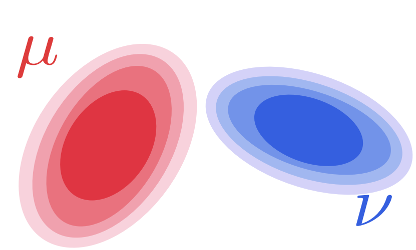
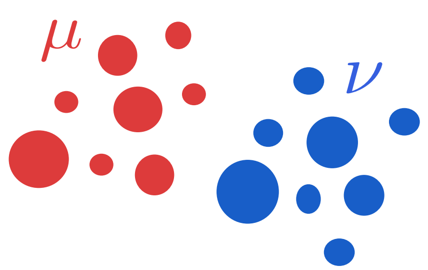
Controlling Wasserstein distances by maximum mean discrepencies with applications to compressive statistical learning

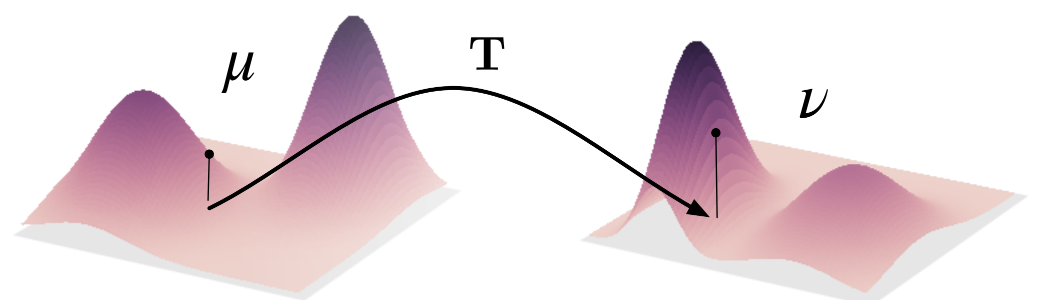
Titouan Vayer

Rémi Gribonval

From Optimal Transport to Maximum Mean Discrepancy

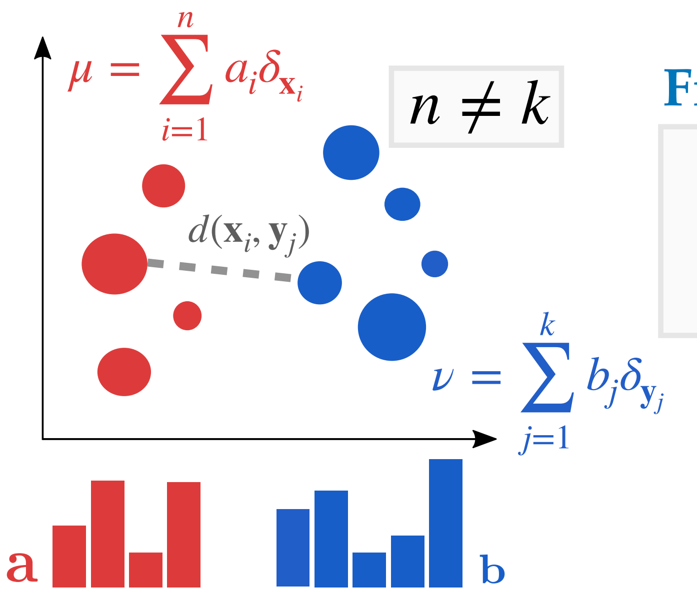






From Wasserstein to MMD

✦ Classical optimal transport (in a nutshell)

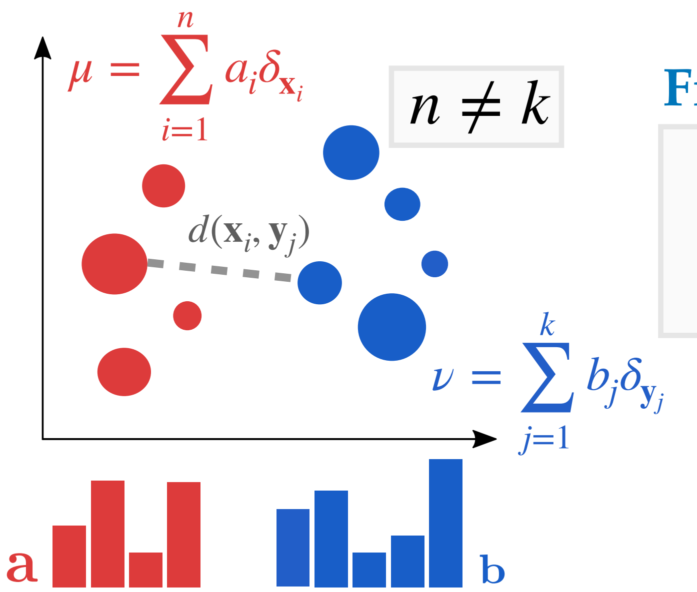


Find plan $T \in \mathbb{R}_+^{n \times k}$

$$\min_T \sum_{ij} d(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$

From Wasserstein to MMD

✦ Classical optimal transport (in a nutshell)



Find plan $T \in \mathbb{R}_+^{n \times k}$

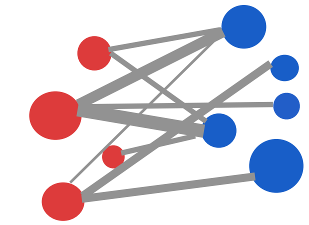
$$\min_T \sum_{ij} d(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$

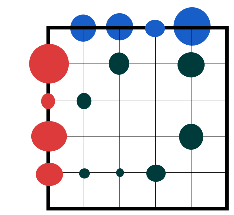
which constraints ?

Coupling $\Pi(a, b)$

$$T 1_k = a$$

$$T^\top 1_n = b$$

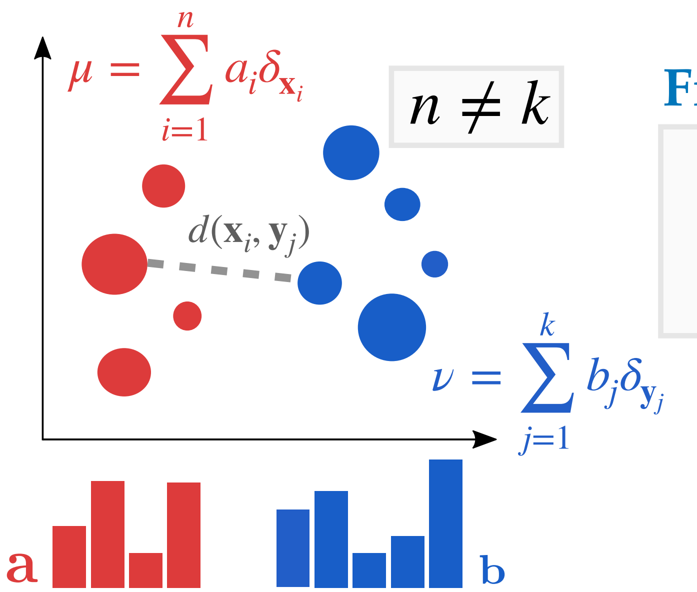




$T \in [0,1]^{n \times k}$

From Wasserstein to MMD

✦ Classical optimal transport (in a nutshell)



$n \neq k$

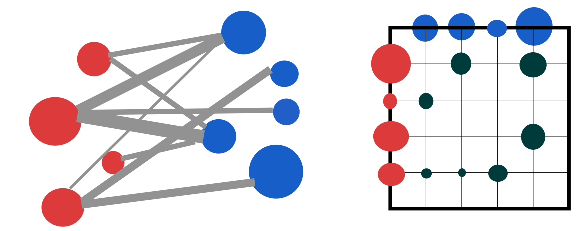
Find plan $T \in \mathbb{R}_+^{n \times k}$

$$\min_T \sum_{ij} d(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$

which constraints ?

Coupling $\Pi(a, b)$

$$T 1_k = a \qquad T^\top 1_n = b$$



$T \in [0,1]^{n \times k}$

✦ Wasserstein distance

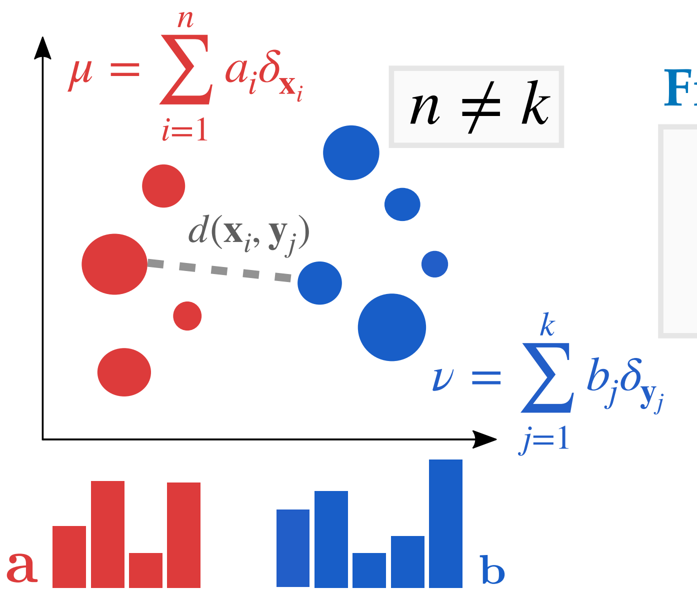
$\mu \in \mathscr{P}(X)$
$\nu \in \mathscr{P}(X)$

$$\mathrm{W}_p(\mu, \nu) = \left(\min_T \int_{X \times X} d(x, y)^p \, \mathrm{d}T(x, y) \right)^{1/p}$$

- It is always **well-defined**
- It is a proper distance on $\mathscr{P}(X)$
- Lifts the geometry of $X \to \mathscr{P}(X)$

From Wasserstein to MMD

✦ Classical optimal transport (in a nutshell)



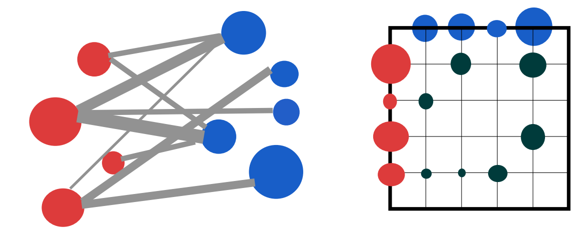
Find plan $T \in \mathbb{R}_+^{n \times k}$

$$\min_T \sum_{ij} d(\mathbf{x}_i, \mathbf{y}_j) T_{ij}$$

which constraints ?

