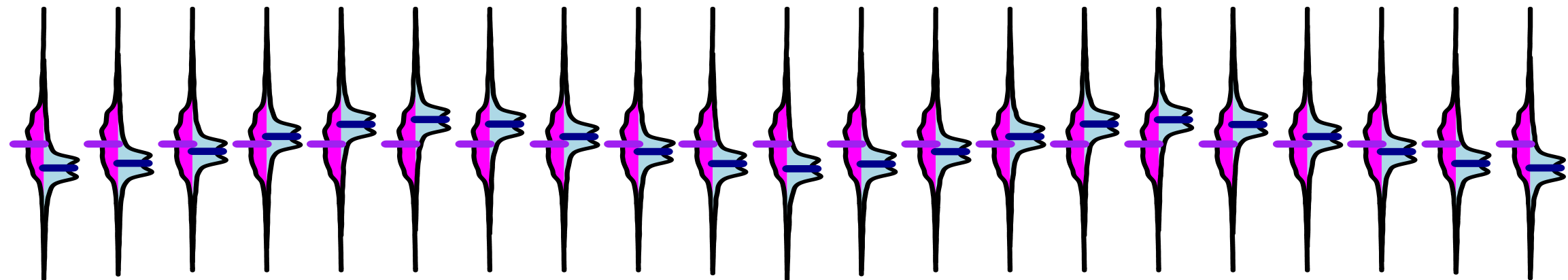
**X1 fixed**



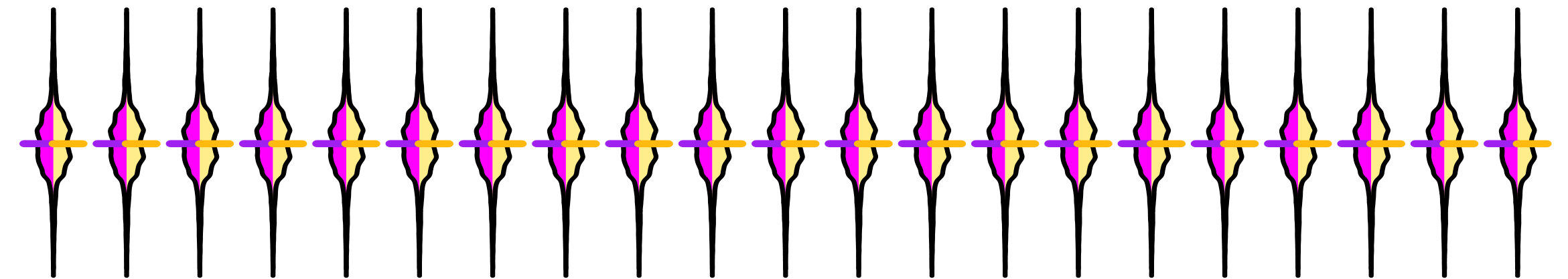
**X3 fixed**

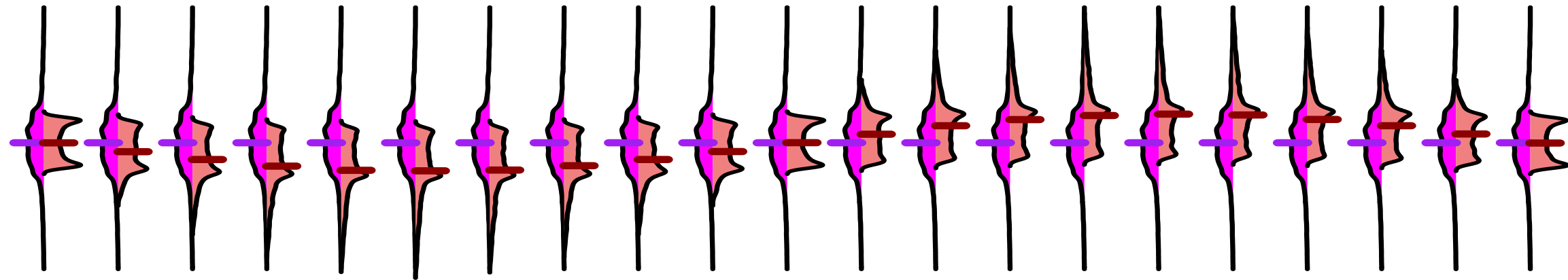
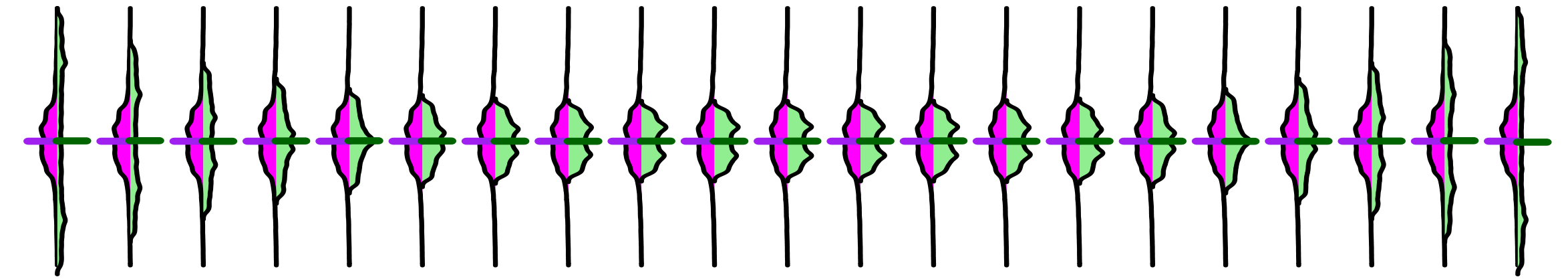
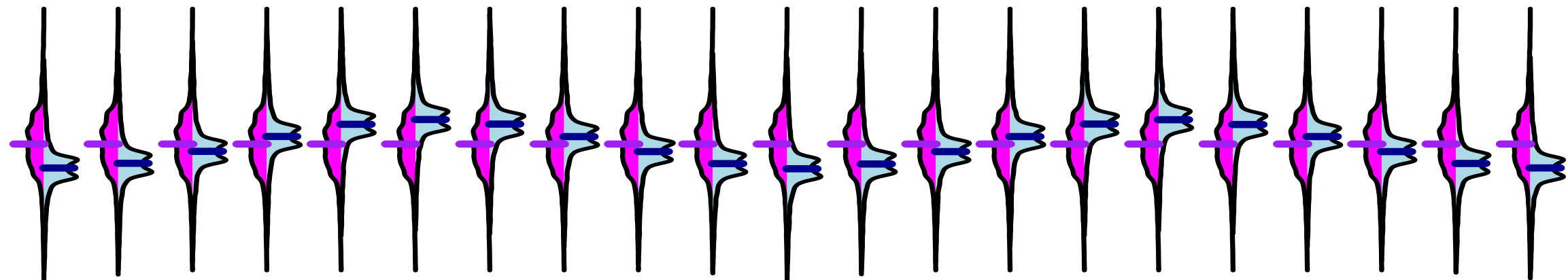
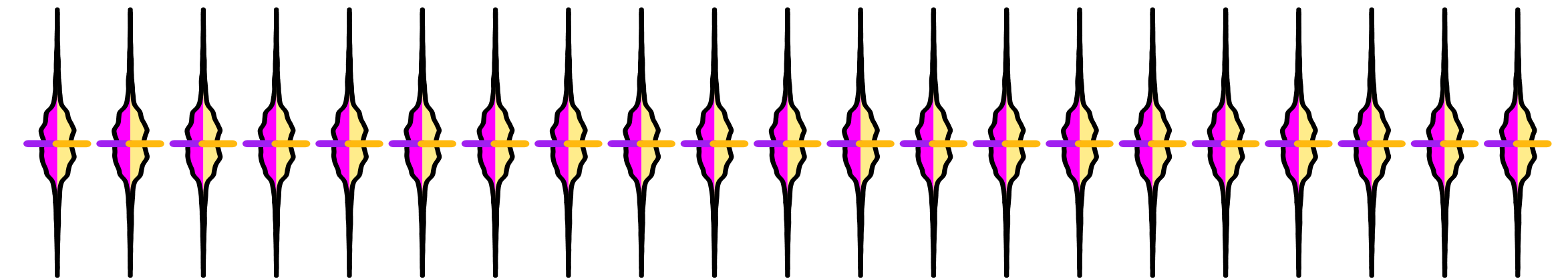


**X2 fixed**



**X4 fixed**



**X1 fixed****X3 fixed****X2 fixed****X4 fixed**

## Moment independent indices

### ➡ Pros

- They account for the whole effect of a parameter on the output distribution
- Density-based (many methods & packages)

### ➡ Cons

- Higher-order indices or outputs implies curse of dimensionality
- No ANOVA (« natural » normalization constant? Separation between interactions & main effects?)

$$\mathcal{S}_{ll'}^{TV} = \int |p_Y(y)p_{X_l}(x)p_{X_{l'}}(x') - p_{X_l, X_{l'}, Y}(x, x', y)| dx dx' dy - \mathcal{S}_l^{TV} - \mathcal{S}_{l'}^{TV}$$

Does this make sense?

# Sensitivity analysis: beyond Sobol' indices

Remember our general GSA setting ?

$$\mathcal{S}_l = \mathbb{E}_{X_l} (d(P_Y, P_{Y|X_l}))$$

Other point of view

$$\begin{aligned}\mathcal{S}_l^{KL} &= \int p_{Y|X_l=x}(y) \ln \left( \frac{p_{Y|X_l=x}(y)}{p_Y(y)} \right) p_{X_l}(x) dx dy \\ &= \int \ln \left( \frac{p_{Y,X_l}(y, x)}{p_Y(y)p_{X_l}(x)} \right) p_{Y,X_l}(y, x) dx dy \\ &= \text{MI}(X_l, Y)\end{aligned}$$

- The KL-based index actually corresponds to the mutual information between one of the inputs and the output, i.e. a measure of their dependence

# Sensitivity analysis: beyond Sobol' indices

Remember our general GSA setting ?

$$\mathcal{S}_l = \mathbb{E}_{X_l} (d(P_Y, P_{Y|X_l}))$$

Other point of view

$$\begin{aligned}\mathcal{S}_l^{KL} &= \int p_{Y|X_l=x}(y) \ln \left( \frac{p_{Y|X_l=x}(y)}{p_Y(y)} \right) p_{X_l}(x) dx dy \\ &= \int \ln \left( \frac{p_{Y,X_l}(y, x)}{p_Y(y)p_{X_l}(x)} \right) p_{Y,X_l}(y, x) dx dy \\ &= \text{MI}(X_l, Y)\end{aligned}$$



- The KL-based index actually corresponds to the mutual information between one of the inputs and the output, i.e. a measure of their dependence

# Sensitivity analysis: beyond Sobol' indices

## HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \text{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Detects independence, with small sample size → **Screening!**
- A kernel for the output just like for the MMD + **now a kernel for the inputs**

Screening can be achieved via statistical tests of independence (De Lozzo & Marrel 2016)



# Sensitivity analysis: beyond Sobol' indices

## HSIC-based sensitivity index

$$S_A^{HS} = \text{HSIC}(\mathbf{X}_A, Y)$$



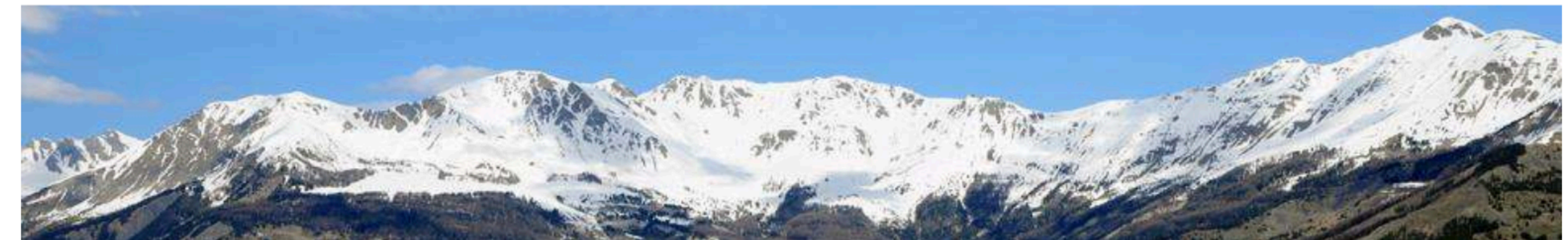
- Already proposed with a hand-made
- Detects independence, with small sample size
- A kernel for the output just like for the input

**ETICS**  
**École Thématique sur les Incertitudes en Calcul Scientifique**  
**Research School on Uncertainty in Scientific Computing**

June 6-10 2016

Centre de séminaire Séolane  
<http://eost.u-strasbg.fr/seolane/>

**Barcelonnette**



Talk of S. D.



# Sensitivity analysis: beyond Sobol' indices

## HSIC-based sensitivity index

$$S_A^{HS} = \text{HSIC}(\mathbf{X}_A, Y)$$



### ETICS 2025

**École Thématique sur les Incertitudes en Calcul Scientifique**  
**Research School on Uncertainty in Scientific Computing**

<https://www.gdr-mascotnum.fr/etics.html>

October, 5-10, VVF Lac Léman Evian-les-Bains, France

<https://www.vvf.fr/villages-vacances/vacances-evian-vvf-villages.html>



Talk of O. Zahm

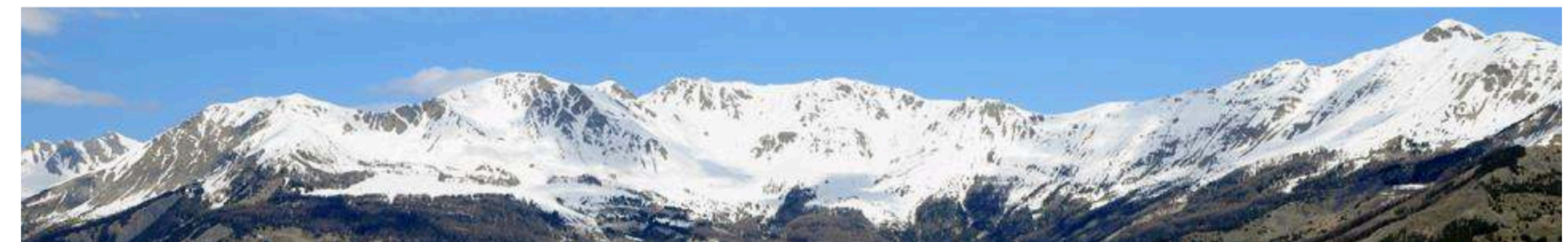
### ETICS

**École Thématique sur les Incertitudes en Calcul Scientifique**  
**Research School on Uncertainty in Scientific Computing**

June 6-10 2016

Centre de séminaire Séolane  
<http://eost.u-strasbg.fr/seolane/>

**Barcelonnette**



Talk of S. D.

# Sensitivity analysis

- **New emerging theme: sensitivity to misspecification of the input distribution**
  - Assess the influence of a perturbation of the input distribution on some quantity of interest of the model output
  - Main question: **define realistic perturbations**



# Sensitivity analysis

- **New emerging theme: sensitivity to misspecification of the input distribution**
  - Assess the influence of a perturbation of the input distribution on some quantity of interest of the model output
  - Main question: **define realistic perturbations**
    - First proposal

$$f_{i\delta} = \underset{\pi}{\operatorname{argmin}} \quad KL(\pi, f_i)$$
$$s.t. \quad \mathbb{E}_{\pi}[\psi_k] = \mathbb{E}_{f_i}[\psi_k] + \delta_k$$
$$k=1, \dots, K$$

where  $\psi_1, \dots, \psi_K$  are  $K$  linear constraints on the modified density, and  $\delta_1, \dots, \delta_K$  are the values for the perturbations.

# Sensitivity analysis

- **New emerging theme: sensitivity to misspecification of the input distribution**

- Assess the influence of a perturbation of the input distribution on some quantity of interest of the model output

**ETICS 2020**

**École Thématique sur les Incertitudes en Calcul Scientifique  
Research School on Uncertainty in Scientific Computing**

October, 4-9, Ile d'Oléron, France - <https://www.caes.cnrs.fr/sejours/la-vieille-perrotine/>



$$KL(\pi, f_i)$$

$+\delta_k$

ie modified density, and  $\delta_1, \dots, \delta_K$

Talk of B. Iooss

**Several appearances in ETICS community**

**2 - Design of experiments**

# Design of experiments

- **Defining a DOE = choosing points in a pre-defined parameter space**
  - Each point will then be evaluated to collect the corresponding value of the outputs of interest (via an experimental protocol, a production process observation, a numerical simulator, ...)
  - In general this evaluation is costly (time/money), which means that the DOE must be carefully chosen
- **Objective: explore the output behavior thanks to a limited number of evaluations**
  - Optimize the information: identify regions of interest (safety, optimization), detect influential parameters, quantify their impact, ...
  - Generate a DOE to build a regression model



# Design of experiments: traditional approaches

- **Family 1: Geometrical criteria**

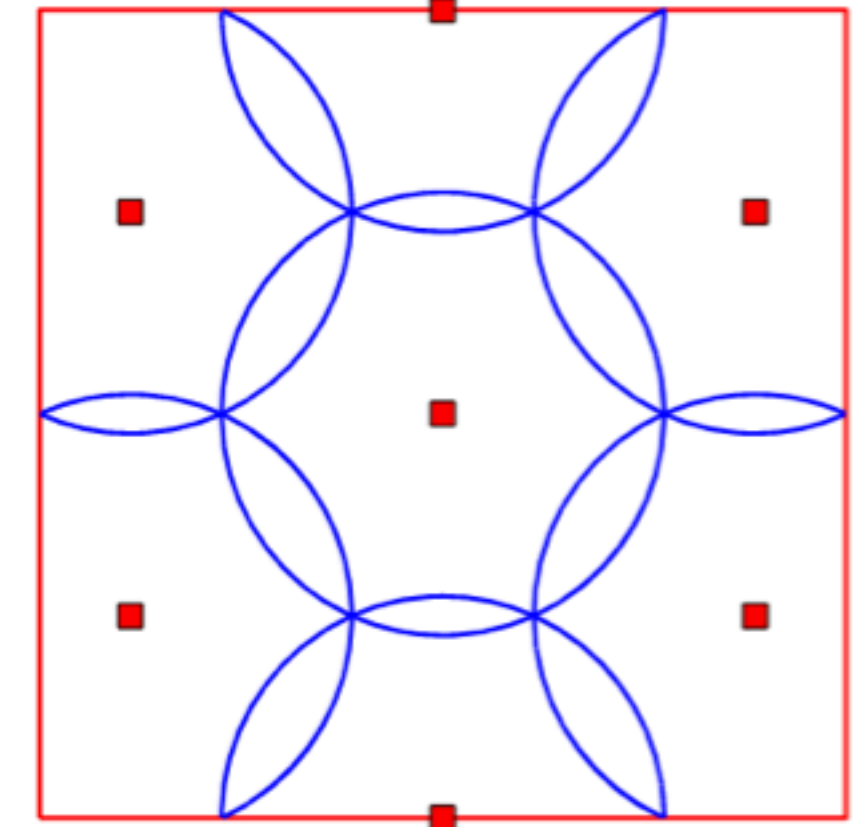
- Minimax DOE

- Minimize the maximal distance between any point in the space and the DOE (i.e. smallest possible holes)

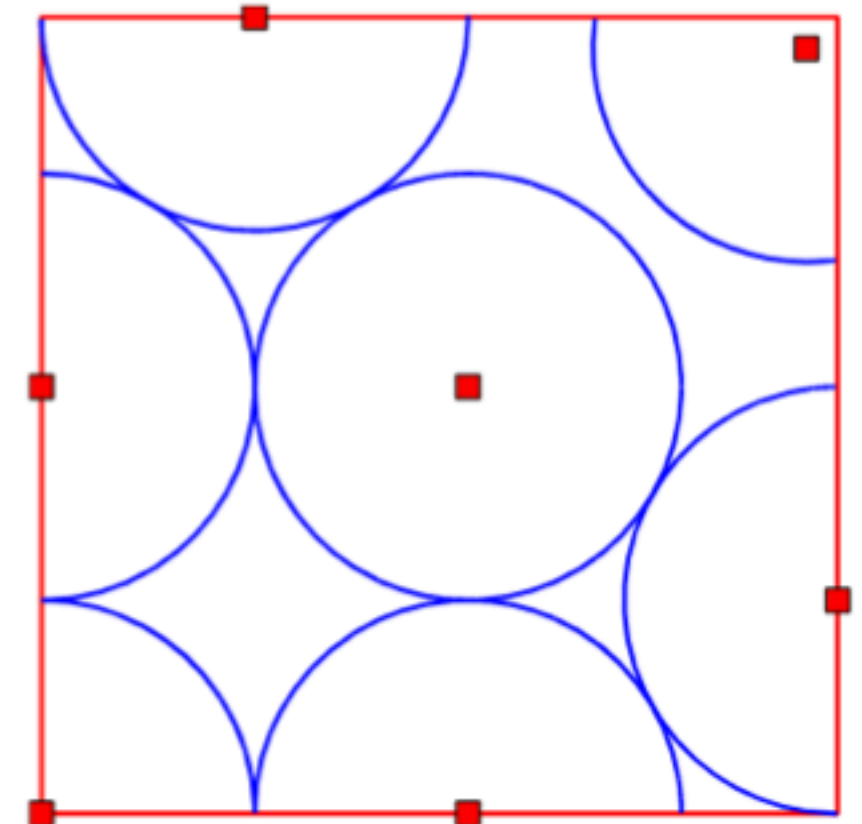
- Maximin DOE

- Maximize the minimal distance between points (i.e. limit cluster effect)

① miniMax  $d = 2, n = 7$   
(radius =  $\phi_{mM}(\mathbf{X}_n)$ )



② Maximin  $d = 2, n = 7$   
(radius =  $\phi_{Mm}(\mathbf{X}_n)/2$ )



Courtesy of L. Pronzato

# Design of experiments: traditional approaches

- **Family 2: Discrepancy criteria**

$$D_n(\mathcal{B}, \mathbf{X}_n) \triangleq \sup_{\mathbb{B} \in \mathcal{B}} \left| \frac{\text{nb. of } \mathbf{x}_i \text{ in } \mathbb{B}}{n} - \text{vol}(\mathbb{B}) \right|$$

with  $\mathcal{B}$  a family of subsets of  $\mathbb{I}_d$  ( $\Rightarrow 0 \leq D_n(\mathcal{B}, \mathbf{X}_n) \leq 1$ )

- Goal: have points as close as possible to the uniform distribution
- Changing  $\mathcal{B}$  yields different discrepancies
- Point of view justified by QMC integration



# Design of experiments: traditional approaches

**ETICS 2017**

**École Thématique sur les Incertitudes en Calcul Scientifique  
Research School on Uncertainty in Scientific Computing**

