Minimax estimation on manifolds via optimal transport

Séminaire des doctorants proba-stats
3 juin 2020

Vincent Divol

DataShape - Inria Saclay

LMO - Université Paris-Sud

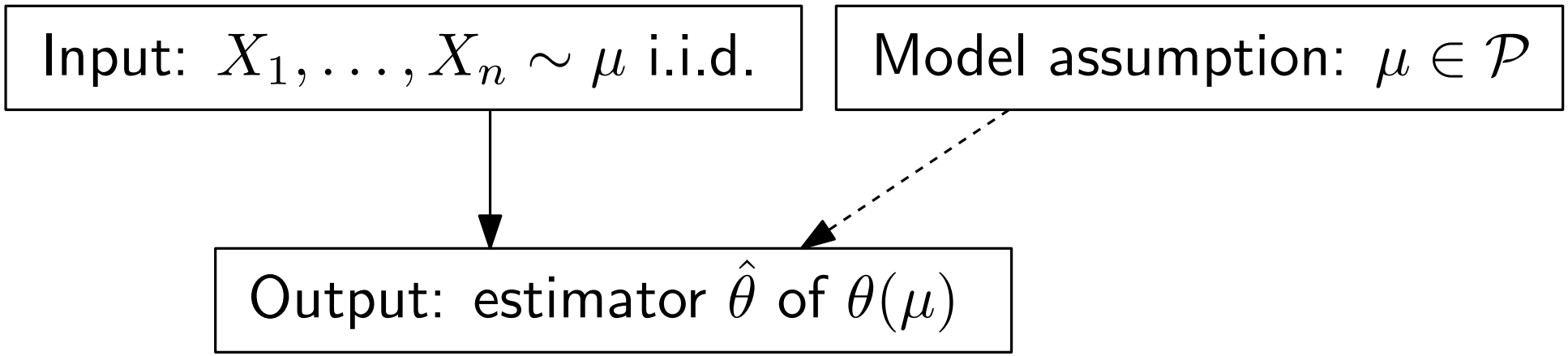
vincent.divol@inria.fr

Density estimation: the basics

A (very) concise summary:

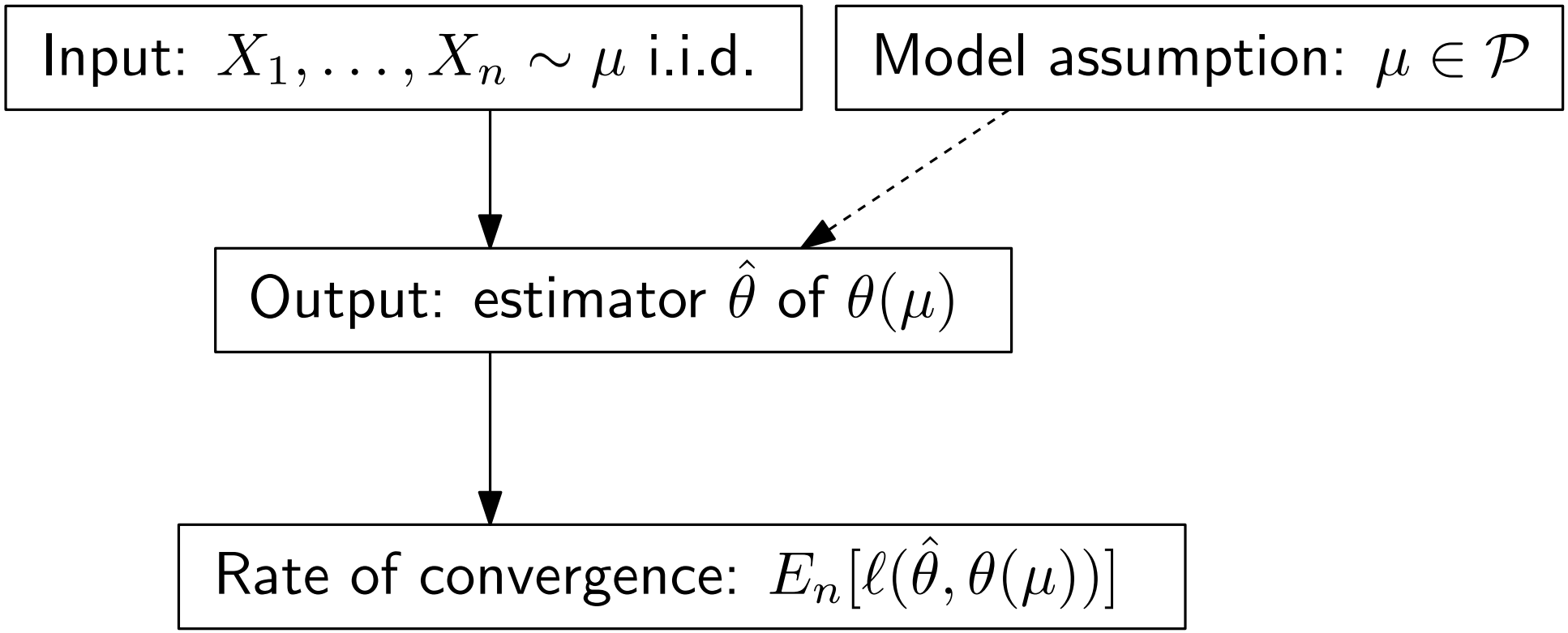
Density estimation: the basics

A (very) concise summary:



