Modern Kernel Methods in Machine Learning: Part II

Danica J. Sutherland (she/her)
Computer Science, University of British Columbia
ETICS "summer" school, Oct 2022

Yesterday, we saw:

- RKHS $\mathcal{H}$ is a function space, $f : \mathcal{X} \to \mathbb{R}$
- Reproducing property: $\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x)$
- Representer theorem:
  $\operatorname{arg\,min} L(f(x_1), \ldots, f(x_n)) + R(\|f\|_{\mathcal{H}}) \in \operatorname{span}\{k(x_i, \cdot)\}_{i=1}^{n}$
- Can use to do kernel ridge regression, SVMs, etc

Today:

- Kernel mean embeddings of distributions
- Gaussian processes and probabilistic numerics
- Kernel approximations, for better computation
- Neural tangent kernels

Today:

- Kernel mean embeddings of distributions
- ~~Gaussian processes and probabilistic numerics~~
- Kernel approximations, for better computation
- Neural tangent kernels

Today:

- Kernel mean embeddings of distributions
- ~~Gaussian processes and probabilistic numerics~~
- ~~Kernel approximations, for better computation~~
- Neural tangent kernels

Today:

- Kernel mean embeddings of distributions
- ~~Gaussian processes and probabilistic numerics~~
- ~~Kernel approximations, for better computation~~
- ~~Neural tangent kernels~~

Mean embeddings of distributions

- Represent point $x \in \mathcal{X}$ as $k(x, \cdot)$: $\quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$

Mean embeddings of distributions

- Represent point $x \in \mathcal{X}$ as $k(x, \cdot)$: $\quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$
- Represent *distribution* $\mathbb{P}$ as $\mu_{\mathbb{P}}$: $\quad \mathbb{E}_{X \sim \mathbb{P}} f(X) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$

Mean embeddings of distributions

- Represent point $x \in \mathcal{X}$ as $k(x, \cdot)$: $\quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$
- Represent *distribution* $\mathbb{P}$ as $\mu_{\mathbb{P}}$: $\quad \mathbb{E}_{X \sim \mathbb{P}} f(X) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$

$$\mathbb{E}_{X \sim \mathbb{P}} f(X) = \mathbb{E}_{X \sim \mathbb{P}} \langle f, k(X, \cdot) \rangle_{\mathcal{H}} = \langle f, \mathbb{E}_{X \sim \mathbb{P}} k(X, \cdot) \rangle_{\mathcal{H}}$$

Mean embeddings of distributions

- Represent point $x \in \mathcal{X}$ as $k(x, \cdot)$: $\quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$
- Represent *distribution* $\mathbb{P}$ as $\mu_{\mathbb{P}}$: $\quad \mathbb{E}_{X \sim \mathbb{P}} f(X) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$

$$\mathbb{E}_{X \sim \mathbb{P}} f(X) = \mathbb{E}_{X \sim \mathbb{P}} \langle f, k(X, \cdot) \rangle_{\mathcal{H}} = \langle f, \mathbb{E}_{X \sim \mathbb{P}} k(X, \cdot) \rangle_{\mathcal{H}}$$

Mean embeddings of distributions

- Represent point $x \in \mathcal{X}$ as $k(x, \cdot)$: $\quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$
- Represent *distribution* $\mathbb{P}$ as $\mu_{\mathbb{P}}$: $\quad \mathbb{E}_{X \sim \mathbb{P}} f(X) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$

$$\mathbb{E}_{X \sim \mathbb{P}} f(X) = \mathbb{E}_{X \sim \mathbb{P}} \langle f, k(X, \cdot) \rangle_{\mathcal{H}} = \langle f, \mathbb{E}_{X \sim \mathbb{P}} k(X, \cdot) \rangle_{\mathcal{H}}$$

  - Last step assumed e.g. $\mathbb{E} \sqrt{k(X, X)} < \infty$

Mean embeddings of distributions

- Represent point $x \in \mathcal{X}$ as $k(x, \cdot)$: $\quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$
- Represent *distribution* $\mathbb{P}$ as $\mu_{\mathbb{P}}$: $\quad \mathbb{E}_{X \sim \mathbb{P}} f(X) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$

$$\mathbb{E}_{X \sim \mathbb{P}} f(X) = \mathbb{E}_{X \sim \mathbb{P}} \langle f, k(X, \cdot) \rangle_{\mathcal{H}} = \langle f, \mathbb{E}_{X \sim \mathbb{P}} k(X, \cdot) \rangle_{\mathcal{H}}$$

  - Last step assumed e.g. $\mathbb{E} \sqrt{k(X, X)} < \infty$
- $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \mathbb{E}_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} k(X, Y)$

Mean embeddings of distributions

- Represent point $x \in \mathcal{X}$ as $k(x, \cdot)$: $\quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$
- Represent *distribution* $\mathbb{P}$ as $\mu_{\mathbb{P}}$: $\quad \mathbb{E}_{X \sim \mathbb{P}} f(X) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$

$$\mathbb{E}_{X \sim \mathbb{P}} f(X) = \mathbb{E}_{X \sim \mathbb{P}} \langle f, k(X, \cdot) \rangle_{\mathcal{H}} = \langle f, \mathbb{E}_{X \sim \mathbb{P}} k(X, \cdot) \rangle_{\mathcal{H}}$$

  - Last step assumed e.g. $\mathbb{E} \sqrt{k(X, X)} < \infty$
- $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \mathbb{E}_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} k(X, Y)$
- Okay. Why?