October 1-6 2017

**Centre IGESA de Porquerolles**

<https://www.iges.fr/les-catalogues-iges/groupe-et-seminaires-2016/>



Talk of L. Pronzato

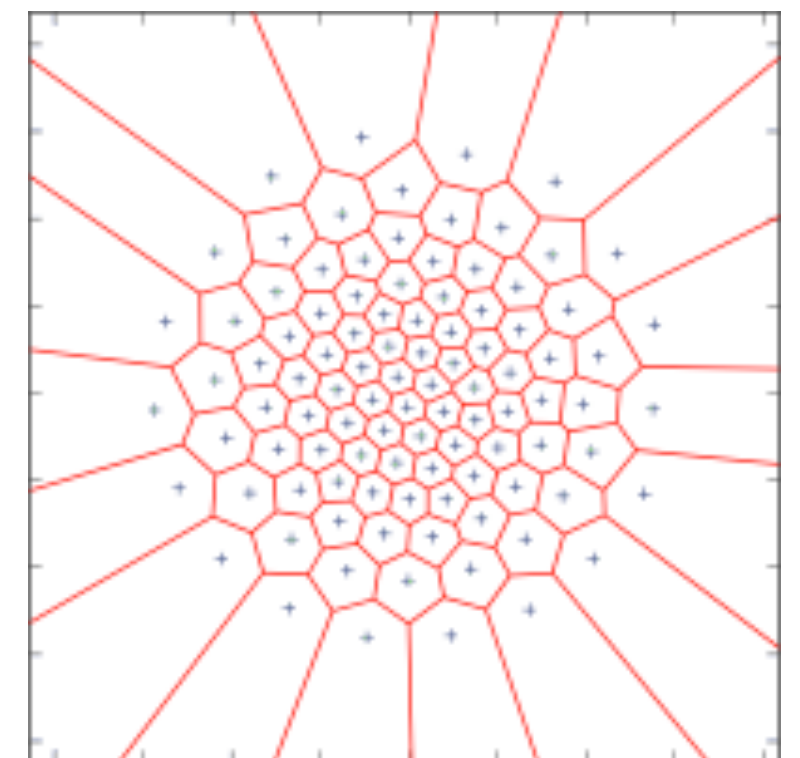
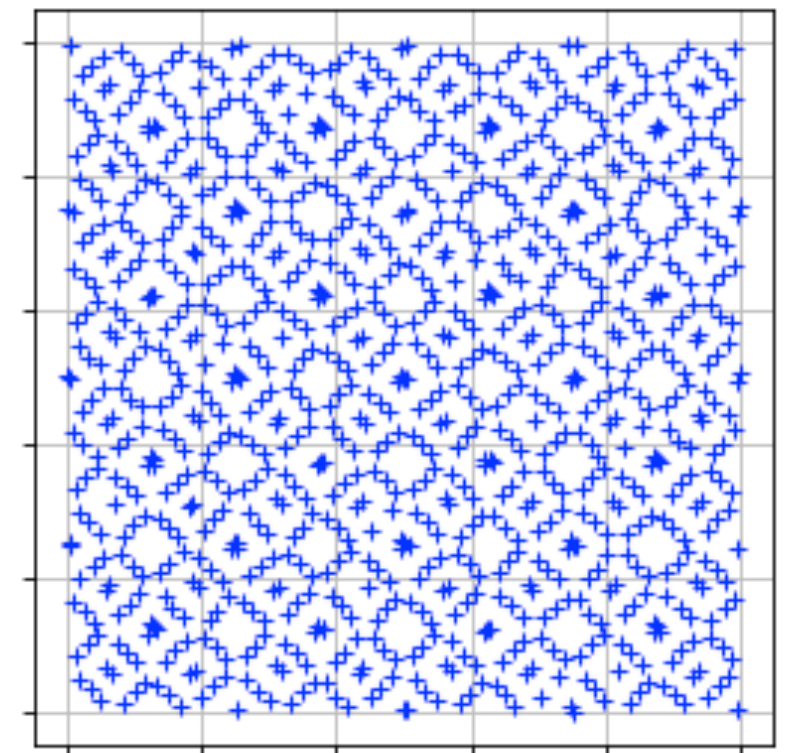
# Design of experiments: Quantization of probability distributions

- **What is quantization?**

- Identify a (small) set of point which represents as well as possible a **target** probability distribution

- **When the target is fully specified**

- **Uniform** on hypercube: literature on (space-filling) design of experiments
  - Quasi Monte-Carlo / Low discrepancy sequences
  - Minimax / Maximin / MaxPro designs
- **Gaussian** (see Pages 2003, extensions to GPs)





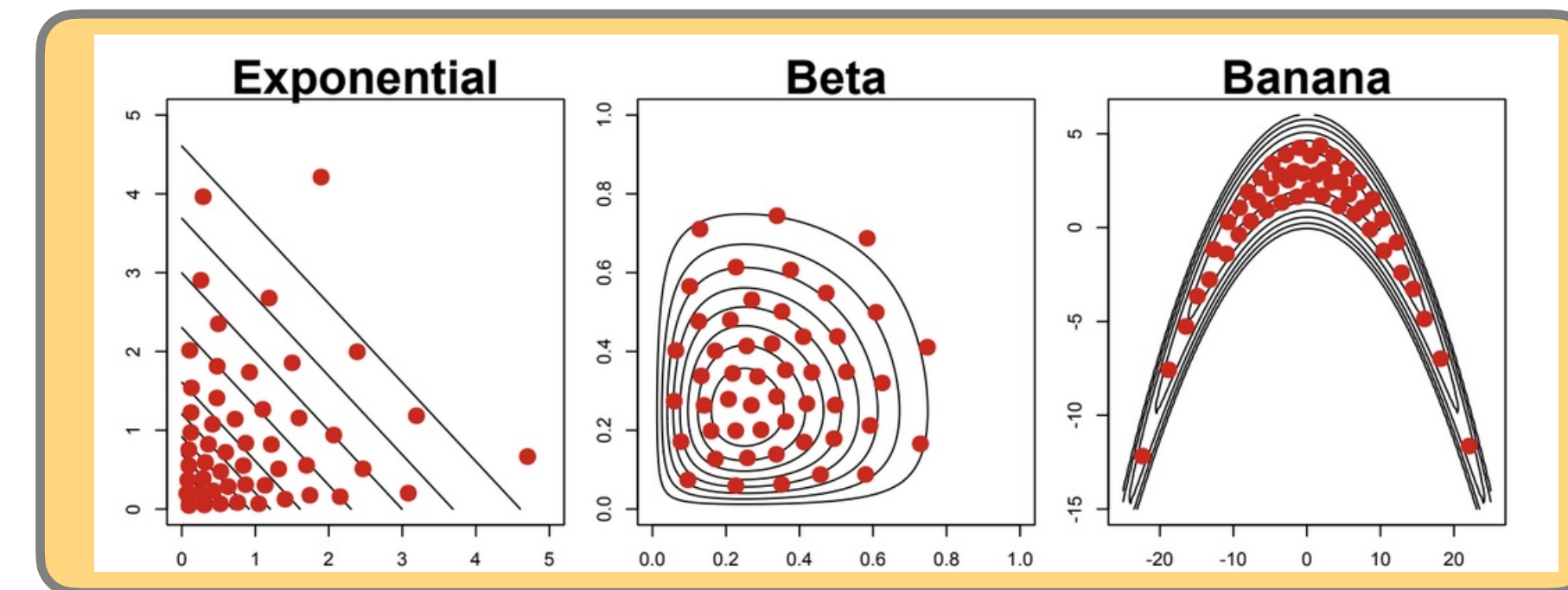
# Design of experiments: Quantization of probability distributions

- More generally, we may encounter situations where the target is

1. Fully specified but neither Uniform nor Gaussian

- e.g. exponential, Beta, ...

Mak &  
Joseph 2018



# Design of experiments: Quantization of probability distributions

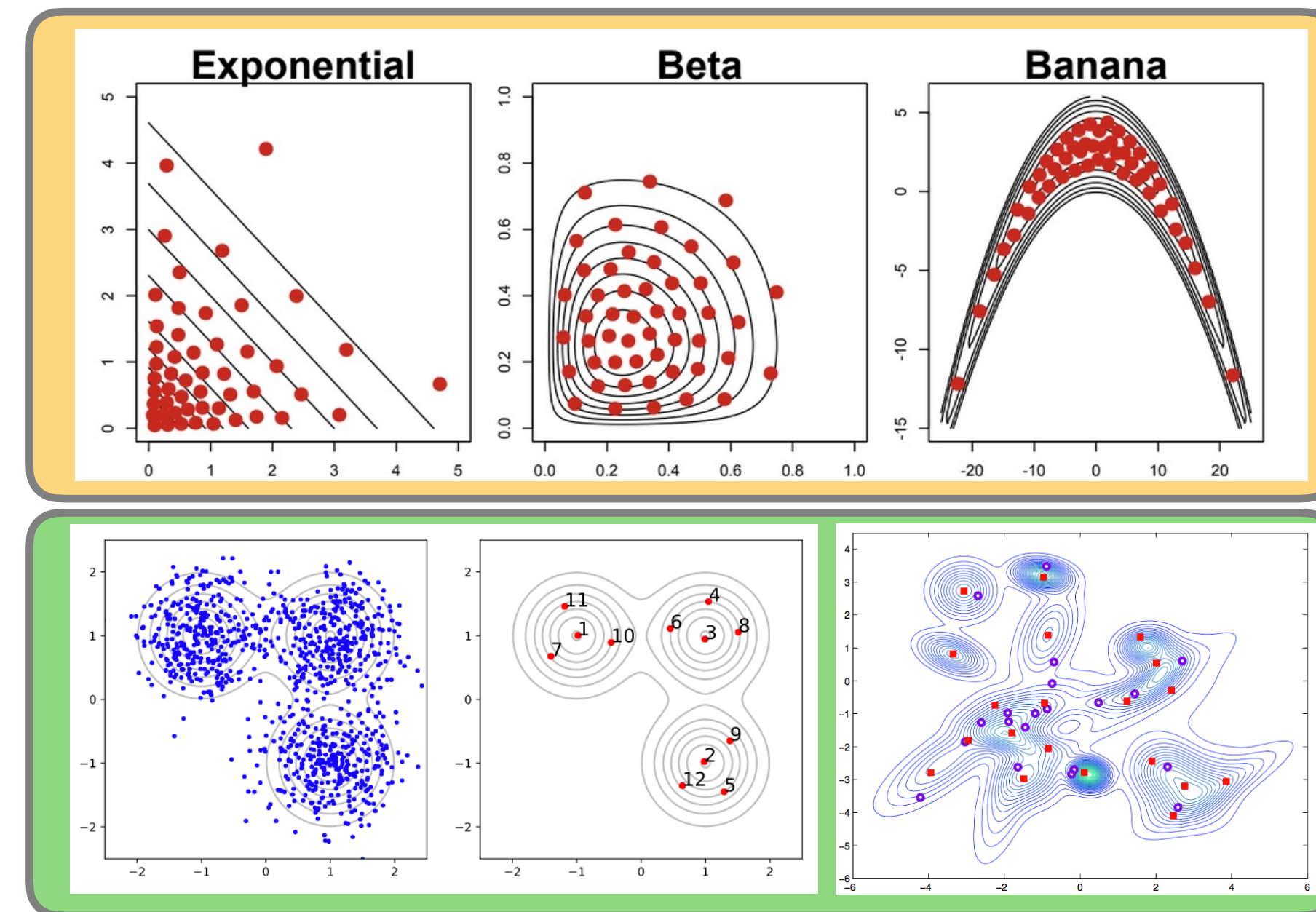
- More generally, we may encounter situations where the target is

1. Fully specified but neither Uniform nor Gaussian

- e.g. exponential, Beta, ...

2. Given as a sample from the target

- This is a **subsampling** problem



Chen et al.  
2010

Teymur et  
al. 2021

# Design of experiments: Quantization of probability distributions

- More generally, we may encounter situations where the target is

1. Fully specified but neither Uniform nor Gaussian

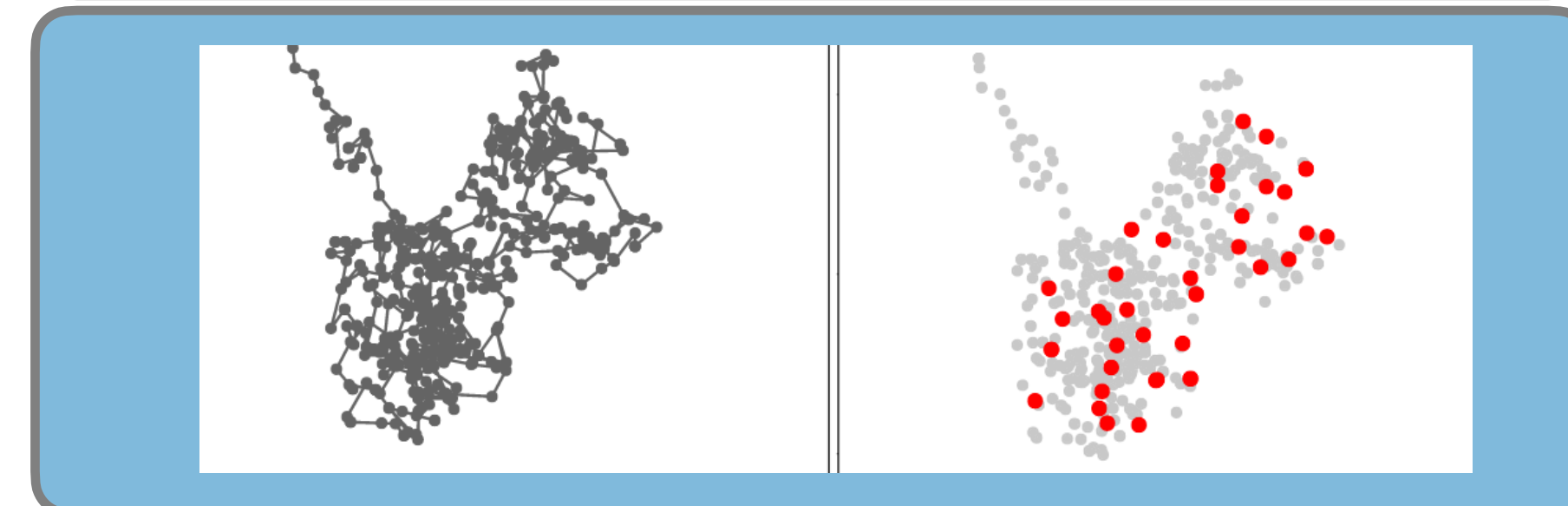
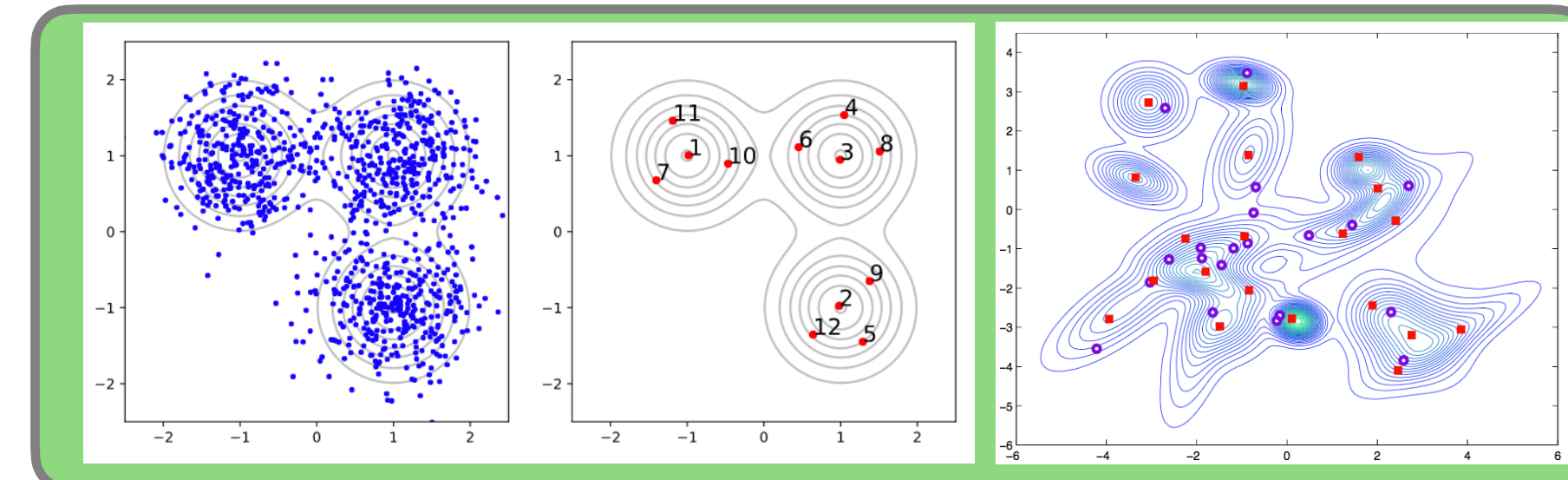
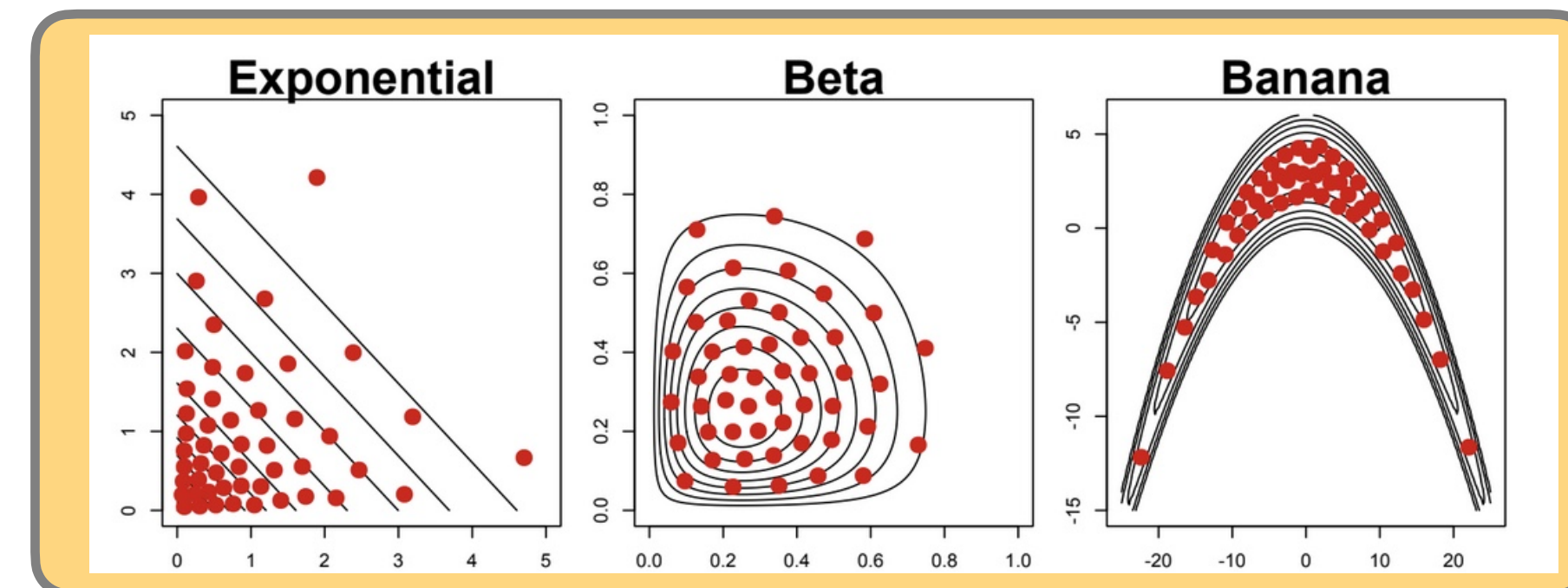
- e.g. exponential, Beta, ...

2. Given as a sample from the target

- This is a **subsampling** problem

3. Given as an approximate sample from the target

- This is a **subsampling** problem with **correction**



Riabiz et al.  
2022



# Design of experiments: Quantization of probability distributions

- **We can rewrite all cases as an optimization problem**

⦿ We seek points  $x_1, \dots, x_n$  leading to an empirical distribution as close as possible to the target  $\mathbb{P}$

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} d \left( \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \mathbb{Q} \right)$$

# Design of experiments: Quantization of probability distributions

- **We can rewrite all cases as an optimization problem**

- ◉ We seek points  $x_1, \dots, x_n$  leading to an empirical distribution as close as possible to the target  $\mathbb{P}$

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} d \left( \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \mathbb{Q} \right)$$

Fully specified

- $\mathbb{Q} = \mathbb{P}$  given
- $\mathcal{X} = \mathbb{R}^d$

# Design of experiments: Quantization of probability distributions

- **We can rewrite all cases as an optimization problem**

- ◉ We seek points  $x_1, \dots, x_n$  leading to an empirical distribution as close as possible to the target  $\mathbb{P}$

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} d \left( \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \mathbb{Q} \right)$$

## Subsampling

- $x_1, \dots, x_N \sim \mathbb{P}$
- $\mathbb{Q} = 1/N \sum_{j=1}^N \delta_{x_j}$  with  $N \gg n$
- $\mathcal{X} = \{x_1, \dots, x_N\}$

# Design of experiments: Quantization of probability distributions

- **We can rewrite all cases as an optimization problem**

- ◉ We seek points  $x_1, \dots, x_n$  leading to an empirical distribution as close as possible to the target  $\mathbb{P}$

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} d \left( \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \mathbb{Q} \right)$$

## Subsampling & correction

- $x_1, \dots, x_N \sim \hat{\mathbb{P}}$  approximation of  $\mathbb{P}$
- $\mathbb{Q} = 1/N \sum_{j=1}^N \delta_{x_j}$  with  $N \gg n$
- $\mathcal{X} = \{x_1, \dots, x_N\}$

# Design of experiments: Quantization of probability distributions

- **We can rewrite all cases as an optimization problem**

- ⦿ We seek points  $x_1, \dots, x_n$  leading to an empirical distribution as close as possible to the target  $\mathbb{P}$

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} d \left( \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \mathbb{Q} \right)$$

- ⦿ Recently in ML, many paper focused on a specific choice of **distance**, based on **kernel embeddings of probability distributions**

- Simple computation with only expectations of kernels
    - A « true » distance if the kernel is characteristic
    - Used also for two-sample tests, independence tests, variable selection, GANs, ...



# Design of experiments: Quantization of probability distributions

- If we plug the MMD in the optimization problem

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} d \left( \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \mathbb{Q} \right)$$

$$\text{MMD}^2(P_1, P_2) = \mathbb{E}_{\xi, \xi'} k_{\mathcal{X}}(\xi, \xi') - 2\mathbb{E}_{\xi, \zeta} k_{\mathcal{X}}(\xi, \zeta) + \mathbb{E}_{\zeta, \zeta'} k_{\mathcal{X}}(\zeta, \zeta')$$

# Design of experiments: Quantization of probability distributions

- If we plug the MMD in the optimization problem

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} d \left( \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \mathbb{Q} \right)$$



$$\text{MMD}^2(P_1, P_2) = \mathbb{E}_{\xi, \xi'} k_{\mathcal{X}}(\xi, \xi') - 2\mathbb{E}_{\xi, \zeta} k_{\mathcal{X}}(\xi, \zeta) + \mathbb{E}_{\zeta, \zeta'} k_{\mathcal{X}}(\zeta, \zeta')$$



# Design of experiments: Quantization of probability distributions

- If we plug the MMD in the optimization problem

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} d \left( \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \mathbb{Q} \right)$$



$\text{MMD}^2(P_1, P_2)$

**ETICS 2021**

**École Thématique sur les Incertitudes en Calcul Scientifique**  
**Research School on Uncertainty in Scientific Computing**

September, 12-17, Keravel resort, Erdeven, France - <https://www.keravelvacances.com/>



Talk of C. Oates

# Design of experiments: Quantization of probability distributions

- If we plug the MMD in the optimization problem

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} d \left( \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \mathbb{Q} \right)$$

$$\text{MMD}^2(P_1, P_2) = \mathbb{E}_{\xi, \xi'} k_{\mathcal{X}}(\xi, \xi') - 2\mathbb{E}_{\xi, \zeta} k_{\mathcal{X}}(\xi, \zeta) + \mathbb{E}_{\zeta, \zeta'} k_{\mathcal{X}}(\zeta, \zeta')$$

1. Fully specified

➡ We can compute both expectations (empirical is easy, theoretical in many cases)

2. Given as a sample from the target

➡ We can compute both empirical expectations

3. Given as an approximate sample from the target

➡ We can compute the empirical expectation **but the second one is biased**  $\left( \hat{\mathbb{P}} \approx \mathbb{P} \right)$



# Design of experiments: Quantization of probability distributions

- If we plug the MMD in the optimization problem

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} d \left( \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \mathbb{Q} \right)$$

$$\text{MMD}^2(P_1, P_2) = \mathbb{E}_{\xi, \xi'} k_{\mathcal{X}}(\xi, \xi') - 2\mathbb{E}_{\xi, \zeta} k_{\mathcal{X}}(\xi, \zeta) + \mathbb{E}_{\zeta, \zeta'} k_{\mathcal{X}}(\zeta, \zeta')$$

1. Fully specified

➡ We can compute both expectations (empirical is easy, theoretical in many cases)

2. Given as a sample from the target

➡ We can compute both empirical expectations

3. Given as an approximate sample from the target

➡ We can compute the empirical expectation **but the second one is biased**  $(\hat{\mathbb{P}} \approx \mathbb{P})$

