Kernel Quantile Embeddings

(and associated probability metrics)

Masha Naslidnyk
University College London

UQSay, 8th of January 2025

Kernel Quantile Embeddings

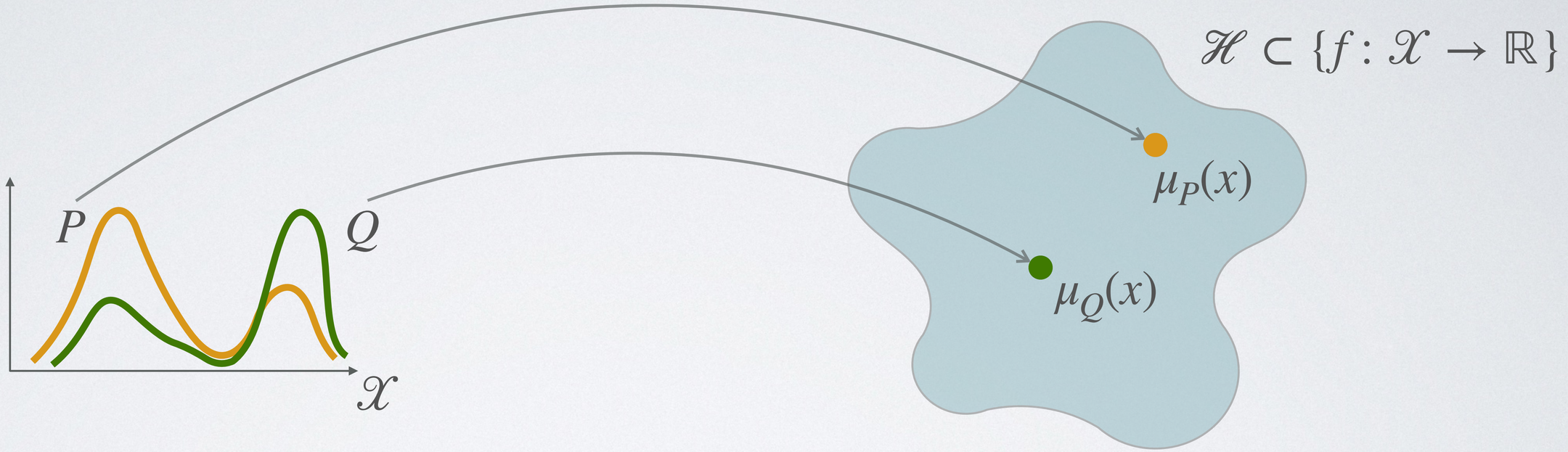
(and associated probability metrics)

Masha Naslidnyk
University College London



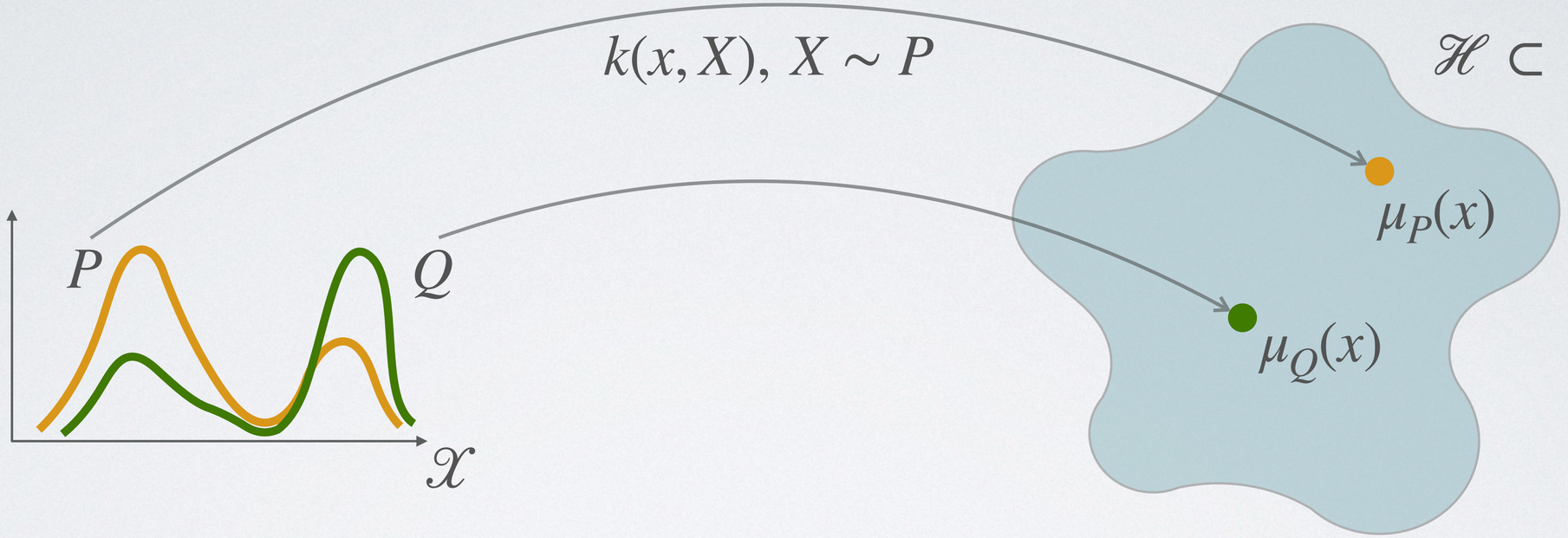
work with Siu Lun (Alan) Chau, François-Xavier Briol, and Krikamol Muandet . ICML 2025.

EMBEDDING DISTRIBUTIONS



Used in: hypothesis testing, parameter estimation, causal inference, feature attribution, learning on distributions…

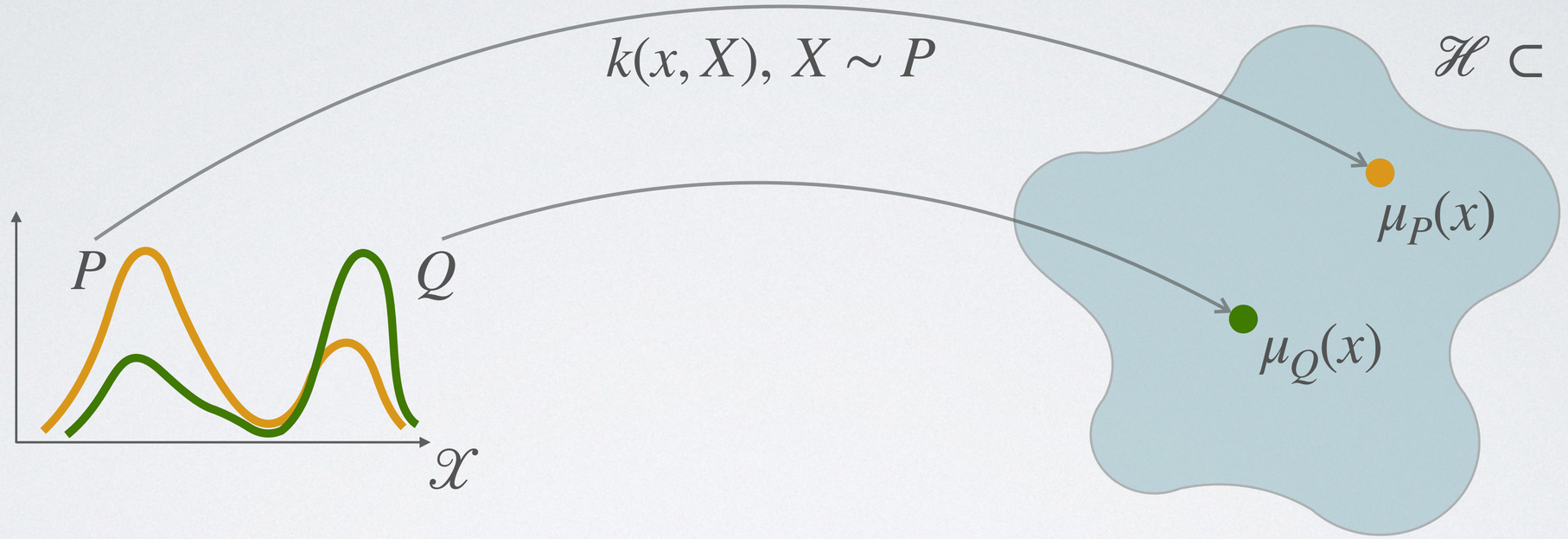
EMBEDDING DISTRIBUTIONS



$$P \mapsto \mu_P(x) = \int_{\mathcal{X}} k(x, x')P(\mathrm{d}x')$$

Used in: hypothesis testing, parameter estimation, causal inference, feature attribution, learning on distributions…