Mean embeddings of distributions

- Represent point $x \in \mathcal{X}$ as $k(x, \cdot)$: $\quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$
- Represent *distribution* $\mathbb{P}$ as $\mu_{\mathbb{P}}$: $\quad \mathbb{E}_{X \sim \mathbb{P}} f(X) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$

$$\mathbb{E}_{X \sim \mathbb{P}} f(X) = \mathbb{E}_{X \sim \mathbb{P}} \langle f, k(X, \cdot) \rangle_{\mathcal{H}} = \langle f, \mathbb{E}_{X \sim \mathbb{P}} k(X, \cdot) \rangle_{\mathcal{H}}$$

  - Last step assumed e.g. $\mathbb{E} \sqrt{k(X, X)} < \infty$
- $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \mathbb{E}_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} k(X, Y)$
- Okay. Why?
  - One reason: ML on distributions [Szabó+ JMLR-16]

Mean embeddings of distributions

- Represent point $x \in \mathcal{X}$ as $k(x, \cdot)$: $\quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$
- Represent *distribution* $\mathbb{P}$ as $\mu_{\mathbb{P}}$: $\quad \mathbb{E}_{X \sim \mathbb{P}} f(X) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$

$$\mathbb{E}_{X \sim \mathbb{P}} f(X) = \mathbb{E}_{X \sim \mathbb{P}} \langle f, k(X, \cdot) \rangle_{\mathcal{H}} = \langle f, \mathbb{E}_{X \sim \mathbb{P}} k(X, \cdot) \rangle_{\mathcal{H}}$$

  - Last step assumed e.g. $\mathbb{E} \sqrt{k(X, X)} < \infty$
- $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \mathbb{E}_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} k(X, Y)$
- Okay. Why?
  - One reason: ML on distributions [Szabó+ JMLR-16]
  - More common reason: comparing distributions

Maximum Mean Discrepancy

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

Maximum Mean Discrepancy

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}\end{aligned}$$

Maximum Mean Discrepancy

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}} f(X) - \mathbb{E}_{Y \sim \mathbb{Q}} f(Y)\end{aligned}$$

- Last line is Integral Probability Metric (IPM) form

Maximum Mean Discrepancy

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}\, f(X) - \mathbb{E}_{Y \sim \mathbb{Q}}\, f(Y)\end{aligned}$$

- Last line is Integral Probability Metric (IPM) form
- f is called "witness function" or "critic": high on $\mathbb{P}$, low on $\mathbb{Q}$

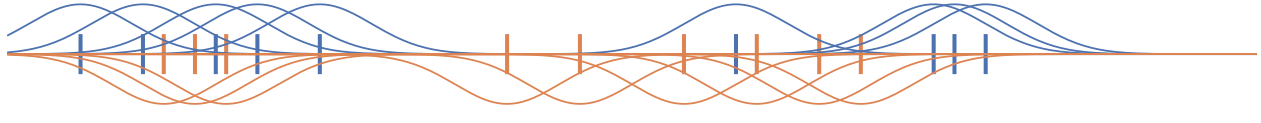
$$f^*(t) \propto \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, k(t, \cdot) \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}\, k(t, X) - \mathbb{E}_{\mathbb{Q}}\, k(t, Y)$$

Maximum Mean Discrepancy

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}\, f(X) - \mathbb{E}_{Y \sim \mathbb{Q}}\, f(Y)\end{aligned}$$

- Last line is Integral Probability Metric (IPM) form
- f is called "witness function" or "critic": high on $\mathbb{P}$, low on $\mathbb{Q}$

$$f^*(t) \propto \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, k(t, \cdot) \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}\, k(t, X) - \mathbb{E}_{\mathbb{Q}}\, k(t, Y)$$

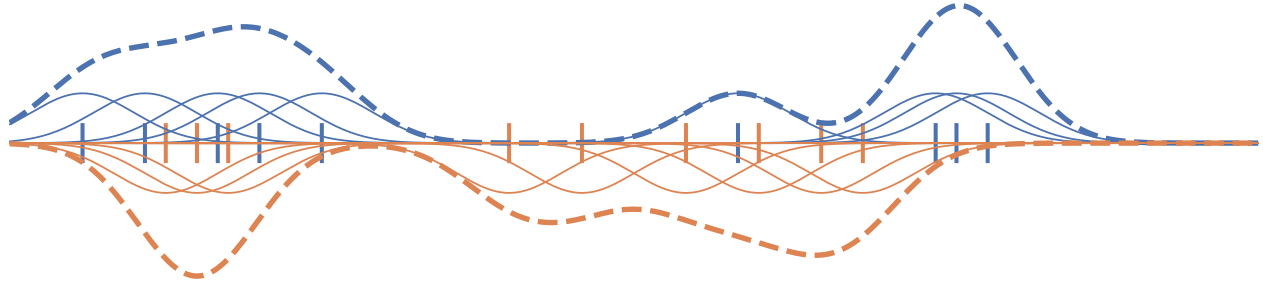


Maximum Mean Discrepancy

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}\, f(X) - \mathbb{E}_{Y \sim \mathbb{Q}}\, f(Y)\end{aligned}$$