Coupling $\Pi(a, b)$

$$T 1_k = a \qquad T^\top 1_n = b$$



$T \in [0,1]^{n \times k}$

✦ Wasserstein distance

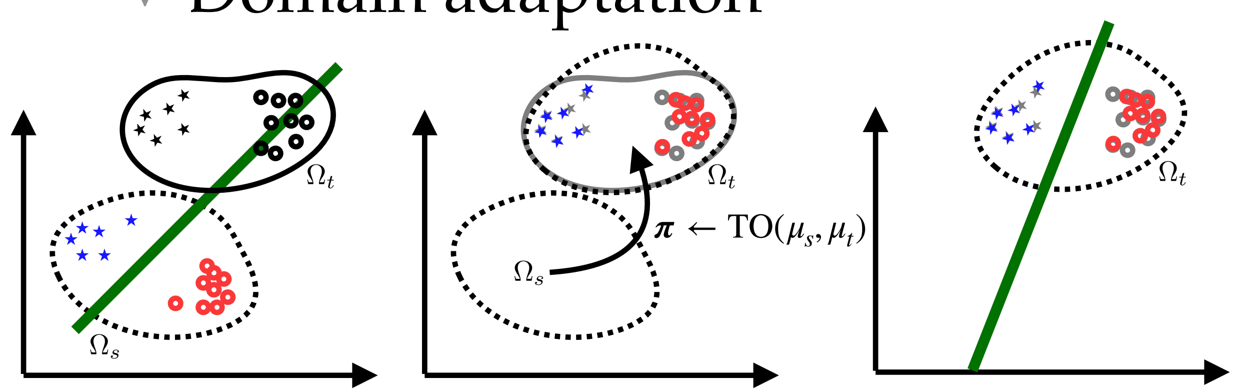
$\mu \in \mathscr{P}(X)$
$\nu \in \mathscr{P}(X)$

$$\mathrm{W}_p(\mu, \nu) = \left(\min_T \int_{X \times X} d(x, y)^p \, \mathrm{d}T(x, y) \right)^{1/p}$$

- It is always **well-defined**
- It is a proper distance on $\mathscr{P}(X)$
- Lifts the geometry of $X \to \mathscr{P}(X)$

✦ In machine learning

- Domain adaptation



- Generative modeling
- Analysis of NN convergence
- ML on graphs, fairness
- And many other …

From Wasserstein to MMD

✦ Kernel theory (in a nutshell)

$\kappa : X \times X \to \mathbb{C}$ a PSD kernel

From Wasserstein to MMD

✦ Kernel theory (in a nutshell)

$\kappa : X \times X \to \mathbb{C}$ a PSD kernel

$\forall x, y\ \kappa(x, y) = \overline{\kappa(y, x)}$

$\forall (x_1, \cdots, x_n),\ K = [\kappa(x_i, x_j)]_{ij}$ is PSD

From Wasserstein to MMD

✦ **Kernel theory (in a nutshell)**

H_κ **the RKHS of** κ

$\kappa : X \times X \to \mathbb{C}$ a PSD kernel $\Longleftrightarrow$ A **Hilbert** space of **functions** from $X \to \mathbb{C}$

From Wasserstein to MMD

✦ **Kernel theory (in a nutshell)**

H_κ **the RKHS of** κ

$\kappa : X \times X \to \mathbb{C}$ a PSD kernel $\iff$ A **Hilbert** space of **functions** from $X \to \mathbb{C}$

- $\forall x,\ \kappa(\,\cdot\,, x) \in H_\kappa$
- $f \in H_\kappa \implies \forall x,\ f(x) = \langle \kappa(\,\cdot\,, x), f \rangle_{H_\kappa}$

$\kappa(x, y) = \langle \kappa(\,\cdot\,, x), \kappa(\,\cdot\,, y) \rangle_{H_\kappa}$

From Wasserstein to MMD

✦ **Kernel theory (in a nutshell)**

H_κ **the RKHS of** κ

$\kappa : X \times X \to \mathbb{C}$ a PSD kernel $\Longleftrightarrow$ A **Hilbert** space of **functions** from $X \to \mathbb{C}$

✦ **Translation invariant kernels**

$$X = \mathbb{R}^d \quad \kappa(x, y) = \kappa_0(x - y)$$

From Wasserstein to MMD

✦ **Kernel theory (in a nutshell)**

H_κ **the RKHS of** κ

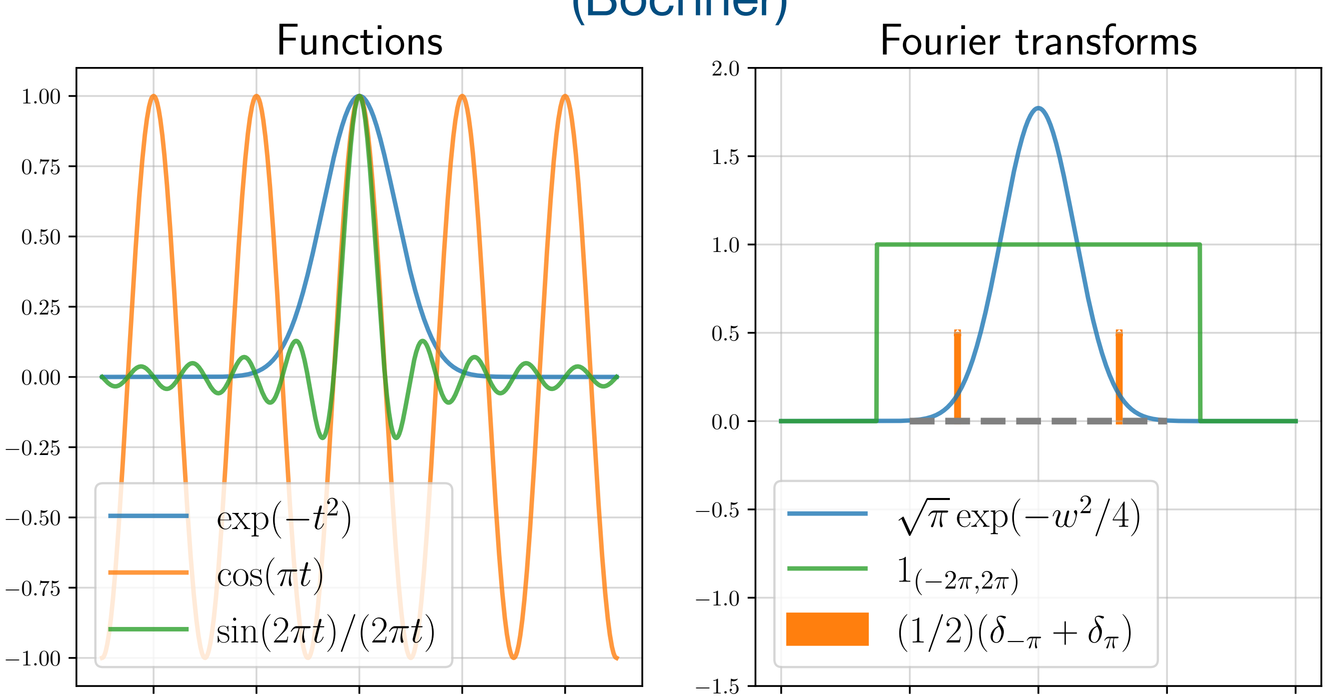
$\kappa : X \times X \to \mathbb{C}$ a PSD kernel $\Longleftrightarrow$ A **Hilbert** space of **functions** from $X \to \mathbb{C}$

✦ **Translation invariant kernels**

$$X = \mathbb{R}^d \quad \kappa(x, y) = \kappa_0(x - y)$$

✦ Is a PSD kernel $\Longleftrightarrow$ $\forall \omega, \widehat{\kappa_0}(\omega) \geq 0$

(Bochner)



From Wasserstein to MMD

✦ Kernel theory (in a nutshell)