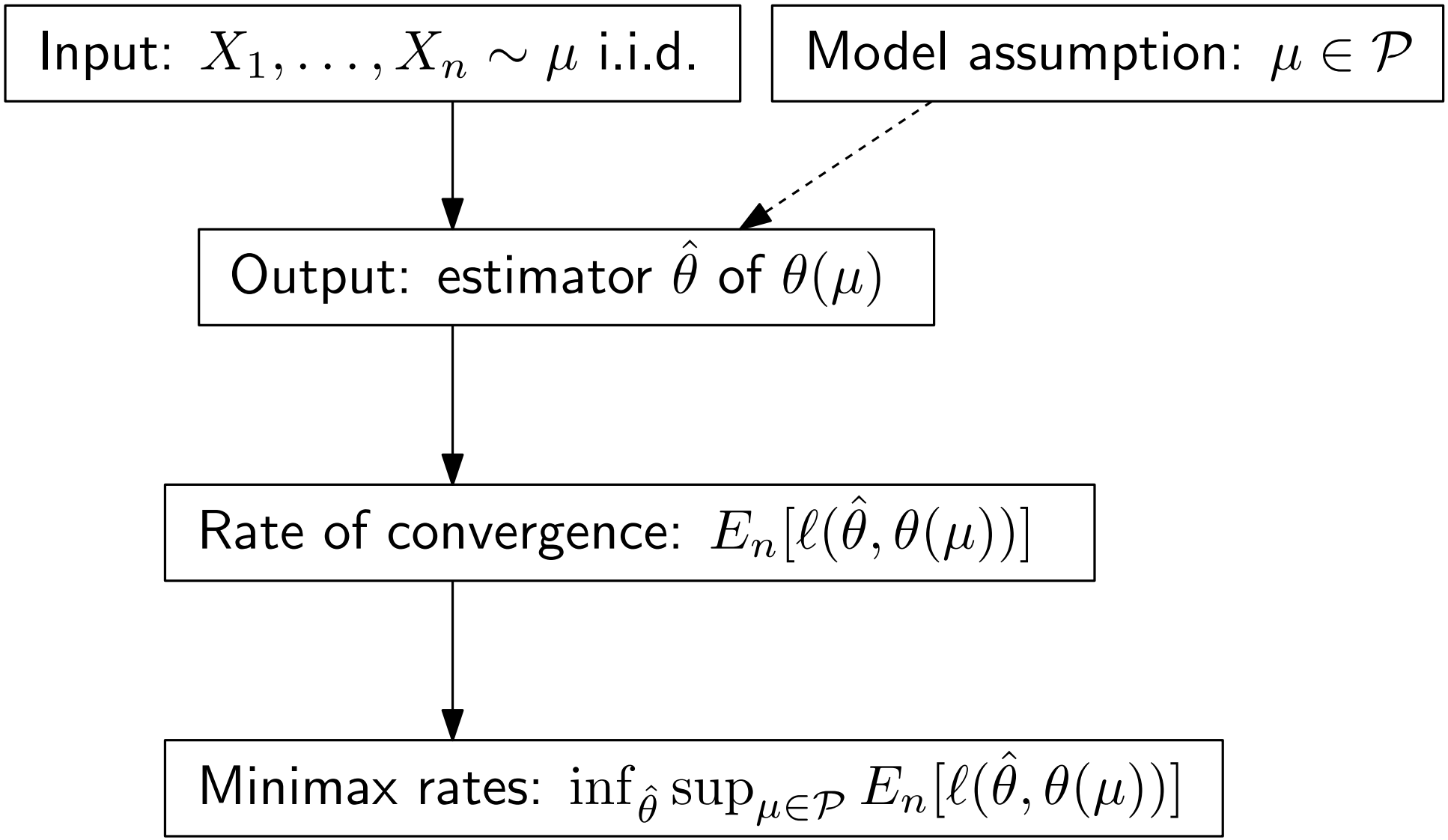
Density estimation: the basics

A (very) concise summary:



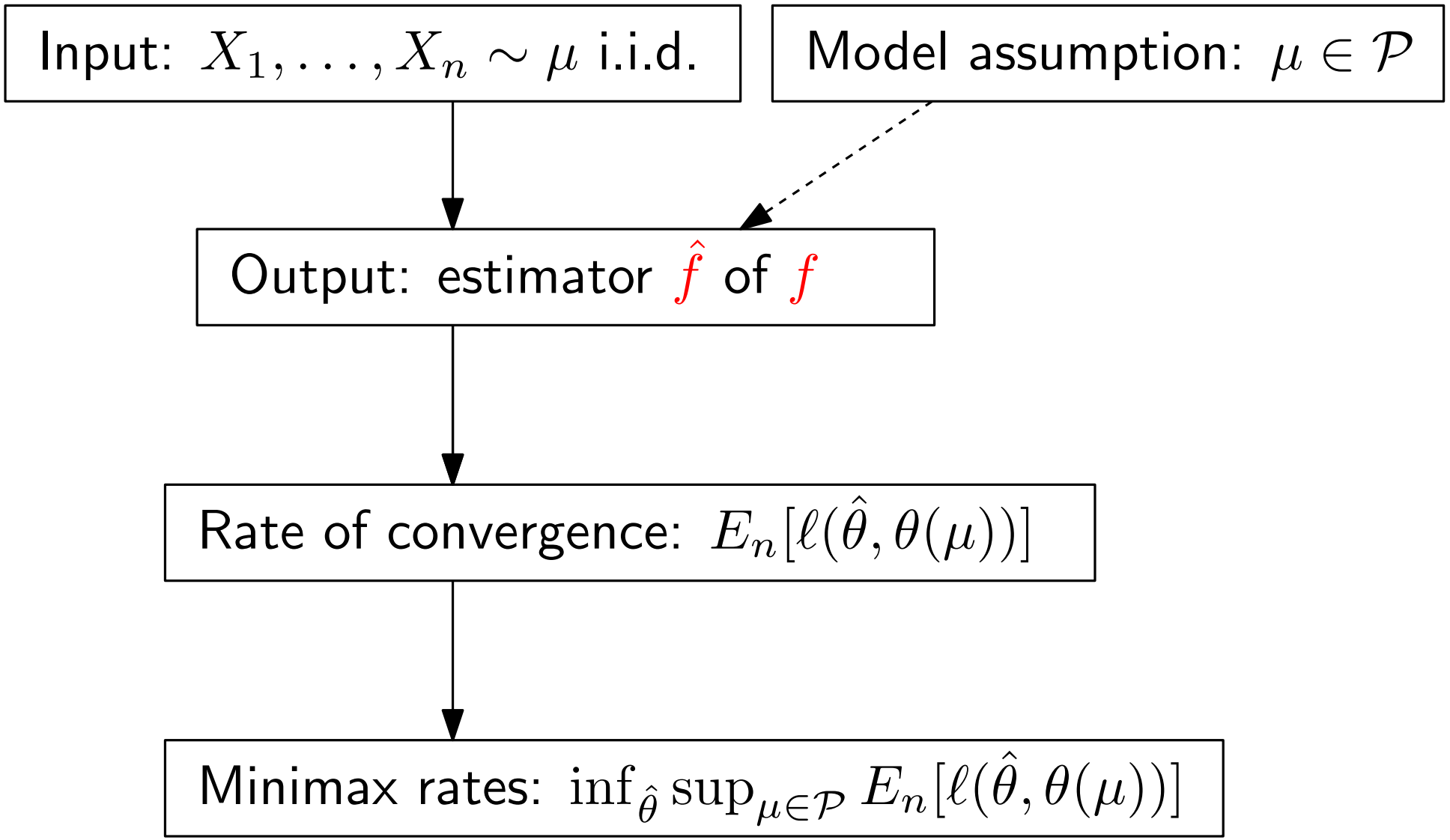
Density estimation: the basics

A (very) concise summary:



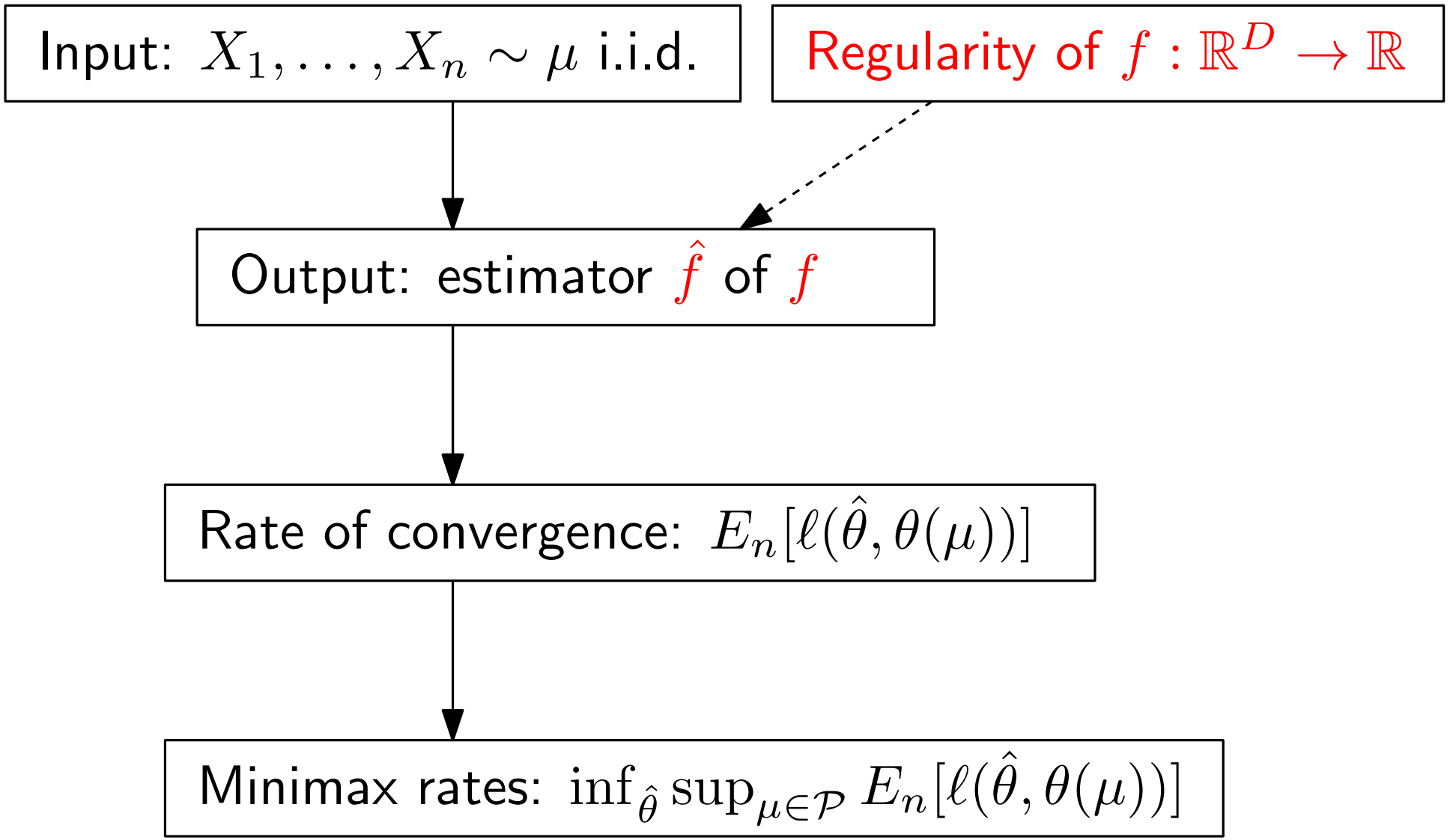
Density estimation: the basics

A (very) concise summary:



Density estimation: the basics

A (very) concise summary:



Density estimation: the basics

Basic example

- $f \in H_2^s(\mathbb{R}^D) = \{f,\ \int |\mathcal{F}(f)(\xi)|^2(1+|\xi|^2)^s d\xi < \infty\}$
- Kernel density estimator: K with Fourier transform κ

$$\hat{f}_h(x) = K_h * \mu_n(x) = \frac{1}{n}\sum_{i=1}^n h^{-D} K\left(\frac{x-X_i}{h}\right)$$