- Last line is Integral Probability Metric (IPM) form
- f is called "witness function" or "critic": high on $\mathbb{P}$, low on $\mathbb{Q}$

$$f^*(t) \propto \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, k(t, \cdot) \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}\, k(t, X) - \mathbb{E}_{\mathbb{Q}}\, k(t, Y)$$

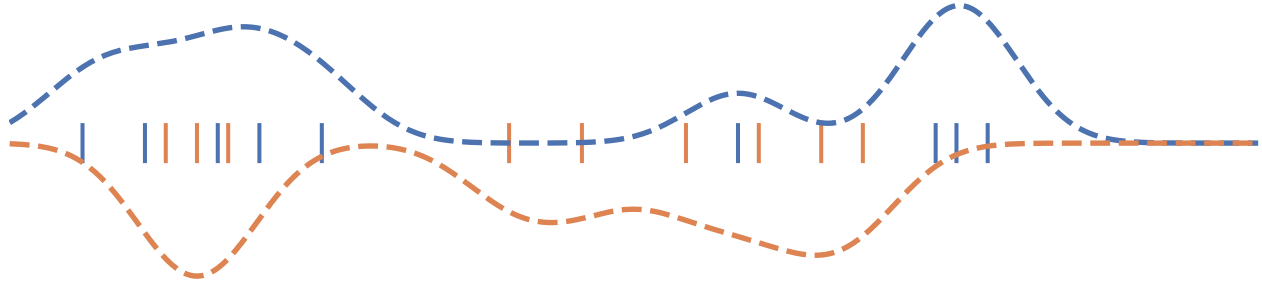


Maximum Mean Discrepancy

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}\, f(X) - \mathbb{E}_{Y \sim \mathbb{Q}}\, f(Y)\end{aligned}$$

- Last line is Integral Probability Metric (IPM) form
- f is called "witness function" or "critic": high on $\mathbb{P}$, low on $\mathbb{Q}$

$$f^*(t) \propto \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, k(t, \cdot) \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}\, k(t, X) - \mathbb{E}_{\mathbb{Q}}\, k(t, Y)$$

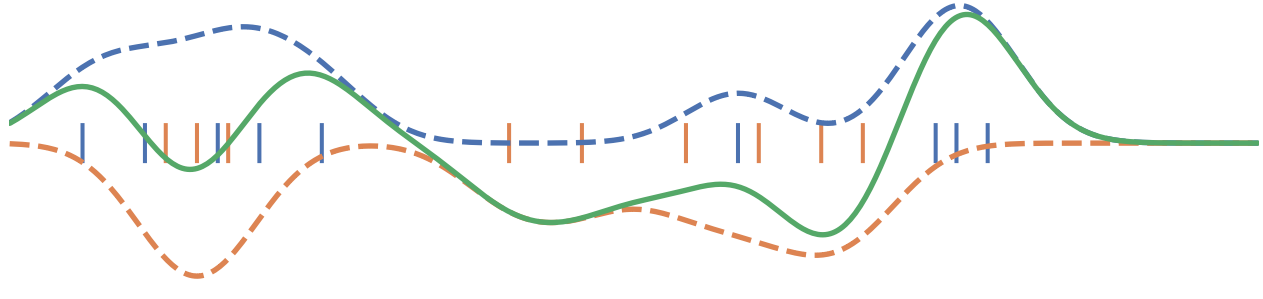


Maximum Mean Discrepancy

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}\, f(X) - \mathbb{E}_{Y \sim \mathbb{Q}}\, f(Y)\end{aligned}$$

- Last line is Integral Probability Metric (IPM) form
- f is called "witness function" or "critic": high on $\mathbb{P}$, low on $\mathbb{Q}$

$$f^*(t) \propto \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, k(t, \cdot) \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}\, k(t, X) - \mathbb{E}_{\mathbb{Q}}\, k(t, Y)$$

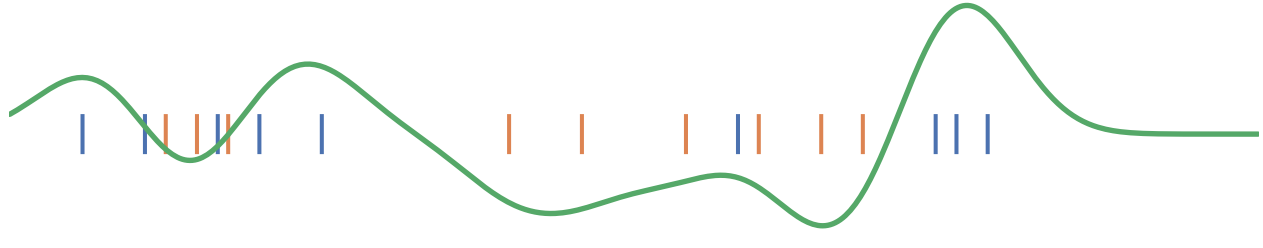


Maximum Mean Discrepancy

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}\, f(X) - \mathbb{E}_{Y \sim \mathbb{Q}}\, f(Y)\end{aligned}$$