H_κ **the RKHS of** κ

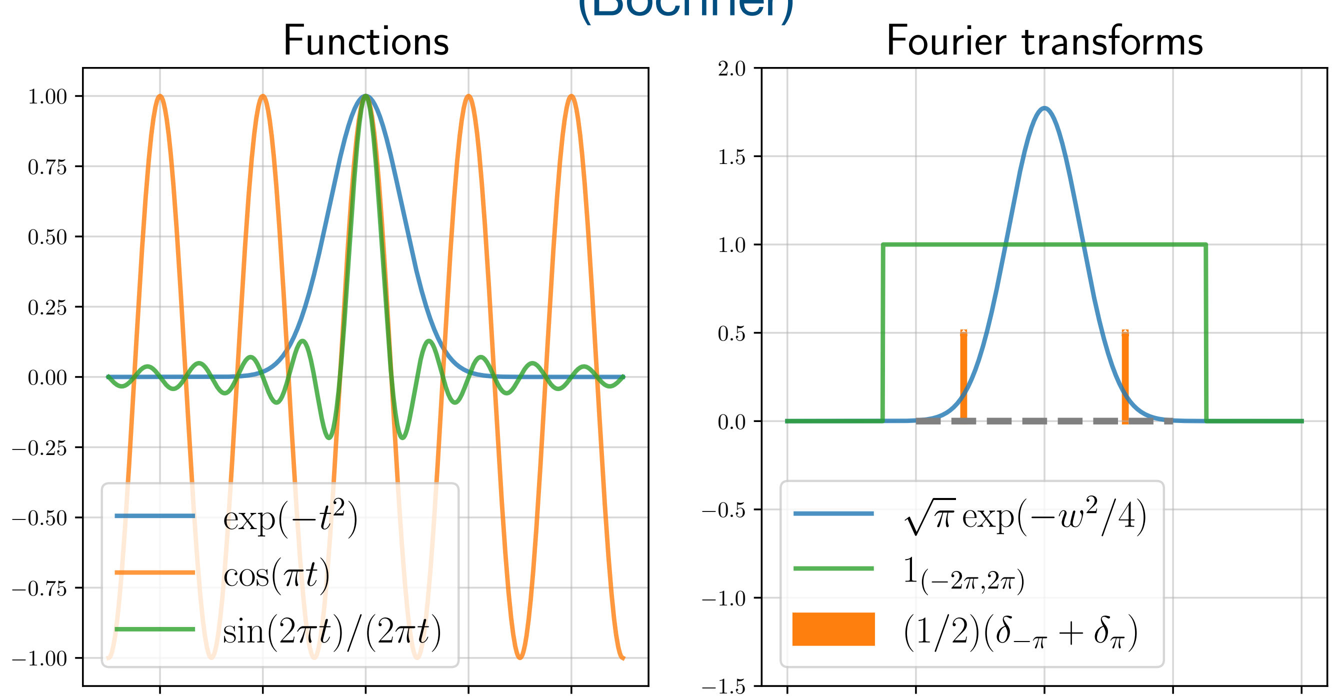
$\kappa : X \times X \to \mathbb{C}$ a PSD kernel $\iff$ A **Hilbert** space of **functions** from $X \to \mathbb{C}$

✦ Translation invariant kernels

$$X = \mathbb{R}^d \quad \kappa(x, y) = \kappa_0(x - y)$$

✦ Is a PSD kernel $\iff$ $\forall \omega, \widehat{\kappa_0}(\omega) \geq 0$

(Bochner)



(Rahimi, 2017)

✦ With the formula:

RFF:

$$\kappa(x, y) = \mathbb{E}_{\omega \sim \Lambda}[e^{-i\langle \omega, x-y \rangle}] \approx \langle \phi(x), \phi(y) \rangle_{\mathbb{R}^m}$$

From Wasserstein to MMD

✦ Kernel theory (in a nutshell)

H_κ **the RKHS of** κ

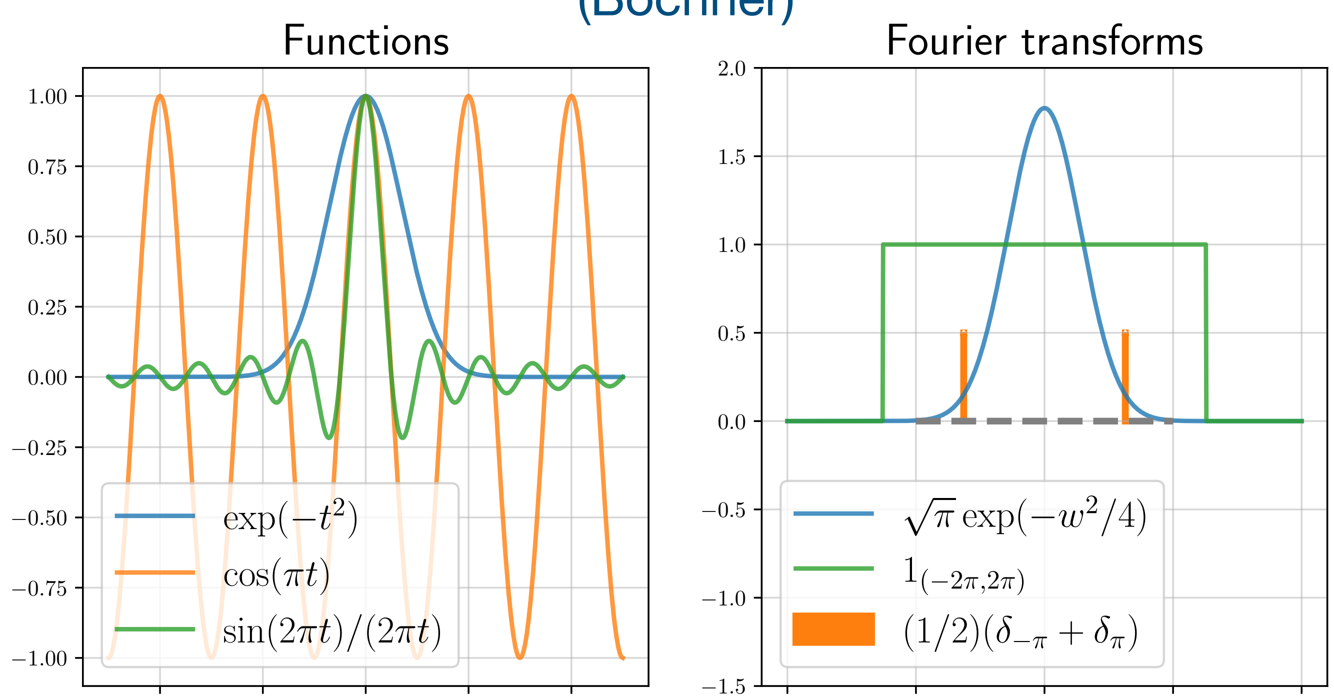
$\kappa : X \times X \to \mathbb{C}$ a PSD kernel $\Longleftrightarrow$ A **Hilbert** space of **functions** from $X \to \mathbb{C}$

✦ Translation invariant kernels

$$X = \mathbb{R}^d \quad \kappa(x, y) = \kappa_0(x - y)$$

✦ Is a PSD kernel $\Longleftrightarrow$ $\forall \omega, \widehat{\kappa_0}(\omega) \geq 0$

(Bochner)



✦ With the formula:

RFF:

$$\kappa(x, y) = \mathbb{E}_{\omega \sim \Lambda}[e^{-i\langle \omega, x-y \rangle}] \approx \langle \phi(x), \phi(y) \rangle_{\mathbb{R}^m}$$

✦ Maximum mean discrepancy

$\mu \in \mathscr{P}(X) \quad \nu \in \mathscr{P}(X)$

$$\mathrm{MMD}_\kappa(\mu, \nu) = \| \int_X \kappa(\cdot, x) \mathrm{d}\mu(x) - \int_X \kappa(\cdot, y) \mathrm{d}\nu(y) \|_{H_\kappa}$$

From Wasserstein to MMD

✦ Kernel theory (in a nutshell)

H_κ **the RKHS of** κ

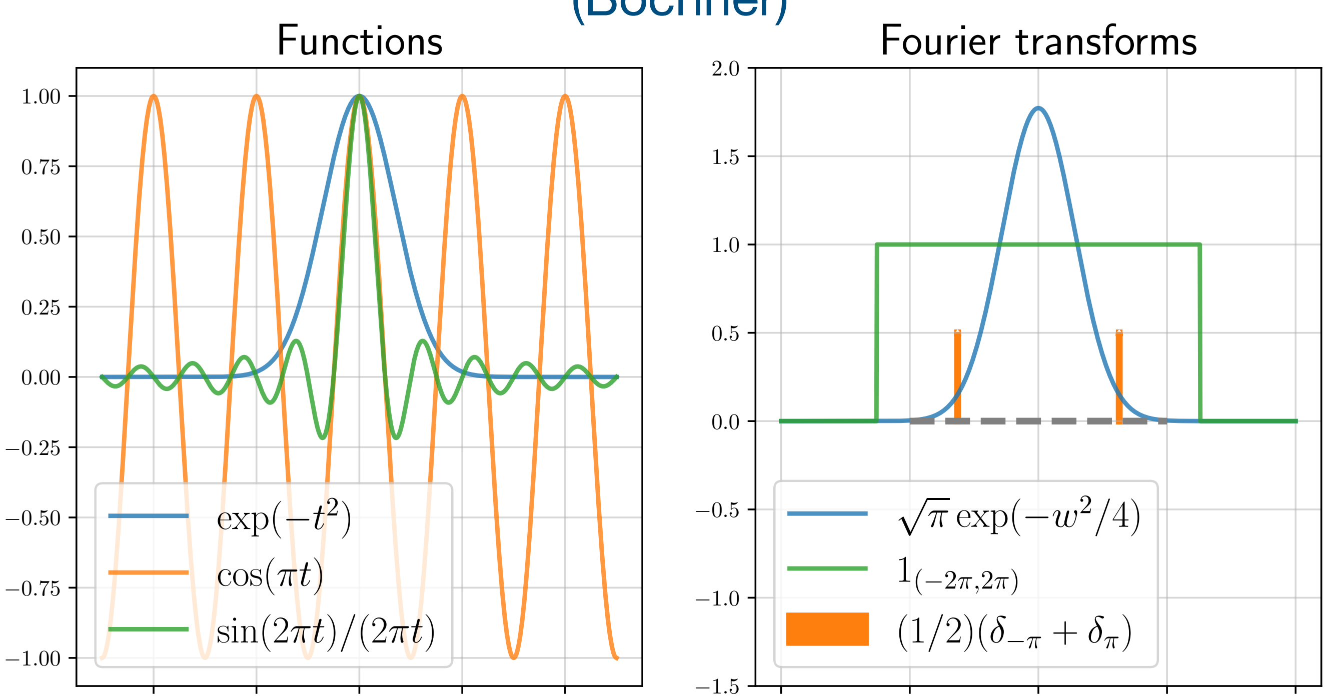
$\kappa : X \times X \to \mathbb{C}$ a PSD kernel $\iff$ A **Hilbert** space of **functions** from $X \to \mathbb{C}$