Density estimation: the basics

Basic example

- $f \in H_2^s(\mathbb{R}^D) = \{f,\ \int |\mathcal{F}(f)(\xi)|^2(1+|\xi|^2)^s d\xi < \infty\}$
- Kernel density estimator: K with Fourier transform κ

$$\hat{f}_h(x) = K_h * \mu_n(x) = \frac{1}{n}\sum_{i=1}^n h^{-D} K\left(\frac{x-X_i}{h}\right)$$

- Expected value: $f_h = E[\hat{f}_h] = K_h * f$

Density estimation: the basics

Basic example

- $f \in H_2^s(\mathbb{R}^D) = \{f,\ \int |\mathcal{F}(f)(\xi)|^2(1+|\xi|^2)^s d\xi < \infty\}$
- Kernel density estimator: K with Fourier transform κ

$$\hat{f}_h(x) = K_h * \mu_n(x) = \frac{1}{n}\sum_{i=1}^n h^{-D} K\left(\frac{x-X_i}{h}\right)$$

- Expected value: $f_h = E[\hat{f}_h] = K_h * f$
- $f(x) - \hat{f}_h(x) = f(x) - f_h(x) + f_h(x) - \hat{f}_h(x)$

Density estimation: the basics

Basic example

- $f \in H_2^s(\mathbb{R}^D) = \{f,\ \int |\mathcal{F}(f)(\xi)|^2(1+|\xi|^2)^s d\xi < \infty\}$
- Kernel density estimator: K with Fourier transform κ

$$\hat{f}_h(x) = K_h * \mu_n(x) = \frac{1}{n}\sum_{i=1}^n h^{-D} K\left(\frac{x-X_i}{h}\right)$$

- Expected value: $f_h = E[\hat{f}_h] = K_h * f$
- $f(x) - \hat{f}_h(x) = \underbrace{f(x) - f_h(x)}_{\text{biais}} + \underbrace{f_h(x) - \hat{f}_h(x)}_{\text{variance}}$

Density estimation: the basics

Basic example

- $f \in H_2^s(\mathbb{R}^D) = \{f,\ \int |\mathcal{F}(f)(\xi)|^2(1+|\xi|^2)^s d\xi < \infty\}$
- Kernel density estimator: K with Fourier transform κ

$$\hat{f}_h(x) = K_h * \mu_n(x) = \frac{1}{n}\sum_{i=1}^n h^{-D} K\left(\frac{x-X_i}{h}\right)$$

- Expected value: $f_h = E[\hat{f}_h] = K_h * f$
- $f(x) - \hat{f}_h(x) = f(x) - f_h(x) + f_h(x) - \hat{f}_h(x)$

biais variance

$$\|f - f_h\|_{L_2} = \|\mathcal{F}(f)(\kappa(h\xi) - 1)\|_{L_2}$$
$$\leq C_0 \|f\|_{H_2^s} h^s$$

Density estimation: the basics

Basic example

- $f \in H_2^s(\mathbb{R}^D) = \{f,\ \int |\mathcal{F}(f)(\xi)|^2(1+|\xi|^2)^s d\xi < \infty\}$
- Kernel density estimator: K with Fourier transform κ

$$\hat{f}_h(x) = K_h * \mu_n(x) = \frac{1}{n}\sum_{i=1}^n h^{-D} K\left(\frac{x-X_i}{h}\right)$$

- Expected value: $f_h = E[\hat{f}_h] = K_h * f$
- $f(x) - \hat{f}_h(x) = f(x) - f_h(x) + f_h(x) - \hat{f}_h(x)$

biais

variance

$$\|f - f_h\|_{L_2} = \|\mathcal{F}(f)(\kappa(h\xi) - 1)\|_{L_2} \leq C_0 \|f\|_{H_2^s} h^s$$

$$E\|f_h - \hat{f}_h\|_{L_2} \leq \frac{C_1}{\sqrt{nh^D}}$$

Density estimation: the basics

Basic example

- $f \in H_2^s(\mathbb{R}^D) = \{f,\ \int |\mathcal{F}(f)(\xi)|^2(1+|\xi|^2)^s d\xi < \infty\}$
- Kernel density estimator: K with Fourier transform κ

$$\hat{f}_h(x) = K_h * \mu_n(x) = \frac{1}{n}\sum_{i=1}^n h^{-D} K\left(\frac{x-X_i}{h}\right)$$

- Expected value: $f_h = E[\hat{f}_h] = K_h * f$
- $f(x) - \hat{f}_h(x) = f(x) - f_h(x) + f_h(x) - \hat{f}_h(x)$

biais

variance

$$\|f - f_h\|_{L_2} = \|\mathcal{F}(f)(\kappa(h\xi)-1)\|_{L_2} \leq C_0 \|f\|_{H_2^s} h^s$$

$$E\|f_h - \hat{f}_h\|_{L_2} \leq \frac{C_1}{\sqrt{nh^D}}$$

- Choose $h \sim n^{-1/(2s+D)} \rightarrow$ Risk $\sim n^{-s/(2s+D)}$.

Density estimation: the basics

Basic example

- $f \in H_2^s(\mathbb{R}^D) = \{f,\ \int |\mathcal{F}(f)(\xi)|^2(1+|\xi|^2)^s d\xi < \infty\}$
- Kernel density estimator: K with Fourier transform κ

$$\hat{f}_h(x) = K_h * \mu_n(x) = \frac{1}{n}\sum_{i=1}^n h^{-D} K\left(\frac{x-X_i}{h}\right)$$

- Expected value: $f_h = E[\hat{f}_h] = K_h * f$
- $f(x) - \hat{f}_h(x) = f(x) - f_h(x) + f_h(x) - \hat{f}_h(x)$

biais

variance

$$\|f - f_h\|_{L_2} = \|\mathcal{F}(f)(\kappa(h\xi)-1)\|_{L_2} \leq C_0 \|f\|_{H_2^s} h^s$$

$$E\|f_h - \hat{f}_h\|_{L_2} \leq \frac{C_1}{\sqrt{nh^D}}$$

- Choose $h \sim n^{-1/(2s+D)} \rightarrow$ Risk $\sim n^{-s/(2s+D)}$.