➡ **Other point of view: do we need to know the target?**

# Quantization with the KSD

- **When the target is not tractable**
  - ◉ Stein's method

# Quantization with the KSD

- **When the target is not tractable**

- Stein's method

- Define an operator  $\mathcal{T}_p$ , that maps functions  $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$  to real-valued functions such that  $\mathbb{E}[\mathcal{T}_p g(X)] = 0$ , with  $X \sim \mathbb{P}$ , for all  $g \in \mathcal{G} = \{g : \mathbb{R}^d \rightarrow \mathbb{R}^d : \sum_{i=1}^d \|g_i\|_{\mathcal{G}} \leq 1\}$
- We assume the probability measure  $\mathbb{P}$  on  $\mathbb{R}^d$  admits a continuously differentiable Lebesgue density  $p \in C^1(\mathbb{R}^d)$ , such that  $\mathbb{E}[\|\nabla \log p(X)\|^2] < \infty$
- The Stein discrepancy is then defined as

$$\text{SD}(\mathbb{P}, \mathbb{P}') = \sup_{g \in \mathcal{G}} \mathbb{E}[(\mathcal{T}_p g)(Z)]$$

where  $Z \sim \mathbb{P}'$

# Quantization with the KSD

- **When the target is not tractable**

- **Kernelized Stein's method**

- Take  $\mathcal{G} = \mathcal{H}_k$  a RKHS with kernel  $k$
- Choose  $\mathcal{T}_p$  as the Langevin operator  $(\mathcal{T}_p g)(x) = \langle g(x), \nabla \log p(x) \rangle + \langle \nabla, g(x) \rangle$
- The Kernel Stein discrepancy (KSD) is given by

$$\text{KSD}^2(\mathbb{P}, \mathbb{P}') = \mathbb{E}[k_p(Z, Z')]$$

where  $Z, Z' \sim \mathbb{P}'$  and  $k_p$  is the Langevin Stein kernel defined from the score function  $s_p(x) = \nabla \log p(x)$  for  $x, x' \in \mathbb{R}^d$ , as

$$k_p(x, x') = \langle \nabla_x, \nabla_{x'} k(x, x') \rangle + \langle s_p(x), \nabla_{x'} k(x, x') \rangle + \langle s_p(x'), \nabla_x k(x, x') \rangle + \langle s_p(x), s_p(x') \rangle k(x, x')$$



# Quantization with the KSD

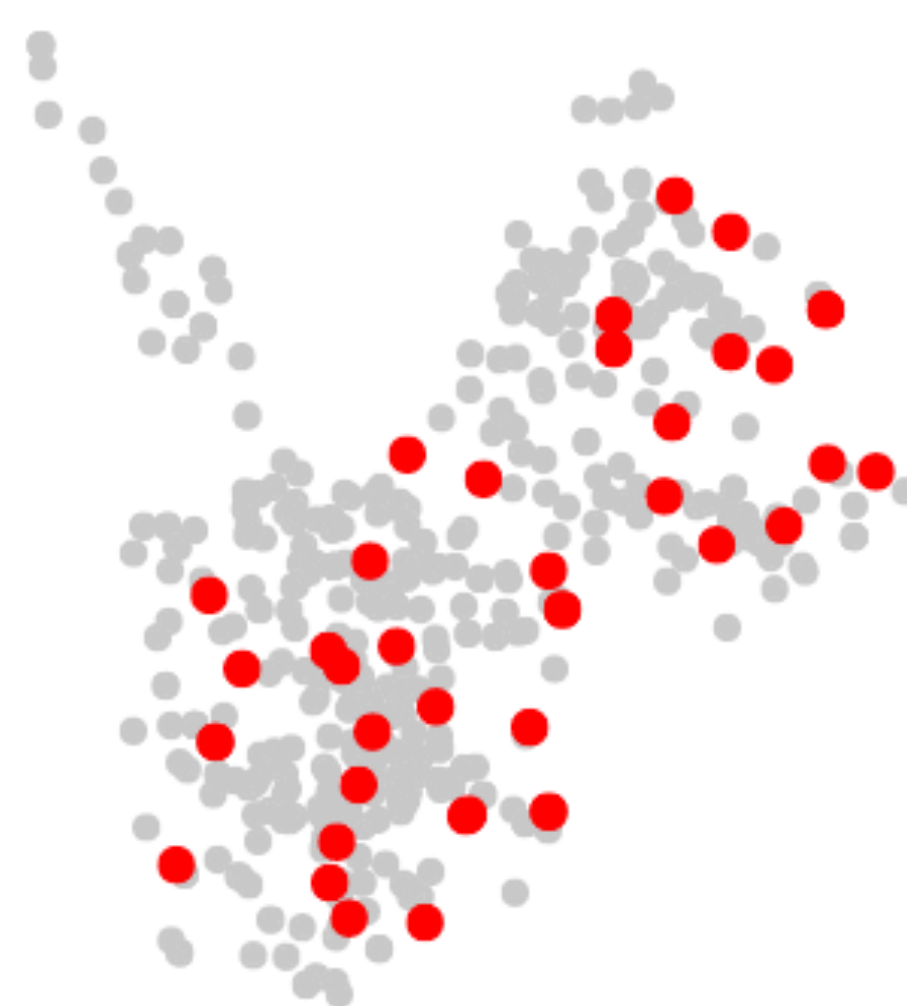
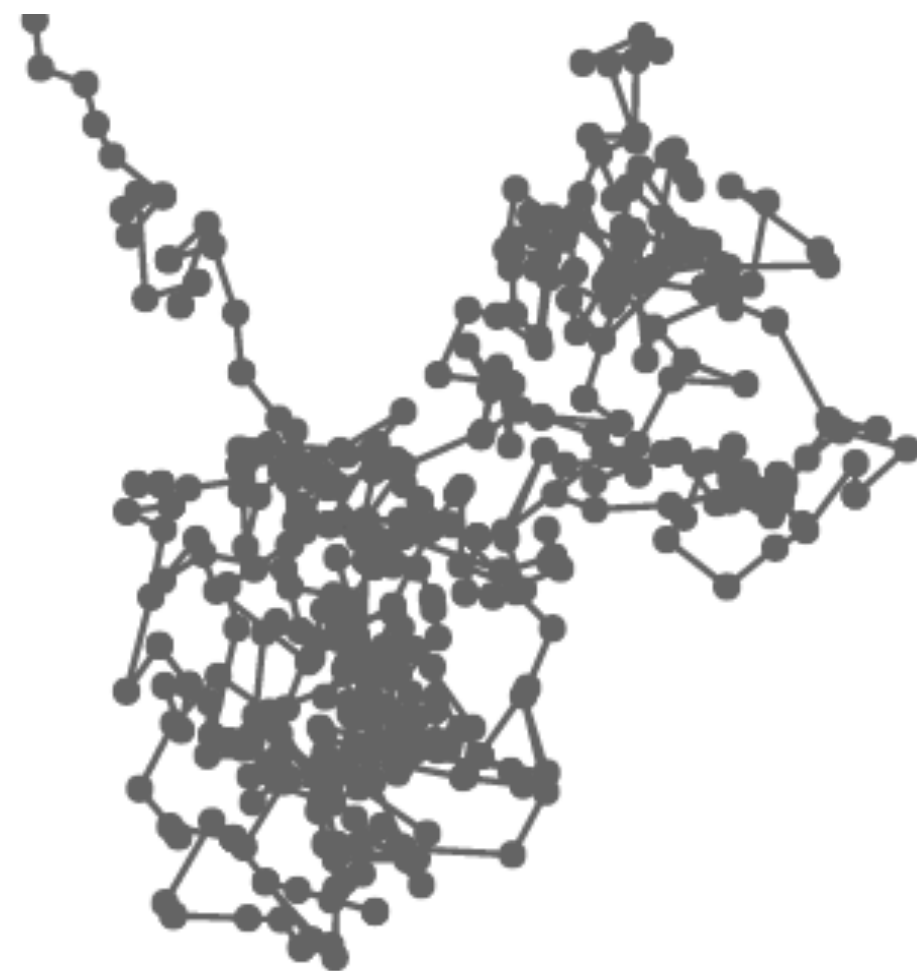
- **Summary**

- ◉ Only requires the score function of the target!
- ◉ This means that we can replace the MMD by the KSD in Case 3 for problems where the score function is known

# Quantization with the KSD

- **Summary**

- ◉ Only requires the score function of the target!
- ◉ This means that we can replace the MMD by the KSD in Case 3 for problems where the score function is known
  - ➔ The KSD is thus popular in **Bayesian inference**, and the sample to correct comes from a MCMC algorithm
  - ➔ This is the so-called **KSD thinning algorithm**



Riabiz et al.  
2022

# Quantization with the KSD

- **KSD thinning**

- ◉ We seek points  $x_1, \dots, x_n$  leading to an empirical distribution as close as possible to the target  $\mathbb{P}$

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} \text{KSD}^2 \left( \mathbb{P}, \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \right)$$

# Quantization with the KSD

- **KSD thinning**

- ◉ We seek points  $x_1, \dots, x_n$  leading to an empirical distribution as close as possible to the target  $\mathbb{P}$

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} \text{KSD}^2 \left( \mathbb{P}, \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \right)$$

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} \frac{1}{n^2} \sum_{i,j}^n k_p(x_i, x_j)$$

# Quantization with the KSD

- **KSD thinning**

- ◉ We seek points  $x_1, \dots, x_n$  leading to an empirical distribution as close as possible to the target  $\mathbb{P}$

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} \text{KSD}^2 \left( \mathbb{P}, \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \right)$$

$$\arg \min_{x_1, \dots, x_n \in \mathcal{X}} \frac{1}{n^2} \sum_{i,j}^n k_p(x_i, x_j)$$

- ◉ Typically solved by greedy algorithm

$$x_t \in \arg \min_{x \in \mathcal{X}} k_p(x, x) + 2 \sum_{j=1}^{t-1} k_p(x, x_j)$$

# KSD pathologies

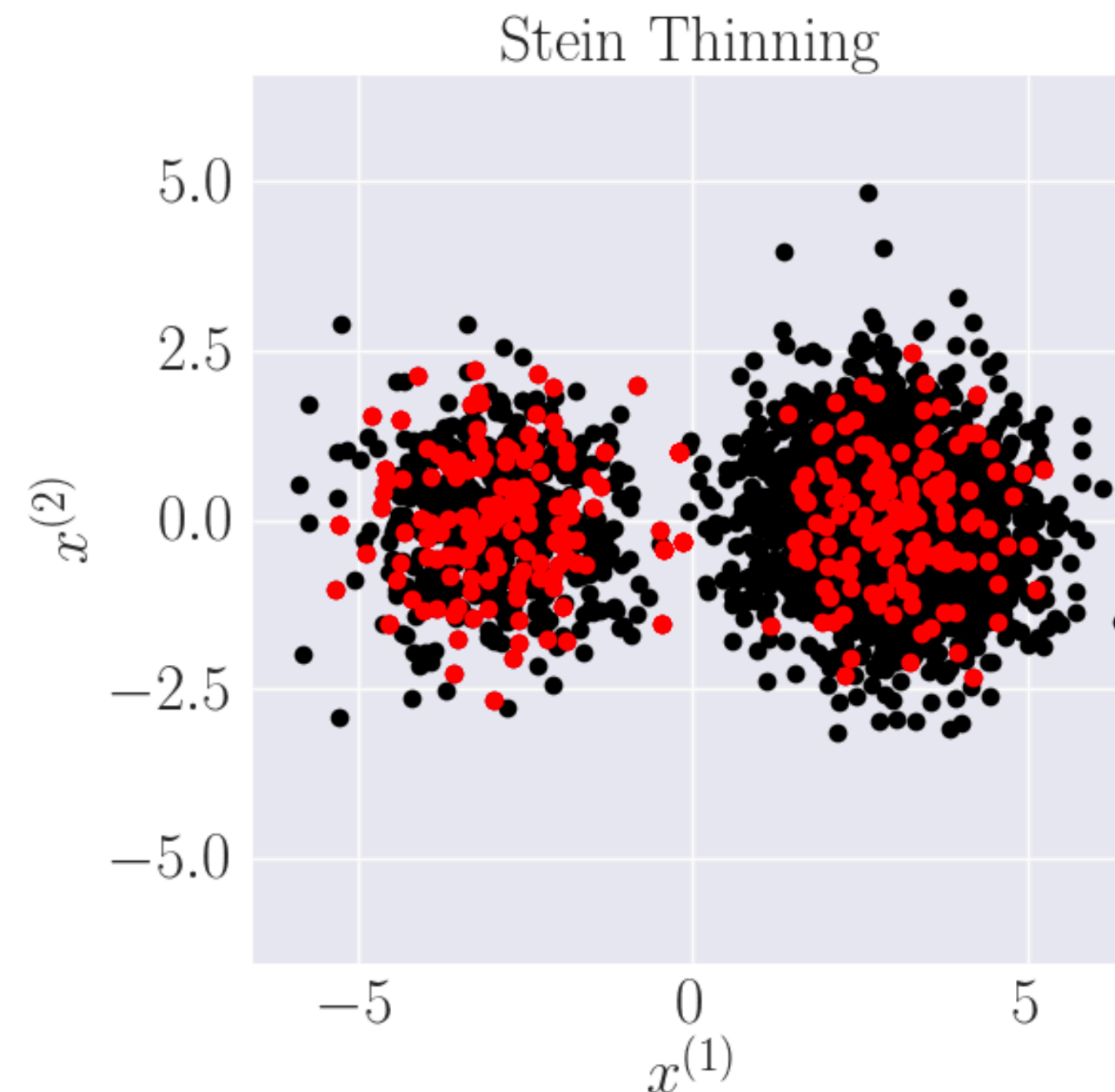
- **Pathology I: mode proportion blindness**
  - ◉ The score function is insensitive to distant mode weights



# KSD pathologies

- **Pathology I: mode proportion blindness**
  - ⦿ The score function is insensitive to distant mode weights

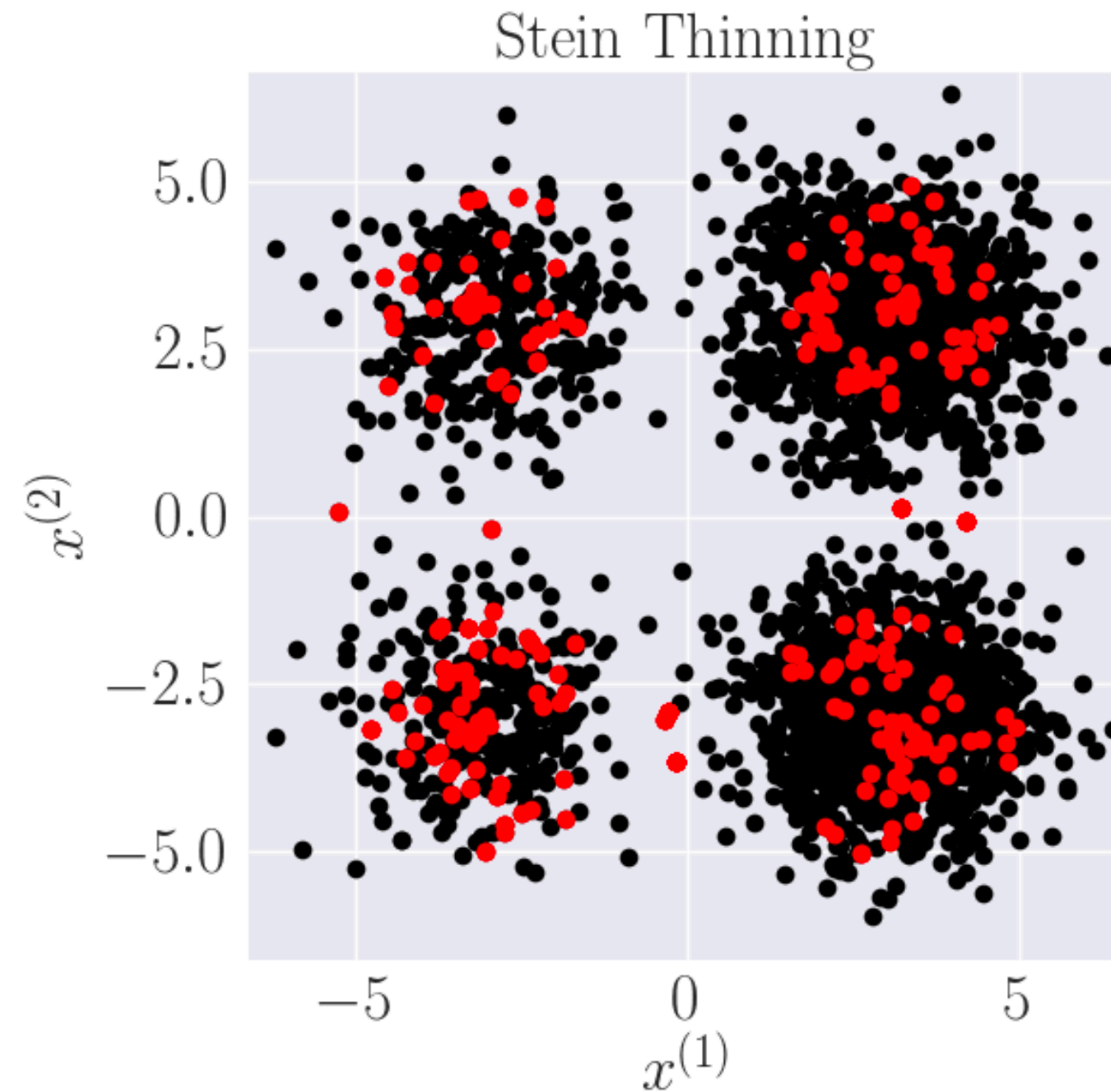
**Example 1.** Let the density  $p$  be a Gaussian mixture model of two components, respectively centered in  $(-\mu, \mathbf{0}_{d-1})$  and  $(\mu, \mathbf{0}_{d-1})$ , of weights  $w$  and  $1 - w$ , and of variance  $\sigma^2 \mathbf{I}_d$ . The initial particles  $\{\mathbf{x}_i\}_{i=1}^n$  are drawn from  $p$ . The KSD thinning algorithm selects  $m < n$  points to approximate  $p$ .



$$d = 2, \mu = 3, \sigma = 1, w = 0.2, n = 3000, m = 300$$

# KSD pathologies

- **Pathology I: mode proportion blindness**
  - The score function is insensitive to distant mode weights



# KSD pathologies

- **Pathology I: mode proportion blindness**
  - ◉ The score function is insensitive to distant mode weights
  - ◉ Observed but quite overlooked in the literature
  - ◉ We proved the following theorem

# KSD pathologies

- **Pathology I: mode proportion blindness**

- ◉ The score function is insensitive to distant mode weights
- ◉ Observed but quite overlooked in the literature
- ◉ We proved the following theorem

**Theorem 2.3.** *Let  $k_p$  be the Stein kernel associated with the radial kernel  $k(\mathbf{x}, \mathbf{x}') = \phi(\|\mathbf{x} - \mathbf{x}'\|_2/\ell)$ , where  $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ ,  $\ell > 0$ , and  $\phi \in \mathcal{C}^2(\mathbb{R}^d)$ , such that  $\phi(z) \rightarrow 0$ ,  $\phi'(z) \rightarrow 0$ , and  $\phi''(z) \rightarrow 0$  for  $z \rightarrow \infty$ . Let  $p$  and  $q$  be two bimodal mixture distributions satisfying Assumptions 2.1 and 2.2, for any  $\eta \in (0, 1)$ . We define  $w^*$  as the optimal mixture weight of  $q$  with respect to the KSD distance, i.e.,  $w^* = \operatorname{argmin}_{w \in [0, 1]} \operatorname{KSD}(\mathbb{P}, \mathbb{Q}_w)$ . Then, for  $\mu$  large enough, we have  $|w^* - \frac{1}{2}| < \frac{\eta}{2(1-\eta)}$ .*

- ◉ **Regardless of the true target weights, the optimal mixture in terms of KSD is 1/2,** whenever the mixture is close to the target, in the distant mode setting

**Assumption 2.2.** For distant bimodal mixture distributions  $q$  and  $p$  satisfying Assumption 2.1, and for  $\eta \in (0, 1)$ , we have  $|\operatorname{KSD}^2(\mathbb{P}, \mathbb{Q}_L)/\operatorname{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - 1| < \eta$ .



# KSD pathologies

- **Pathology II: spurious minimum**
  - ◉ The KSD selects samples concentrated in regions of low probability

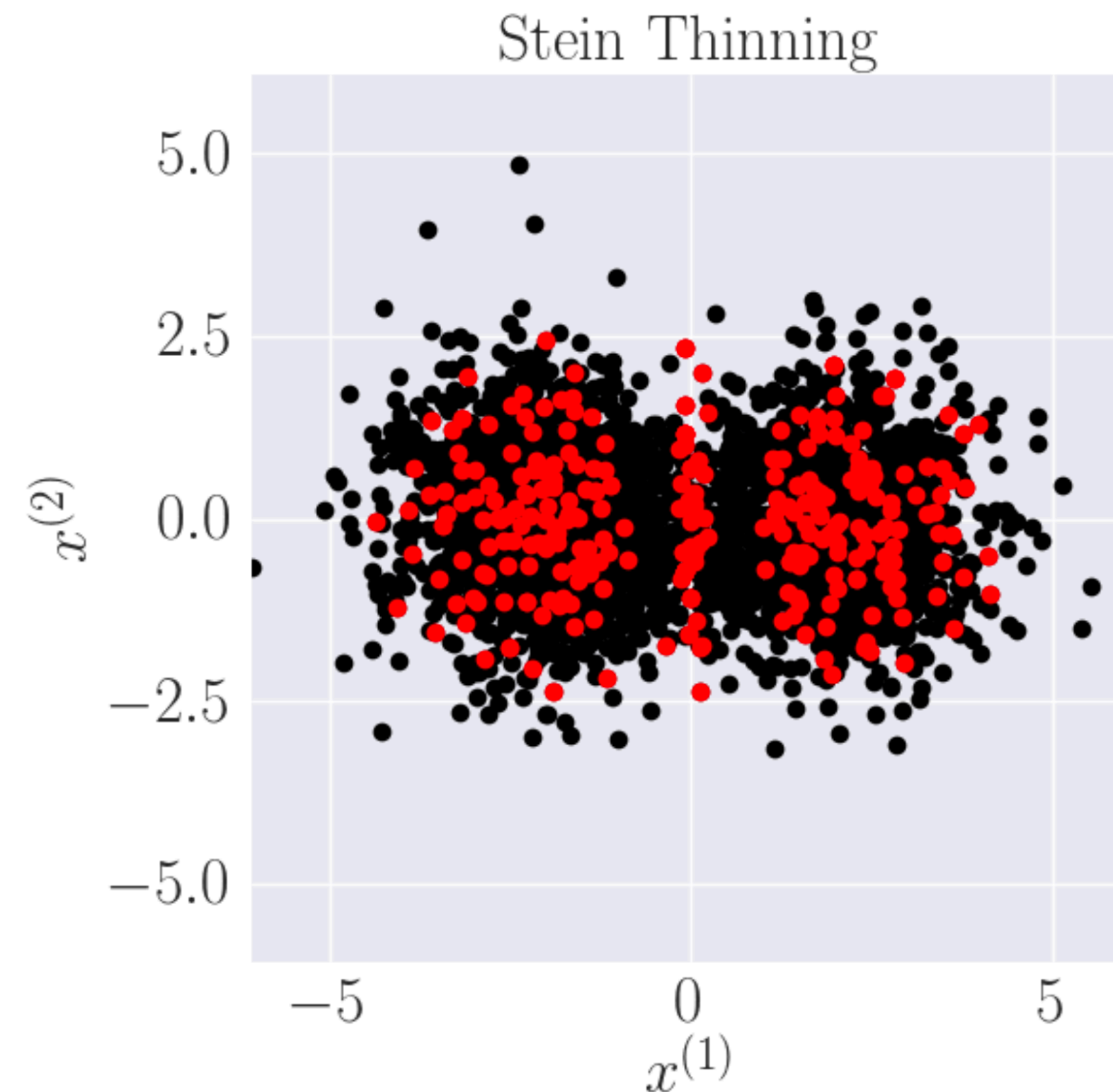


# KSD pathologies

- **Pathology II: spurious minimum**

- The KSD selects samples concentrated in regions of low probability

**Example 1.** Let the density  $p$  be a Gaussian mixture model of two components, respectively centered in  $(-\mu, \mathbf{0}_{d-1})$  and  $(\mu, \mathbf{0}_{d-1})$ , of weights  $w$  and  $1 - w$ , and of variance  $\sigma^2 \mathbf{I}_d$ . The initial particles  $\{\mathbf{x}_i\}_{i=1}^n$  are drawn from  $p$ . The KSD thinning algorithm selects  $m < n$  points to approximate  $p$ .