✦ Translation invariant kernels

$$X = \mathbb{R}^d \quad \kappa(x, y) = \kappa_0(x - y)$$

✦ Is a PSD kernel $\iff$ $\forall \omega, \widehat{\kappa_0}(\omega) \geq 0$

(Bochner)



✦ With the formula:

RFF:

$$\kappa(x, y) = \mathbb{E}_{\omega\sim\Lambda}[e^{-i\langle\omega, x-y\rangle}] \approx \langle\phi(x), \phi(y)\rangle_{\mathbb{R}^m}$$

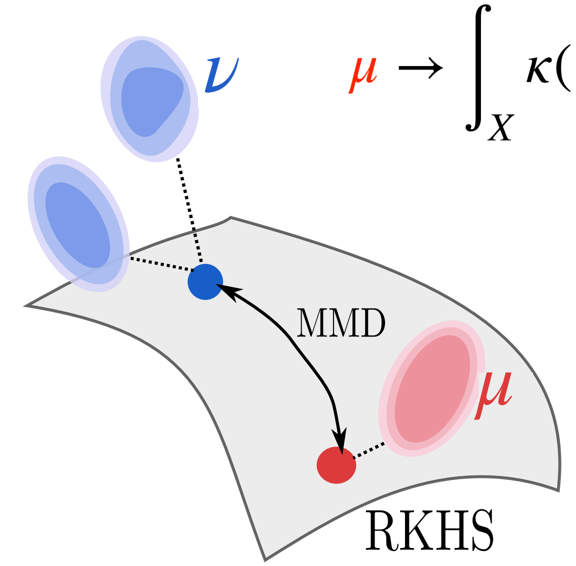
✦ Maximum mean discrepancy

$\mu \in \mathscr{P}(X) \quad \nu \in \mathscr{P}(X)$

$$\mathrm{MMD}_\kappa(\mu, \nu) = \left\| \int_X \kappa(\cdot, x)\mathrm{d}\mu(x) - \int_X \kappa(\cdot, y)\mathrm{d}\nu(y) \right\|_{H_\kappa}$$

✦ Distance in the embedding

$$\mu \to \int_X \kappa(\cdot, x)\mathrm{d}\mu(x)$$



From Wasserstein to MMD

✦ Kernel theory (in a nutshell)

H_κ **the RKHS of** κ

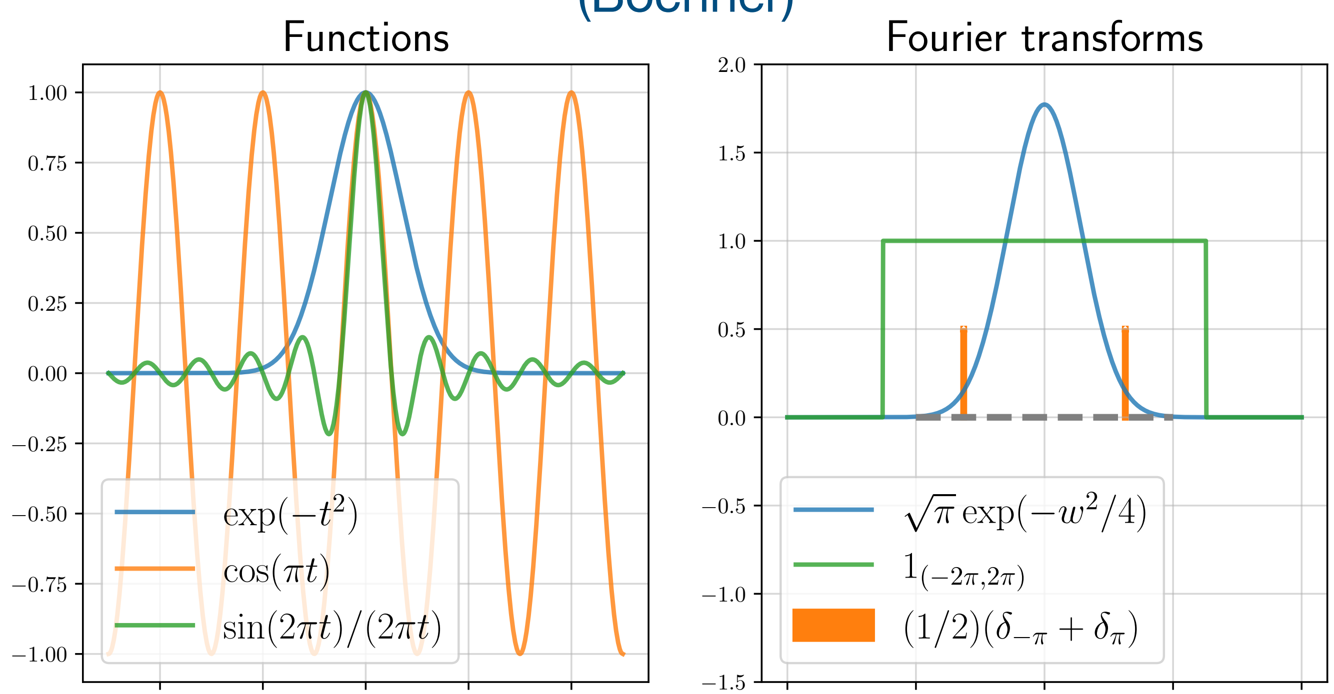
$\kappa : X \times X \to \mathbb{C}$ a PSD kernel $\iff$ A **Hilbert** space of **functions** from $X \to \mathbb{C}$

✦ Translation invariant kernels

$$X = \mathbb{R}^d \quad \kappa(x, y) = \kappa_0(x - y)$$

✦ Is a PSD kernel $\iff$ $\forall \omega, \widehat{\kappa_0}(\omega) \geq 0$

(Bochner)



✦ With the formula:

$$\kappa(x, y) = \mathbb{E}_{\omega \sim \Lambda}[e^{-i\langle \omega, x-y\rangle}] \approx \langle \phi(x), \phi(y)\rangle_{\mathbb{R}^m}$$

✦ Maximum mean discrepancy

$\mu \in \mathscr{P}(X) \quad \nu \in \mathscr{P}(X)$

$$\mathrm{MMD}_\kappa(\mu, \nu) = \left\| \int_X \kappa(\cdot, x)\mathrm{d}\mu(x) - \int_X \kappa(\cdot, y)\mathrm{d}\nu(y) \right\|_{H_\kappa}$$

$\gamma \in \mathscr{M}(X)$

$$\|\gamma\|_\kappa := \left(\int_{X \times X} \kappa(x, y)\ \mathrm{d}\gamma(x)\mathrm{d}\gamma(y) \right)^{1/2}$$

✦ Semi-norm on $\mathscr{M}(X)$

✦ Alternative formula:

$$\mathrm{MMD}_\kappa(\mu, \nu) = \|\mu - \nu\|_\kappa$$

Are they both equivalent ?

$$\forall \mu, \nu : C_1 \cdot \mathrm{W}_p(\mu, \nu) \leq \mathrm{MMD}(\mu, \nu) \leq C_2 \cdot \mathrm{W}_p(\mu, \nu)$$

Controlling MMDs by Wasserstein distances

✦ **Can we find** $C > 0$ **such that**

$$(\star) \quad \forall \mu, \nu : \mathrm{MMD}_\kappa(\mu, \nu) \leq C \cdot \mathrm{W}_p(\mu, \nu)$$

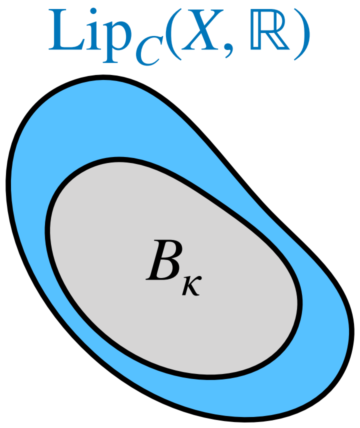
Controlling MMDs by Wasserstein distances

✦ **Can we find $C > 0$ such that**

$$(\star) \quad \forall \mu, \nu : \mathrm{MMD}_\kappa(\mu, \nu) \leq C \cdot \mathrm{W}_p(\mu, \nu)$$

A characterization $\exists C > 0$ such that $(\star)$

$$\iff \forall x, y \sqrt{\kappa(x,x) + \kappa(y,y) - 2\kappa(x,y)} \leq C \cdot d(x,y)$$



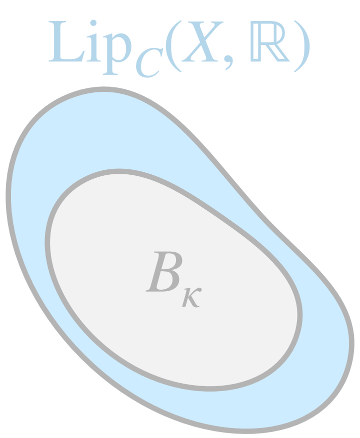
Controlling MMDs by Wasserstein distances

✦ **Can we find** $C > 0$ **such that**

$$(\star) \quad \forall \mu, \nu : \mathrm{MMD}_\kappa(\mu, \nu) \le C \cdot \mathrm{W}_p(\mu, \nu)$$

A characterization $\quad \exists C > 0$ such that $(\star)$

$$\iff \forall x, y \ \sqrt{\kappa(x,x) + \kappa(y,y) - 2\kappa(x,y)} \le C \cdot d(x,y)$$



✦ **Corollary for TI kernels** $\quad \kappa(x,y) = \kappa_0(x-y) \quad d(x,y) = \|x-y\|_2$

$(\star)$ **always holds** with $C = \kappa_0(0)\sqrt{\lambda_{\max}(-\nabla^2[\kappa_0](0))}$

- The MMD is a weaker notion of metric for smooth TI kernel
- This direction is easy !!