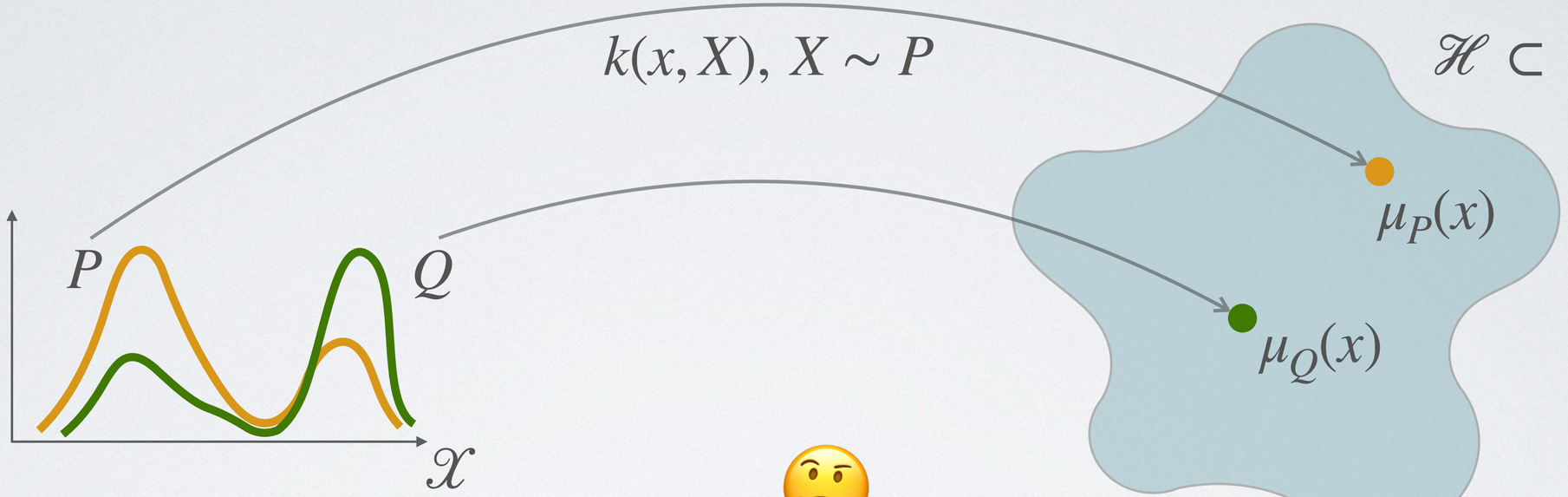
EMBEDDING DISTRIBUTIONS



$$P \mapsto \mu_P(x) = \int_{\mathscr{X}} k(x, x') P(\mathrm{d}x')$$

$\mathrm{MMD}(P, Q) = \|\mu_P - \mu_Q\|_{\mathscr{H}}$ is a distance for mean-characteristic k

EMBEDDING DISTRIBUTIONS

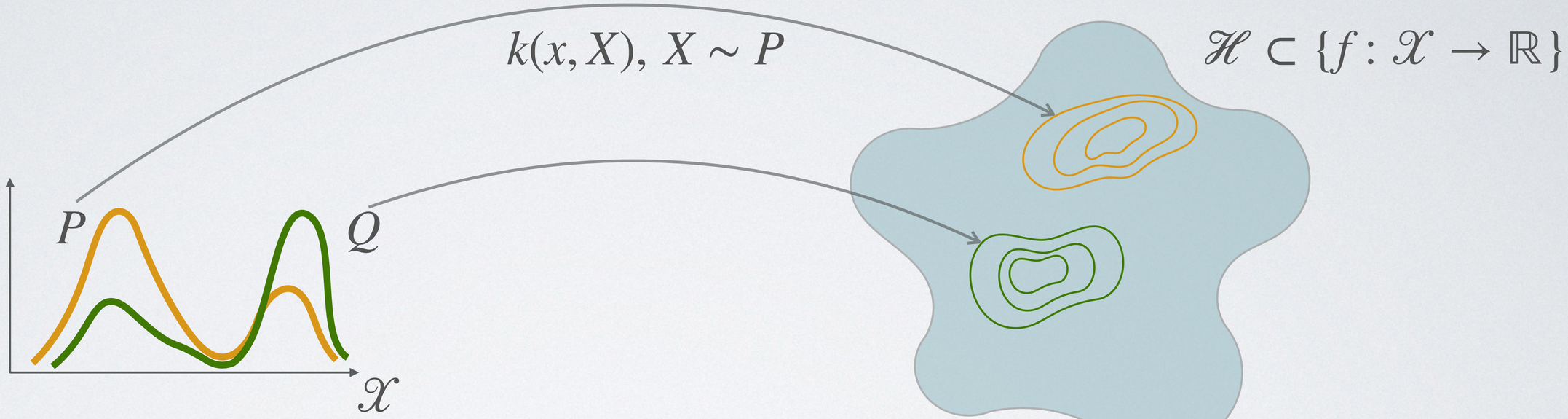


…alternatives?

$$P \mapsto \mu_P(x) = \int_{\mathcal{X}} k(x, x') P(\mathrm{d}x')$$

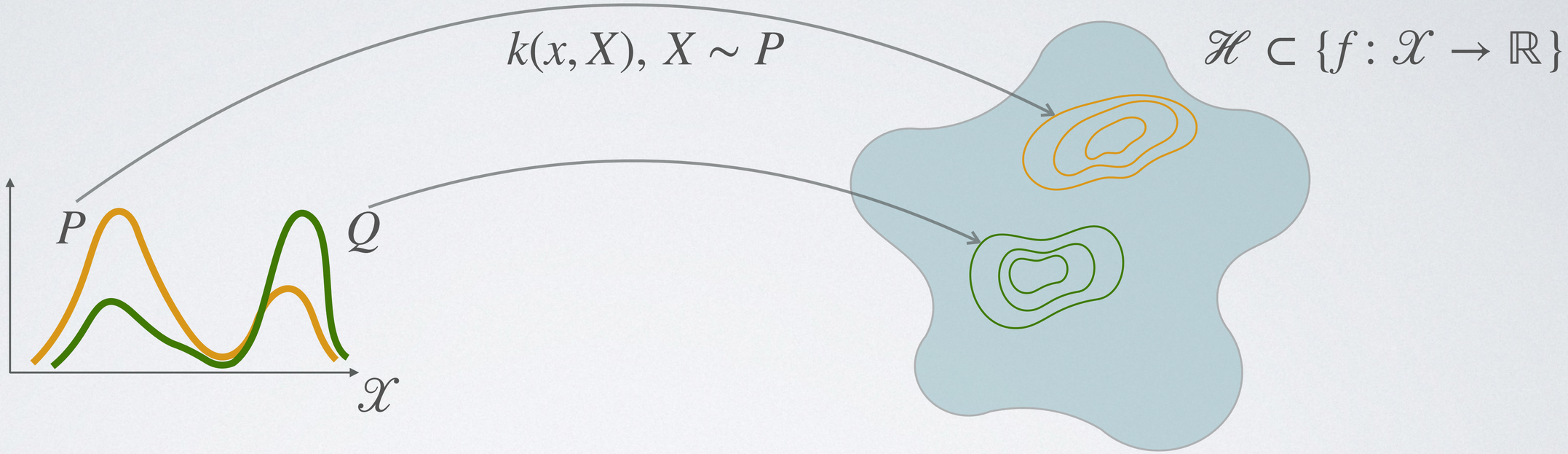
$\mathrm{MMD}(P, Q) = \|\mu_P - \mu_Q\|_{\mathscr{H}}$ is a distance for *mean-characteristic* k