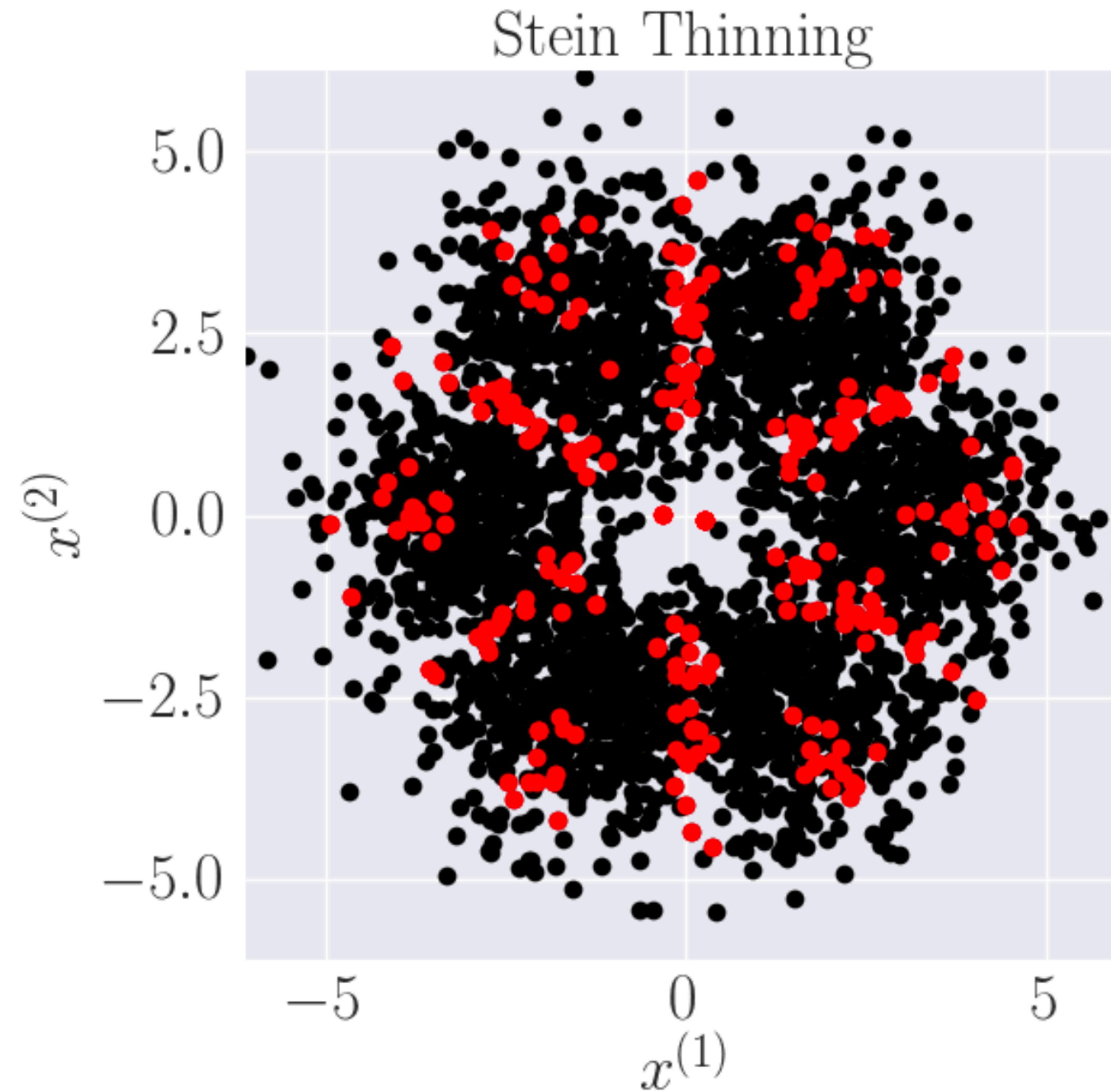


$$d = 2, \mu = 2, \sigma = 1, w = 0.5, n = 3000, m = 300$$

# KSD pathologies

- **Pathology II: spurious minimum**

- The KSD selects samples concentrated in regions of low probability





# KSD pathologies

- **Pathology II: spurious minimum**

- The KSD selects samples concentrated in regions of low probability
- Also observed but quite overlooked in the literature
- We proved the following theorem

**Theorem 2.4** (KSD spurious minimum). *Let  $k_p$  be the Stein kernel associated with the IMQ kernel with  $\ell > 0$ ,  $\beta \in (0, 1)$ , and  $c = 1$ . Let  $\{\mathbf{x}_i\}_{i=1}^m \subset \mathcal{M}_{s_0} = \{\mathbf{x} \in \mathbb{R}^d : \|s_p(\mathbf{x})\|_2 \leq s_0\}$  be a fixed set of points of empirical measure  $\mathbb{Q}_m = \frac{1}{m} \sum_{i=1}^m \delta(\mathbf{x}_i)$ , with  $s_0 \geq 0$  and  $m \geq 2$ . We have  $\text{KSD}^2(\mathbb{P}, \mathbb{Q}_m) < \mathbb{E}[\text{KSD}^2(\mathbb{P}, \mathbb{P}_m)]$ , if the score threshold  $s_0$  and the sample size  $m$  are small enough to satisfy  $m < 1 + (\mathbb{E}[\|s_p(\mathbf{X})\|_2^2] - s_0^2) / (2\beta d / \ell^2 + 2\beta s_0 / \ell + s_0^2)$ .*

**Corollary 2.5** (Low KSD samples at density minimum). *Let  $k_p$  be the Stein kernel associated with the IMQ kernel with  $\ell > 0$ ,  $\beta \in (0, 1)$ , and  $c = 1$ . Let  $p$  be a density with at least one local minimum or saddle point. For  $m \geq 2$ , if  $\{\mathbf{x}_i\}_{i=1}^m \subset \mathbb{R}^d$  is a set of points, all located at local minimum or saddle points of  $p$ , then we have  $\text{KSD}^2(\mathbb{P}, \mathbb{Q}_m) < \mathbb{E}[\text{KSD}^2(\mathbb{P}, \mathbb{P}_m)]$ , if  $m < 1 + \frac{\ell^2}{2\beta d} \mathbb{E}[\|s_p(\mathbf{X})\|_2^2]$ .*

- **Samples in low score regions have a better KSD than samples from the true target**

# KSD pathologies - Regularized KSD

- **Pathology I: mode proportion blindness**

- ◉ The score function is insensitive to distant mode weights
- ◉ We propose **entropic regularization** to lessen this phenomenon

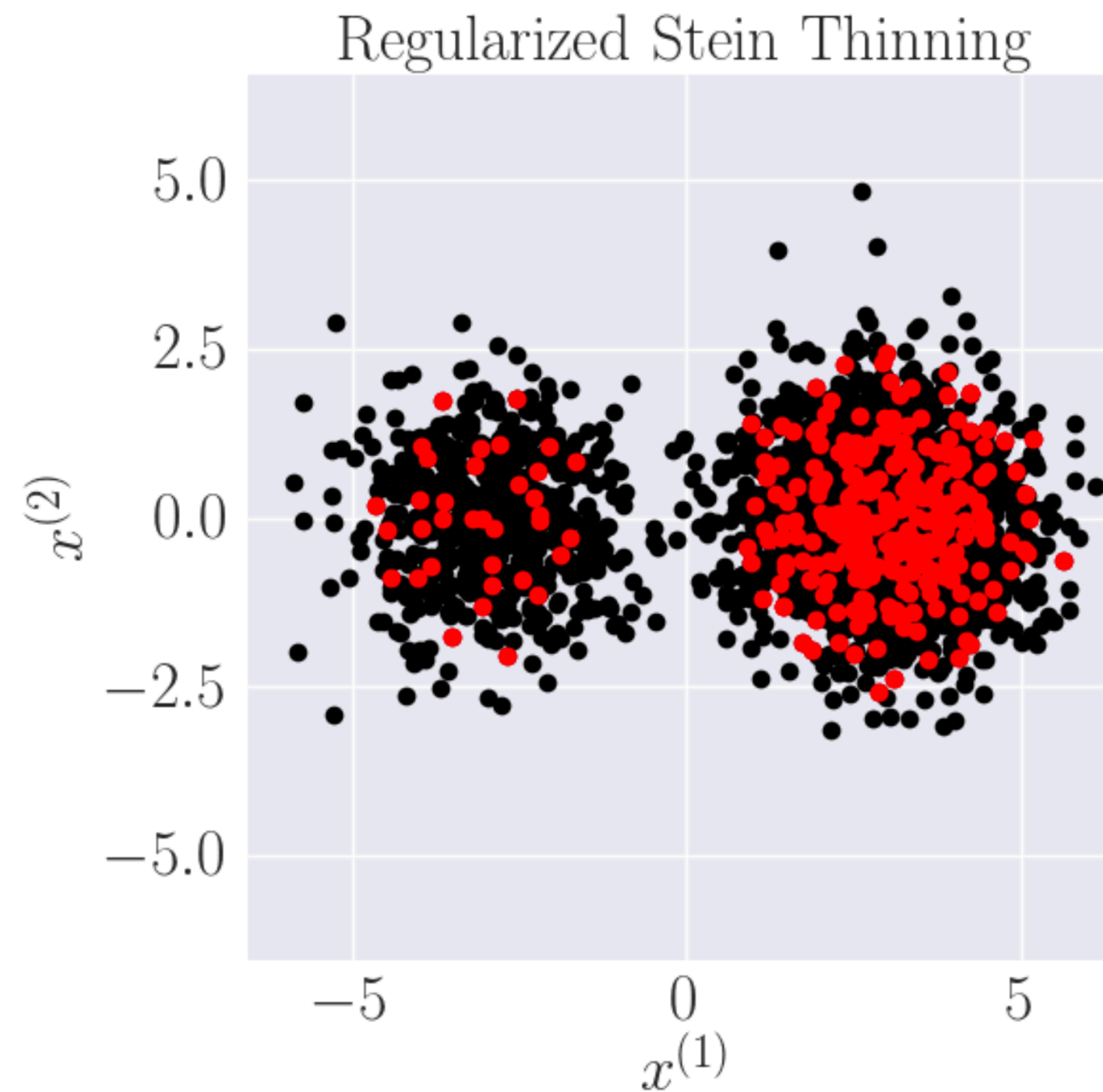
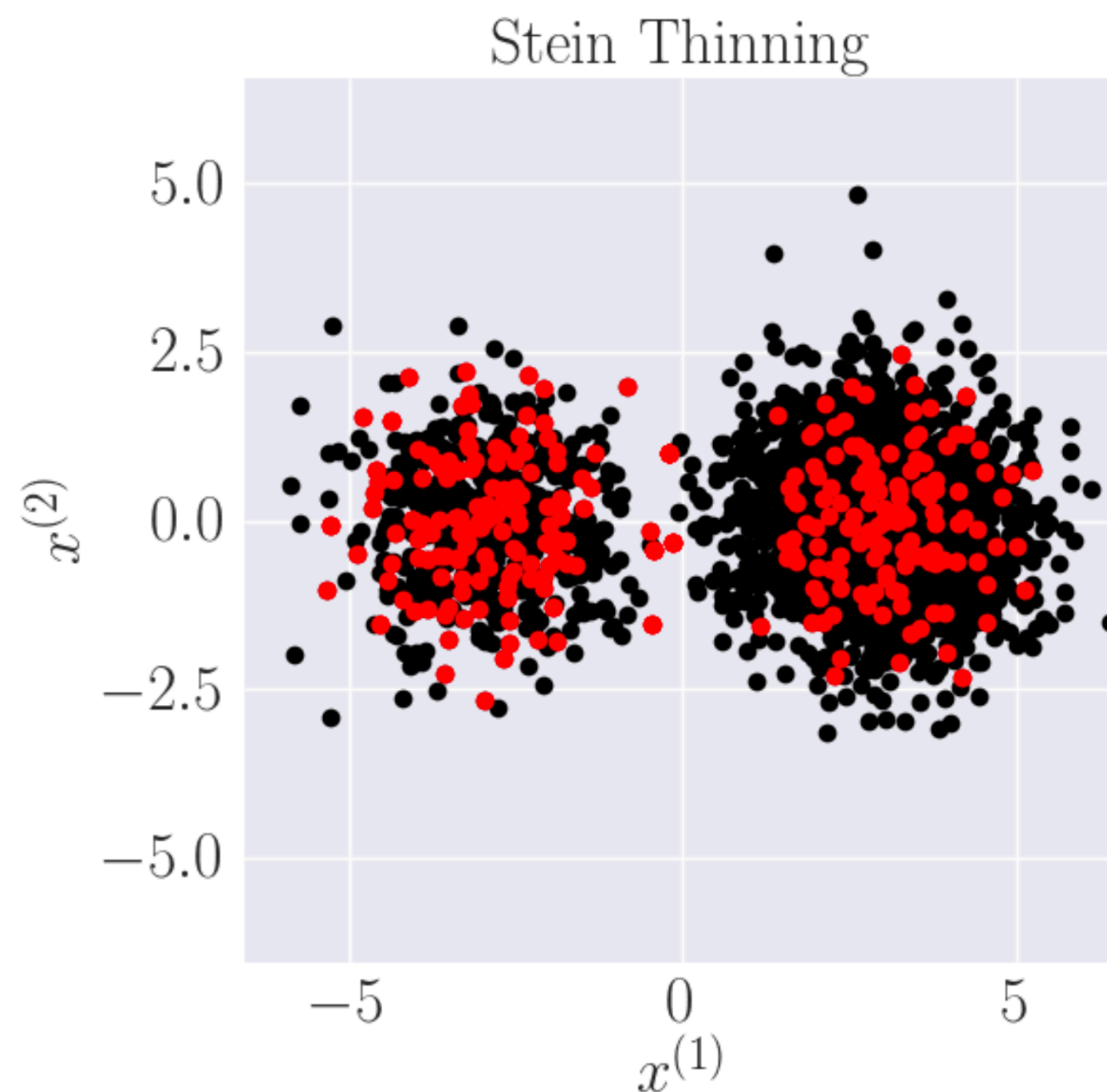
$$\text{KSD}_\lambda^2(\mathbb{P}, \mathbb{P}') = \mathbb{E}[k_p(Z, Z')] - \lambda \mathbb{E}[\log p(Z)]$$

- ◉ The second term takes higher values in modes with smaller probability
- ◉ It is known up to an additive constant in the Bayesian setting, but greedy selection of particles used in practice does not need it

# KSD pathologies - Regularized KSD

- **Pathology I: mode proportion blindness**
  - The score function is insensitive to distant mode weights
  - We propose **entropic regularization** to lessen this phenomenon

$$\text{KSD}_\lambda^2(\mathbb{P}, \mathbb{P}') = \mathbb{E}[k_p(Z, Z')] - \lambda \mathbb{E}[\log p(Z)]$$

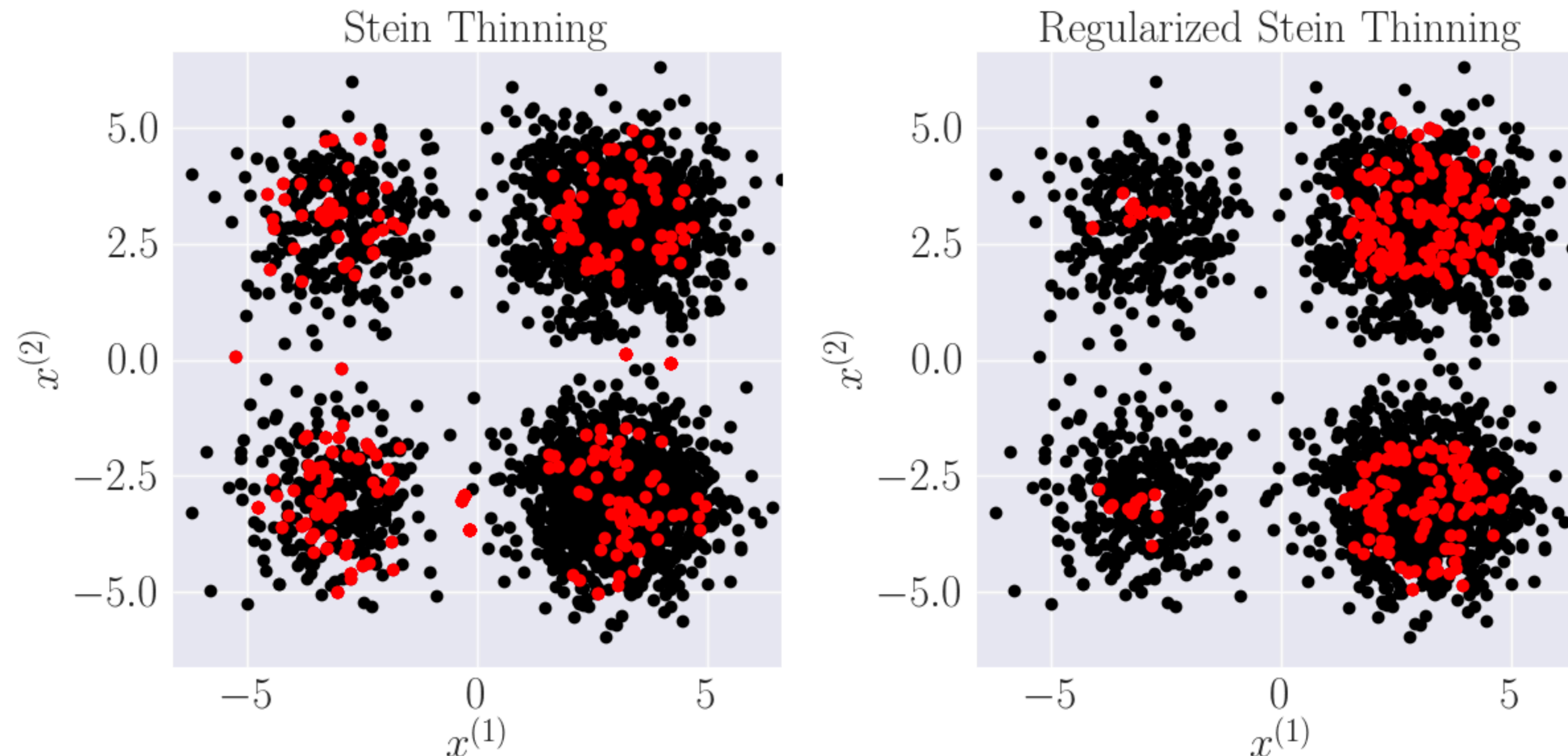




# KSD pathologies - Regularized KSD

- **Pathology I: mode proportion blindness**
  - The score function is insensitive to distant mode weights
  - We propose **entropic regularization** to lessen this phenomenon

$$\text{KSD}_\lambda^2(\mathbb{P}, \mathbb{P}') = \mathbb{E}[k_p(Z, Z')] - \lambda \mathbb{E}[\log p(Z)]$$



# KSD pathologies - Regularized KSD

- **Pathology I: mode proportion blindness**
  - ◉ The score function is insensitive to distant mode weights
  - ◉ We propose **entropic regularization** to lessen this phenomenon

$$\text{KSD}_\lambda^2(\mathbb{P}, \mathbb{P}') = \mathbb{E}[k_p(Z, Z')] - \lambda \mathbb{E}[\log p(Z)]$$

**Theorem 3.2.** *Let  $k_p$  be the Stein kernel associated with the radial kernel  $k(\mathbf{x}, \mathbf{x}') = \phi(\|\mathbf{x} - \mathbf{x}'\|_2/\ell)$ , where  $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ ,  $\ell > 0$ , and  $\phi \in \mathcal{C}^2(\mathbb{R}^d)$ . Let  $p$  and  $q$  be two bimodal mixture distributions satisfying Assumption 2.1. We define  $w_\lambda^*$  as the optimal mixture weight of  $q$  with respect to the entropic regularized KSD distance, i.e.,  $w_\lambda^* = \underset{w \in [0,1]}{\operatorname{argmin}} \text{KSD}_\lambda(\mathbb{P}, \mathbb{Q}_w)$ . If  $\mathbb{E}[\log(p(\mathbf{Z}_L))] \neq \mathbb{E}[\log(p(\mathbf{Z}_R))]$  where  $\mathbf{Z}_L \sim \mathbb{Q}_L$  and  $\mathbf{Z}_R \sim \mathbb{Q}_R$ , it exists  $\lambda \in \mathbb{R}$  such that  $w_\lambda^* = w_p$ .*

- ◉ There is a  $\lambda$  such that the true proportion is recovered

# KSD pathologies - Regularized KSD

- **Pathology II: spurious minimum**

- ◉ The KSD selects samples concentrated in regions of low probability
- ◉ We propose a **Laplacian correction** to lessen this phenomenon

$$\text{L-KSD}^2(\mathbb{P}, \mathbb{P}'_m) = \frac{1}{m^2} \sum_{i \neq j}^m k_p(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^m [k_p(x_i, x_i) + \Delta^+ \log p(x_i)]$$

- ◉ Since Pathology II is caused by the weaknesses of the diagonal terms, which favor samples concentrated in stationary points
- ◉ We thus penalize them more heavily with the positive values of the Laplacian of the density (since they are located in areas of convexity of the density)

$$\Delta^+ f(\mathbf{x}) = \sum_{j=1}^d \left( \partial^2 f(\mathbf{x}) / \partial x^{(j)2} \right)^+$$

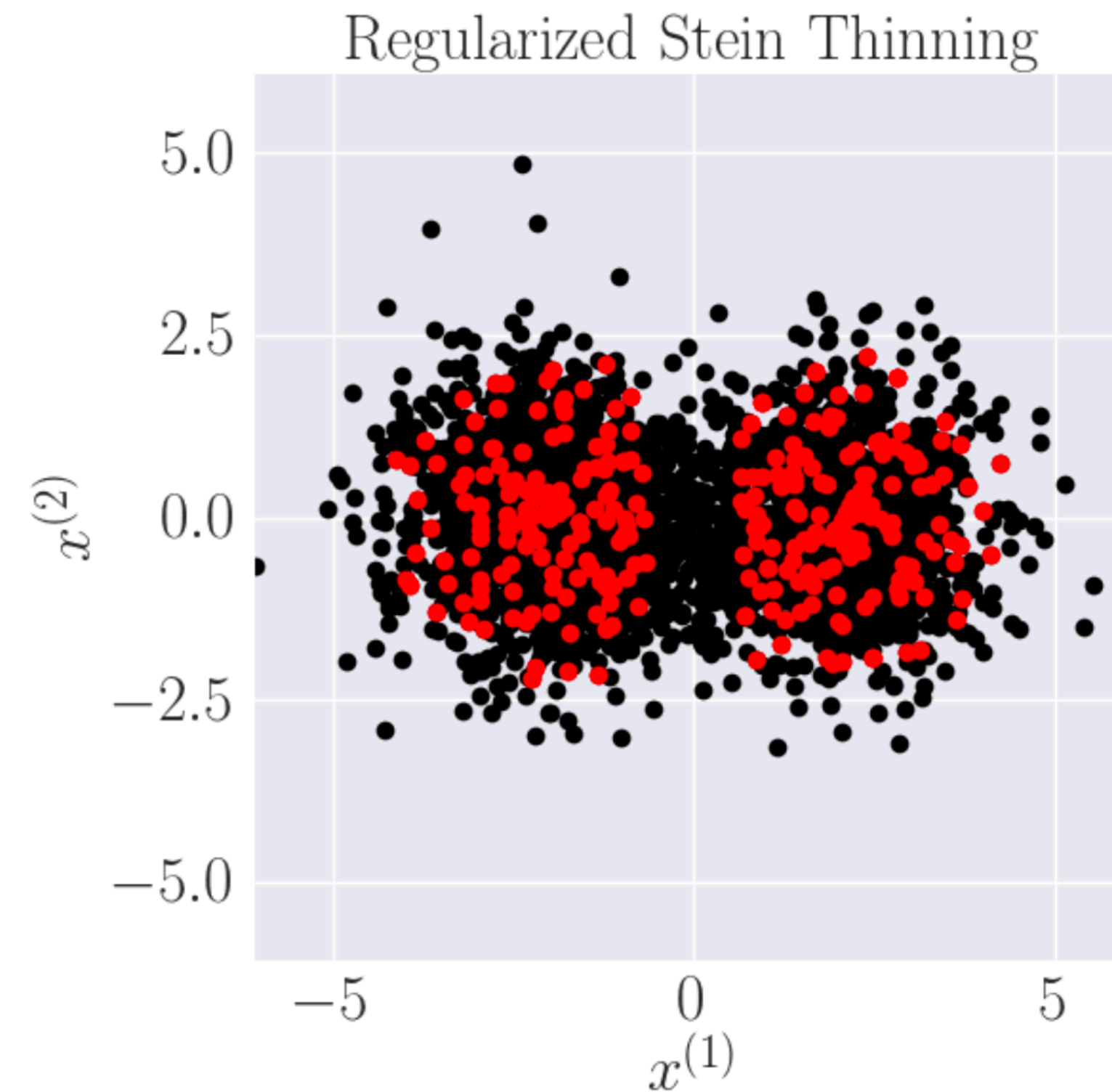
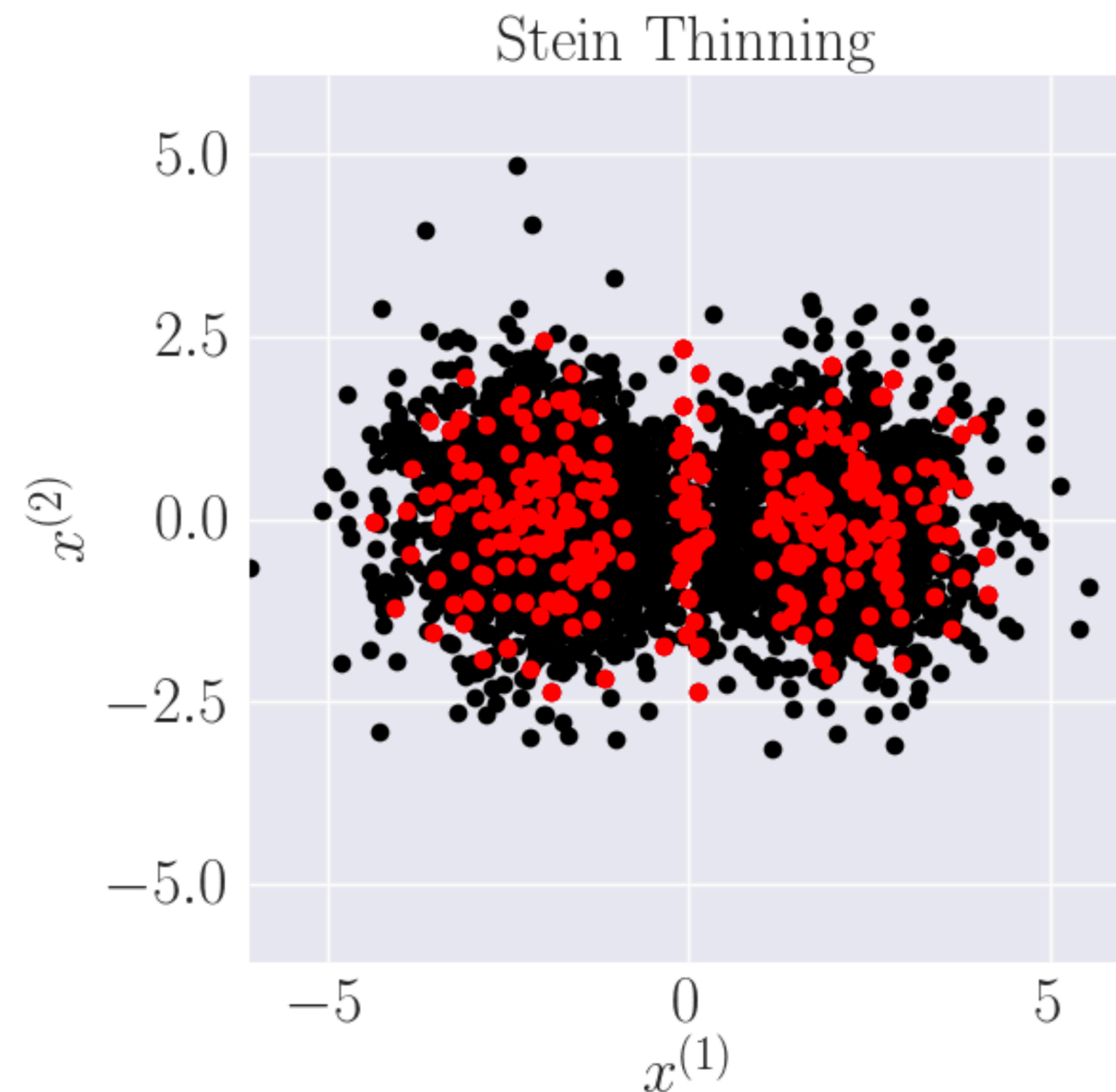