Controlling Wasserstein distances by MMDs

- **Can we find** $C > 0$ **such that**

$$(\star\star)\quad \forall \mu, \nu : \mathrm{W}_p(\mu, \nu) \leq C \cdot \mathrm{MMD}_\kappa(\mu, \nu)$$

Controlling Wasserstein distances by MMDs

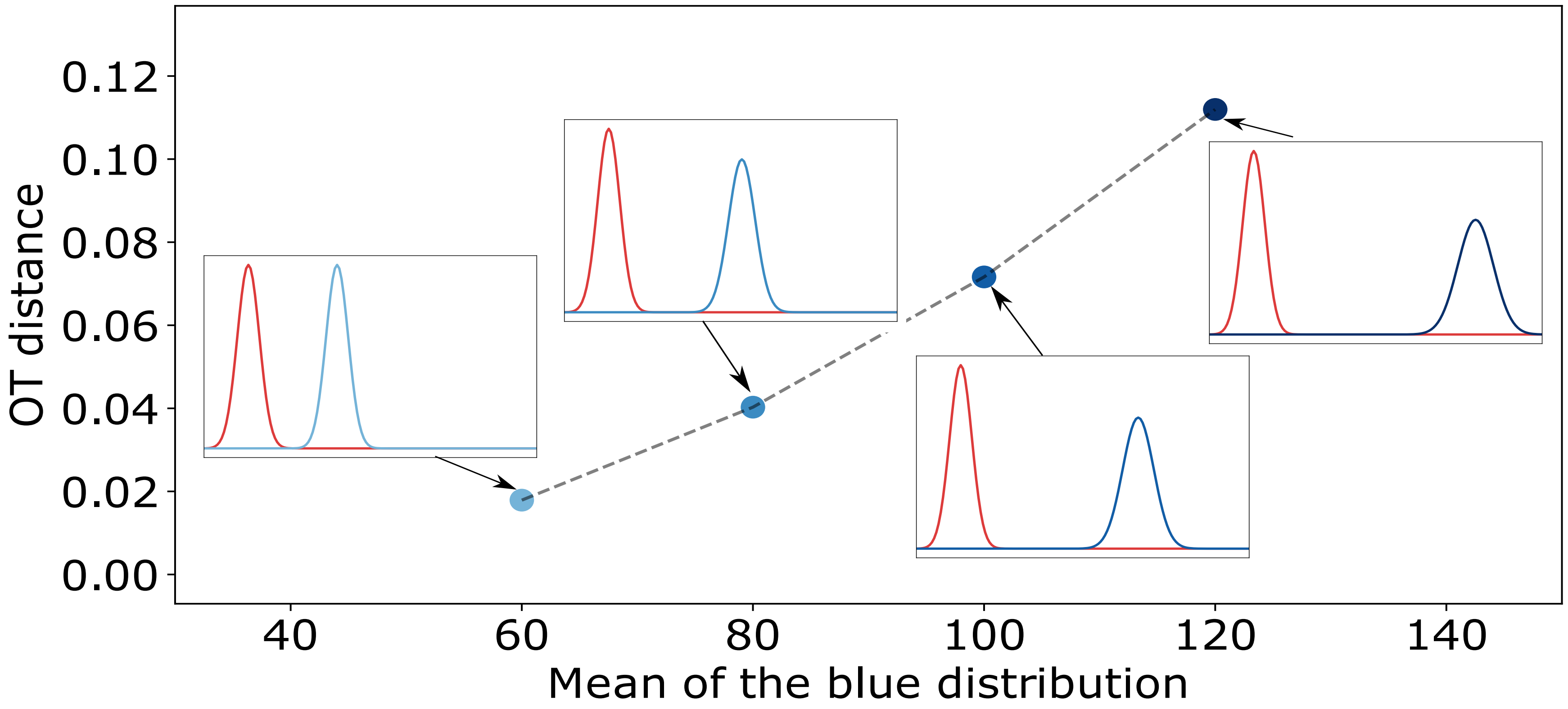
✦ **Can we find** $C > 0$ **such that**

$$(\star\star)\ \ \forall \mu, \nu : \mathrm{W}_p(\mu, \nu) \leq C \cdot \mathrm{MMD}_\kappa(\mu, \nu)$$

Not without any assumption !!

✦ e.g : if κ bounded then $\forall \mu, \nu \in \mathscr{P}(X)\ \ \mathrm{MMD}_\kappa(\mu, \nu) \leq \mathrm{cte}$

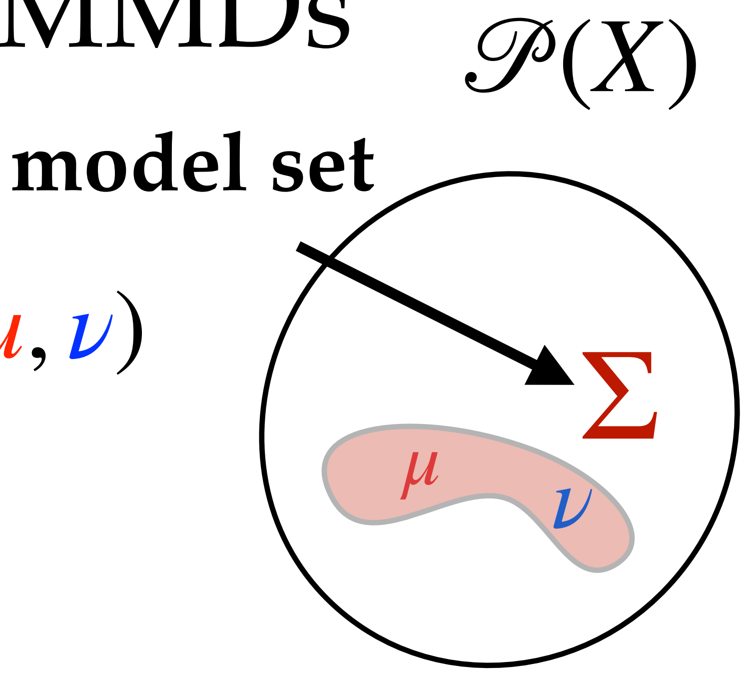
✦ but not Wasserstein distances !



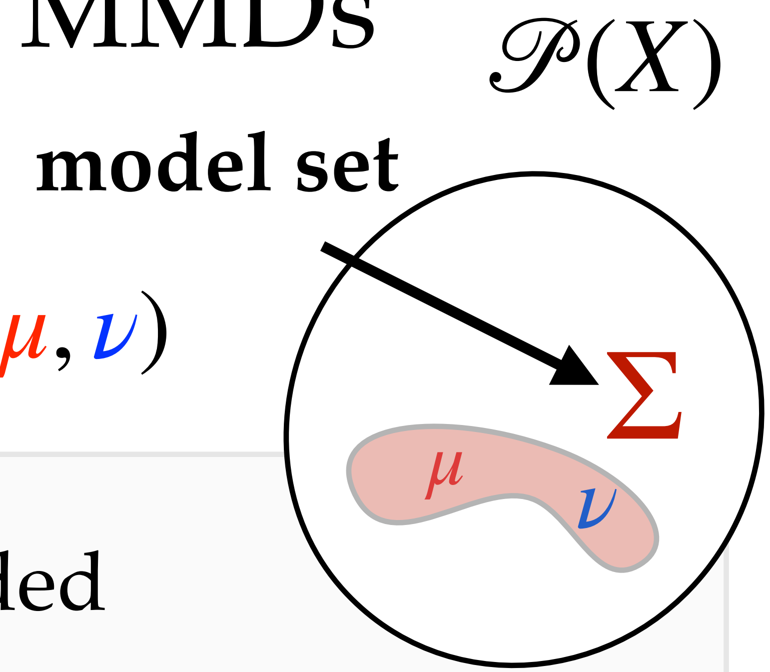
Controlling Wasserstein distances by MMDs

- **Can we find** $C_\Sigma > 0$ **such that**

$$(\star\star\star)\ \forall \mu, \nu \in \Sigma : \mathrm{W}_p(\mu, \nu) \leq C_\Sigma \cdot \mathrm{MMD}_\kappa(\mu, \nu)$$