CONTRIBUTIONS



We propose: $P \mapsto$ *kernel quantiles* of P under k

CONTRIBUTIONS



We propose: $P \mapsto$ *kernel quantiles* of P under k

\+ *kernel quantile discrepancies* that are:

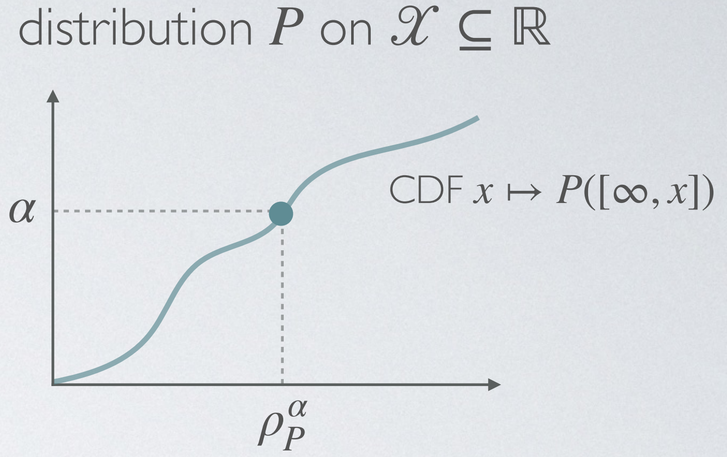
- distances under mild conditions,
- connected to Wasserstein distances,
- can be estimated at near-linear cost.

METHODOLOGY

QUANTILES

- When $\mathcal{X} \subseteq \mathbb{R}$, the α-quantiles of P is

$$\rho_P^\alpha := \inf\{x \in \mathcal{X} : P([-\infty, x]) \geq \alpha\}$$



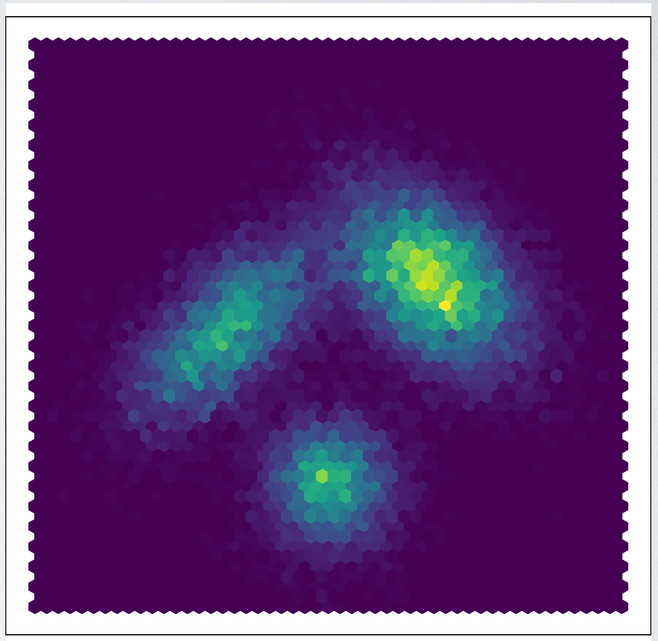
QUANTILES

- When $\mathcal{X} \subseteq \mathbb{R}$, the α-quantiles of P is

$$\rho_P^\alpha := \inf\{x \in \mathcal{X} : P([-\infty, x]) \geq \alpha\}$$

- When $\mathcal{X} \subseteq \mathbb{R}^d$, there are multiple extensions. We consider directional quantiles.

distribution P on $\mathcal{X} \subseteq \mathbb{R}^2$



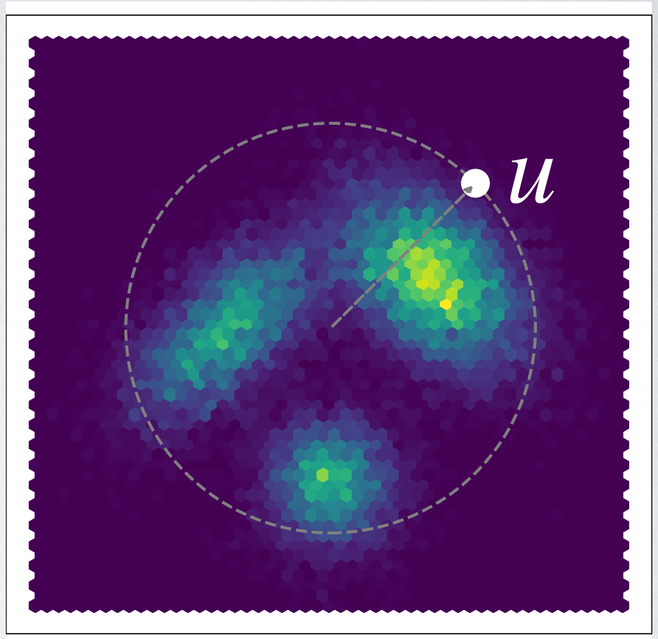
QUANTILES

- When $\mathcal{X} \subseteq \mathbb{R}$, the α-quantiles of P is

$$\rho_P^\alpha := \inf\{x \in \mathcal{X} : P([-\infty, x]) \geq \alpha\}$$

- When $\mathcal{X} \subseteq \mathbb{R}^d$, there are multiple extensions. We consider directional quantiles.

distribution P on $\mathcal{X} \subseteq \mathbb{R}^2$



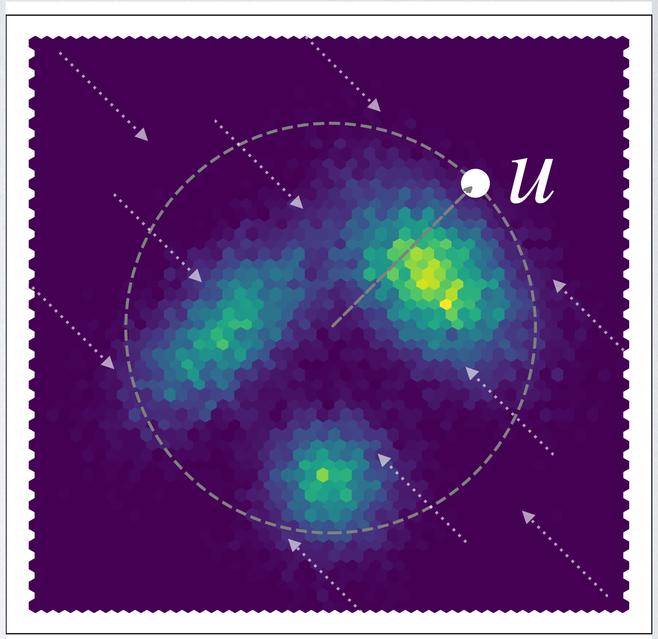
QUANTILES

- When $\mathcal{X} \subseteq \mathbb{R}$, the α-quantiles of P is

$$\rho_P^\alpha := \inf\{x \in \mathcal{X} : P([-\infty, x]) \geq \alpha\}$$

- When $\mathcal{X} \subseteq \mathbb{R}^d$, there are multiple extensions. We consider directional quantiles.