# KSD pathologies - Regularized KSD

- **Pathology II: spurious minimum**

- The KSD selects samples concentrated in regions of low probability
- We propose a **Laplacian correction** to lessen this phenomenon

$$\text{L-KSD}^2(\mathbb{P}, \mathbb{P}'_m) = \frac{1}{m^2} \sum_{i \neq j}^m k_p(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^m [k_p(x_i, x_i) + \Delta^+ \log p(x_i)]$$

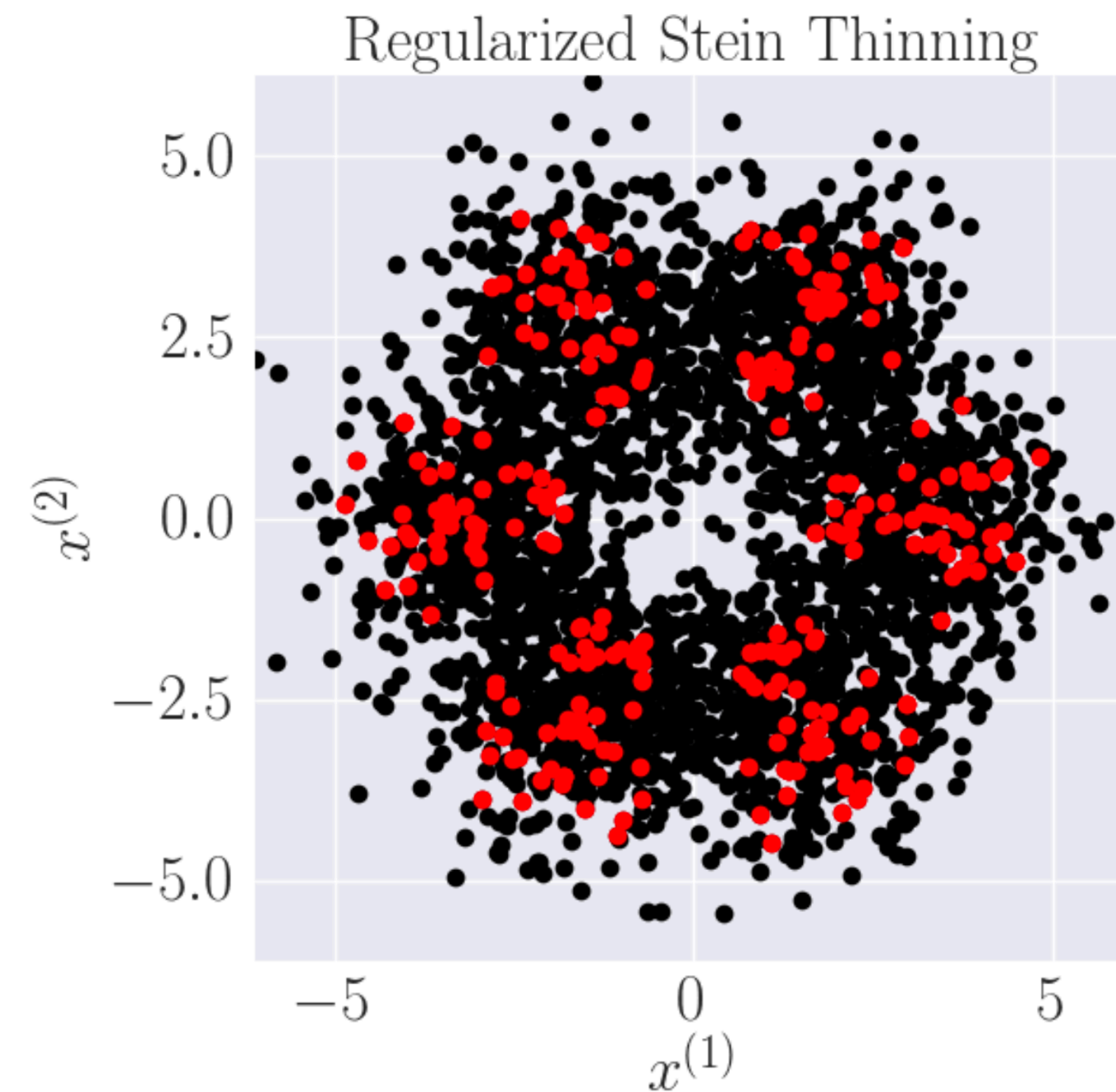
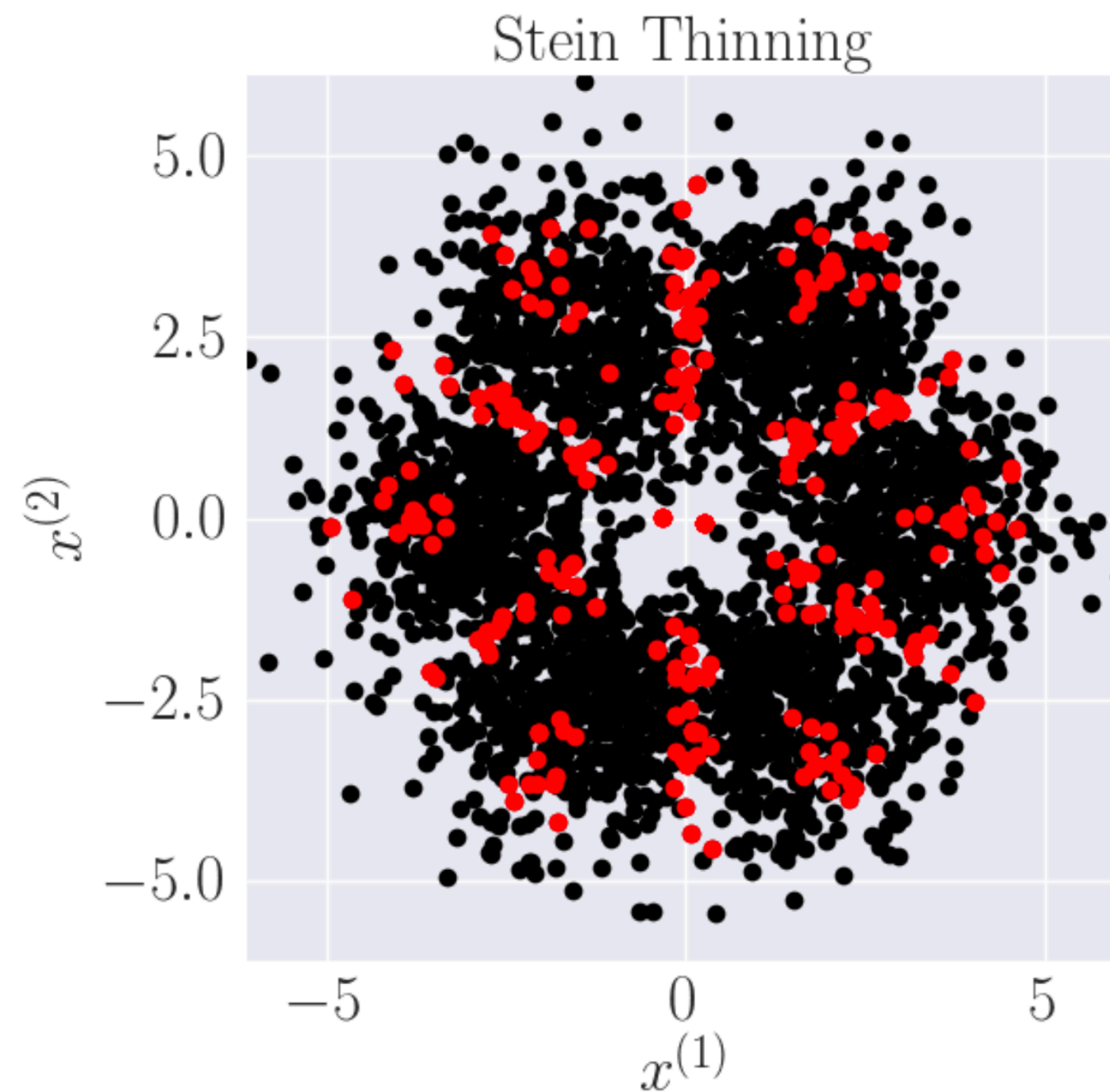


# KSD pathologies - Regularized KSD

- **Pathology II: spurious minimum**

- The KSD selects samples concentrated in regions of low probability
- We propose a **Laplacian correction** to lessen this phenomenon

$$\text{L-KSD}^2(\mathbb{P}, \mathbb{P}'_m) = \frac{1}{m^2} \sum_{i \neq j}^m k_p(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^m [k_p(x_i, x_i) + \Delta^+ \log p(x_i)]$$



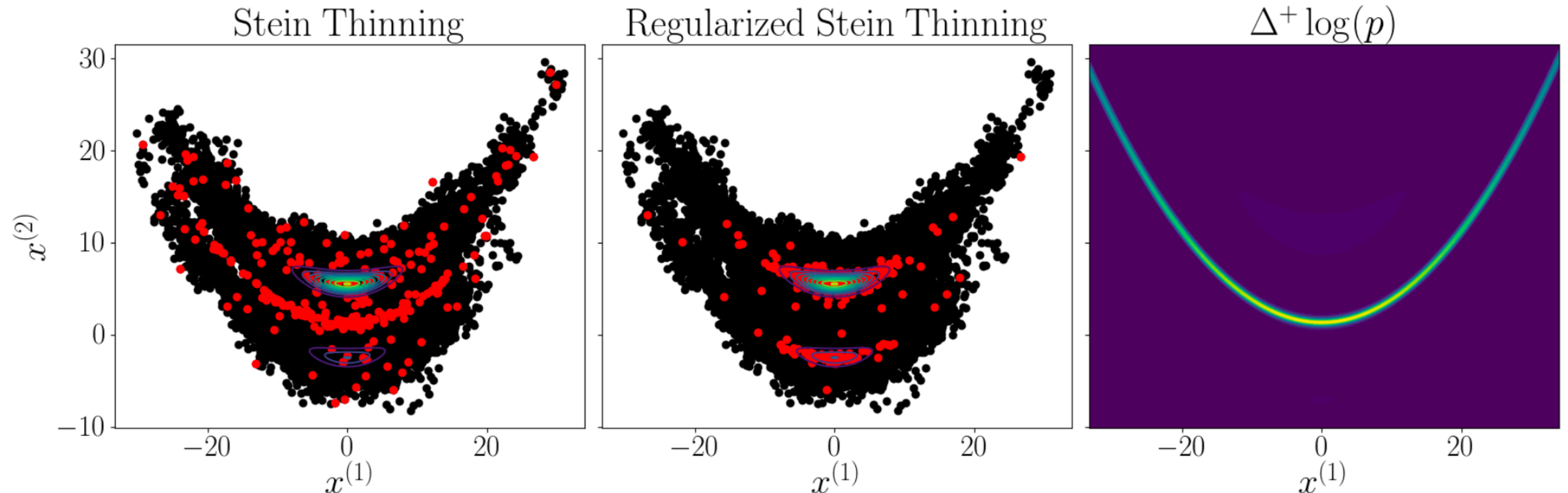


# KSD pathologies - Regularized KSD

- **Pathology II: spurious minimum**

- ◉ The KSD selects samples concentrated in regions of low probability
- ◉ We propose a **Laplacian correction** to lessen this phenomenon

$$\text{L-KSD}^2(\mathbb{P}, \mathbb{P}'_m) = \frac{1}{m^2} \sum_{i \neq j}^m k_p(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^m [k_p(x_i, x_i) + \Delta^+ \log p(x_i)]$$



# KSD pathologies - Regularized KSD

- **Pathology II: spurious minimum**

- ◉ The KSD selects samples concentrated in regions of low probability
- ◉ We propose a **Laplacian correction** to lessen this phenomenon

$$\text{L-KSD}^2(\mathbb{P}, \mathbb{P}'_m) = \frac{1}{m^2} \sum_{i \neq j}^m k_p(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^m [k_p(x_i, x_i) + \Delta^+ \log p(x_i)]$$

**Theorem 3.3.** *Let  $k_p$  be the Stein kernel associated with the IMQ kernel with  $\ell > 0$ ,  $\beta \in (0, 1)$ , and  $c = 1$ . For  $m \geq 2$ , let  $\{\mathbf{x}_i\}_{i=1}^m \subset \mathbb{R}^d$  be a set of points concentrated at  $\mathbf{x}_0$ , a local minimum or saddle point of  $p$ , and of empirical measure  $\mathbb{Q}_m$ . Then, we have  $\text{L-KSD}^2(\mathbb{P}, \mathbb{Q}_m) > \mathbb{E}[\text{L-KSD}^2(\mathbb{P}, \mathbb{P}_m)]$ , if the density at  $\mathbf{x}_0$  satisfies  $p(\mathbf{x}_0) < \Delta^+ p(\mathbf{x}_0) / (\mathbb{E}[\|s_p(\mathbf{X})\|_2^2] + \mathbb{E}[\Delta^+ \log p(\mathbf{X})])$ .*

- ◉ Points with low score are not interesting candidates with respect to the L-KSD

# KSD pathologies - Regularized KSD

- **We also keep the central convergence result of KSD thinning**

- ◉ Riabiz et al. 2022: for a distantly dissipative target distribution and if the sample candidates are generated by a MCMC algorithm, samples generated by KSD thinning converge almost surely towards the target

- ◉ We extend their result to our regularized KSD, with the additional assumption

$$\lambda_m = o(\log m/m)$$

- ◉ This gives a rule of thumb for the choice of the penalty intensity, which works surprisingly well in all our experiments :

$$\lambda = 1/m$$

Bénard, C., Staber, B., & Da Veiga, S. (2023). Kernel Stein Discrepancy thinning: a theoretical perspective of pathologies and a practical fix with regularization. *Neurips 2023*

**Several appearances in ETICS community**

**3 - Optimal transport & the Wasserstein**



# Optimal transport: everywhere in ETICS

## ETICS 2019

École Thématique sur les Incertitudes en Calcul Scientifique  
Research School on Uncertainty in Scientific Computing

September, 22-27, Fréjus, France - <https://www.caes.cnrs.fr/sejours/la-villa-clythia>



Talk of Y. Marzouk

## ETICS 2023

École Thématique sur les Incertitudes en Calcul Scientifique  
Research School on Uncertainty in Scientific Computing

<https://www.gdr-mascotnum.fr/etics.html>

October, 8-13, [VVF Lège Cap Ferret](#), France



Talk of R. Carpintero-Perez



## ETICS 2022

École Thématique sur les Incertitudes en Calcul Scientifique  
Research School on Uncertainty in Scientific Computing

<https://www.gdr-mascotnum.fr/etics.html>

October, 2-7, Belhambra, Belgodère Golfe de Lozari , France -  
<https://www.belambra.com/club-belgodere-golfe-de-lozari/summer>



Talks of G. Peyré & M. Il Idrissi

## ETICS 2024

École Thématique sur les Incertitudes en Calcul Scientifique  
Research School on Uncertainty in Scientific Computing

<https://www.gdr-mascotnum.fr/etics.html>

September, 22-27, VVF, France  
<https://www.vvf.fr/villages-vacances/vacances-saissac-vvf-villages.html>



Talk of R. Carpintero-Perez



# Optimal transport: focus on building kernels

- **Goal: build regression models with highly structured inputs**

- 3D meshes / graphs

- Point clouds

ETICS 2023  
École Thématique sur les Incertitudes en Calcul Scientifique  
Research School on Uncertainty in Scientific Computing  
<https://www.gdr-mascotnum.fr/etics.html>

October, 8-13, [VVF Lège Cap Ferret](#), France



Talk of R.  
Carpintero-Perez

ETICS 2022  
École Thématique sur les Incertitudes en Calcul Scientifique  
Research School on Uncertainty in Scientific Computing  
<https://www.gdr-mascotnum.fr/etics.html>

October, 2-7, Belhambra, Belgodère Golfe de Lozari, France -  
<https://www.belambra.com/club-belgodere-golfe-de-lozari/summer>

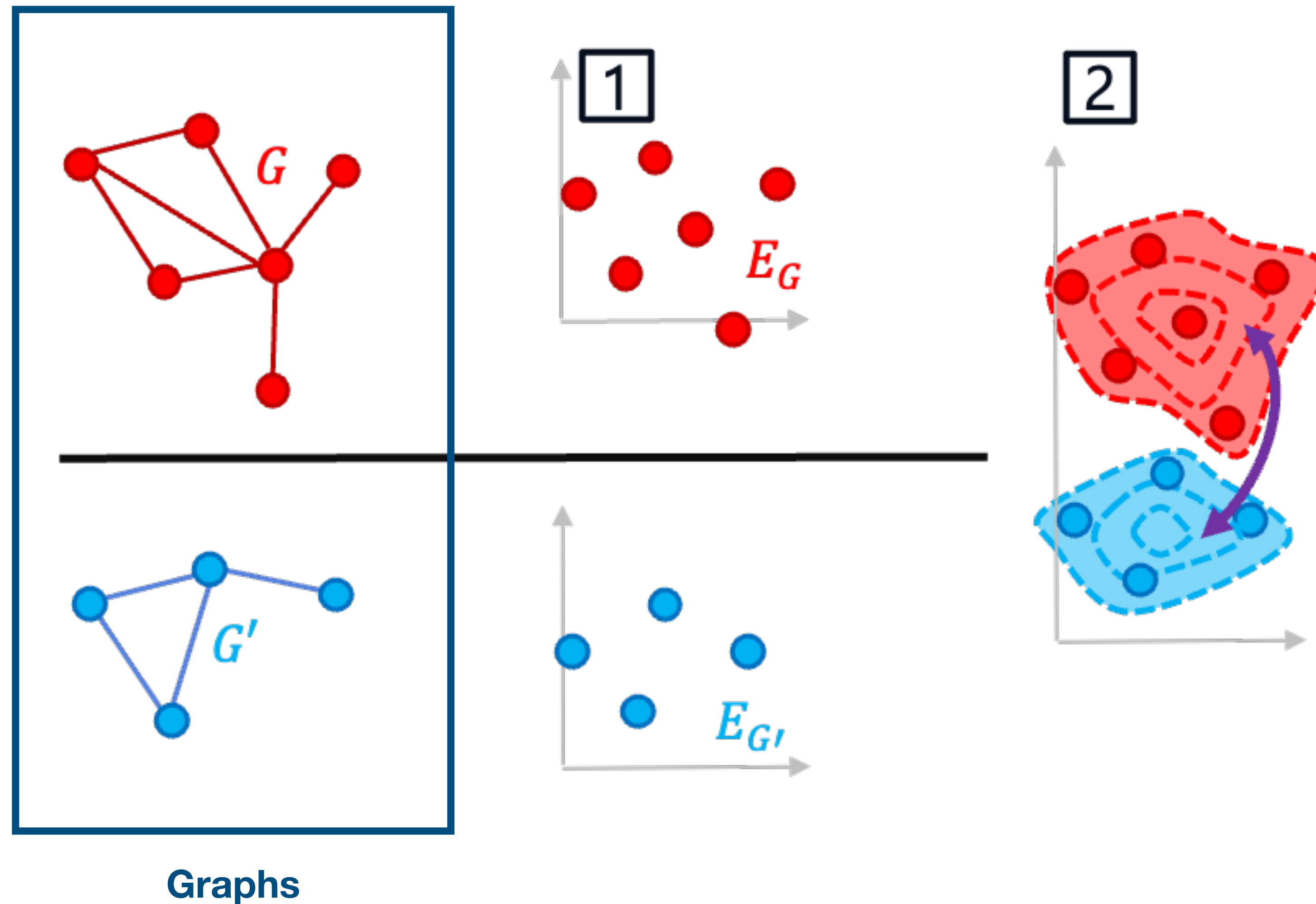


Talk of B. Sow

- **If using kernel methods, « just » need to design the kernel**

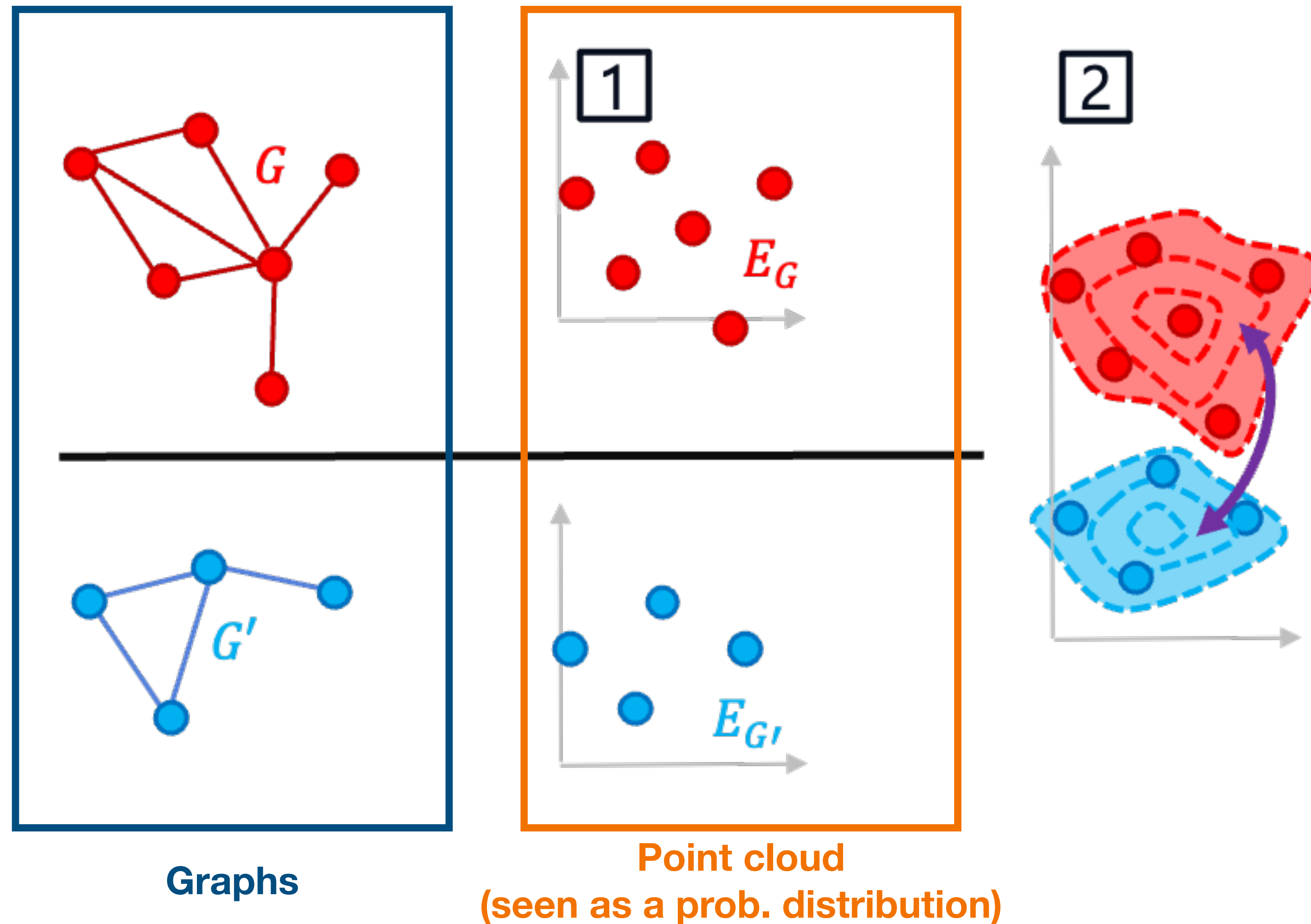
# Optimal transport: focus on building kernels