Curse of dimensionality (and what to do about it)

Structural assumption on the signal

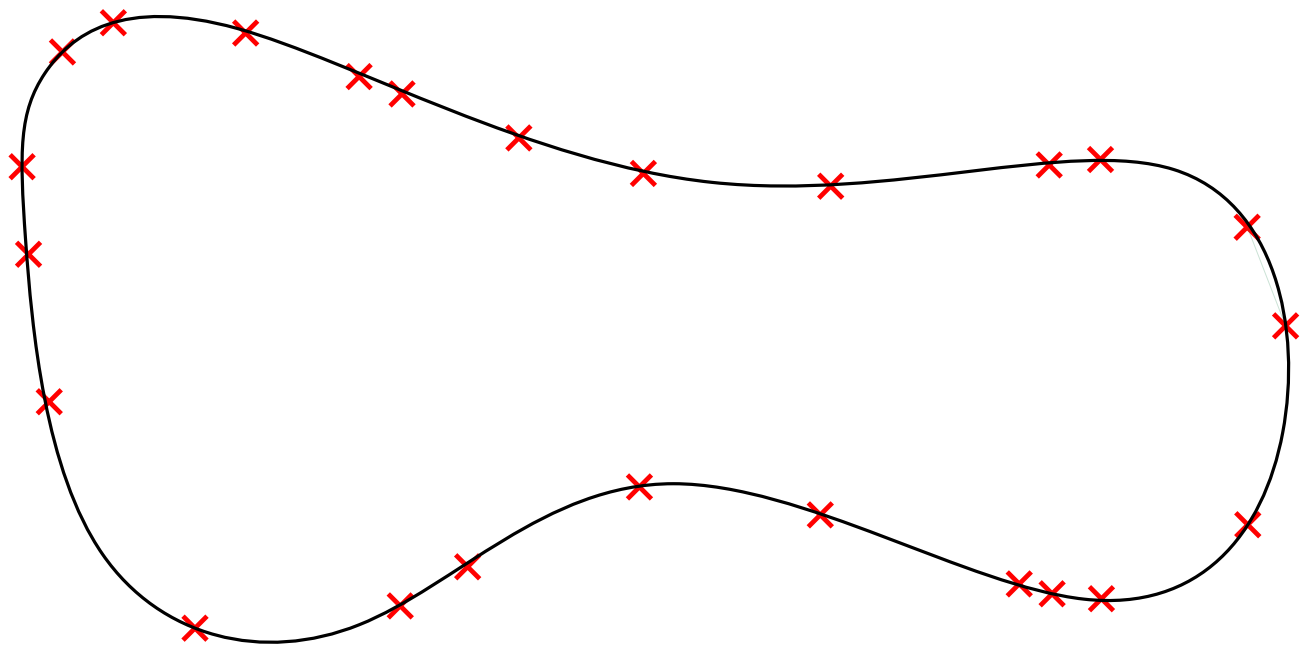
- Multi-index model, sparsity, small dimensional support, etc.

Curse of dimensionality (and what to do about it)

Structural assumption on the signal

- Multi-index model, sparsity, small dimensional support, etc.

  $\rightarrow$ Manifold assumption: μ on $\mathbb{R}^D$ is supported on some unknown manifold M of dimension $d \ll D$.

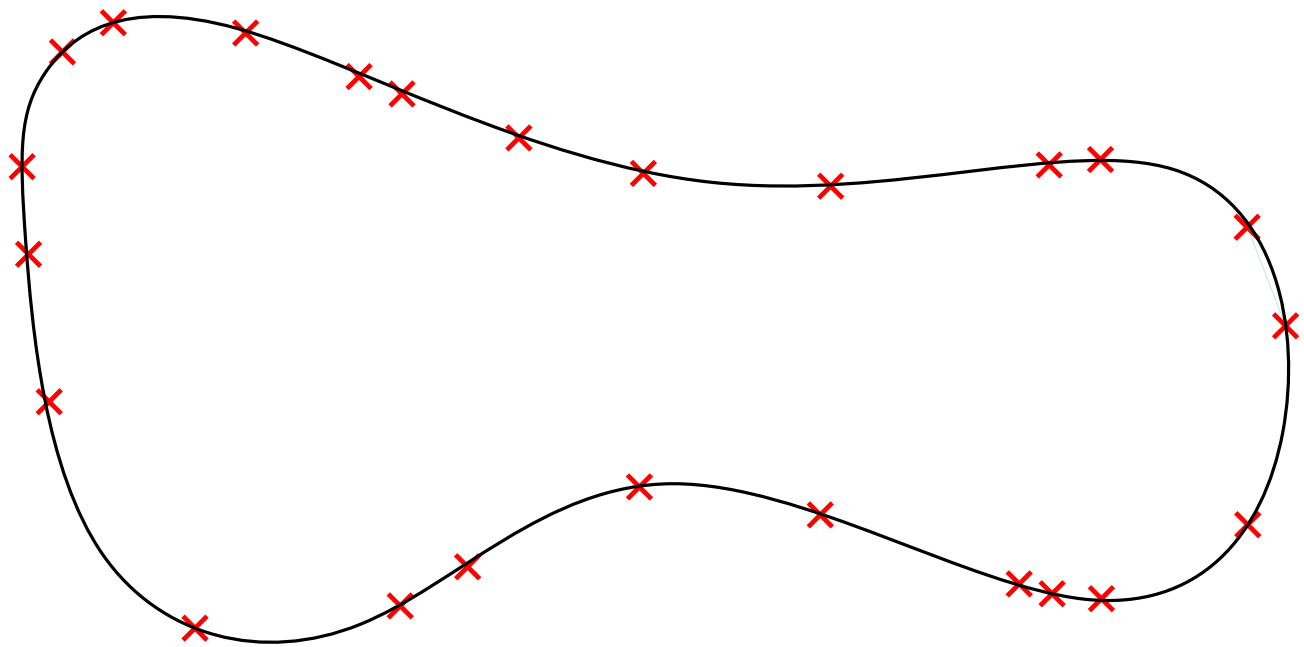


Curse of dimensionality (and what to do about it)

Structural assumption on the signal

- Multi-index model, sparsity, small dimensional support, etc.