- Last line is Integral Probability Metric (IPM) form
- f is called "witness function" or "critic": high on $\mathbb{P}$, low on $\mathbb{Q}$

$$f^*(t) \propto \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, k(t, \cdot) \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}\, k(t, X) - \mathbb{E}_{\mathbb{Q}}\, k(t, Y)$$

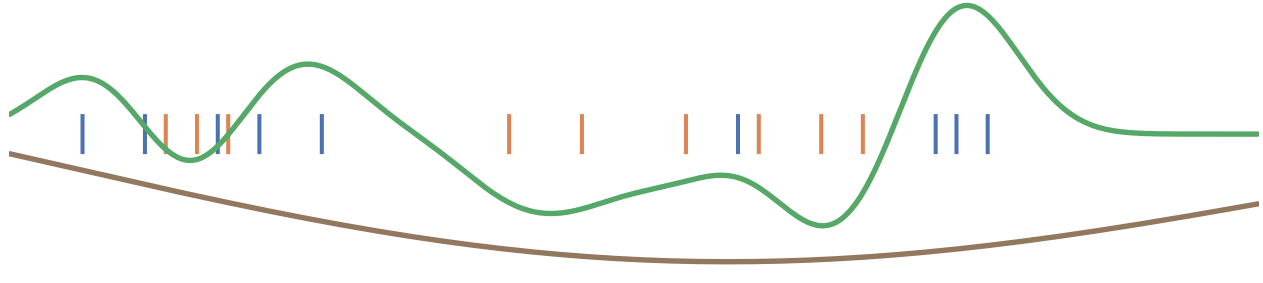


Maximum Mean Discrepancy

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}\, f(X) - \mathbb{E}_{Y \sim \mathbb{Q}}\, f(Y)\end{aligned}$$

- Last line is Integral Probability Metric (IPM) form
- f is called "witness function" or "critic": high on $\mathbb{P}$, low on $\mathbb{Q}$

$$f^*(t) \propto \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, k(t, \cdot) \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}\, k(t, X) - \mathbb{E}_{\mathbb{Q}}\, k(t, Y)$$

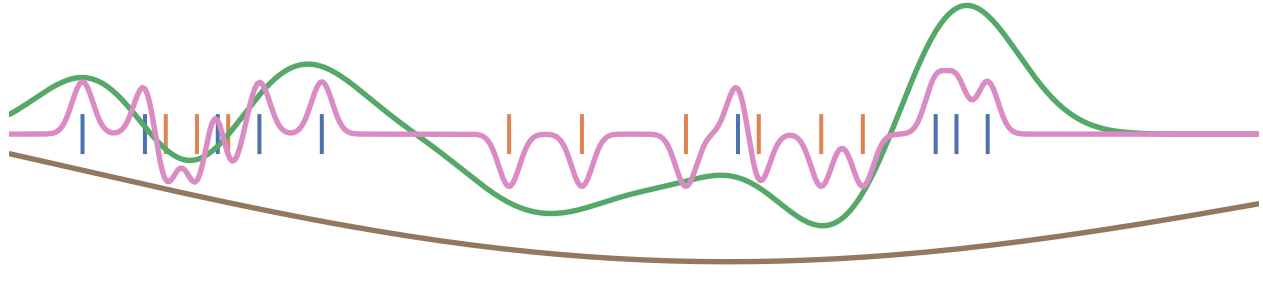


Maximum Mean Discrepancy

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}\, f(X) - \mathbb{E}_{Y \sim \mathbb{Q}}\, f(Y)\end{aligned}$$

- Last line is Integral Probability Metric (IPM) form
- f is called "witness function" or "critic": high on $\mathbb{P}$, low on $\mathbb{Q}$

$$f^*(t) \propto \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, k(t, \cdot) \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}\, k(t, X) - \mathbb{E}_{\mathbb{Q}}\, k(t, Y)$$



MMD properties

- $\mathrm{MMD}(\mathbb{P}, \mathbb{P}) = 0$, symmetry, triangle inequality

MMD properties

- $\mathrm{MMD}(\mathbb{P}, \mathbb{P}) = 0$, symmetry, triangle inequality
- If k is *characteristic*, then $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$
  - i.e. $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective

MMD properties

- $\mathrm{MMD}(\mathbb{P}, \mathbb{P}) = 0$, symmetry, triangle inequality
- If k is *characteristic*, then $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$
  - i.e. $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective
  - Makes MMD a metric on probability distributions

MMD properties

- $\mathrm{MMD}(\mathbb{P}, \mathbb{P}) = 0$, symmetry, triangle inequality
- If $\boldsymbol{k}$ is *characteristic*, then $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$
  - i.e. $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective
  - Makes MMD a metric on probability distributions
  - Universal => characteristic