distribution P on $\mathcal{X} \subseteq \mathbb{R}^2$



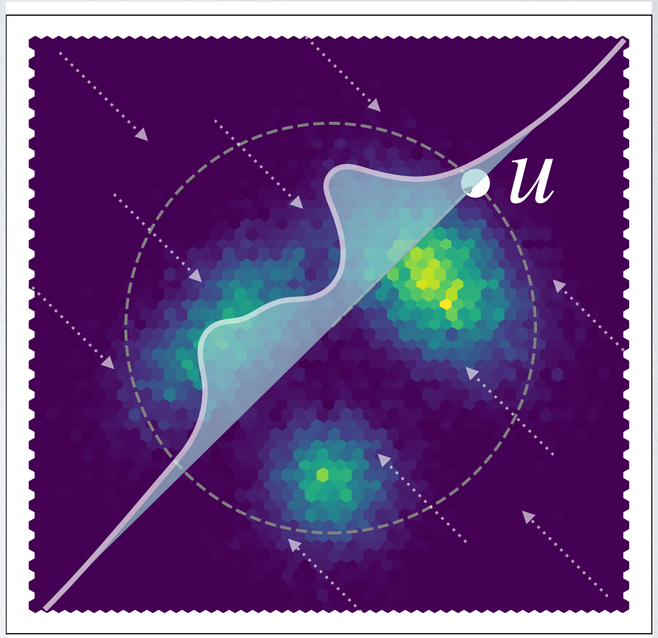
QUANTILES

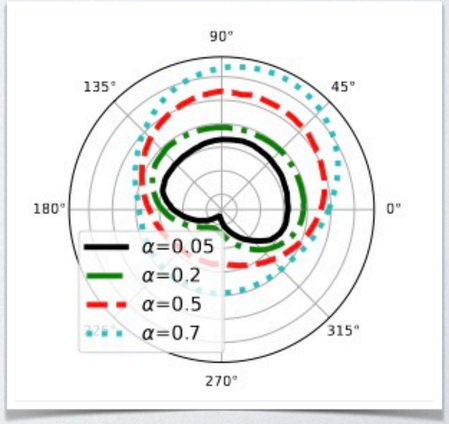
- When $\mathcal{X} \subseteq \mathbb{R}$, the α-quantiles of P is

$$\rho_P^\alpha := \inf\{x \in \mathcal{X} : P([-\infty, x]) \geq \alpha\}$$

- When $\mathcal{X} \subseteq \mathbb{R}^d$, there are multiple extensions. We consider directional quantiles.

distribution P on $\mathcal{X} \subseteq \mathbb{R}^2$





QUANTILES

- When $\mathcal{X} \subseteq \mathbb{R}$, the α-quantiles of P is

$$\rho_P^\alpha := \inf\{x \in \mathcal{X} : P([-\infty, x]) \geq \alpha\}$$

- When $\mathcal{X} \subseteq \mathbb{R}^d$, there are multiple extensions. We consider directional quantiles.

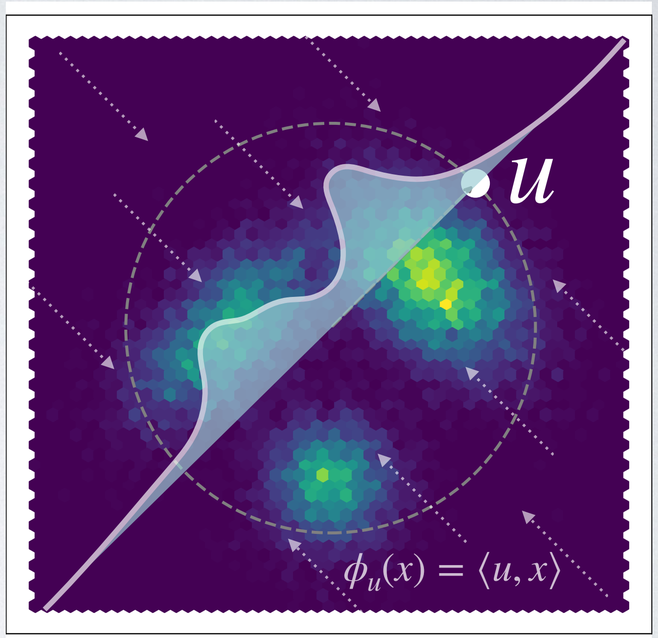
direction $u \in S^{d-1}$

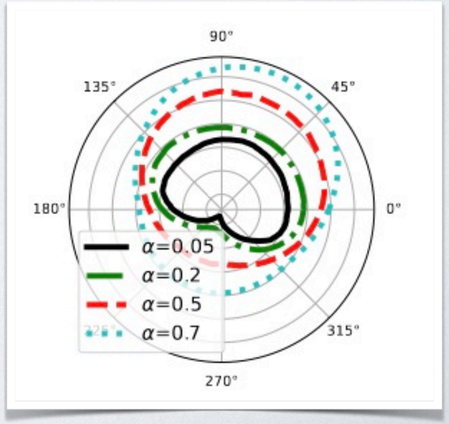
$$\rho_P^{\alpha,u} := \rho_{\phi_u \# P}^\alpha$$

projection map $\phi_u(x) = \langle u, x \rangle$

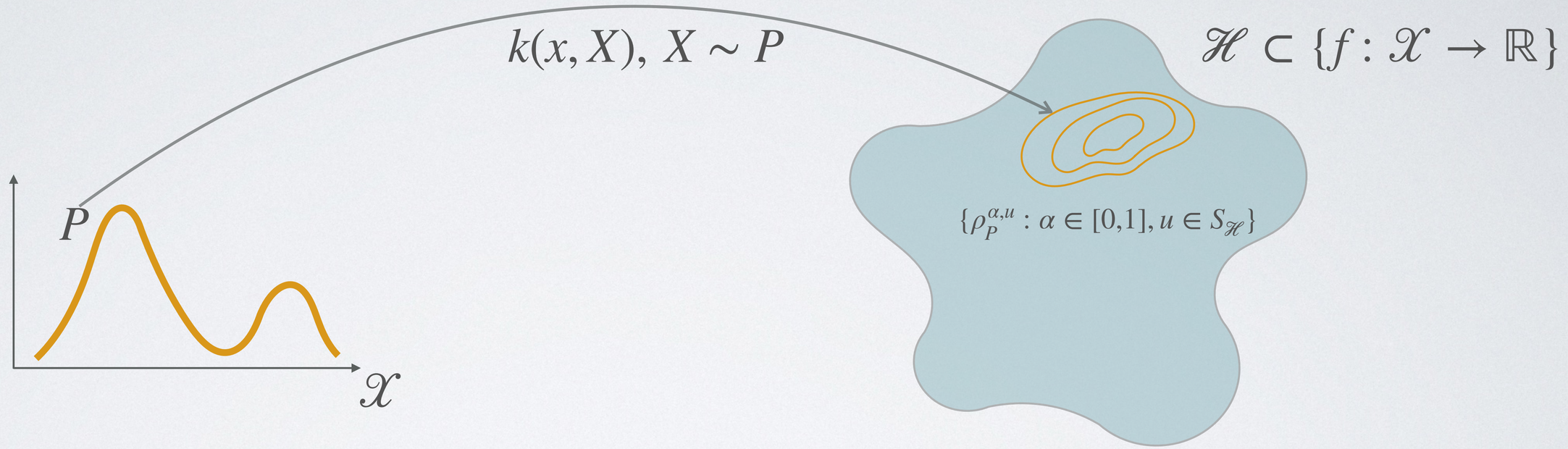
- extends naturally to infinite-dimensional Hilbert spaces!