- General methodology for a kernel between graphs



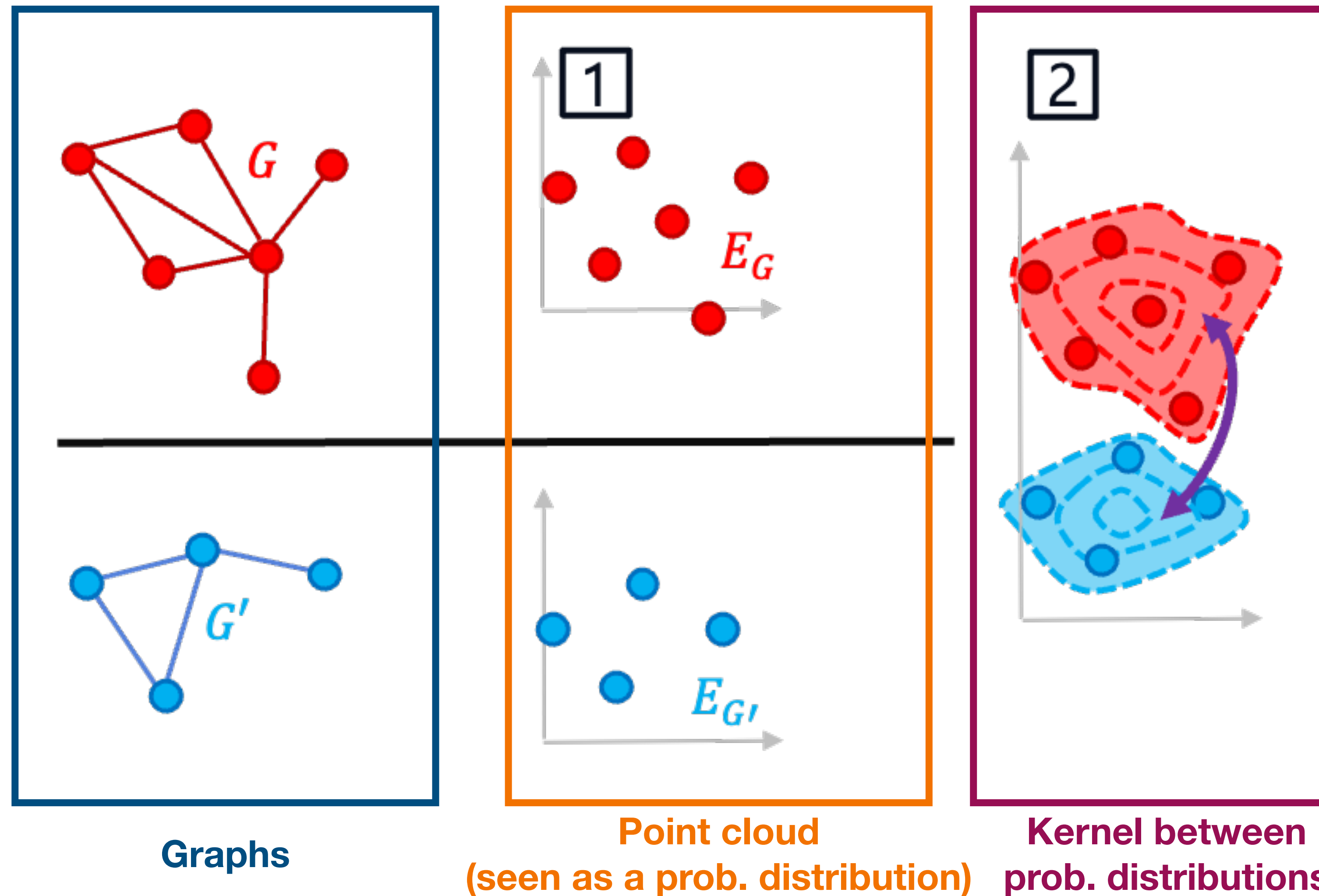
# Optimal transport: focus on building kernels

- General methodology for a kernel between graphs



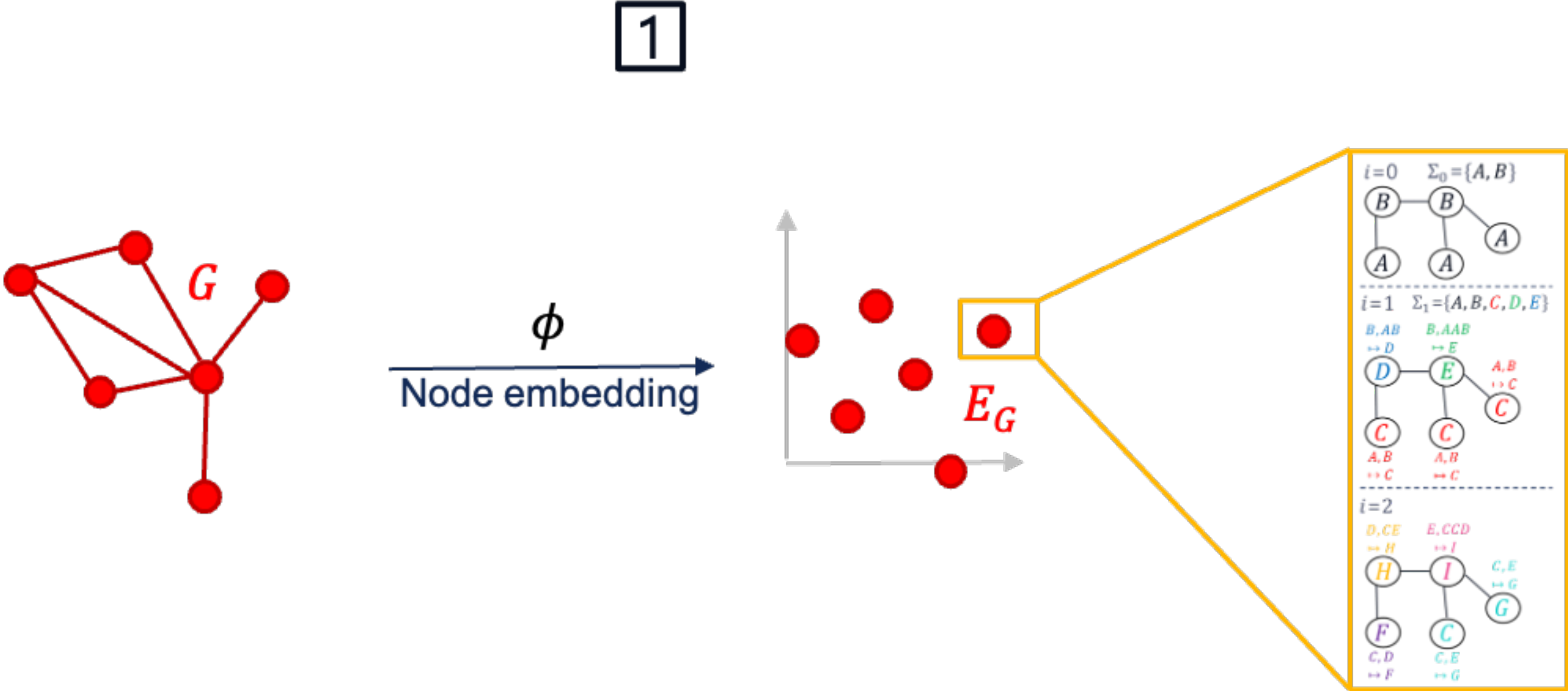
# Optimal transport: focus on building kernels

- General methodology for a kernel between graphs



# Optimal transport: focus on building kernels

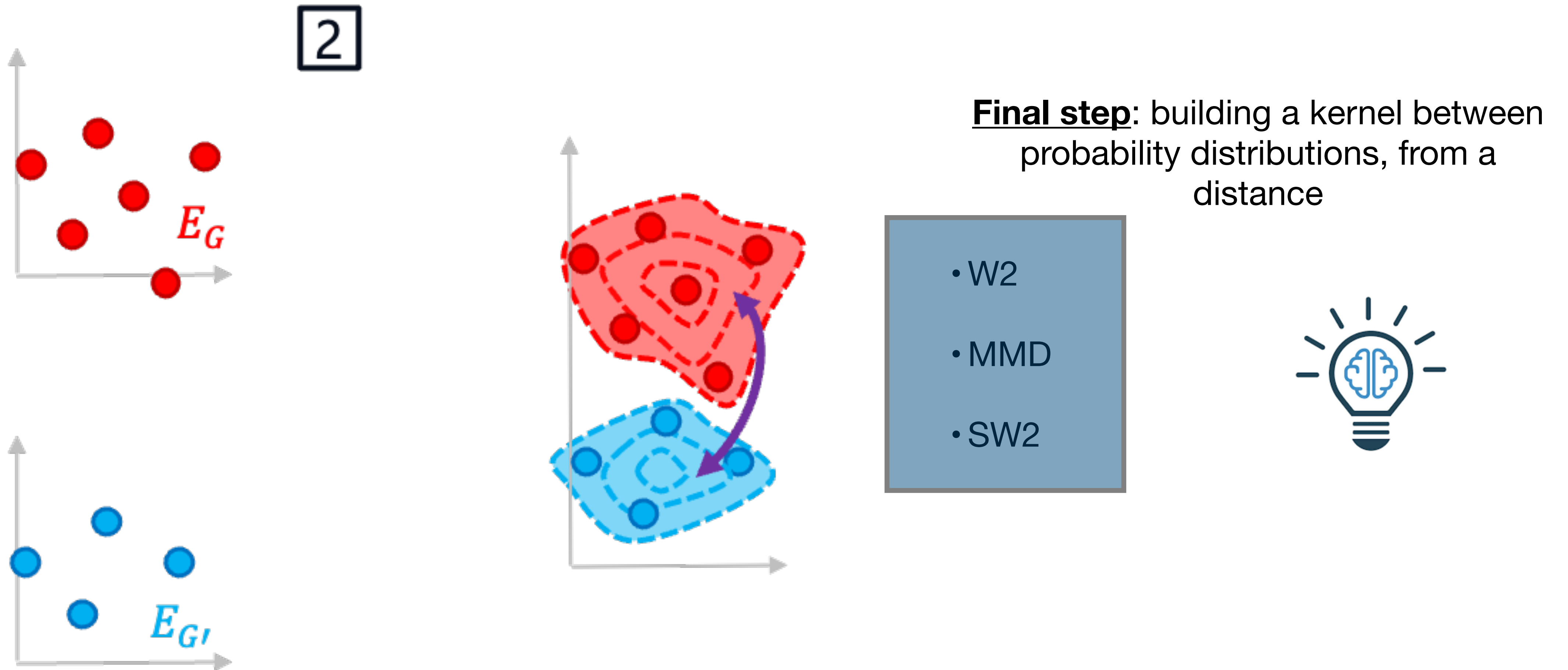
- General methodology for a kernel between graphs





# Optimal transport: focus on building kernels

- General methodology for a kernel between graphs



# Optimal transport: focus on building kernels

- The last ingredient is to define a (sdp) kernel to compare probability distributions

- Kernel based on Wasserstein distance ( $W_2$ )

$$k_{W_2}(P, Q) = \exp(-\gamma W_2(P, Q))$$

sdp for any power between 0 and 2  
and for **one-dimensional distributions only**

Complexity  $\mathcal{O}(n \log n)$

Peyré & Cuturi (2019)

- Kernel based on Maximum Mean Discrepancy (MMD)

$$k_{\text{MMD}}(P, Q) = \exp(-\gamma \text{MMD}^2(P, Q))$$

sdp for **any distributions**

Complexity  $\mathcal{O}(n^2)$

Song (2008)

- Kernel based on Sliced-Wasserstein distance ( $SW_2$ )

$$k_{SW_2}(P, Q) = \exp(-\gamma SW_2(P, Q))$$

sdp for any power between 0 and 2  
and for **any distributions**

Complexity  $\mathcal{O}(R n \log n)$

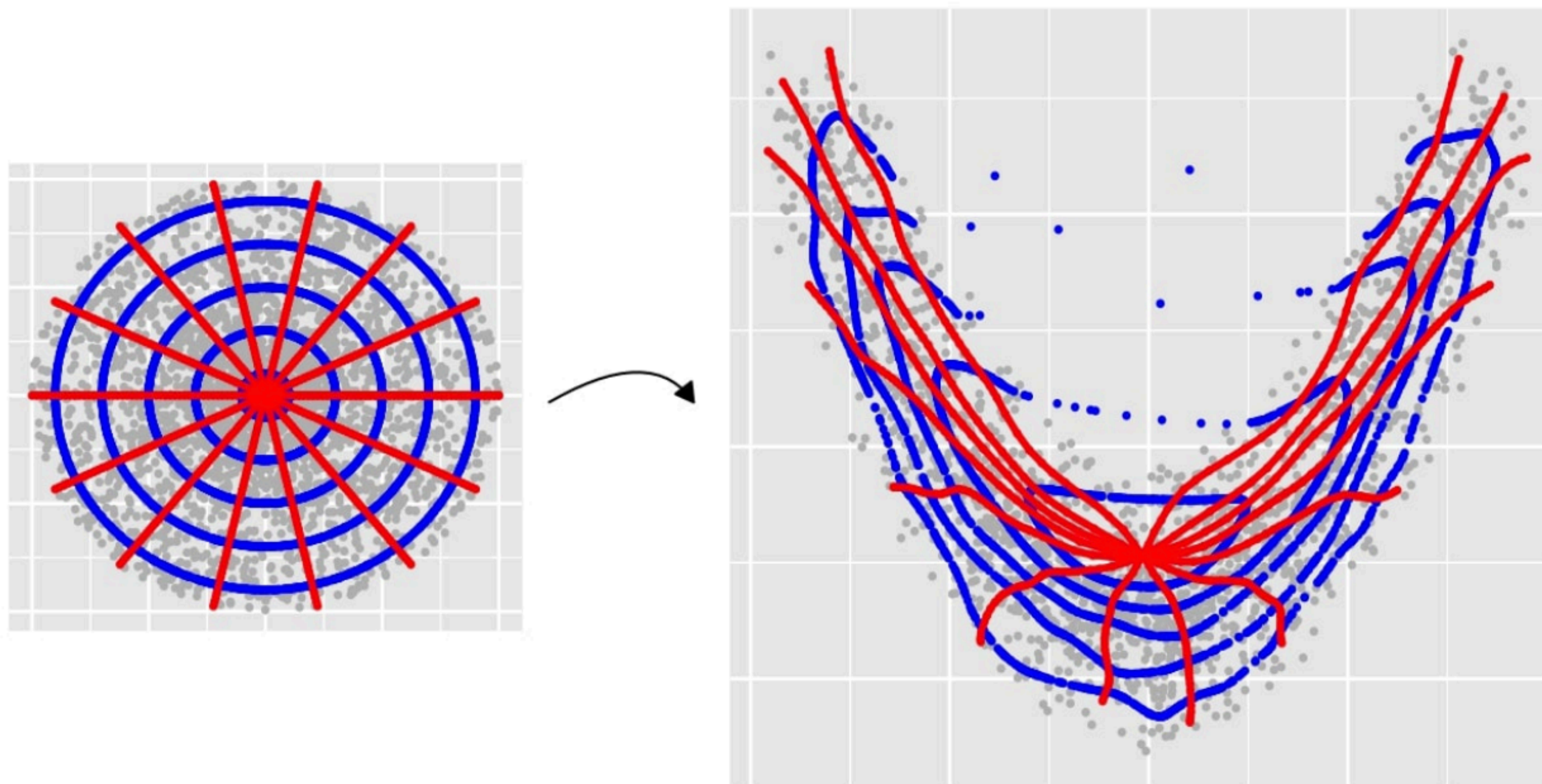
Meunier et al. (2022)

# **Selected topics for future directions**

**Selected topics for future directions**  
**1- Other usage of OT**

# Multivariate quantiles

- **Recent framework (Hallin et al. 2021, Ghosal and Sen 2022)**
  - Step 1: choose a reference measure, with natural ordering
  - Step 2: transport your multivariate distribution towards the reference



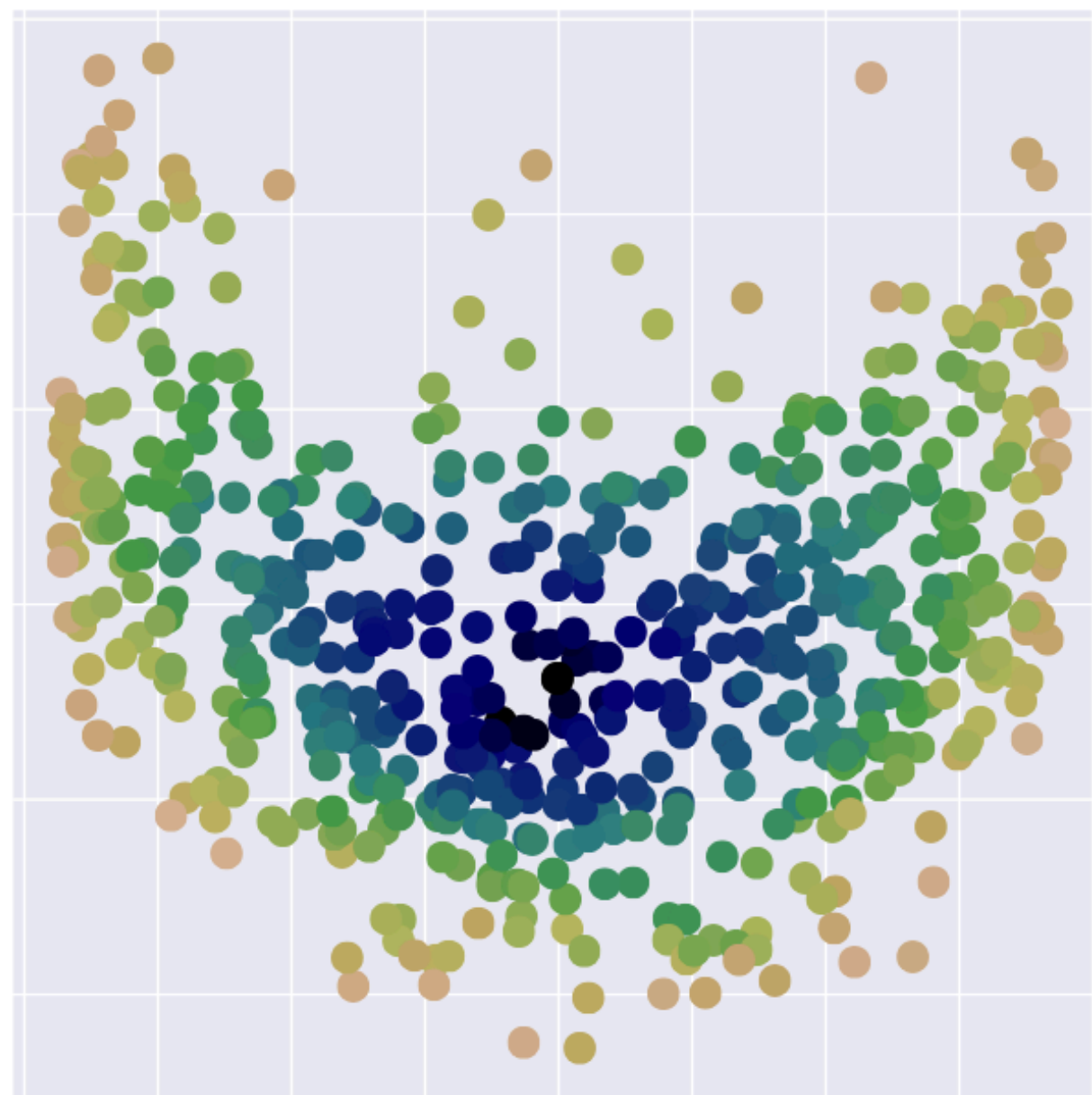
Thurin 2024

FIGURE 3 – (Gauche) quantiles d'une loi de référence et (droite) quantiles de Monge-Kantorovich d'une loi discrète  $\nu$  obtenus par  $\mathbf{Q}_{\#}\mu = \nu$ .