MMD properties

- $\mathbf{MMD}(\mathbb{P}, \mathbb{P}) = 0$, symmetry, triangle inequality
- If $\boldsymbol{k}$ is *characteristic*, then $\mathbf{MMD}(\mathbb{P}, \mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$
  - i.e. $\mathbb{P} \mapsto \boldsymbol{\mu}_{\mathbb{P}}$ is injective
  - Makes MMD a metric on probability distributions
  - Universal => characteristic
- Linear kernel: $\mathbf{MMD}(\mathbb{P}, \mathbb{Q}) = \|\boldsymbol{\mu}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{Q}}\|_{\mathcal{H}}$ is just Euclidean distance between means

Application: Kernel Herding

- Want a "super-sample" from $\mathbb{P}$: $\frac{1}{n}\sum_i f(X_i) \approx \mathbb{E}\, f(X)$

Application: Kernel Herding

- Want a "super-sample" from $\mathbb{P}$: $\frac{1}{n}\sum_i f(X_i) \approx \mathbb{E}\, f(X)$
- If $f \in \mathcal{H}$, error $\leq \|f\|_{\mathcal{H}}\, \mathrm{MMD}(\mathbb{P}, \frac{1}{n}\sum_{i=1}^{T} \delta_{X_i})$

Application: Kernel Herding

- Want a "super-sample" from $\mathbb{P}$: $\frac{1}{n}\sum_i f(X_i) \approx \mathbb{E}\, f(X)$
- If $f \in \mathcal{H}$, error $\leq \|f\|_{\mathcal{H}}\, \mathbf{MMD}(\mathbb{P}, \frac{1}{n}\sum_{i=1}^{T} \delta_{X_i})$
- Greedily minimize the MMD:

$$X_{T+1} \in \operatorname*{arg\,min}_{X \in \mathcal{X}} \mathbb{E}_{X' \sim \mathbb{P}}\, k(X, X') - \frac{1}{T+1}\sum_{i=1}^{T} k(X, X_i)$$

Application: Kernel Herding

- Want a "super-sample" from $\mathbb{P}$: $\frac{1}{n}\sum_i f(X_i) \approx \mathbb{E}\, f(X)$
- If $f \in \mathcal{H}$, error $\leq \|f\|_{\mathcal{H}} \operatorname{MMD}(\mathbb{P}, \frac{1}{n}\sum_{i=1}^{T} \delta_{X_i})$
- Greedily minimize the MMD:

$$X_{T+1} \in \operatorname*{arg\,min}_{X \in \mathcal{X}} \mathbb{E}_{X' \sim \mathbb{P}}\, k(X, X') - \frac{1}{T+1}\sum_{i=1}^{T} k(X, X_i)$$

- Get $\mathcal{O}(1/T)$ approximation instead of $\mathcal{O}(1/\sqrt{T})$ with random samples

A g

- Want a "supe $\approx \mathbb{E}\, f(X)$
- If $f \in \mathcal{H}$, erro $X_i)$
- Greedily mini