distribution P on $\mathcal{X} \subseteq \mathbb{R}^2$

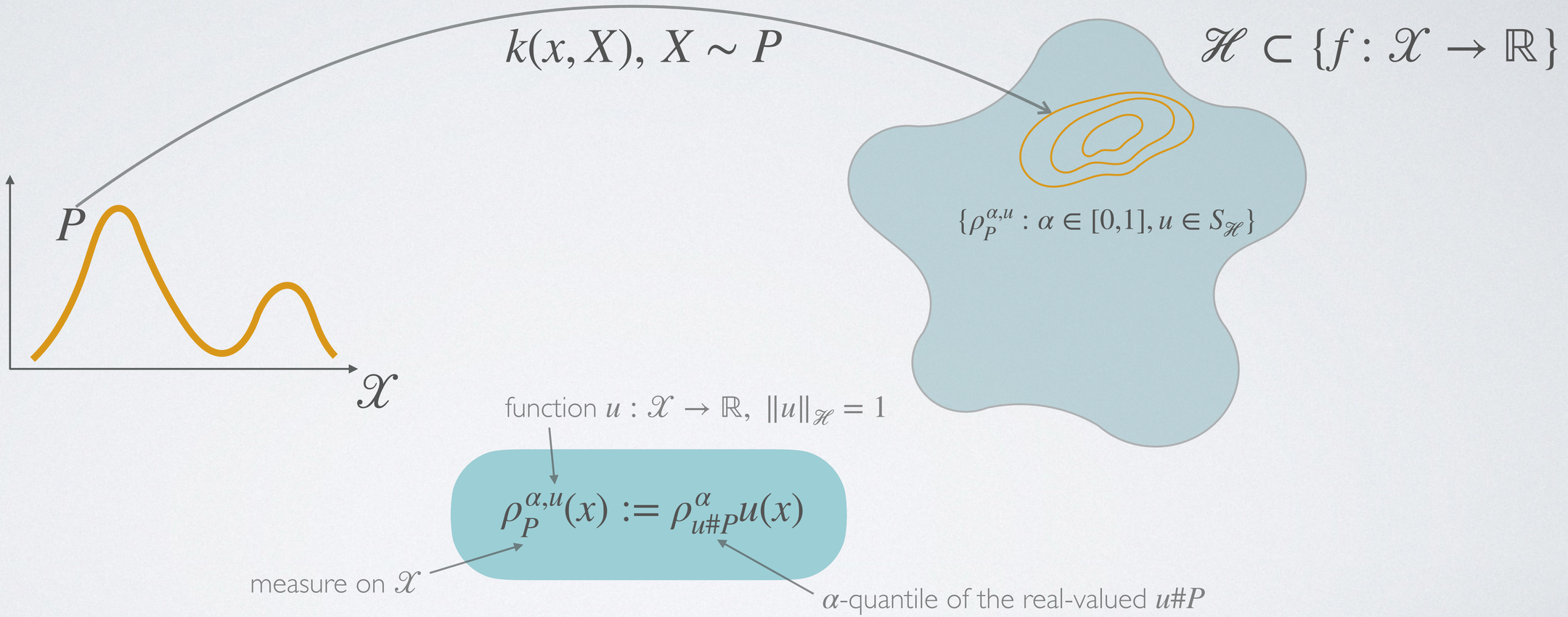




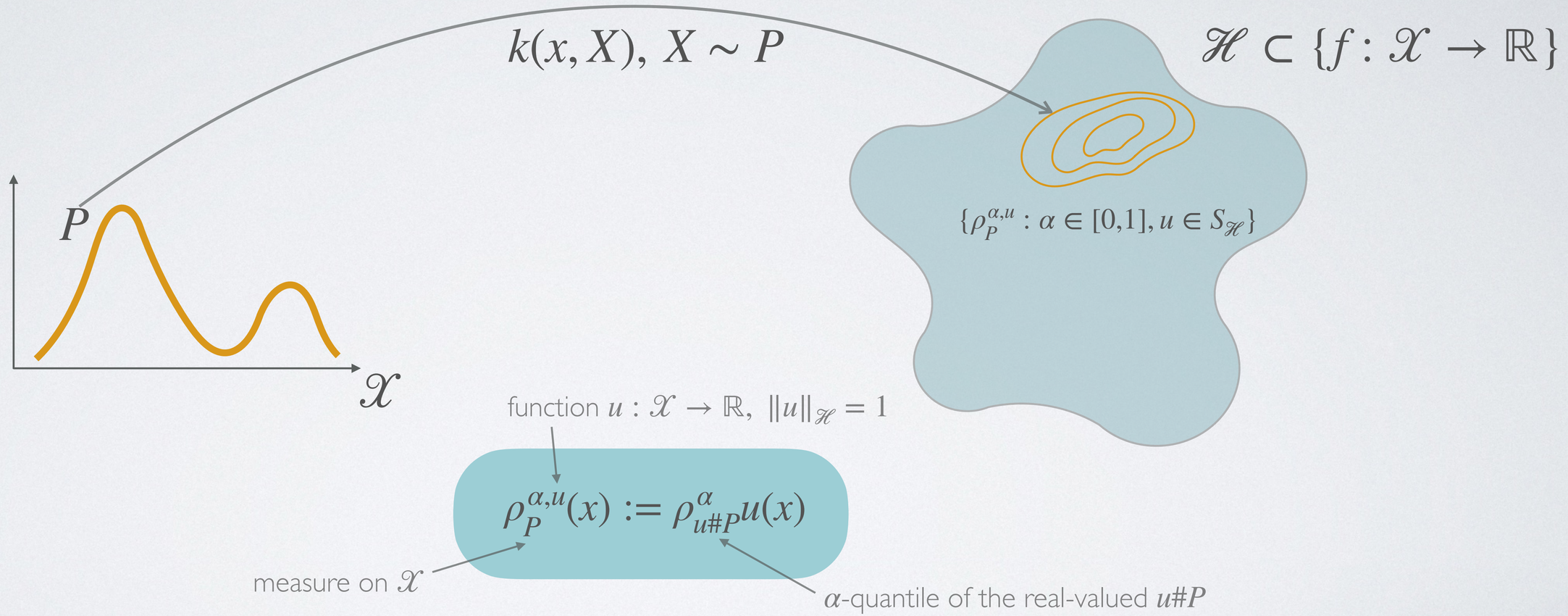
KERNEL QUANTILES



KERNEL QUANTILES



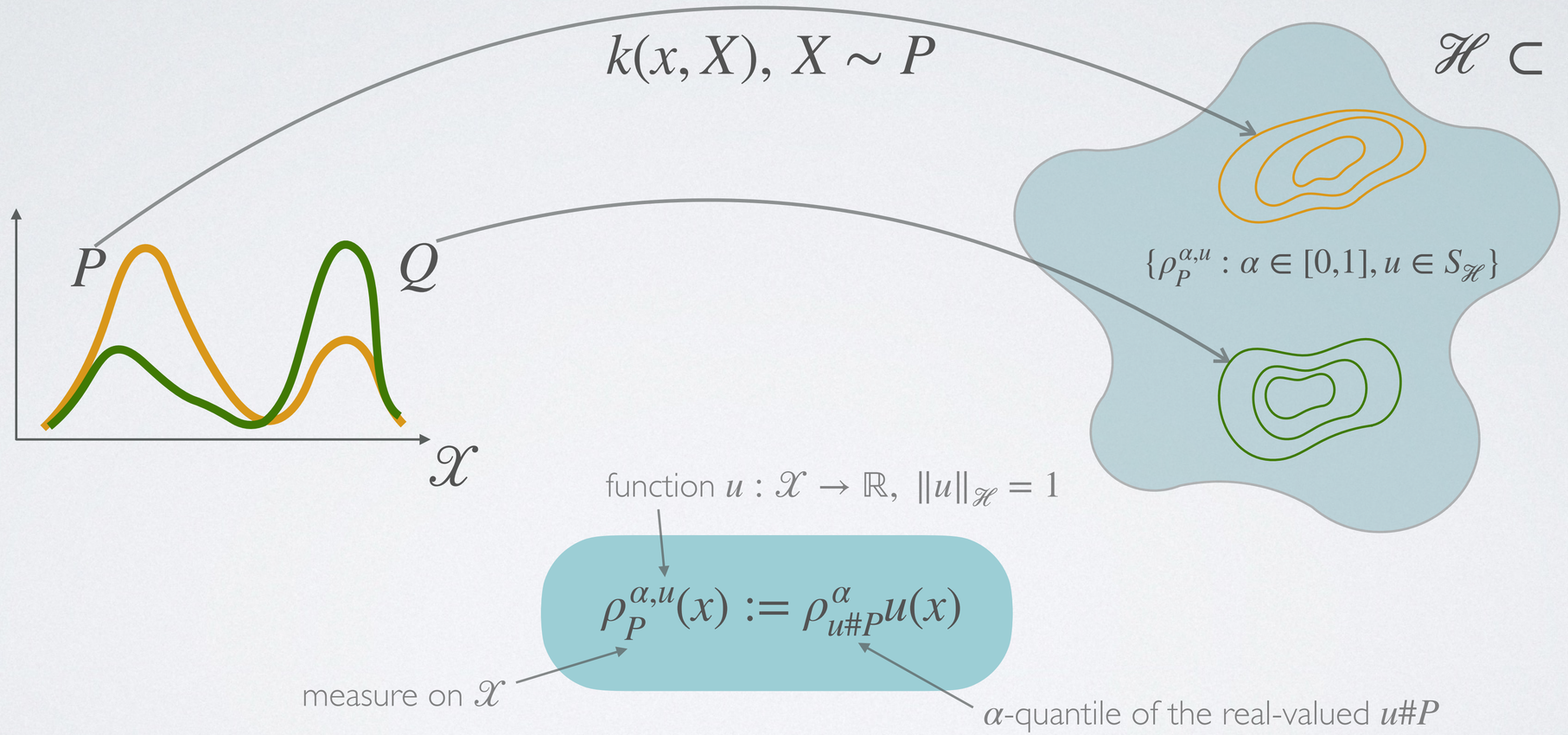
KERNEL QUANTILES



require no structure in $\mathscr{X}$—other than there being a kernel $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$!

KERNEL QUANTILES

…injective?



require no structure in $\mathscr{X}$—other than there being a kernel $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$!

QUANTILE-CHARACTERISTIC KERNELS

- $\mathcal{X}$ is Hausdorff, separable, and σ-compact,
- the kernel k is continuous, and separating on $\mathcal{X}$,

$$x \neq y \Rightarrow k(x, \cdot) \neq k(y, \cdot)$$

$\Rightarrow$

k is quantile-characteristic

($P \mapsto \{\rho_P^{\alpha,u} : \alpha \in [0,1],\ u \in S_{\mathcal{H}}\}$ is injective)

QUANTILE-CHARACTERISTIC KERNELS

- $\mathcal{X}$ is Hausdorff, separable, and σ-compact,
- the kernel k is continuous, and separating on $\mathcal{X}$,

$$x \neq y \Rightarrow k(x, \cdot) \neq k(y, \cdot)$$

$\Rightarrow$

k is quantile-characteristic

$(P \mapsto \{\rho_P^{\alpha,u} : \alpha \in [0,1],\ u \in S_{\mathcal{H}}\}$ is injective)

very mild assumptions!

- "k is mean-characteristic" is more challenging to establish…