Controlling Wasserstein distances by MMDs



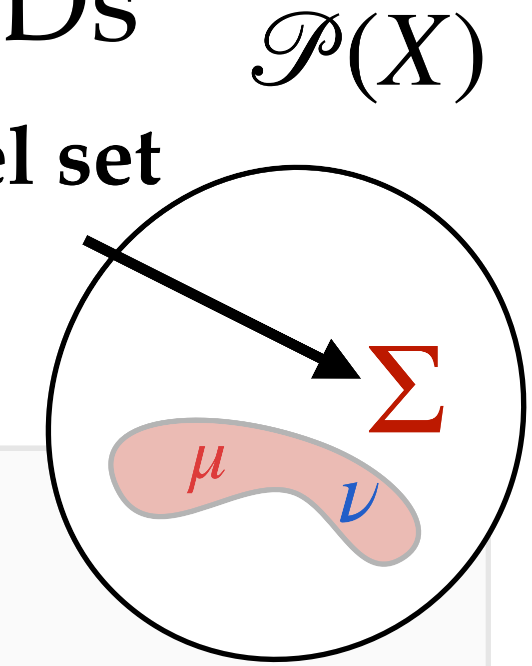
✦ **Can we find $C_\Sigma > 0$ such that**

$$(\star\star\star)\ \forall \mu, \nu \in \Sigma : \mathrm{W}_p(\mu, \nu) \leq C_\Sigma \cdot \mathrm{MMD}_\kappa(\mu, \nu)$$

✦ If κ bounded then necessarily Σ must be bounded

$$(\star\star\star) \implies \sup_{\mu,\nu\in\Sigma} \|\,\mathrm{mean}(\mu) - \mathrm{mean}(\nu)\|_2 < +\infty$$

Controlling Wasserstein distances by MMDs



✦ **Can we find** $C_\Sigma > 0$ **such that**

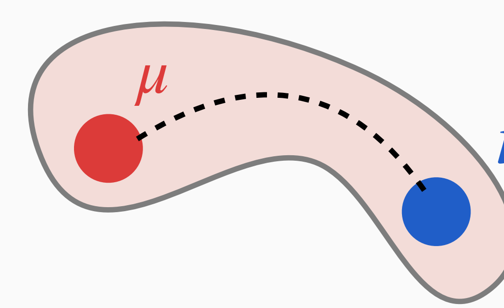
$$(\star\star\star)\ \forall \mu, \nu \in \Sigma : \mathrm{W}_p(\mu,\nu) \leq C_\Sigma \cdot \mathrm{MMD}_\kappa(\mu,\nu)$$

✦ If κ bounded then necessarily Σ must be bounded

$$(\star\star\star) \implies \sup_{\mu,\nu\in\Sigma} \|\mathrm{mean}(\mu) - \mathrm{mean}(\nu)\|_2 < +\infty$$

✦ If Σ contains $[\mu, \nu]$ with $\mathrm{supp}(\mu) \cap \mathrm{supp}(\nu) = \varnothing$

$(\star\star\star)$ impossible with $p > 1$



We must find a « larger » definition

Controlling Wasserstein distances by MMDs

✦ **Definition: embeddability** $\Sigma \subset \mathscr{P}(X), \delta \in [0,1]$

(Σ, W_p) is $(\kappa, \delta) -$ embeddable when

$$\exists C > 0,\ \forall \mu, \nu \in \Sigma : \mathrm{W}_p(\mu, \nu) \leq C \cdot \mathrm{MMD}_\kappa^\delta(\mu, \nu)$$

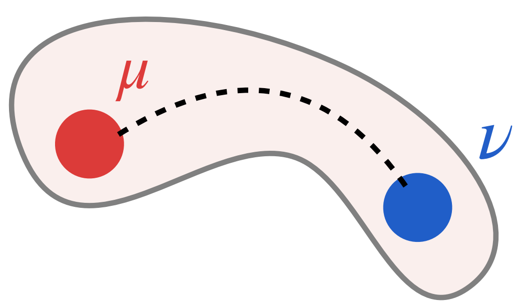
Controlling Wasserstein distances by MMDs

✦ **Definition: embeddability** $\Sigma \subset \mathscr{P}(X), \delta \in [0,1]$

$$(\Sigma, \mathrm{W}_p) \text{ is } (\kappa, \delta) - \text{embeddable when}$$

$$\exists C > 0, \forall \mu, \nu \in \Sigma : \mathrm{W}_p(\mu, \nu) \leq C \cdot \mathrm{MMD}_\kappa^\delta(\mu, \nu)$$

✦ **Some necessary conditions:**

✦ If 

$(\kappa, \delta) - \text{embeddable} \implies \delta \leq 1/p$

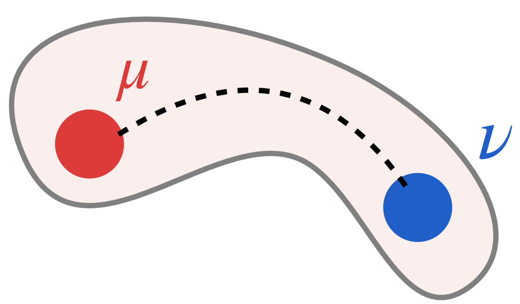
Controlling Wasserstein distances by MMDs

✦ **Definition: embeddability** $\Sigma \subset \mathscr{P}(X), \delta \in [0,1]$

(Σ, W_p) is (κ, δ) − embeddable when

$$\exists C > 0,\ \forall \mu, \nu \in \Sigma : \mathrm{W}_p(\mu, \nu) \leq C \cdot \mathrm{MMD}_\kappa^\delta(\mu, \nu)$$

✦ **Some necessary conditions:**

✦ If 

(κ, δ) − embeddable $\Longrightarrow$ $\delta \leq 1/p$

✦ If $\Sigma = \{\mu \in \mathscr{P}(X) : \mu(B(0,R)) = 1\}$ and κ bounded

$$(\kappa, \delta) - \text{embeddable} \Longrightarrow \delta \leq 2/d$$

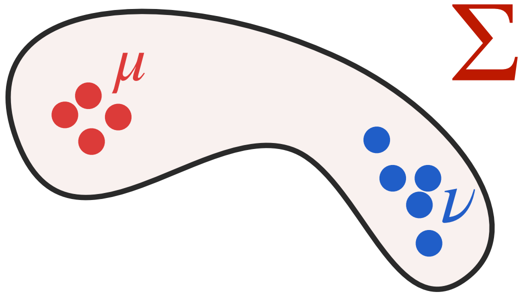
Controlling Wasserstein distances by MMDs

✦ **Definition: embeddability** $\Sigma \subset \mathscr{P}(X), \delta \in [0,1]$

(Σ, W_p) is $(\kappa, \delta) -$ embeddable when

$$\exists C > 0, \forall \mu, \nu \in \Sigma : \mathrm{W}_p(\mu, \nu) \leq C \cdot \mathrm{MMD}_\kappa^\delta(\mu, \nu)$$

✦ **Some necessary conditions:**



Σ discrete distributions **with K atoms**

$$\Sigma = \{ \sum_{i=1}^{K} a_i \delta_{\mathbf{x}_i} : \mathbf{x}_i \in B(0,R) \}$$

$\kappa(x, y) = \kappa_0(x - y)$ with κ_0 smooth

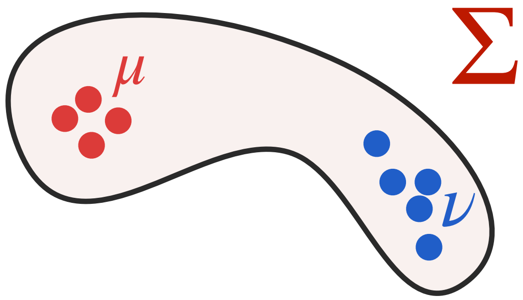
Controlling Wasserstein distances by MMDs

✦ **Definition: embeddability** $\Sigma \subset \mathscr{P}(X), \delta \in [0,1]$

(Σ, W_p) is (κ, δ) – embeddable when

$$\exists C > 0, \forall \mu, \nu \in \Sigma : \mathrm{W}_p(\mu, \nu) \leq C \cdot \mathrm{MMD}_\kappa^\delta(\mu, \nu)$$

✦ **Some necessary conditions:**



Σ discrete distributions **with K atoms**

$$\Sigma = \{ \sum_{i=1}^{K} a_i \delta_{\mathbf{x}_i} : \mathbf{x}_i \in B(0,R) \}$$

$\kappa(x, y) = \kappa_0(x - y)$ with κ_0 smooth

$$\boxed{(\kappa, \delta) - \text{embeddable} \implies \delta \leq 2/K}$$

Controlling Wasserstein distances by MMDs

✦ **Definition: embeddability** $\Sigma \subset \mathscr{P}(X), \delta \in [0,1]$

$$(\Sigma, \mathrm{W}_p) \text{ is } (\kappa, \delta)-\text{embeddable when}$$

$$\exists C > 0,\ \forall \mu, \nu \in \Sigma : \mathrm{W}_p(\mu, \nu) \leq C \cdot \mathrm{MMD}_\kappa^\delta(\mu, \nu)$$

We **cannot** control Wass by MMD **uniformly over all discrete distrib.** (even in a compact) **for a smooth TI kernel**

Controlling Wasserstein distances by MMDs

✦ **Some positive results**

✦ Distrib. with density in Sobolev space + bounded moments

$$\underbrace{\int |\partial^{|s|} f(\mathbf{x})|^2 \mathrm{d}\mathbf{x}}_{} \qquad \int \|\mathbf{x}\|_2^r \mathrm{d}\mu(\mathbf{x})$$

$$\Sigma = \{\mu = f\mathrm{d}\mathbf{x} : \|f\|_s \leq B, \mathrm{M}_r(\mu) \leq M\}$$

Controlling Wasserstein distances by MMDs

✦ Some positive results

- Distrib. with density in Sobolev space + bounded moments

$\int |\partial^{|s|} f(\mathbf{x})|^2 \mathrm{d}\mathbf{x}$ $\qquad$ $\int \|\mathbf{x}\|_2^r \mathrm{d}\mu(\mathbf{x})$