$$X_{T+1} \in \arg \max_{X \in \mathcal{X}} \sum_{i=1}^{T} k(X, X_i)$$

- Get $\mathcal{O}(1/T)$ $\sqrt{T})$ with random samples

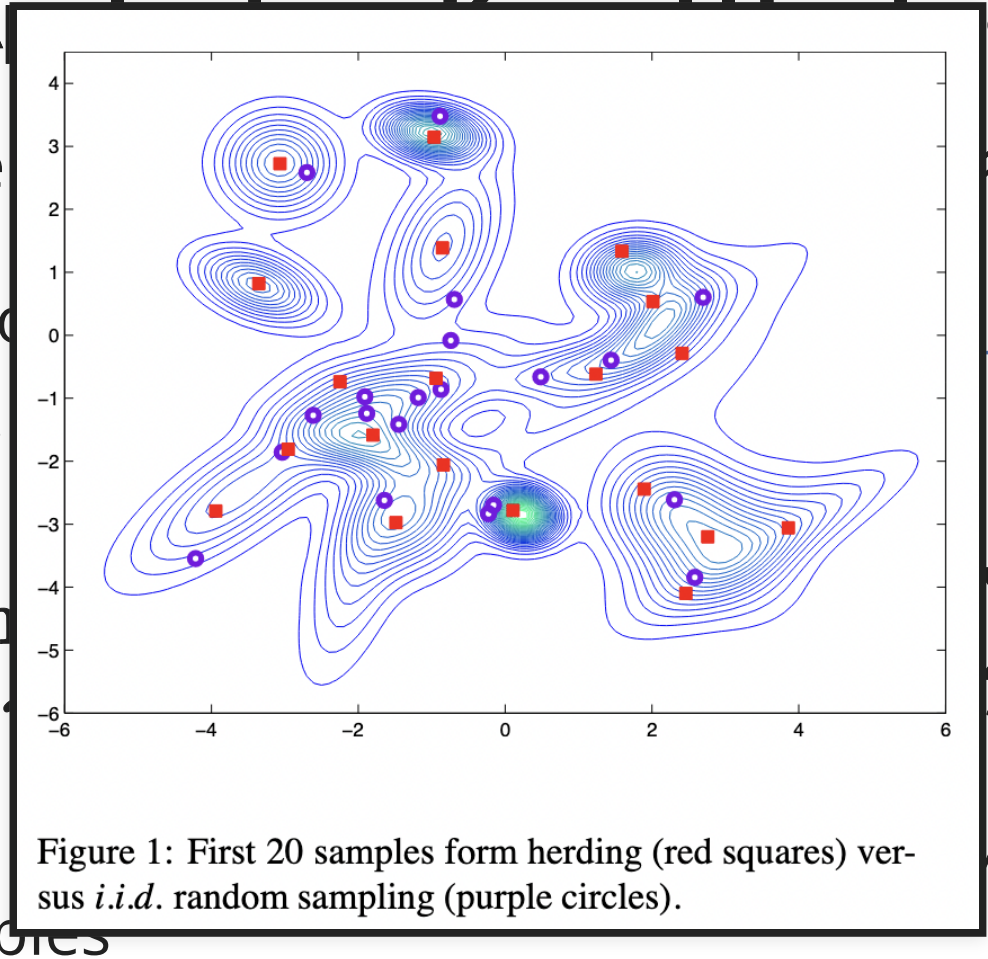


Figure 1: First 20 samples form herding (red squares) versus *i.i.d.* random sampling (purple circles).

Estimating MMD from samples

$$\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\substack{X, X' \sim \mathbb{P} \\ Y, Y' \sim \mathbb{Q}}} \left[k(X, X') - 2k(X, Y) + k(Y, Y') \right]$$

Estimating MMD from samples

$$\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\substack{X, X' \sim \mathbb{P} \\ Y, Y' \sim \mathbb{Q}}} \left[k(X, X') - 2k(X, Y) + k(Y, Y') \right]$$

$$\widehat{\mathrm{MMD}}_k^2(X, Y) = \mathrm{mean}(K_{XX}) + \mathrm{mean}(K_{YY}) - 2\,\mathrm{mean}(K_{XY})$$

Estimating MMD from samples

$$\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\substack{X, X' \sim \mathbb{P} \\ Y, Y' \sim \mathbb{Q}}} \left[k(X, X') - 2k(X, Y) + k(Y, Y') \right]$$

$$\widehat{\mathrm{MMD}}_k^2(X, Y) = \mathrm{mean}(K_{XX}) + \mathrm{mean}(K_{YY}) - 2\,\mathrm{mean}(K_{XY})$$

K_{XX}

Estimating MMD from samples

$$\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\substack{X, X' \sim \mathbb{P} \\ Y, Y' \sim \mathbb{Q}}} \left[k(X, X') - 2k(X, Y) + k(Y, Y') \right]$$

$$\widehat{\mathrm{MMD}}_k^2(X, Y) = \mathrm{mean}(K_{XX}) + \mathrm{mean}(K_{YY}) - 2\,\mathrm{mean}(K_{XY})$$

K_{XX}

K_{YY}

Estimating MMD from samples

$$\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\substack{X, X' \sim \mathbb{P} \\ Y, Y' \sim \mathbb{Q}}} [k(X, X') - 2k(X, Y) + k(Y, Y')]$$

$$\widehat{\mathrm{MMD}}_k^2(X, Y) = \mathrm{mean}(K_{XX}) + \mathrm{mean}(K_{YY}) - 2\,\mathrm{mean}(K_{XY})$$

K_{XX}

K_{YY}

K_{XY}

MMD vs other distances

- MMD has easy $\mathcal{O}(n^2)$ estimator
  - *block* or *incomplete* estimators are $\mathcal{O}(n^\alpha)$ for $\alpha \in [1, 2]$, but noisier

MMD vs other distances

- MMD has easy $\mathcal{O}(n^2)$ estimator
  - *block* or *incomplete* estimators are $\mathcal{O}(n^\alpha)$ for $\alpha \in [1, 2]$, but noisier
- For bounded kernel, $\mathcal{O}_p(1/\sqrt{n})$ estimation error