MEAN- VS QUANTILE-CHARACTERISTIC

k is mean-characteristic $\Rightarrow$ k is quantile-characteristic

$\not\Leftarrow$

MEAN- VS QUANTILE-CHARACTERISTIC

k is mean-characteristic $\Rightarrow$ k is quantile-characteristic

$\not\Leftarrow$

Implication: for *any* discrepancy $D(P, Q)$ that aggregates k-quantiles injectively,

$\mathrm{MMD}(P, Q) > 0 \Rightarrow D(P, Q) > 0$

$\not\Leftarrow$

KERNEL QUANTILE DISCREPANCIES

Recipe:
1. for every u, aggregate the u-direction quantiles,
2. aggregate over u.

DIRECTIONAL DIFFERENCES

(aggregating over quantiles)

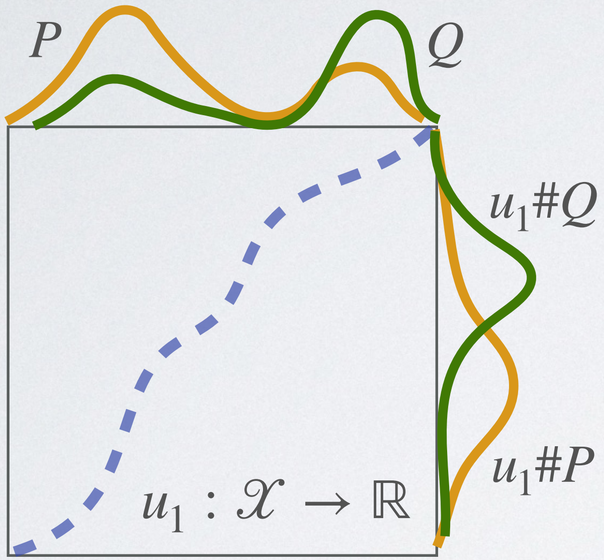
importance measure on quantiles

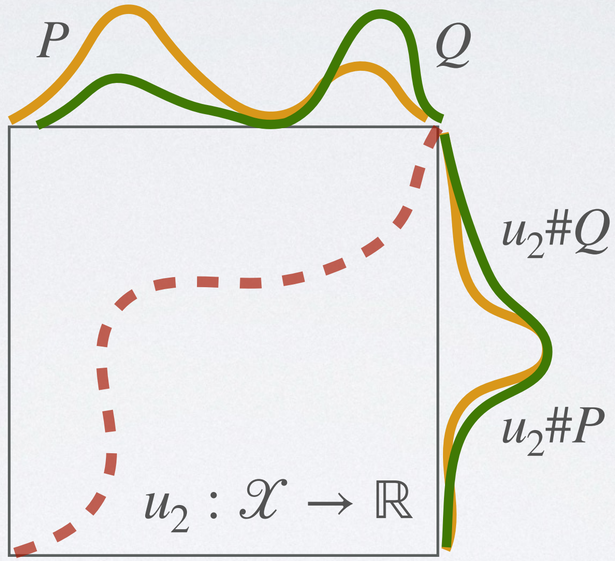
$$\tau_p(P, Q; \nu, u) = \left(\int_0^1 \left\| \rho_P^{\alpha,u} - \rho_Q^{\alpha,u} \right\|_{\mathscr{H}}^p \nu(\mathrm{d}\alpha) \right)^{1/p}$$

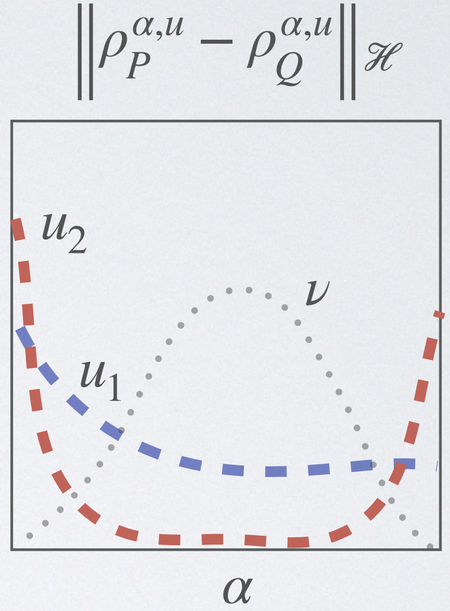
direction, $\|u\|_{\mathscr{H}} = 1$

DIRECTIONAL DIFFERENCES

(aggregating over quantiles)







importance measure on quantiles

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

direction, $\|u\|_{\mathcal{H}} = 1$

KERNEL QUANTILE DISCREPANCIES

aggregate the directional differences $\tau_p(P, Q; \nu, u) = \left(\int_0^1 \left\| \rho_P^{\alpha,u} - \rho_Q^{\alpha,u} \right\|_{\mathscr{H}}^p \nu(\mathrm{d}\alpha) \right)^{1/p}$ over $u \in S_{\mathscr{H}}$:

KERNEL QUANTILE DISCREPANCIES

aggregate the directional differences $\tau_p(P, Q; \nu, u) = \left(\int_0^1 \left\| \rho_P^{\alpha,u} - \rho_Q^{\alpha,u} \right\|_{\mathscr{H}}^p \nu(\mathrm{d}\alpha) \right)^{1/p}$ over $u \in S_{\mathscr{H}}$:

L^p aggregation:

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

measure over $S_{\mathscr{H}}$

KERNEL QUANTILE DISCREPANCIES

aggregate the directional differences $\tau_p(P, Q; \nu, u) = \left(\int_0^1 \left\| \rho_P^{\alpha,u} - \rho_Q^{\alpha,u} \right\|_{\mathscr{H}}^p \nu(\mathrm{d}\alpha) \right)^{1/p}$ over $u \in S_{\mathscr{H}}$:

L^p aggregation:

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

measure over $S_{\mathscr{H}}$

L^∞ aggregation:

$$\text{sup-KQD}_p(P, Q; \nu) = \left(\sup_{u \in S_{\mathscr{H}}} \left[\tau_p^p \left(P, Q; \nu, u \right) \right] \right)^{1/p}$$

KERNEL QUANTILE DISCREPANCIES

…are distances?

aggregate the directional differences $\tau_p(P, Q; \nu, u) = \left(\int_0^1 \left\| \rho_P^{\alpha,u} - \rho_Q^{\alpha,u} \right\|_{\mathscr{H}}^p \nu(\mathrm{d}\alpha) \right)^{1/p}$ over $u \in S_{\mathscr{H}}$:

L^p aggregation:

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

measure over $S_{\mathscr{H}}$

L^∞ aggregation:

$$\text{sup-KQD}_p(P, Q; \nu) = \left(\sup_{u \in S_{\mathscr{H}}} \left[\tau_p^p \left(P, Q; \nu, u \right) \right] \right)^{1/p}$$

KERNEL QUANTILE DISCREPANCIES

Recall:

- $\mathcal{X}$ is Hausdorff, separable, and σ-compact,
- the kernel k is continuous, and separating on $\mathcal{X}$, $x \neq y \Rightarrow k(x, \cdot) \neq k(y, \cdot)$

$\Rightarrow$

k is quantile-characteristic

KERNEL QUANTILE DISCREPANCIES

Recall:

- $\mathcal{X}$ is Hausdorff, separable, and σ-compact,
- the kernel k is continuous, and separating on $\mathcal{X}$,
  $x \neq y \Rightarrow k(x, \cdot) \neq k(y, \cdot)$

$\Rightarrow$

k is quantile-characteristic

When the quantiles are aggregated "invectively", we get a distance!

KERNEL QUANTILE DISCREPANCIES

- $\mathcal{X}$ is Hausdorff, separable, and σ-compact,
- the kernel k is continuous, and separating on $\mathcal{X}$, $x \neq y \Rightarrow k(x, \cdot) \neq k(y, \cdot)$
- ν has full support on $[0,1]$,

$\Rightarrow$

$$\text{sup-KQD}_p(P, Q; \nu) = \left(\sup_{u \in S_{\mathcal{H}}} \left[\tau_p^p \left(P, Q; \nu, u \right) \right] \right)^{1/p}$$

is a distance on the set of Borel probability measures on $\mathcal{X}$

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

KERNEL QUANTILE DISCREPANCIES

- $\mathcal{X}$ is Hausdorff, separable, and σ-compact,
- the kernel k is continuous, and separating on $\mathcal{X}$, $x \neq y \Rightarrow k(x, \cdot) \neq k(y, \cdot)$
- ν has full support on $[0,1]$,

$\Rightarrow$

$$\text{sup-KQD}_p(P, Q; \nu) = \left(\sup_{u \in S_{\mathcal{H}}} \left[\tau_p^p \left(P, Q; \nu, u \right) \right] \right)^{1/p}$$

is a distance on the set of Borel probability measures on $\mathcal{X}$

Further,

- γ has full support on the unit sphere $S_{\mathcal{H}}$,

$\Rightarrow$

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

is as well!

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

CONNECTIONS

$$\text{e-KQD}_p(P, Q; \nu, \gamma) = \left(\mathbb{E}_{u \sim \gamma} \left[\int_0^1 \left\| \rho_P^{\alpha,u} - \rho_Q^{\alpha,u} \right\|_{\mathscr{H}}^p \nu(\mathrm{d}\alpha) \right] \right)^{1/p}$$

$$\text{sup-KQD}_p(P, Q; \nu) = \left(\sup_{u \in S_{\mathscr{H}}} \left[\int_0^1 \left\| \rho_P^{\alpha,u} - \rho_Q^{\alpha,u} \right\|_{\mathscr{H}}^p \nu(\mathrm{d}\alpha) \right] \right)^{1/p}$$

- when $\mathscr{X} \subseteq \mathbb{R}^d$, $k(x, y)$ is linear, ν and γ are uniform, recover (max-)sliced Wasserstein.

CONNECTIONS

$$\text{e-KQD}_p(P, Q; \nu, \gamma) = \left(\mathbb{E}_{u\sim\gamma} \left[\int_0^1 \left\| \rho_P^{\alpha,u} - \rho_Q^{\alpha,u} \right\|_{\mathscr{H}}^p \nu(\mathrm{d}\alpha) \right] \right)^{1/p}$$

$$\text{sup-KQD}_p(P, Q; \nu) = \left(\sup_{u\in S_{\mathscr{H}}} \left[\int_0^1 \left\| \rho_P^{\alpha,u} - \rho_Q^{\alpha,u} \right\|_{\mathscr{H}}^p \nu(\mathrm{d}\alpha) \right] \right)^{1/p}$$

- when $\mathcal{X} \subseteq \mathbb{R}^d$, $k(x, y)$ is linear, ν and γ are uniform, recover (max-)sliced Wasserstein.
- generally, when ν is uniform, KQD are "kernelised" (max-)sliced Wasserstein.
  - kernel SW is new;
  - kernel max-SW was introduced in [1]

[1] Wang et al., Two-Sample Test with Kernel Projected Wasserstein Distance. AISTATS 2022.