  $\rightarrow$ Manifold assumption: μ on $\mathbb{R}^D$ is supported on some unknown manifold M of dimension $d \ll D$.

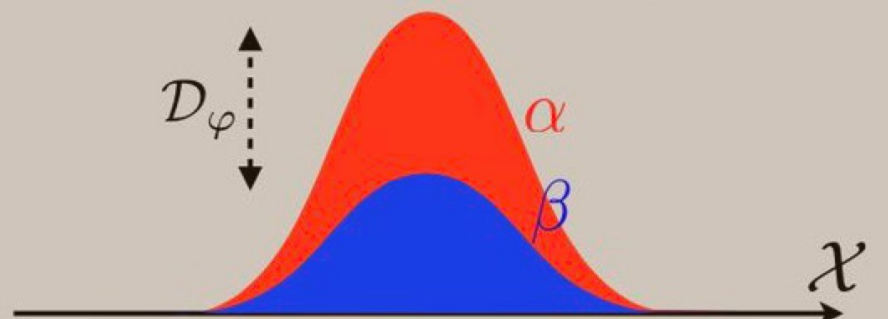


Loss?

Distances between probability measures

Csiszár divergences:

$$\mathcal{D}_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\beta$$



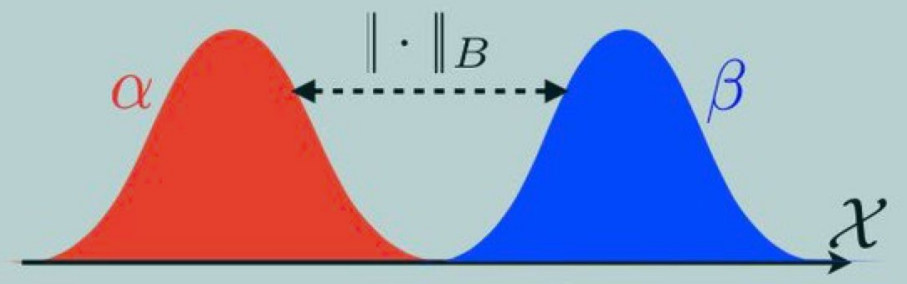
Strong topology

φ convex, $\varphi(1) = 0$.

$$\varphi(r) = \begin{cases} r\log(r) & \to \text{KL} \\ |r-1| & \to \text{TV} \\ |\sqrt{r}-1|^2 & \to \text{Hellinger} \\ \dots \end{cases}$$

Dual norms:

$$\|\alpha - \beta\|_B \overset{\text{def.}}{=} \max_{f \in B} \int_{\mathcal{X}} f(x)(\mathrm{d}\alpha(x) - \mathrm{d}\beta(x))$$



Weak topology

$$B = \{f \,;\, \|f\|_B \leqslant 1\}$$

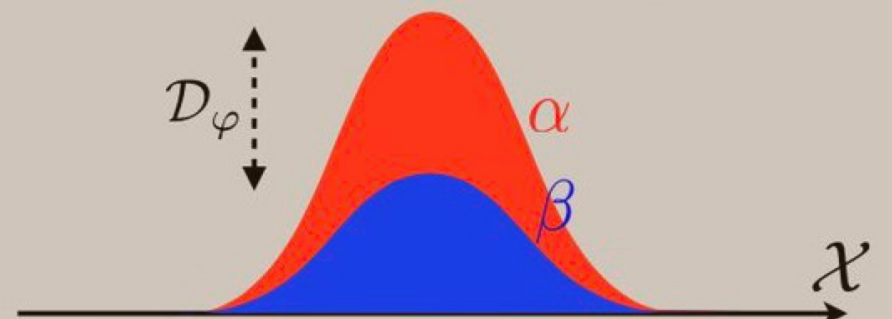
$$\|f\|_B = \begin{cases} \|\nabla f\|_\infty & \to W_1 \\ \|f\|_\infty + \|\nabla f\|_\infty & \to \text{flat} \\ \|\nabla^k f\|_2 & \to \text{MMD} \\ \dots \end{cases}$$

[Source: Gabriel Peyré]

Distances between probability measures

Csiszár divergences:

$$\mathcal{D}_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\beta$$



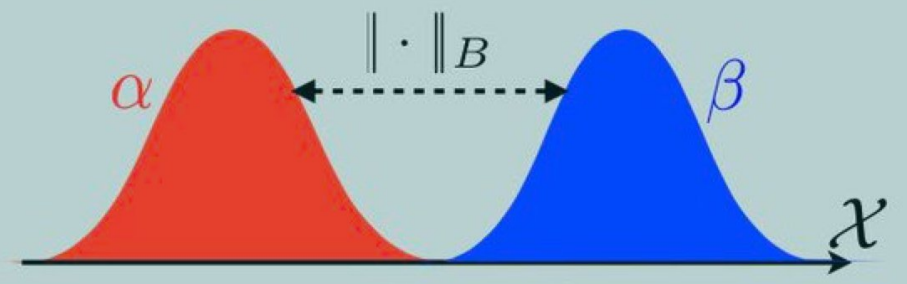
Strong topology

φ convex, $\varphi(1) = 0$.

$$\varphi(r) = \begin{cases} r\log(r) & \to \text{KL} \\ |r-1| & \to \text{TV} \\ |\sqrt{r}-1|^2 & \to \text{Hellinger} \\ \dots \end{cases}$$

Dual norms:

$$\|\alpha - \beta\|_B \overset{\text{def.}}{=} \max_{f \in B} \int_{\mathcal{X}} f(x)(\mathrm{d}\alpha(x) - \mathrm{d}\beta(x))$$