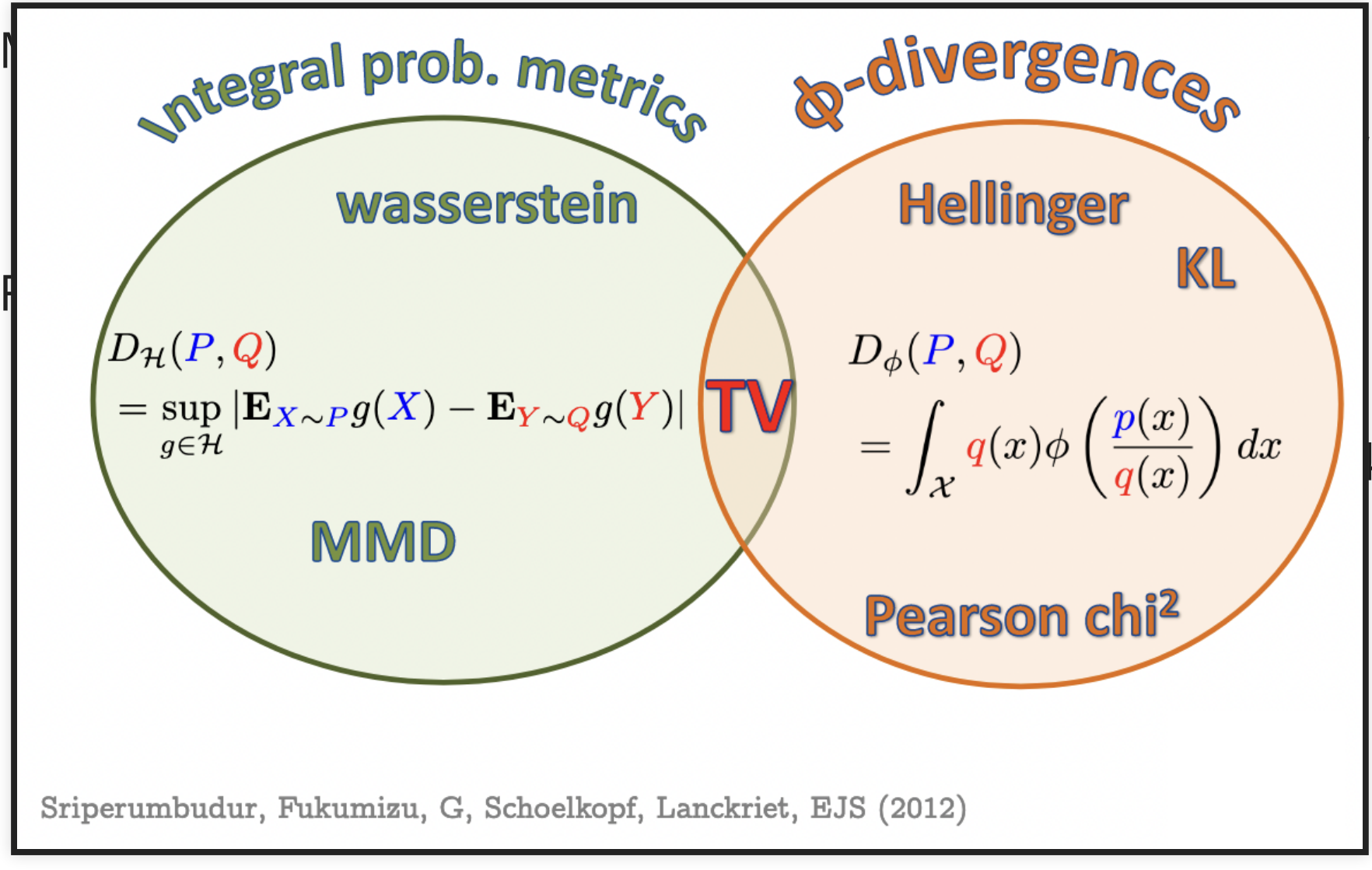
MMD vs other distances

- MMD has easy $\mathcal{O}(n^2)$ estimator
  - *block* or *incomplete* estimators are $\mathcal{O}(n^\alpha)$ for $\alpha \in [1, 2]$, but noisier
- For bounded kernel, $\mathcal{O}_p(1/\sqrt{n})$ estimation error
  - Independent of data dimension!

MMD vs other distances

- MMD has easy $\mathcal{O}(n^2)$ estimator
  - *block* or *incomplete* estimators are $\mathcal{O}(n^\alpha)$ for $\alpha \in [1, 2]$, but noisier
- For bounded kernel, $\mathcal{O}_p(1/\sqrt{n})$ estimation error
  - Independent of data dimension!
  - But, no free lunch...the *value* of the MMD generally shrinks with growing dimension, so constant $\mathcal{O}_p(1/\sqrt{n})$ error gets worse relatively

MMD vs other distances



Application: Two-sample testing

- Given samples from two unknown distributions

$$X \sim \mathbb{P} \qquad Y \sim \mathbb{Q}$$

- Question: is $\mathbb{P} = \mathbb{Q}$?

Application: Two-sample testing

- Given samples from two unknown distributions

$$X \sim \mathbb{P} \qquad Y \sim \mathbb{Q}$$

- Do smokers/non-smokers get different cancers?

Application: Two-sample testing

- Given samples from two unknown distributions

$$X \sim \mathbb{P} \qquad Y \sim \mathbb{Q}$$

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?

Application: Two-sample testing

- Given samples from two unknown distributions

$$X \sim \mathbb{P} \qquad Y \sim \mathbb{Q}$$

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?
- When does my laser agree with the one on Mars?

Application: Two-sample testing

- Given samples from two unknown distributions

$$X \sim \mathbb{P} \qquad Y \sim \mathbb{Q}$$

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?
- When does my laser agree with the one on Mars?
- Are storms in the 2000s different from storms in the 1800s?

Application: Two-sample testing

- Given samples from two unknown distributions

$$X \sim \mathbb{P} \qquad Y \sim \mathbb{Q}$$

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?
- When does my laser agree with the one on Mars?
- Are storms in the 2000s different from storms in the 1800s?
- Does presence of this protein affect DNA binding? [MMDiff2]

Application: Two-sample testing

- Given samples from two unknown distributions

$$X \sim \mathbb{P} \qquad Y \sim \mathbb{Q}$$

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?
- When does my laser agree with the one on Mars?
- Are storms in the 2000s different from storms in the 1800s?
- Does presence of this protein affect DNA binding? [MMDiff2]
- Do these `dob` and `birthday` columns mean the same thing?