ESTIMATORS AND EMPIRICAL STUDY

ESTIMATOR

(for e-KQD)

- Let $\{x_j\}_{j=1}^n \sim P, \{y_j\}_{j=1}^n \sim Q$. Denote $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i'}$ $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$

ESTIMATOR

(for e-KQD)

- Let $\{x_j\}_{j=1}^n \sim P, \{y_j\}_{j=1}^n \sim Q$. Denote $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \quad Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$
- Pick γ on $S_{\mathscr{H}}$, and sample $\{u_i\}_{i=1}^l \sim \gamma$.

$$\text{e-KQD}_p(P_n, Q_n; \nu, \gamma_l) = \frac{1}{ln}\sum_{i=1}^{l}\sum_{j=1}^{n}\left(\left[u_i(x_{1:n})\right]_j - \left[u_i(y_{1:n})\right]_j\right)^p f_\nu\left(\left\lceil\frac{j}{n}\right\rceil\right)$$

j'th smallest element of $\{u_i(x_1), \ldots u_i(x_n)\}$

ESTIMATOR

(for e-KQD)

- Let $\{x_j\}_{j=1}^n \sim P, \{y_j\}_{j=1}^n \sim Q$. Denote $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \quad Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$
- Pick γ on $S_{\mathscr{H}}$, and sample $\{u_i\}_{i=1}^l \sim \gamma$.

$$\text{e-KQD}_p(P_n, Q_n; \nu, \gamma_l) = \frac{1}{ln}\sum_{i=1}^{l}\sum_{j=1}^{n}\left(\left[u_i(x_{1:n})\right]_j - \left[u_i(y_{1:n})\right]_j\right)^p f_\nu\left(\left\lceil\frac{j}{n}\right\rceil\right)$$

j'th smallest element of $\{u_i(x_1), \ldots u_i(x_n)\}$

- This is a consistent estimator; more in the paper.

ESTIMATOR

(for e-KQD)

- Let $\{x_j\}_{j=1}^n \sim P, \{y_j\}_{j=1}^n \sim Q$. Denote $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \quad Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$
- Pick γ on $S_{\mathscr{H}}$, and sample $\{u_i\}_{i=1}^l \sim \gamma$. 🤔

measure on $S_{\mathscr{H}}$. No uniform when $\dim(\mathscr{H}) = \infty$…

$$\text{e-KQD}_p(P_n, Q_n; \nu, \gamma_l) = \frac{1}{ln}\sum_{i=1}^{l}\sum_{j=1}^{n}\left(\left[u_i(x_{1:n})\right]_j - \left[u_i(y_{1:n})\right]_j\right)^p f_\nu\left(\left\lceil\frac{j}{n}\right\rceil\right)$$

j'th smallest element of $\{u_i(x_1), \ldots u_i(x_n)\}$

- This is a consistent estimator; more in the paper.

ESTIMATOR

(for e-KQD)

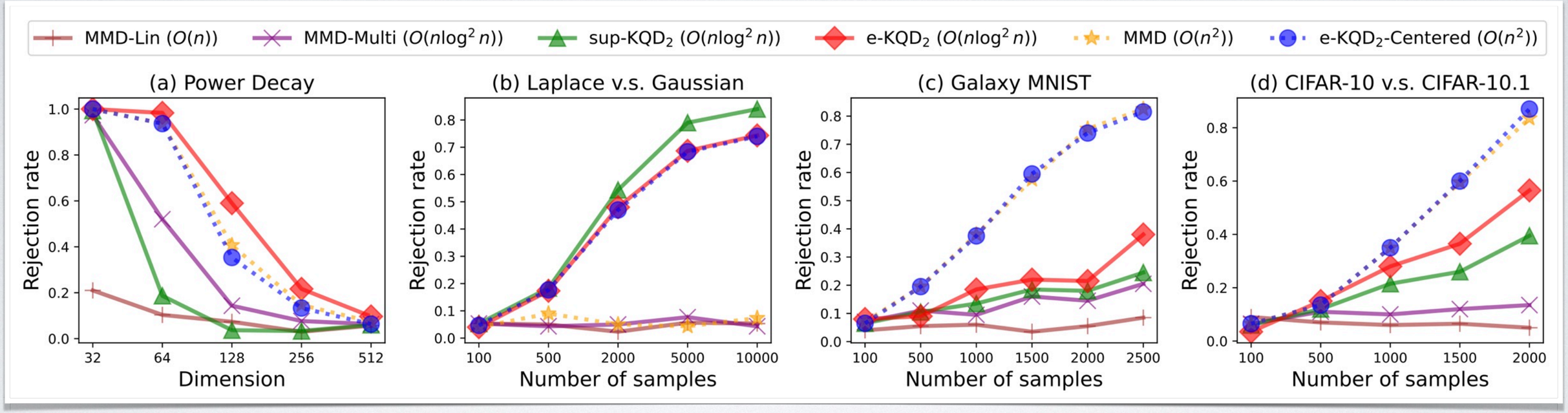
- Let $\{x_j\}_{j=1}^n \sim P, \{y_j\}_{j=1}^n \sim Q$. Denote $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \quad Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$
- Pick γ on $S_{\mathscr{H}}$, and sample $\{u_i\}_{i=1}^l \sim \gamma$.

$$\text{e-KQD}_p(P_n, Q_n; \nu, \gamma_l) = \frac{1}{ln}\sum_{i=1}^{l}\sum_{j=1}^{n}\left(\left[u_i(x_{1:n})\right]_j - \left[u_i(y_{1:n})\right]_j\right)^p f_\nu\left(\left\lceil\frac{j}{n}\right\rceil\right)$$

j'th smallest element of $\{u_i(x_1), \ldots u_i(x_n)\}$

- This is a consistent estimator; more in the paper.
- A projected Gaussian γ gives an $O(n\log^2 n)$ estimator when $l = \log n$.

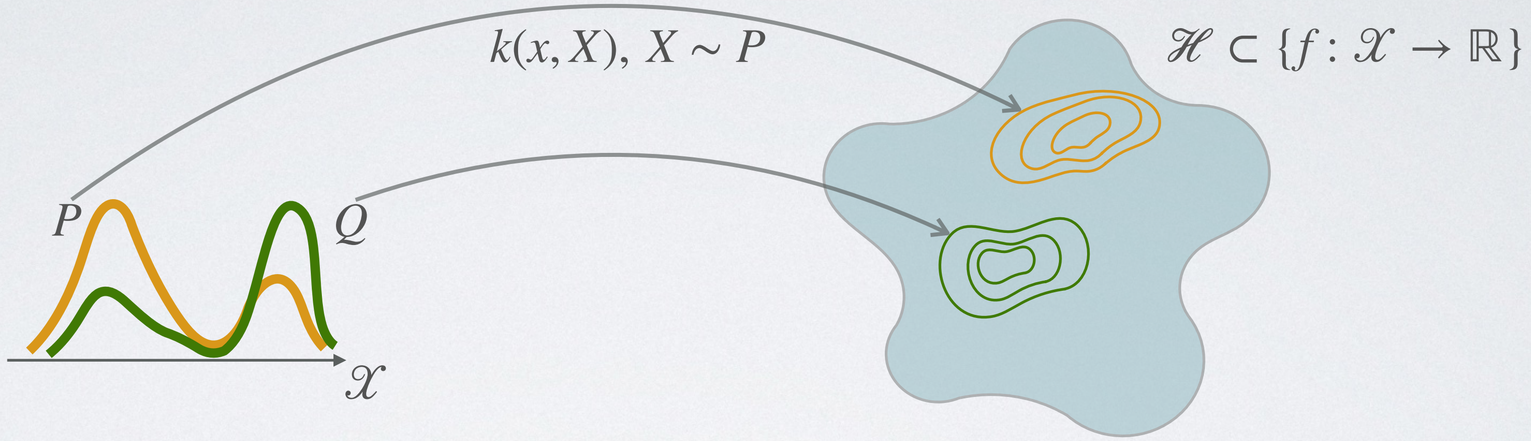
TWO-SAMPLE TESTING



- Within a comparable budget, KQD is competitive with MMD (for mean-characteristic kernels).
- KQD can be a distance for non-mean-characteristic kernels.
- Centered KQD offers the best of both worlds (KQD and MMD).

CONCLUSION

AND FUTURE WORK



- proposed *kernel quantile embeddings* (KQEs) and *kernel quantile discrepancies* (KQDs).
- established that KQD is a distance under conditions strictly more general than those needed for MMD.
- proposed an efficient estimator for e-KQD and demonstrate its effectiveness on two-sample testing problems.
- **next up**: kernel selection; conditional KQEs; applications beyond two-sample testing.