Weak topology

$$B = \{f \; ; \; \|f\|_B \leqslant 1\}$$

$$\|f\|_B = \begin{cases} \|\nabla f\|_\infty & \to W_1 \\ \|f\|_\infty + \|\nabla f\|_\infty & \to \text{flat} \\ \|\nabla^k f\|_2 & \to \text{MMD} \\ \dots \end{cases}$$

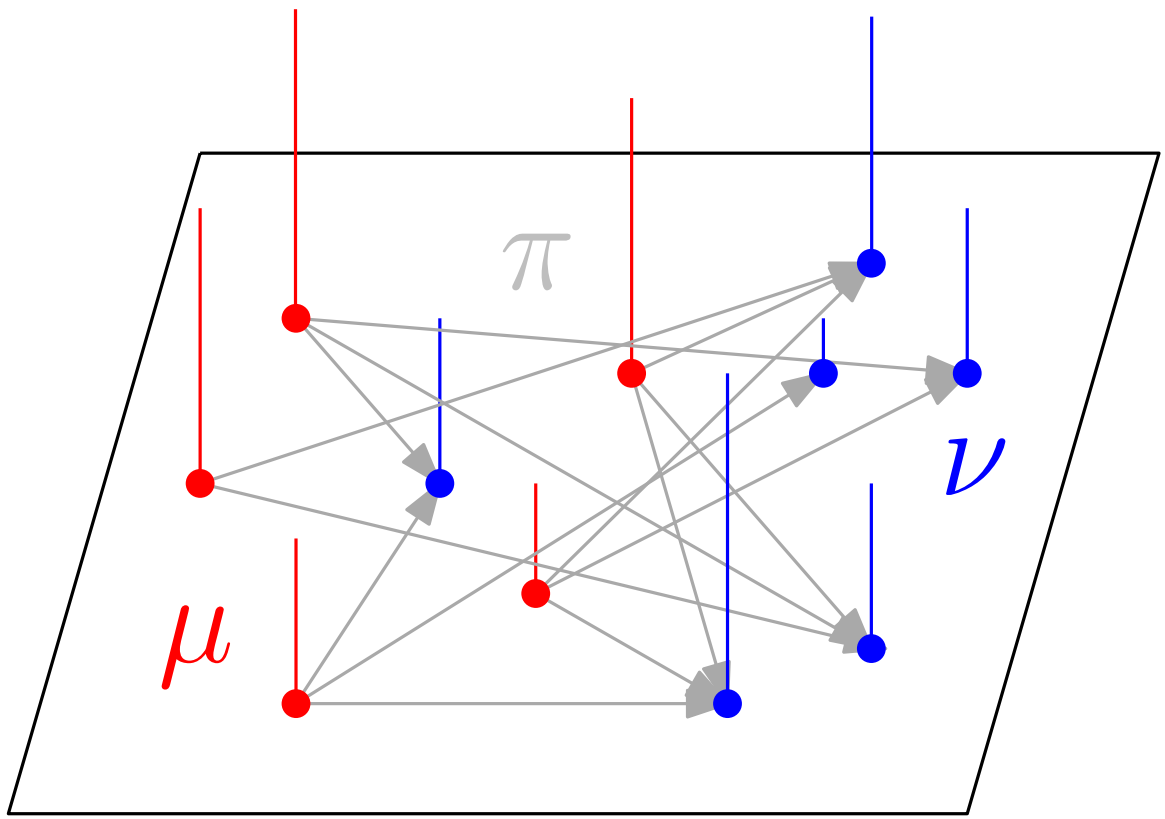
[Source: Gabriel Peyré]

Distances between probability measures

Wasserstein distances:

- Let μ, ν be probability measures on $\mathbb{R}^D$.
- Let $\Pi(\mu, \nu)$ be the set of couplings between μ and ν.

$$W_2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}^D \times \mathbb{R}^D} |x - y|^2 d\pi(x, y) \right)^{1/2}$$



Distances between probability measures

Wasserstein distances:

- Let μ, ν be probability measures on $\mathbb{R}^D$.
- Let $\Pi(\mu, \nu)$ be the set of couplings between μ and ν.

$$W_2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}^D \times \mathbb{R}^D} |x - y|^2 d\pi(x, y) \right)^{1/2}$$