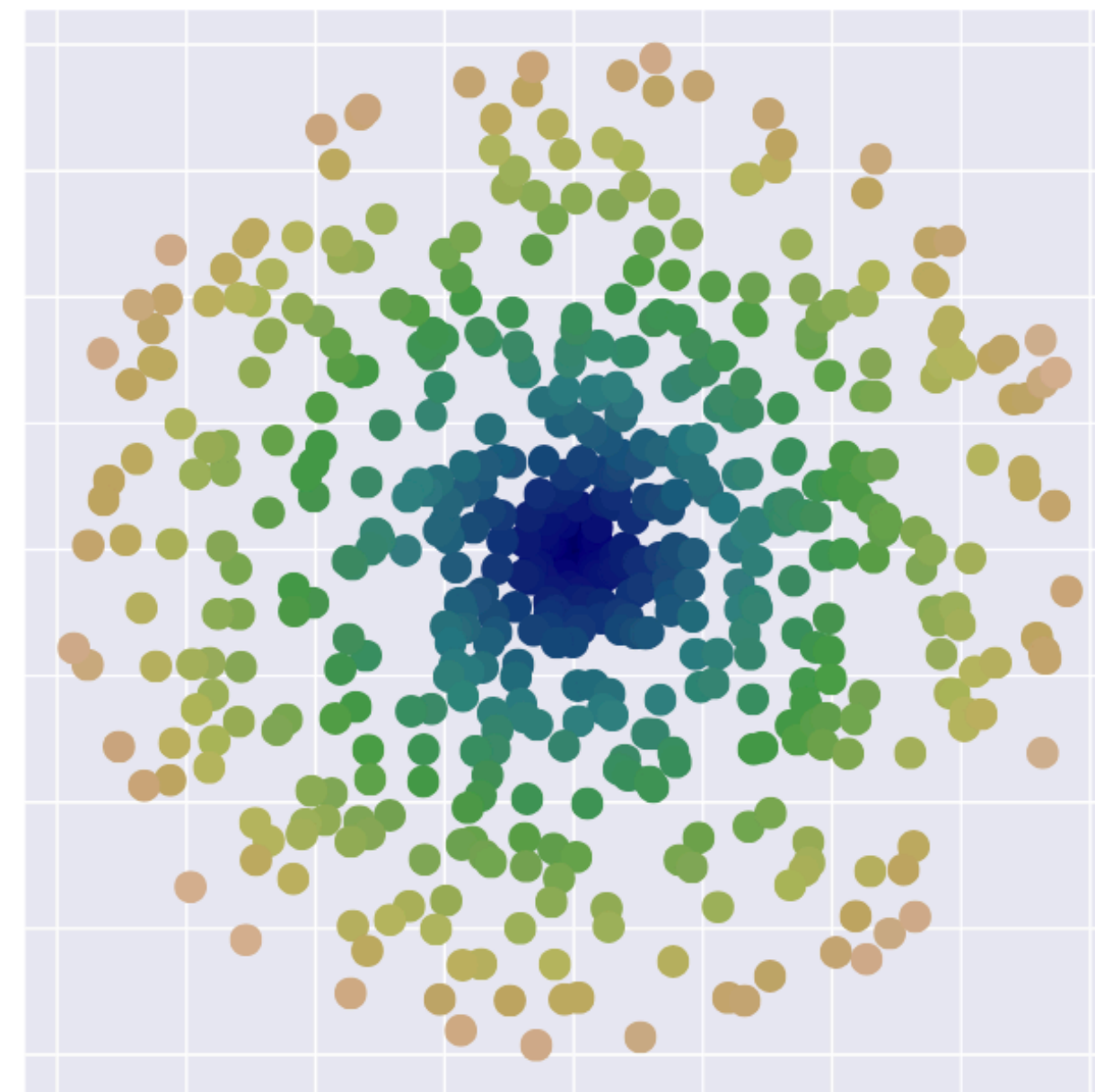


# Multivariate quantiles

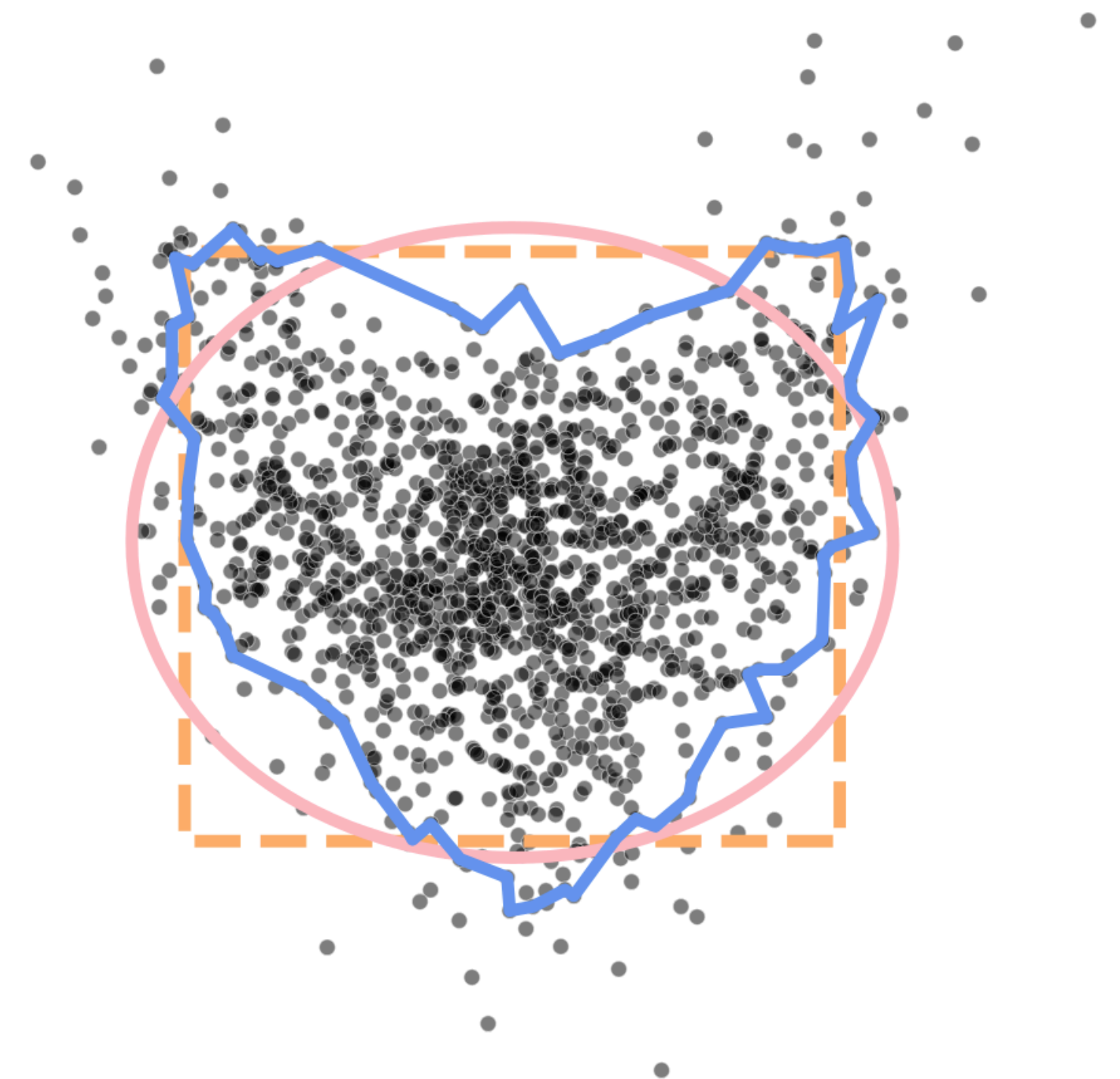
- **Recent framework (Hallin et al. 2021, Ghosal and Sen 2022)**
  - Step 1: choose a reference measure, with natural ordering
  - Step 2: transport your multivariate distribution towards the reference
- **Recently used for multivariate conformal prediction (Thurin et al. 2025)**



(a) Multivariate scores  $\{S_i\}_{i=1}^n$



(b) Reference rank vectors  $\{U_i\}_{i=1}^n$



**Selected topics for future directions**  
**2- Links between distances**

# Links between distances

- Recent results to link HSIC and MI

$$\sup_{x \in \mathcal{X}, y \in \mathcal{Y}} h((x, y), (x, y)) \leq \nu^2$$

Assumption that the HSIC kernel is bounded

**Result 1:** links between HSIC and TV

$$\sqrt{\text{HSIC}(X, Y)} \leq \sup_{f: \|f\|_\infty \leq \nu} \mathbb{E}_{(X, Y) \sim \mathbb{P}_{xy}} [f(X, Y)] - \mathbb{E}_{\substack{X \sim \mathbb{P}_x \\ Y' \sim \mathbb{P}_y}} [f(X, Y')] = 2\nu \text{TV}(\mathbb{P}_{xy}, \mathbb{P}_x \times \mathbb{P}_y)$$

Wang & Tay 2023, Xu et al. 2025

# Links between distances

- Recent results to link HSIC and MI

$$\sup_{x \in \mathcal{X}, y \in \mathcal{Y}} h((x, y), (x, y)) \leq \nu^2$$

Assumption that the HSIC kernel is bounded

**Result 1:** links between HSIC and TV

$$\sqrt{\text{HSIC}(X, Y)} \leq \sup_{f: \|f\|_\infty \leq \nu} \mathbb{E}_{(X, Y) \sim \mathbb{P}_{xy}} [f(X, Y)] - \mathbb{E}_{\substack{X \sim \mathbb{P}_x \\ Y' \sim \mathbb{P}_y}} [f(X, Y')] = 2\nu \text{TV}(\mathbb{P}_{xy}, \mathbb{P}_x \times \mathbb{P}_y)$$

Wang & Tay 2023, Xu et al. 2025

$$\frac{1}{2\nu^2} \text{HSIC}(X, Y) \leq \text{I}(X; Y)$$

Xu et al. 2025

$$-\log \left( 1 - \frac{1}{4\nu^2} \text{HSIC}(X, Y) \right) \leq \text{I}(X; Y)$$

Allain et al. 2025, Xu et al. 2025

# Links between distances

- Recent results to link HSIC and MI

$$\sup_{x \in \mathcal{X}, y \in \mathcal{Y}} h((x, y), (x, y)) \leq \nu^2$$

Assumption that the HSIC kernel is bounded

**Result 1:** links between HSIC and TV

$$\sqrt{\text{HSIC}(X, Y)} \leq \sup_{f: \|f\|_\infty \leq \nu} \mathbb{E}_{(X, Y) \sim \mathbb{P}_{xy}} [f(X, Y)] - \mathbb{E}_{\substack{X \sim \mathbb{P}_x \\ Y' \sim \mathbb{P}_y}} [f(X, Y')] = 2\nu \text{TV}(\mathbb{P}_{xy}, \mathbb{P}_x \times \mathbb{P}_y)$$

Wang & Tay 2023, Xu et al. 2025

$$\frac{1}{2\nu^2} \text{HSIC}(X, Y) \leq \text{I}(X; Y)$$

Xu et al. 2025

$$-\log \left( 1 - \frac{1}{4\nu^2} \text{HSIC}(X, Y) \right) \leq \text{I}(X; Y)$$

Allain et al. 2025, Xu et al. 2025



**Use for CP!**

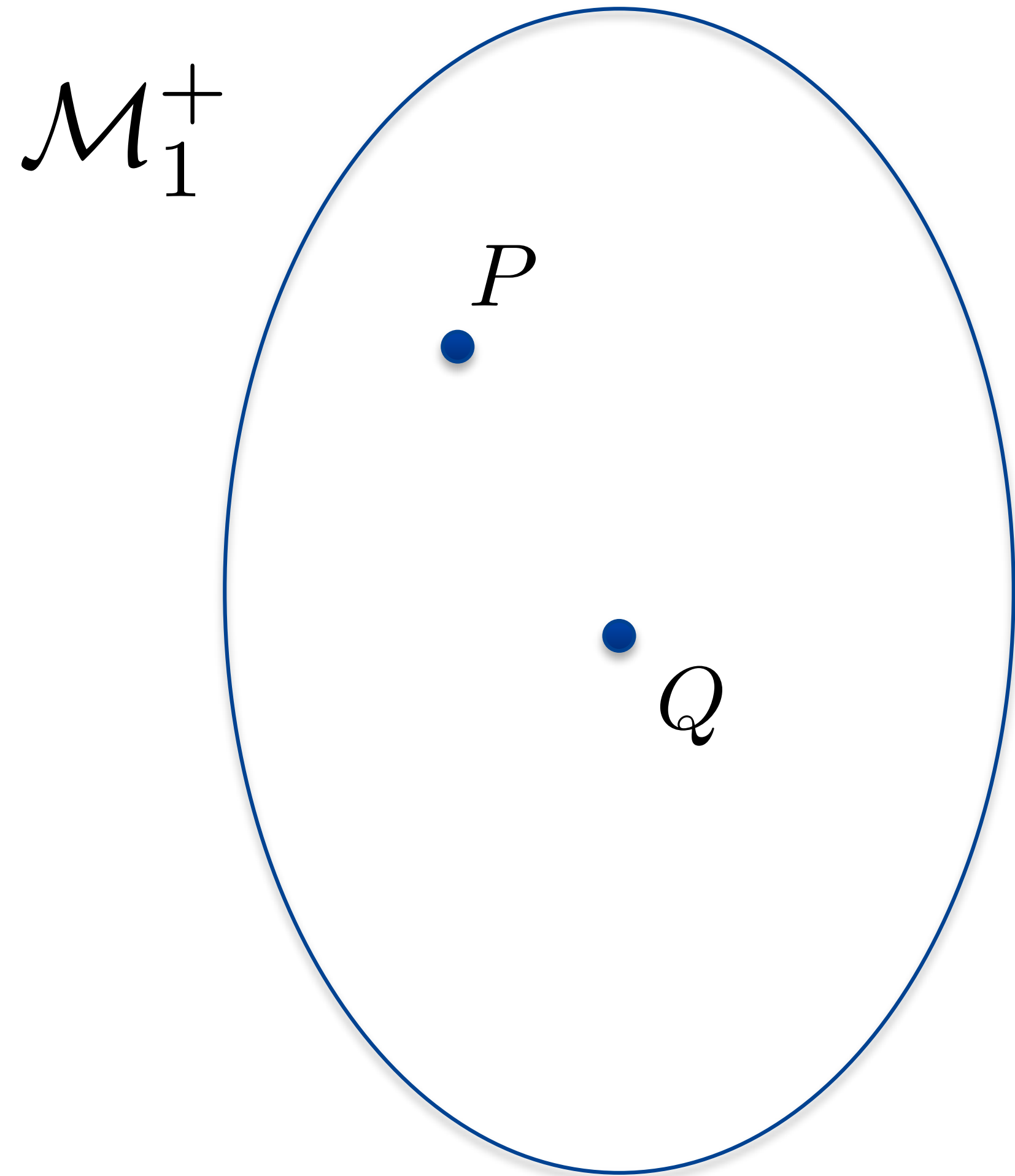


**Selected topics for future directions**

- 3- More kernels!**

# Kernel mean embedding becomes kernel quantile embedding

- Remember mean embedding?

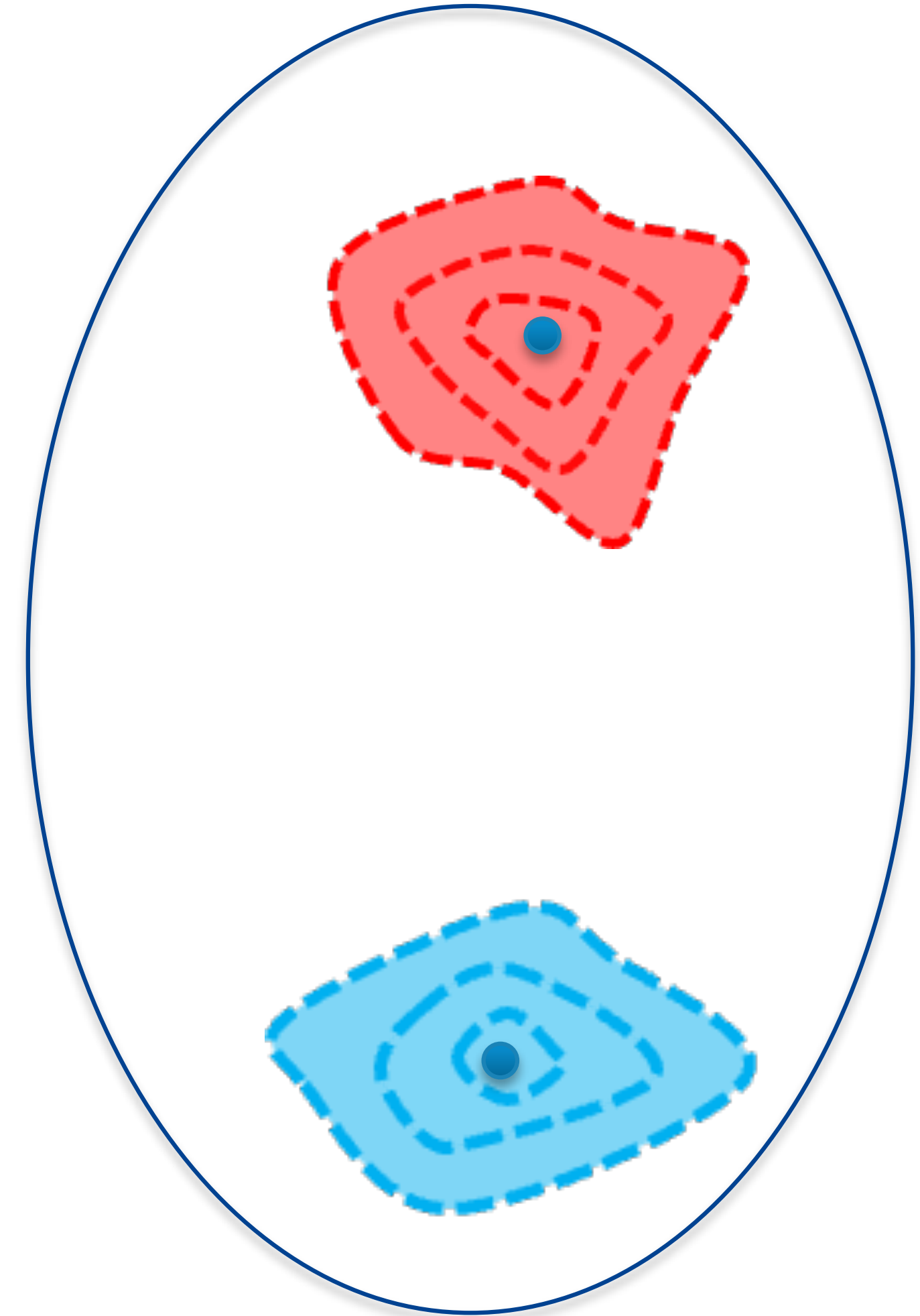
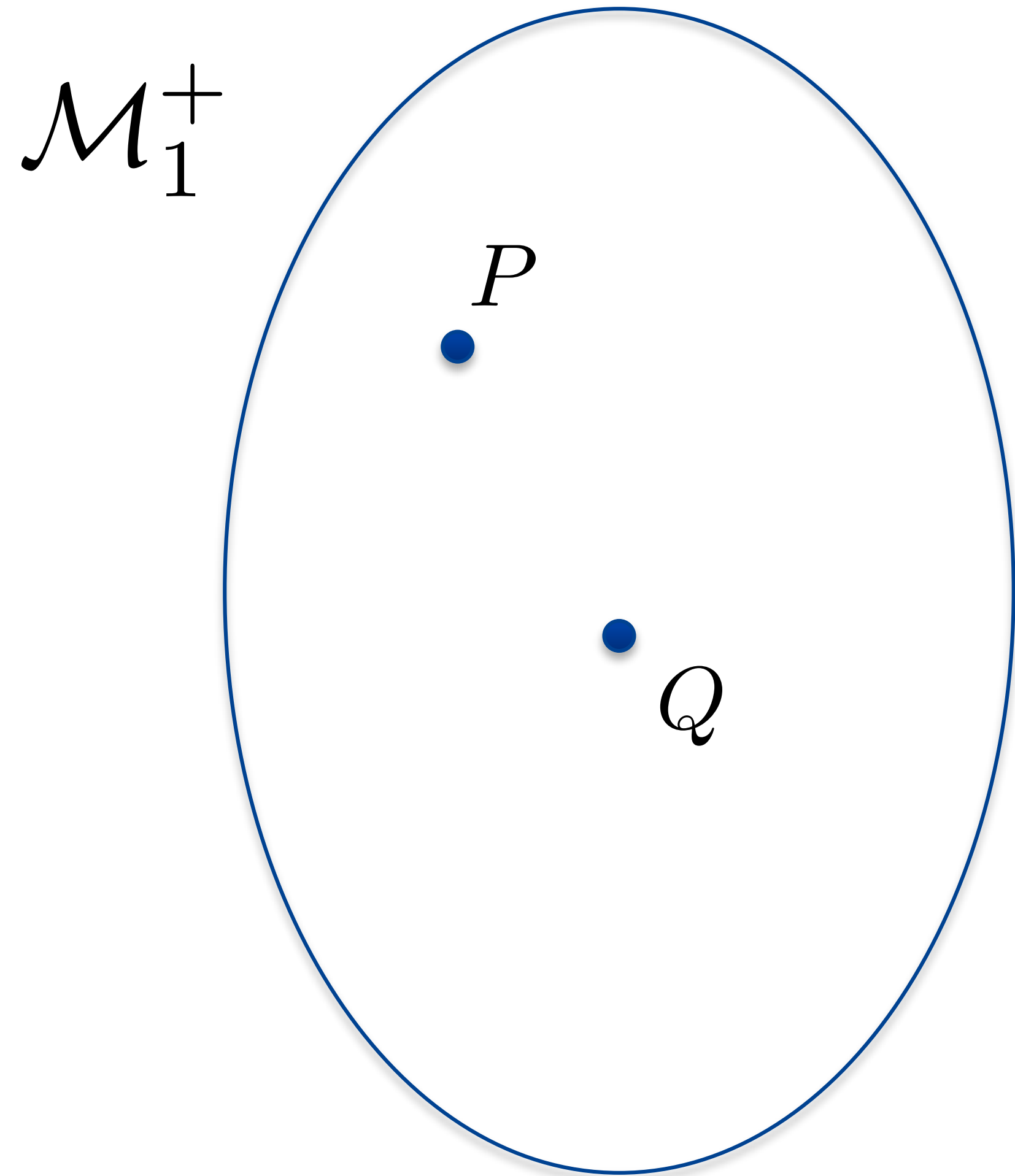


A diagram showing a large blue oval representing a set. Inside the oval, there are two blue dots. The upper dot is associated with the equation  $\mu_P = \mathbb{E}_P (k(X, \cdot))$  and the lower dot is associated with the equation  $\mu_Q = \mathbb{E}_Q (k(X, \cdot))$ .

$$\mu_P = \mathbb{E}_P (k(X, \cdot))$$
$$\mu_Q = \mathbb{E}_Q (k(X, \cdot))$$

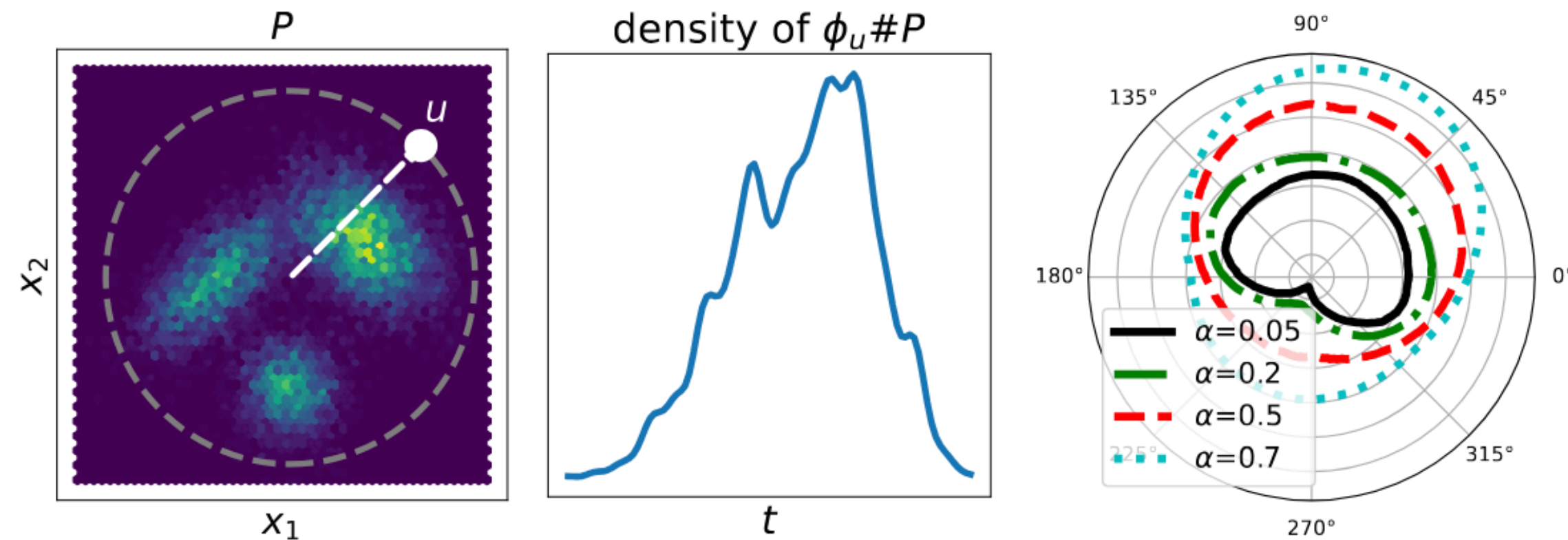
# Kernel mean embedding becomes kernel quantile embedding

- Why focus on the mean?