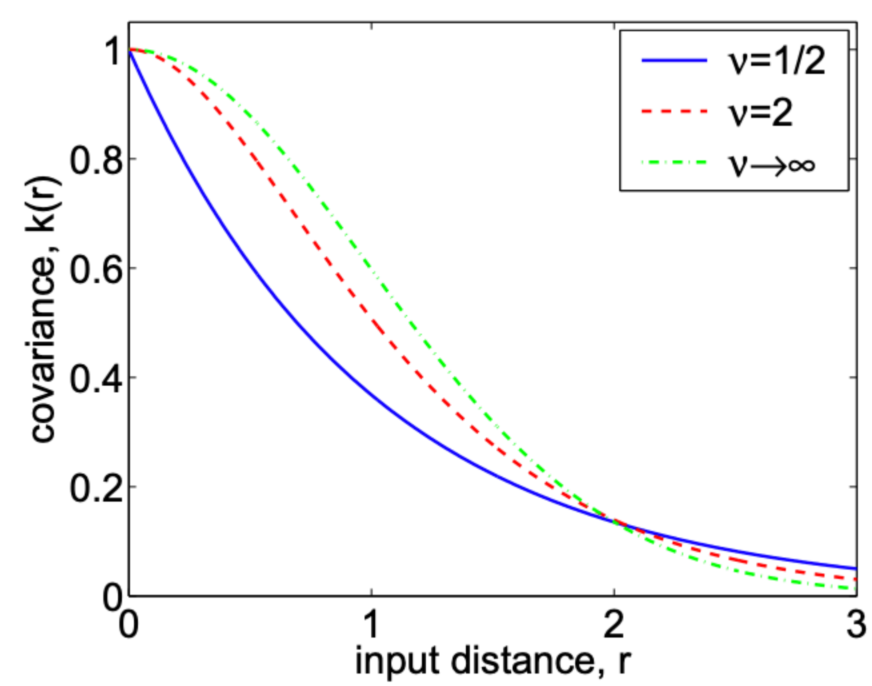
$$\Sigma = \{\mu = f\mathrm{d}\mathbf{x} : \|f\|_s \leq B, \mathrm{M}_r(\mu) \leq M\}$$

- Any TI, PSD kernels $\kappa(x, y) = \kappa_0(x - y)$ + some regularity

~~Gaussian~~

$$1/\widehat{\kappa_0}(\omega) = O_{\omega \to +\infty}(\|\omega\|^{2s})$$

Matérn class, splines, polyharmonic curves



Controlling Wasserstein distances by MMDs

✦ Some positive results

- Distrib. with density in Sobolev space + bounded moments

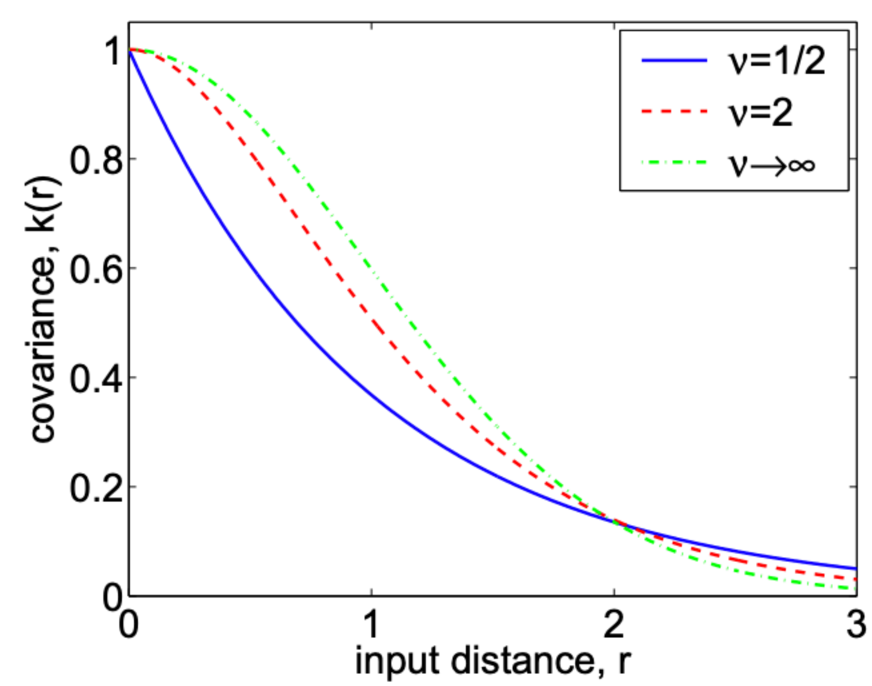
$$\int |\partial^{|s|} f(\mathbf{x})|^2 \mathrm{d}\mathbf{x} \qquad \int \|\mathbf{x}\|_2^r \mathrm{d}\mu(\mathbf{x})$$

$$\Sigma = \{\mu = f\mathrm{d}\mathbf{x} : \|f\|_s \leq B, \mathrm{M}_r(\mu) \leq M\}$$

- Any TI, PSD kernels $\kappa(x, y) = \kappa_0(x - y)$ + some regularity

~~Gaussian~~ $\quad 1/\widehat{\kappa_0}(\omega) = O_{\omega \to +\infty}(\|\omega\|^{2s})$

Matérn class, splines, polyharmonic curves



$$(\Sigma, \mathrm{W}_p) \text{ is } (\kappa, \delta) - \text{embeddable}$$

$$\text{with } \delta = \frac{r - p}{p(d + 2r)}$$

Controlling Wasserstein distances by MMDs

✦ Some positive results

✦ Distrib. with density in Sobolev space + bounded moments

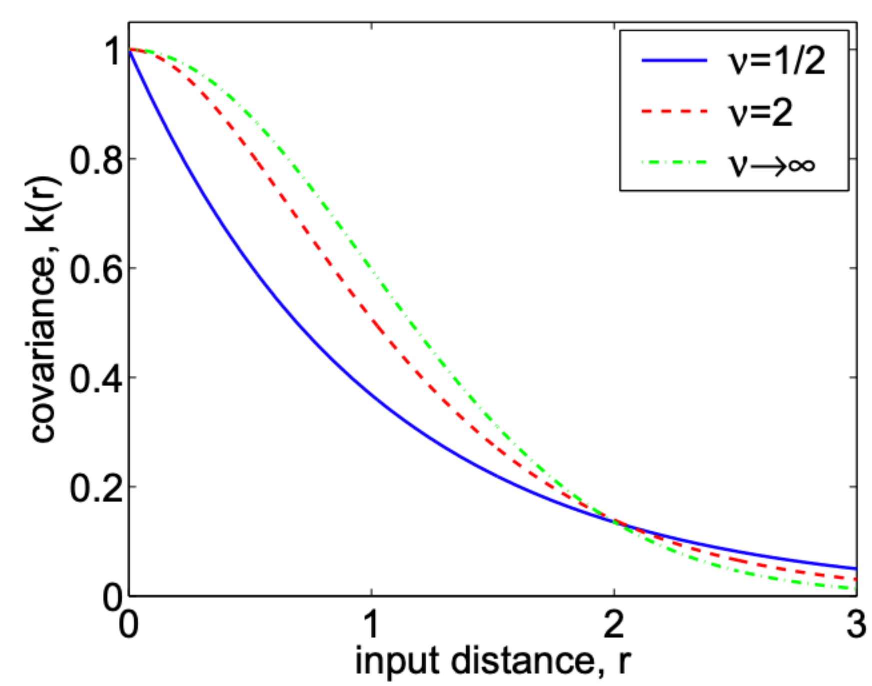
$$\underbrace{\text{density in Sobolev space}}_{\int |\partial^{|s|} f(\mathbf{x})|^2 \mathrm{d}\mathbf{x}} \qquad \underbrace{\text{bounded moments}}_{\int \|\mathbf{x}\|_2^r \mathrm{d}\mu(\mathbf{x})}$$

$$\Sigma = \{\mu = f\mathrm{d}\mathbf{x} : \|f\|_s \leq B, \mathrm{M}_r(\mu) \leq M\}$$

✦ Any TI, PSD kernels $\kappa(x, y) = \kappa_0(x - y)$ + some regularity

~~Gaussian~~ $\qquad 1/\widehat{\kappa_0}(\omega) = O_{\omega \to +\infty}(\|\omega\|^{2s})$

Matérn class, splines, polyharmonic curves



(Σ, W_p) is (κ, δ) − embeddable

with $\delta \approx \dfrac{1}{2p}$ **(r big)**

Controlling Wasserstein distances by MMDs

- **Some positive results**
  - Distrib. with smooth densities + compact support
  - Any TI, PSD kernels $\kappa(x, y) = \kappa_0(x - y)$ + Matérn class

$$(\Sigma, \mathrm{W}_p) \text{ is } (\kappa, \delta)-\text{embeddable with } \delta = 1/2p$$