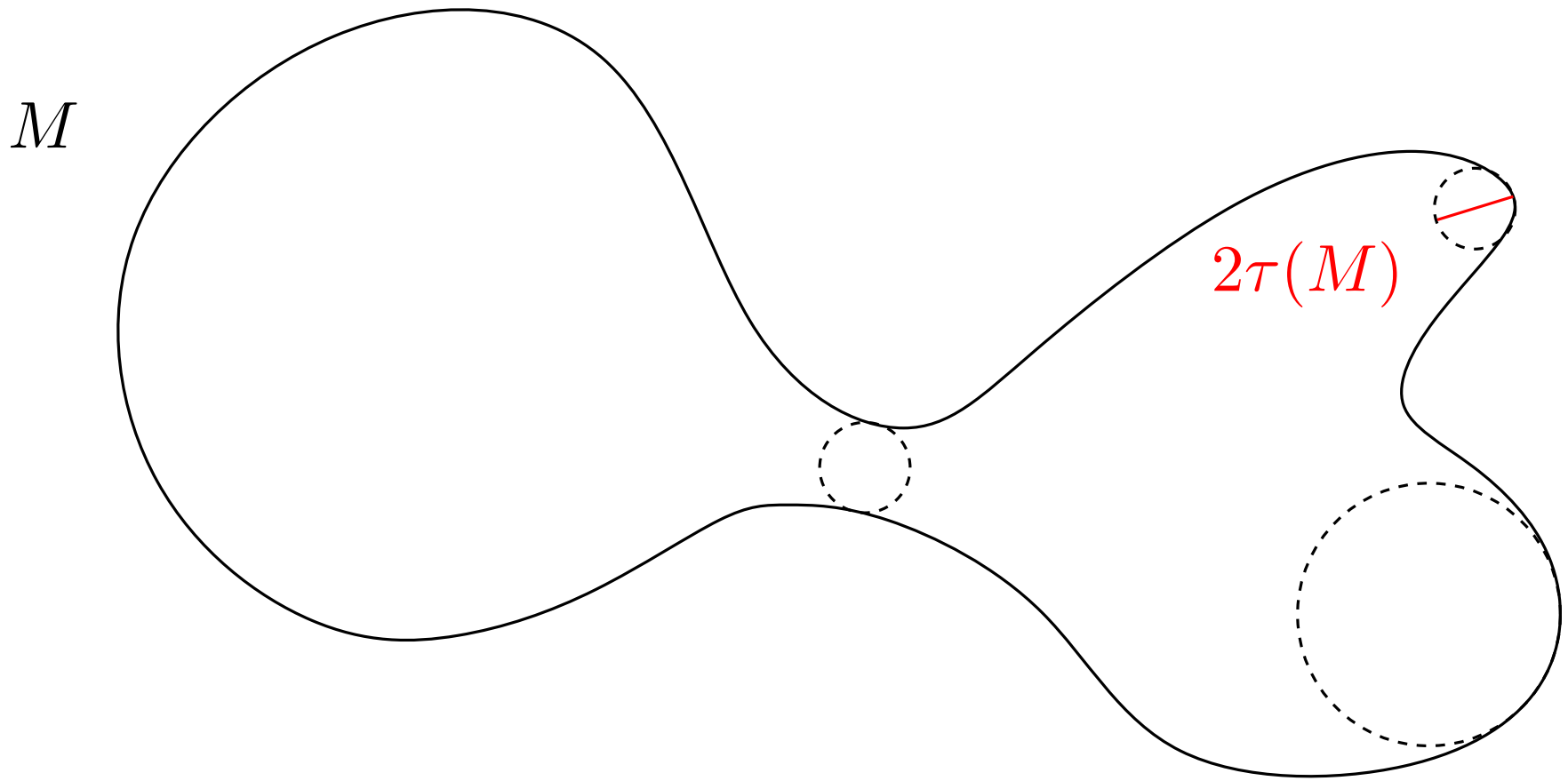
Key observation:

If μ, ν have densities f, g on a manifold M with $f, g \geq a$, then $$W_2(\mu, \nu) \leq a^{-1/2} \|f - g\|_{H_2^{-1}(M)}$$

- In $1D$: distance between c.d.f.
- For $p = 1$, $W_1(\mu, \nu) = \sup\{(\mu - \nu)(f),\ f\ 1 - \text{Lip}\}$.

Model assumptions

- $M \subset \mathbb{R}^D$ is a d-dimensional with reach $\tau(M) \geq \tau_{\min}$.
- μ has a density f on M in $H_2^s(M)$ with $f \geq a > 0$.



Model assumptions

- $M \subset \mathbb{R}^D$ is a d-dimensional with reach $\tau(M) \geq \tau_{\min}$.
- μ has a density f on M in $H_2^s(M)$ with $f \geq a > 0$.

Goal:

Given $X_1, \ldots, X_n \sim \mu$, produce an estimator $\hat{\mu}$ of μ with $E_n W_2(\hat{\mu}, \mu)$ small.

Model assumptions

- $M \subset \mathbb{R}^D$ is a d-dimensional with reach $\tau(M) \geq \tau_{\min}$.
- μ has a density f on M in $H_2^s(M)$ with $f \geq a > 0$.

Goal:

Given $X_1, \ldots, X_n \sim \mu$, produce an estimator $\hat{\mu}$ of μ with $E_n W_2(\hat{\mu}, \mu)$ small.

- Kernel estimators still work! Let $\hat{\mu}_h$ have a density $\hat{f}_h$ on M, given by

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n h^{-d} K\left(\frac{x - X_i}{h}\right) \quad \text{for } x \in M$$

Model assumptions

Bias	Variance
$\|f - f_h\|_{H_2^{-1}(M)} \leq Ch^{s+1}$	$E\|\hat{f}_h - f_h\|_{H_2^{-1}(M)} \leq \frac{C_1 h}{\sqrt{nh^d}}$

- Choose $h \sim n^{-1/(2s+d)}$

$$E_n W_2(\hat{\mu}_h, \mu) \lesssim n^{-(s+1)/(2s+d)}$$

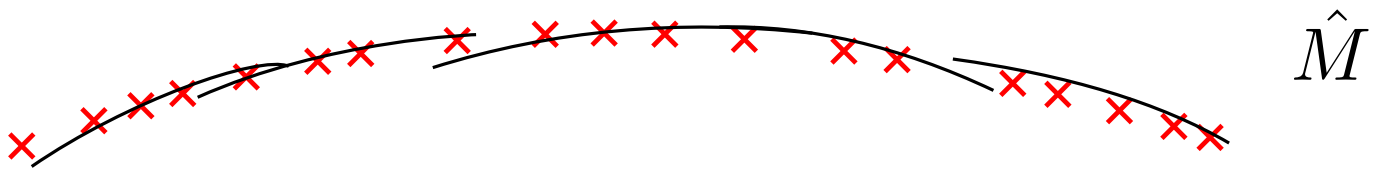
Model assumptions

Bias	Variance
$\|f - f_h\|_{H_2^{-1}(M)} \leq Ch^{s+1}$	$E\|\hat{f}_h - f_h\|_{H_2^{-1}(M)} \leq \frac{C_1 h}{\sqrt{nh^d}}$

- Choose $h \sim n^{-1/(2s+d)}$

$$E_n W_2(\hat{\mu}_h, \mu) \lesssim n^{-(s+1)/(2s+d)}$$

- $\hat{\mu}_h$ is supported on M... which is unknown!
  $\rightarrow$ Use a manifold estimator $\hat{M}$ in a preprocessing step



Take-home message

- To avoid the curse of dimensionality, a structural assumption has to be made on the signal to be estimate.
- Signal on d-manifold $M \subset \mathbb{R}^D \rightarrow$ rates of estimation depending on d and not D.

  $\rightarrow$ via kernel estimation on an estimated manifold