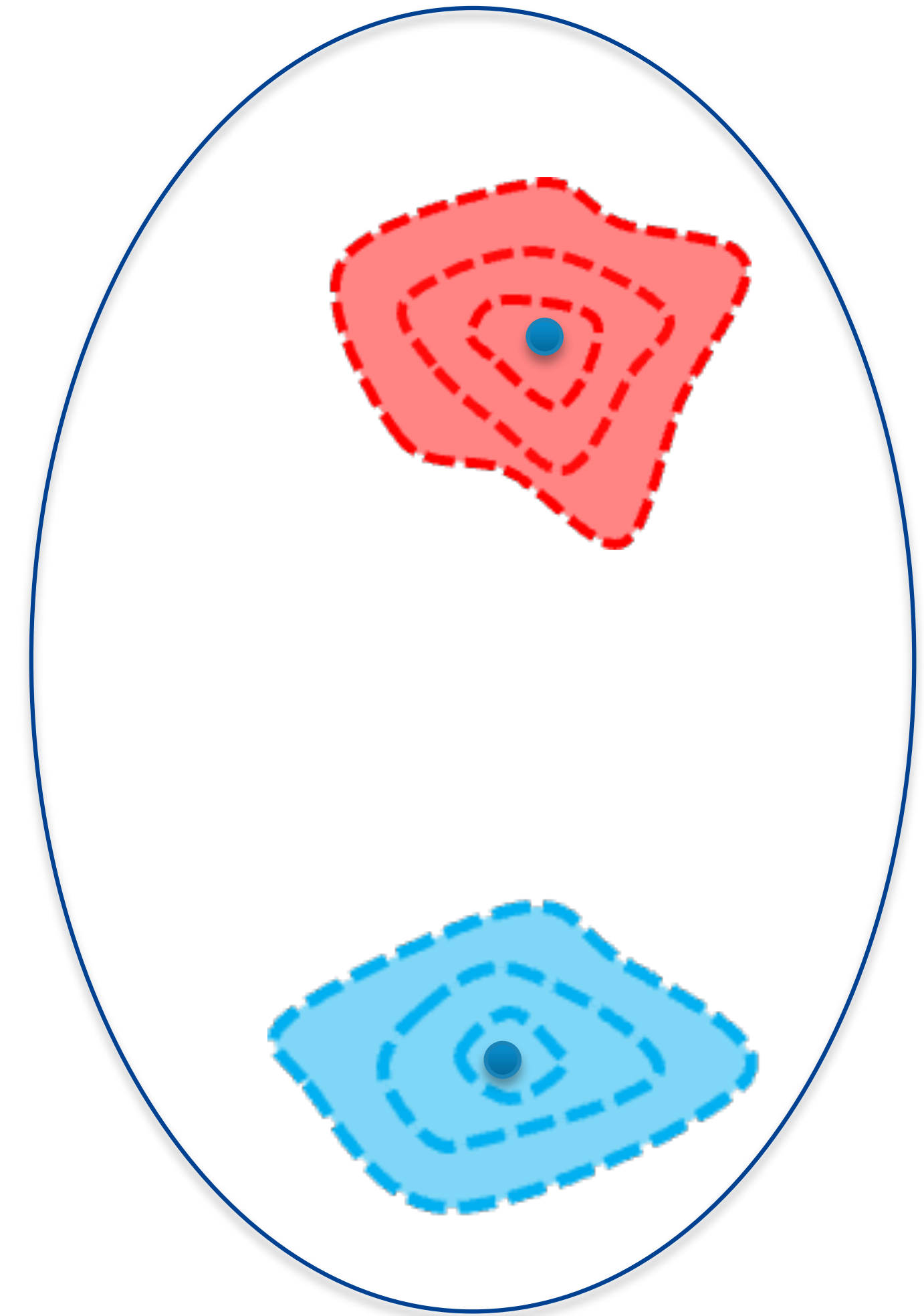


# Kernel mean embedding becomes kernel quantile embedding

- Directional quantiles (Kong & Mizera 2012)

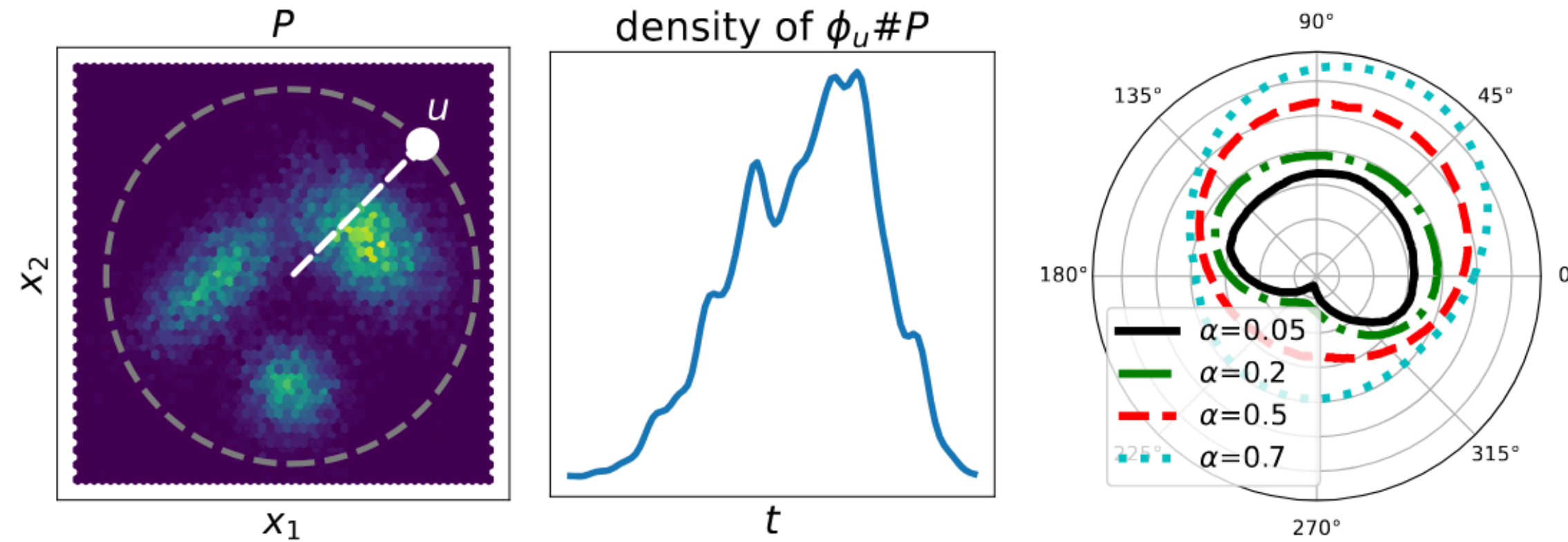


$$\rho_P^{\alpha,u} := \rho_{\phi_u \# P}^{\alpha} u, \quad \phi_u(y) = \langle u, y \rangle$$



# Kernel mean embedding becomes kernel quantile embedding

- **Directional quantiles (Kong & Mizera 2012)**



$$\rho_P^{\alpha,u} := \rho_{\phi_u \# P}^{\alpha} u, \quad \phi_u(y) = \langle u, y \rangle$$

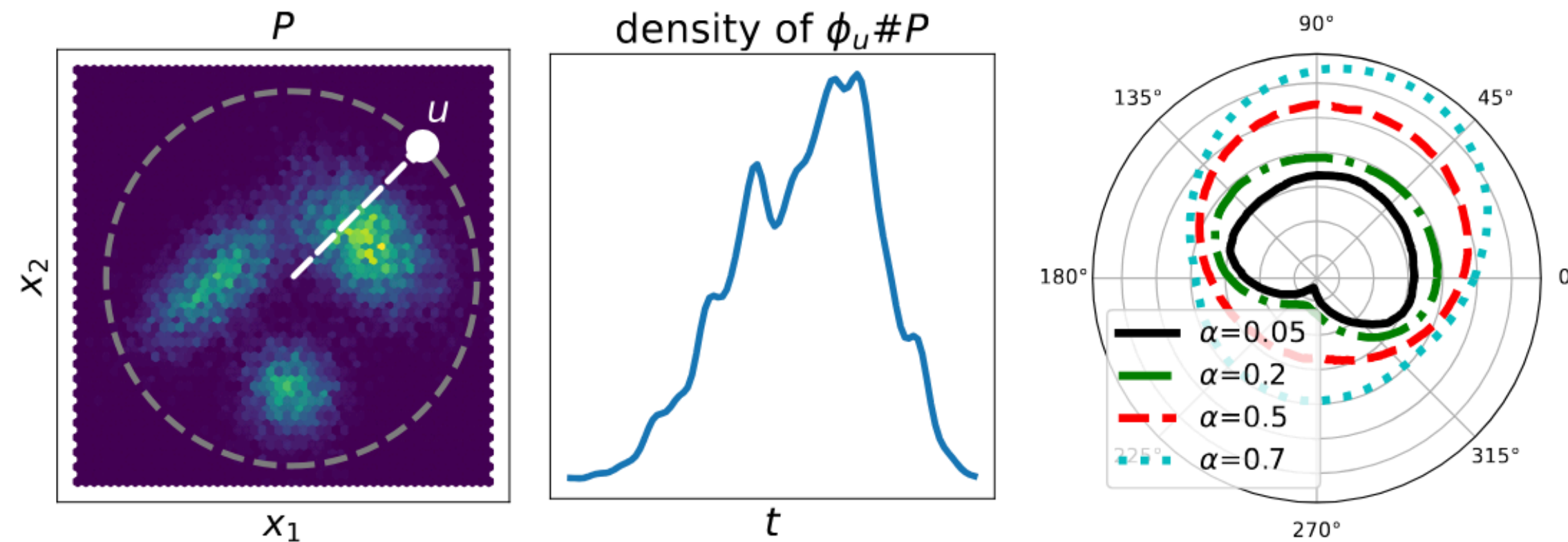
$$\tau_p(P, Q; \nu, u) = \left( \int_0^1 \left\| \rho_P^{\alpha,u} - \rho_Q^{\alpha,u} \right\|_{\mathcal{H}}^p \nu(d\alpha) \right)^{1/p}$$

$$\text{e-KQD}_p(P, Q; \nu, \gamma) = \left( \mathbb{E}_{u \sim \gamma} \left[ \tau_p^p(P, Q; \nu, u) \right] \right)^{1/p}$$



# Kernel mean embedding becomes kernel quantile embedding

- Directional quantiles (Kong & Mizera 2012)



$$\rho_P^{\alpha,u} := \rho_{\phi_u \# P}^\alpha u, \quad \phi_u(y) = \langle u, y \rangle$$

$$\tau_p(P, Q; \nu, u) = \left( \int_0^1 \left\| \rho_P^{\alpha,u} - \rho_Q^{\alpha,u} \right\|_{\mathcal{H}}^p \nu(d\alpha) \right)^{1/p}$$



$$\text{e-KQD}_p(P, Q; \nu, \gamma) = \left( \mathbb{E}_{u \sim \gamma} \left[ \tau_p^p(P, Q; \nu, u) \right] \right)^{1/p}$$

Kernel Quantile Discrepancy (KQD) - Naslidnyk et al. 2025

$\gamma, \nu$  uniform + linear kernel = SW2!

**Selected topics for future directions**

**4- Distributionally robust ML**

# Distributionally robust ML

- Optimal UQ

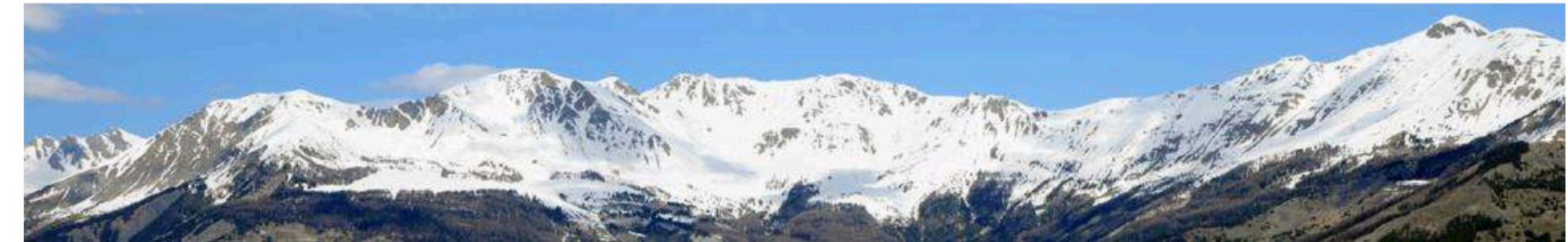
## ETICS

**École Thématique sur les Incertitudes en Calcul Scientifique**  
**Research School on Uncertainty in Scientific Computing**

June 6-10 2016

Centre de séminaire Séolane  
<http://eost.u-strasbg.fr/seolane/>

**Barcelonnette**



Talk of T. Sullivan

## ETICS 2018

**École Thématique sur les Incertitudes en Calcul Scientifique**  
**Research School on Uncertainty in Scientific Computing**



June 3-8 2018

<http://www.sb-roscoff.fr/>



Talk of M. Keller & J. Stenger



# Distributionally robust ML

- **Optimal UQ**

## Principle

Find optimal bounds for a quantity of interest  $Q(\mu^\dagger)$ , functional of an uncertain probability measure  $\mu^\dagger$ , known only to lie in some subset  $\mathcal{A}$  of  $\mathcal{M}_1(\mathcal{X})$  :

$$\underline{Q}(\mathcal{A}) \leq Q(\mu^\dagger) \leq \overline{Q}(\mathcal{A}),$$

with :

- $\underline{Q}(\mathcal{A}) = \inf_{\mu \in \mathcal{A}} Q(\mu)$
- $\overline{Q}(\mathcal{A}) = \sup_{\mu \in \mathcal{A}} Q(\mu)$
- $\mathcal{A} = \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \Phi_j(\mu) \leq c_j, j = 1, \dots, N\}$  the *admissible* subset,

## ETICS

École Thématique sur les Incertitudes en Calcul Scientifique  
Research School on Uncertainty in Scientific Computing

June 6-10 2016

Centre de séminaire Séolane  
<http://eost.u-strasbg.fr/seolane/>

Barcelonnette



Talk of T. Sullivan

## ETICS 2018

École Thématique sur les Incertitudes en Calcul Scientifique  
Research School on Uncertainty in Scientific Computing



CNRS UPMC  
Station Biologique  
Roscoff

June 3-8 2018

<http://www.sb-roscoff.fr/>



Talk of M. Keller & J. Stenger



# Distributionally robust ML

- **Optimal UQ**

## Principle

Find optimal bounds for a quantity of interest  $Q(\mu^\dagger)$ , functional of an uncertain probability measure  $\mu^\dagger$ , known only to lie in some subset  $\mathcal{A}$  of  $\mathcal{M}_1(\mathcal{X})$  :

$$\underline{Q}(\mathcal{A}) \leq Q(\mu^\dagger) \leq \overline{Q}(\mathcal{A}),$$

with :

- $\underline{Q}(\mathcal{A}) = \inf_{\mu \in \mathcal{A}} Q(\mu)$
- $\overline{Q}(\mathcal{A}) = \sup_{\mu \in \mathcal{A}} Q(\mu)$
- $\mathcal{A} = \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \Phi_j(\mu) \leq c_j, j = 1, \dots, N\}$  the *admissible* subset,

## Theorem (Measure affine functionals over generalized moment classes)

**If :**

- $Q(\mu)$  is measure affine (e.g.  $Q(\mu) := \mathbb{E}_\mu[q]$ ,  $q$  bounded above or below)
- $\mathcal{A} = \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \mathbb{E}_\mu[\varphi_j] \leq c_j, j = 1, \dots, N\}$  for measurable functions  $\varphi_j$
- $\mathcal{A}_\Delta = \{\mu \in \mathcal{A} \mid \mu = \sum_{i=1}^N w_i \delta_{x_i}\}$  extremal admissible probability measures

**Then :**

- $\underline{Q}(\mathcal{A}) = \underline{Q}(\mathcal{A}_\Delta)$  ;  $\overline{Q}(\mathcal{A}) = \overline{Q}(\mathcal{A}_\Delta)$

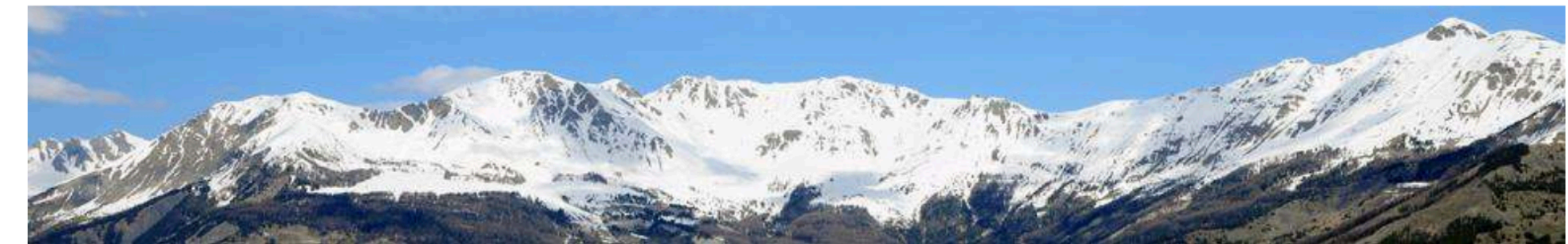
ETICS

École Thématique sur les Incertitudes en Calcul Scientifique  
Research School on Uncertainty in Scientific Computing

June 6-10 2016

Centre de séminaire Séolane  
<http://eost.u-strasbg.fr/seolane/>

Barcelonnette



Talk of T. Sullivan

ETICS 2018

École Thématique sur les Incertitudes en Calcul Scientifique  
Research School on Uncertainty in Scientific Computing



CNRS UPMC  
Station Biologique  
Roscoff

June 3-8 2018

<http://www.sb-roscoff.fr/>



Talk of M. Keller & J. Stenger



# Distributionally robust ML

## • Optimal UQ

Staib & Jegelka 2019

### Principle

Find optimal bounds for a quantity of interest  $Q(\mu^\dagger)$ , functional of an uncertain probability measure  $\mu^\dagger$ , known only to lie in some subset  $\mathcal{A}$  of  $\mathcal{M}_1(\mathcal{X})$  :

$$\underline{Q}(\mathcal{A}) \leq Q(\mu^\dagger) \leq \overline{Q}(\mathcal{A}),$$

with :

- $\underline{Q}(\mathcal{A}) = \inf_{\mu \in \mathcal{A}} Q(\mu)$
- $\overline{Q}(\mathcal{A}) = \sup_{\mu \in \mathcal{A}} Q(\mu)$
- $\mathcal{A} = \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \Phi_j(\mu) \leq c_j, j = 1, \dots, N\}$  the *admissible* subset,

$$\mathbb{E}_{x \sim \mathbb{Q}}[\ell_f(x)].$$

### Theorem (Measure affine functionals over generalized moment classes)

If :

- $Q(\mu)$  is measure affine (e.g.  $Q(\mu) := \mathbb{E}_\mu[q]$ ,  $q$  bounded above or below)
- $\mathcal{A} = \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \mathbb{E}_\mu[\varphi_j] \leq c_j, j = 1, \dots, N\}$  for measurable functions  $\varphi_j$
- $\mathcal{A}_\Delta = \{\mu \in \mathcal{A} \mid \mu = \sum_{i=1}^N w_i \delta_{x_i}\}$  extremal admissible probability measures

Then :

- $\underline{Q}(\mathcal{A}) = \underline{Q}(\mathcal{A}_\Delta) ; \quad \overline{Q}(\mathcal{A}) = \overline{Q}(\mathcal{A}_\Delta)$

# Distributionally robust ML

- **Optimal UQ**

## Principle

Find optimal bounds for a quantity of interest  $Q(\mu^\dagger)$ , functional of an uncertain probability measure  $\mu^\dagger$ , known only to lie in some subset  $\mathcal{A}$  of  $\mathcal{M}_1(\mathcal{X})$  :

$$\underline{Q}(\mathcal{A}) \leq Q(\mu^\dagger) \leq \overline{Q}(\mathcal{A}),$$

with :

- $\underline{Q}(\mathcal{A}) = \inf_{\mu \in \mathcal{A}} Q(\mu)$
- $\overline{Q}(\mathcal{A}) = \sup_{\mu \in \mathcal{A}} Q(\mu)$
- $\mathcal{A} = \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \Phi_j(\mu) \leq c_j, j = 1, \dots, N\}$  the *admissible* subset,

## Theorem (Measure affine functionals over generalized moment classes)

**If :**

- $Q(\mu)$  is measure affine (e.g.  $Q(\mu) := \mathbb{E}_\mu[q]$ ,  $q$  bounded above or below)
- $\mathcal{A} = \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \mathbb{E}_\mu[\varphi_j] \leq c_j, j = 1, \dots, N\}$  for measurable functions  $\varphi_j$
- $\mathcal{A}_\Delta = \{\mu \in \mathcal{A} \mid \mu = \sum_{i=1}^N w_i \delta_{x_i}\}$  extremal admissible probability measures

**Then :**

- $\underline{Q}(\mathcal{A}) = \underline{Q}(\mathcal{A}_\Delta)$  ;  $\overline{Q}(\mathcal{A}) = \overline{Q}(\mathcal{A}_\Delta)$

Staib & Jegelka 2019

sup

$$\mathbb{E}_{x \sim \mathbb{Q}}[\ell_f(x)].$$

# Distributionally robust ML

## • Optimal UQ

### Principle

Find optimal bounds for a quantity of interest  $Q(\mu^\dagger)$ , functional of an uncertain probability measure  $\mu^\dagger$ , known only to lie in some subset  $\mathcal{A}$  of  $\mathcal{M}_1(\mathcal{X})$  :

$$\underline{Q}(\mathcal{A}) \leq Q(\mu^\dagger) \leq \overline{Q}(\mathcal{A}),$$

with :

- $\underline{Q}(\mathcal{A}) = \inf_{\mu \in \mathcal{A}} Q(\mu)$
- $\overline{Q}(\mathcal{A}) = \sup_{\mu \in \mathcal{A}} Q(\mu)$
- $\mathcal{A} = \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \Phi_j(\mu) \leq c_j, j = 1, \dots, N\}$  the *admissible* subset,

### Theorem (Measure affine functionals over generalized moment classes)

If :

- $Q(\mu)$  is measure affine (e.g.  $Q(\mu) := \mathbb{E}_\mu[q]$ ,  $q$  bounded above or below)
- $\mathcal{A} = \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \mathbb{E}_\mu[\varphi_j] \leq c_j, j = 1, \dots, N\}$  for measurable functions  $\varphi_j$
- $\mathcal{A}_\Delta = \{\mu \in \mathcal{A} \mid \mu = \sum_{i=1}^N w_i \delta_{x_i}\}$  extremal admissible probability measures

Then :

- $\underline{Q}(\mathcal{A}) = \underline{Q}(\mathcal{A}_\Delta)$  ;  $\overline{Q}(\mathcal{A}) = \overline{Q}(\mathcal{A}_\Delta)$

Staib & Jegelka 2019

$$\sup_{\mathbb{Q}: d_{\text{MMD}}(\mathbb{Q}, \hat{\mathbb{P}}_n) \leq \epsilon} \mathbb{E}_{x \sim \mathbb{Q}}[\ell_f(x)],$$





# Conclusion

● Comparing probability distributions has been a key ingredient in many ETICS courses and talks since the beginning, sometimes hidden

➡ Were you aware of that?

● This is a very active research area in machine learning, and we should follow the current developments with care!

X <sub>1</sub>	X <sub>2</sub>	U <sub>1</sub>	Y
0.47	-1.47	red	-1.5
0.52	-0.79	green	0.20
0.11	-2.67	green	0.48
0.75	0.43	blue	1.82
0.11	1.91	red	-4.2
0.96	2.92	blue	2.34
0.64	0.33	blue	4.51
0.01	2.14	red	-3.7
0.15	1.39	green	0.86
0.63	-1.93	red	-2.9

Table 3: Original dataset.











X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	Y
0.47	-1.47		-1.5
0.52	-0.79		0.20
0.11	-2.67		0.48
0.75	0.43		1.82
0.11	1.91		-4.2
0.96	2.92		2.34
0.64	0.33		4.51
0.01	2.14		-3.7
0.15	1.39		0.86
0.63	-1.93		-2.9

Table 4: Distributional encoding.

X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	Y <sub>1</sub>	Y <sub>2</sub>
0.47	-1.47			-1.5	5.67
0.52	-0.79			0.20	-0.89
0.11	-2.67			0.48	-3.65
0.75	0.43			1.82	7.34
0.11	1.91			-4.2	6.32
0.96	2.92			2.34	4.28
0.64	0.33			4.51	10.12
0.01	2.14			-3.7	7.98
0.15	1.39			0.86	0.73
0.63	-1.93			-2.9	9.21

Table 7: Multi 1D-Distrib. encoding.











X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	Y <sub>1</sub>	Y <sub>2</sub>
0.47	-1.47		-1.5	5.67
0.52	-0.79		0.20	-0.89
0.11	-2.67		0.48	-3.65
0.75	0.43		1.82	7.34
0.11	1.91		-4.2	6.32
0.96	2.92		2.34	4.28
0.64	0.33		4.51	10.12
0.01	2.14		-3.7	7.98
0.15	1.39		0.86	0.73
0.63	-1.93		-2.9	9.21

Table 8: 2D-Distrib. encoding.