Controlling Wasserstein distances by MMDs

✦ Some positive results

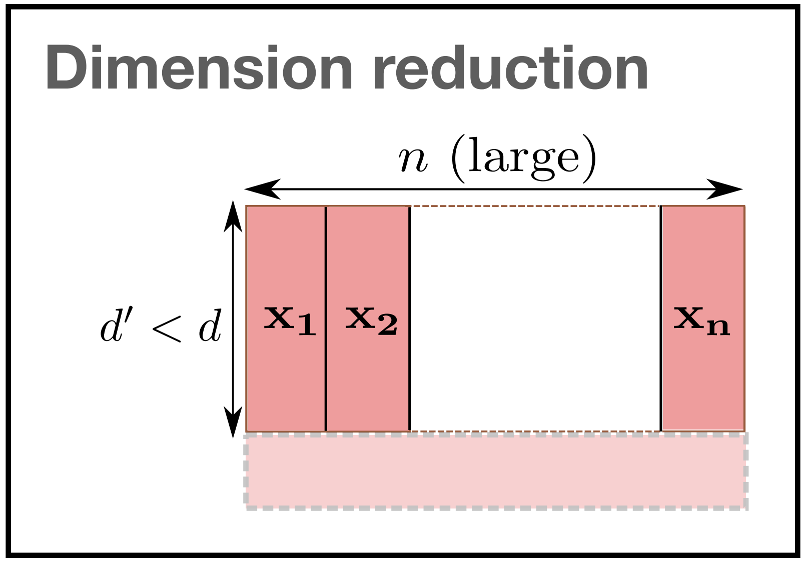
- Distrib. with smooth densities + compact support
- Any TI, PSD kernels $\kappa(x, y) = \kappa_0(x - y)$ + Matérn class

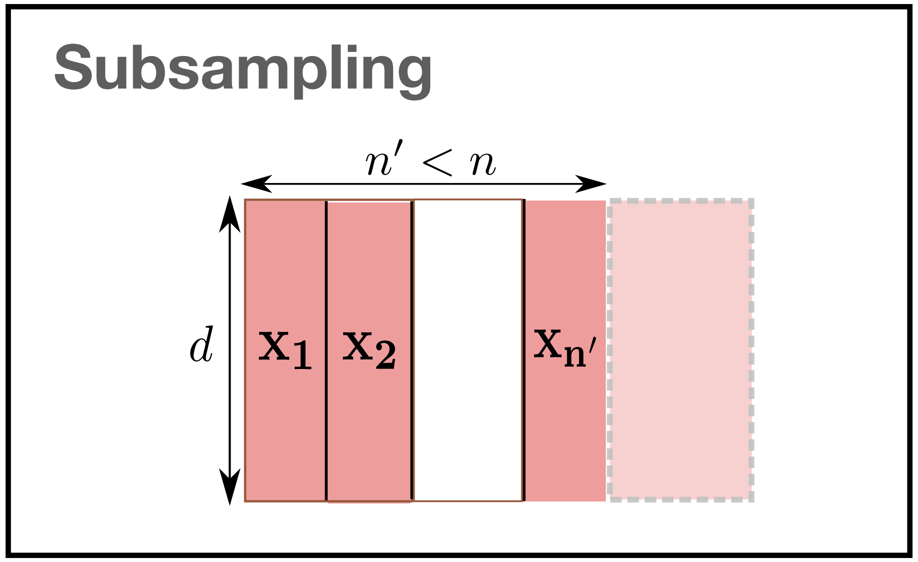
$$(\Sigma, \mathrm{W}_p) \text{ is } (\kappa, \delta)-\text{embeddable with } \delta = 1/2p$$

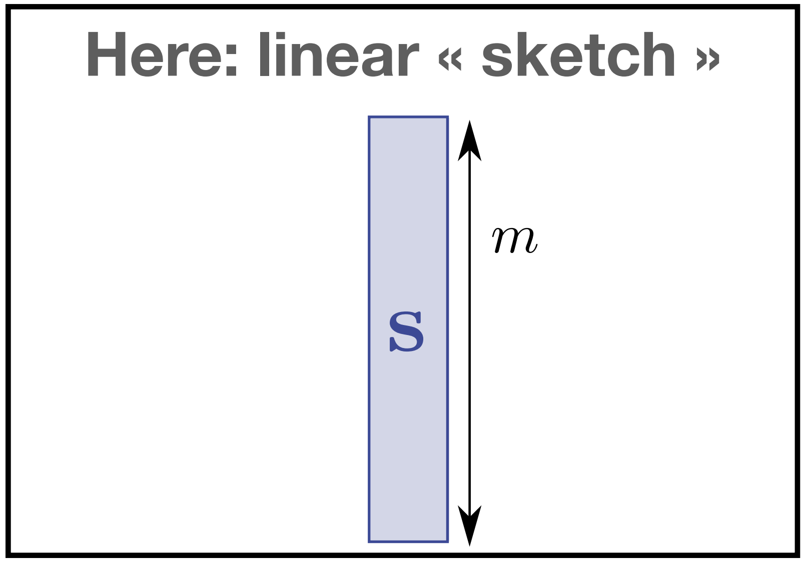
✦ Other results

- Larger class of distrib. / kernels if we allow an error $\eta > 0$
- For unbounded + conditionally PSD kernels (Chafaï, 2016)
- Other connections (Modeste, 2022), (Goldfeld, 2020)

Motivations: compressive learning

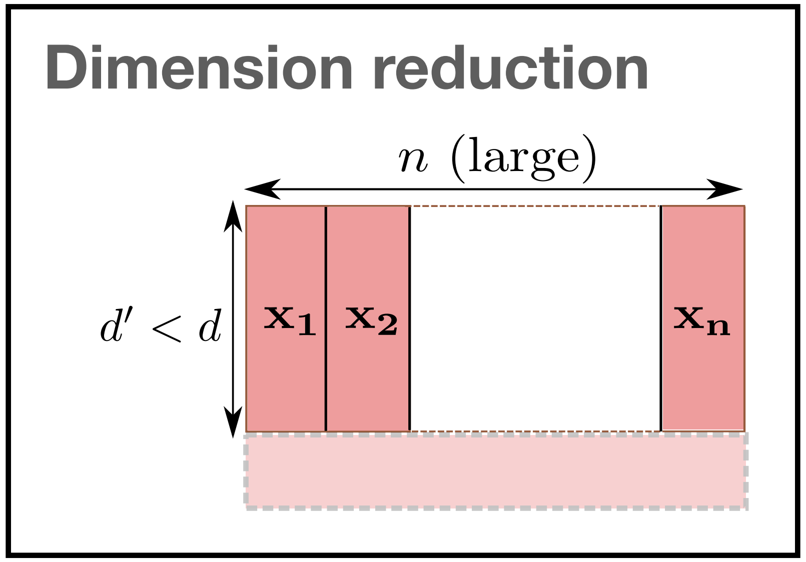


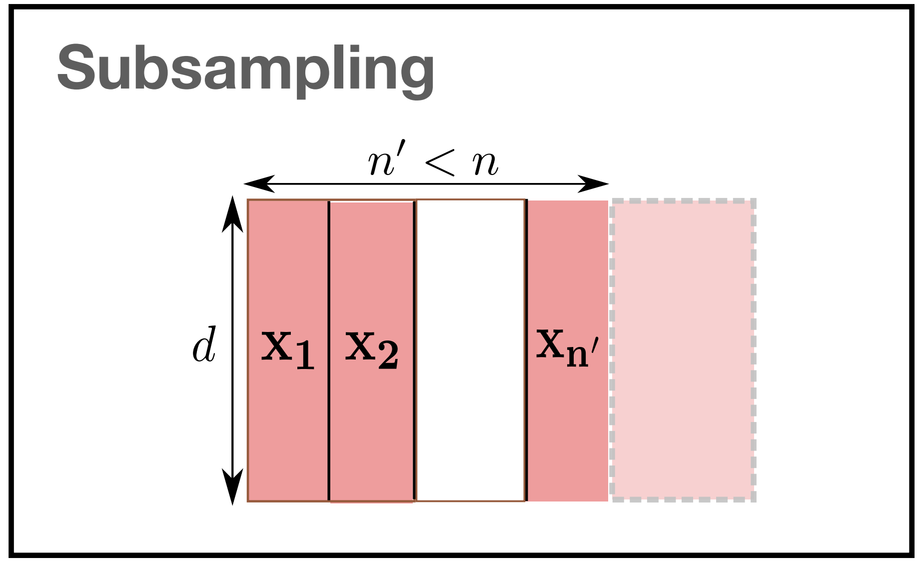


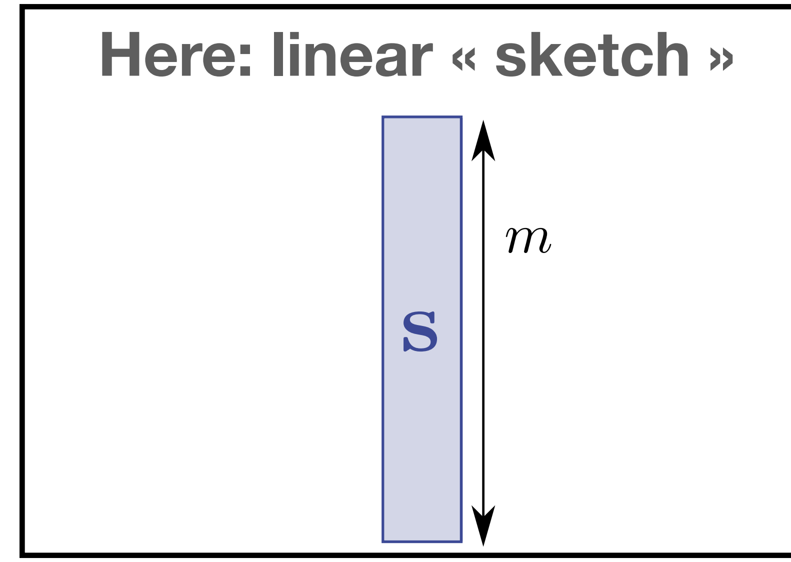


[Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, Yann Traonmilin, Antoine Chatalic, Vincent Schellekens, Laurent Jacques…]

Motivations: compressive learning







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