Application: Two-sample testing

- Given samples from two unknown distributions

$$X \sim \mathbb{P} \qquad Y \sim \mathbb{Q}$$

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?
- When does my laser agree with the one on Mars?
- Are storms in the 2000s different from storms in the 1800s?
- Does presence of this protein affect DNA binding? [MMDiff2]
- Do these `dob` and `birthday` columns mean the same thing?
- Does my generative model $\mathbb{Q}_\theta$ match $\mathbb{P}_{\mathbf{data}}$?

Application: Two-sample testing

- Given samples from two unknown distributions

$$X \sim \mathbb{P} \qquad Y \sim \mathbb{Q}$$

- Question: is $\mathbb{P} = \mathbb{Q}$?

Application: Two-sample testing

- Given samples from two unknown distributions

$$X \sim \mathbb{P} \qquad Y \sim \mathbb{Q}$$

- Question: is $\mathbb{P} = \mathbb{Q}$?
- Hypothesis testing approach:

$$H_0 : \mathbb{P} = \mathbb{Q} \qquad H_1 : \mathbb{P} \neq \mathbb{Q}$$

Application: Two-sample testing

- Given samples from two unknown distributions

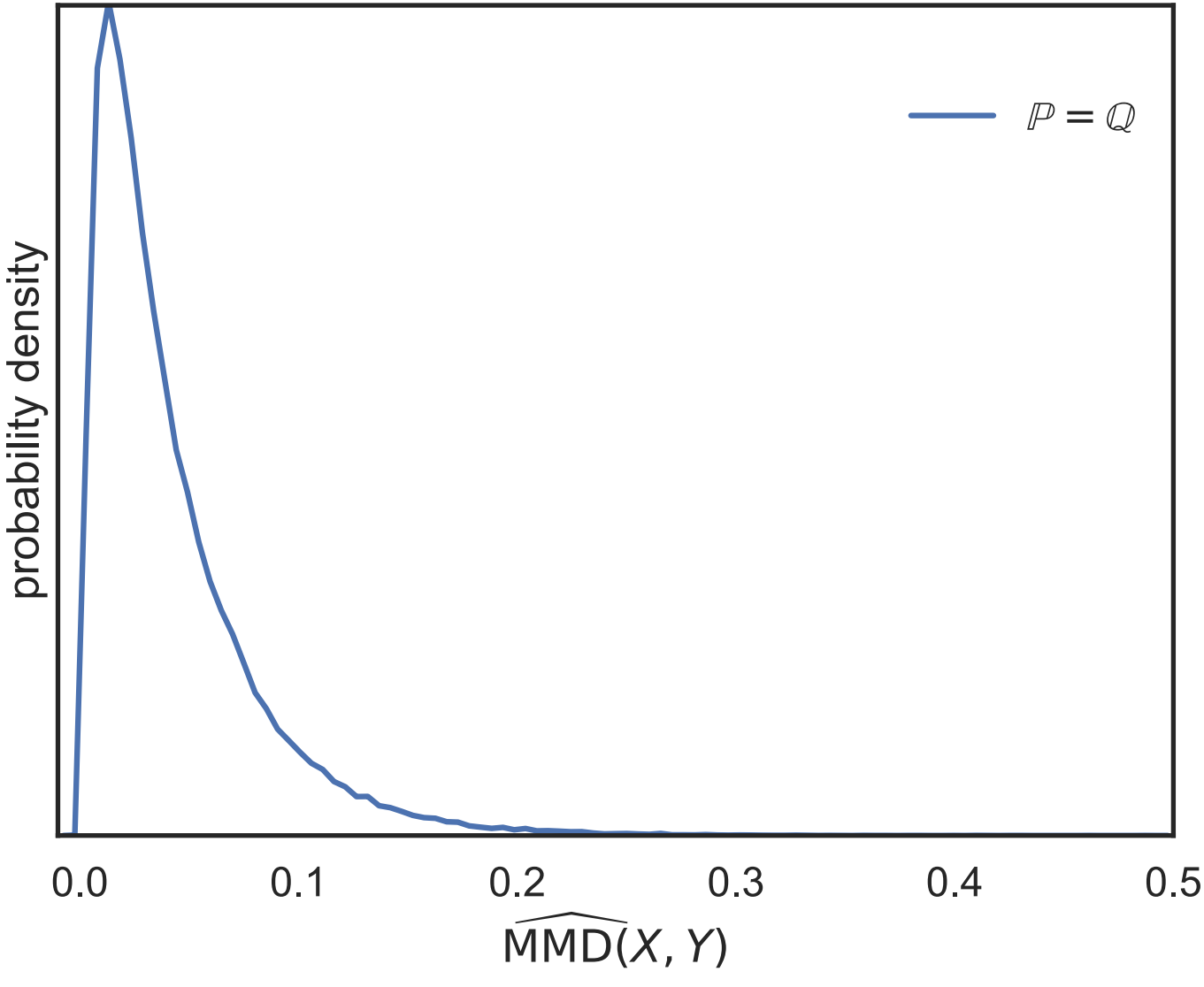
$$X \sim \mathbb{P} \qquad Y \sim \mathbb{Q}$$

- Question: is $\mathbb{P} = \mathbb{Q}$?
- Hypothesis testing approach:

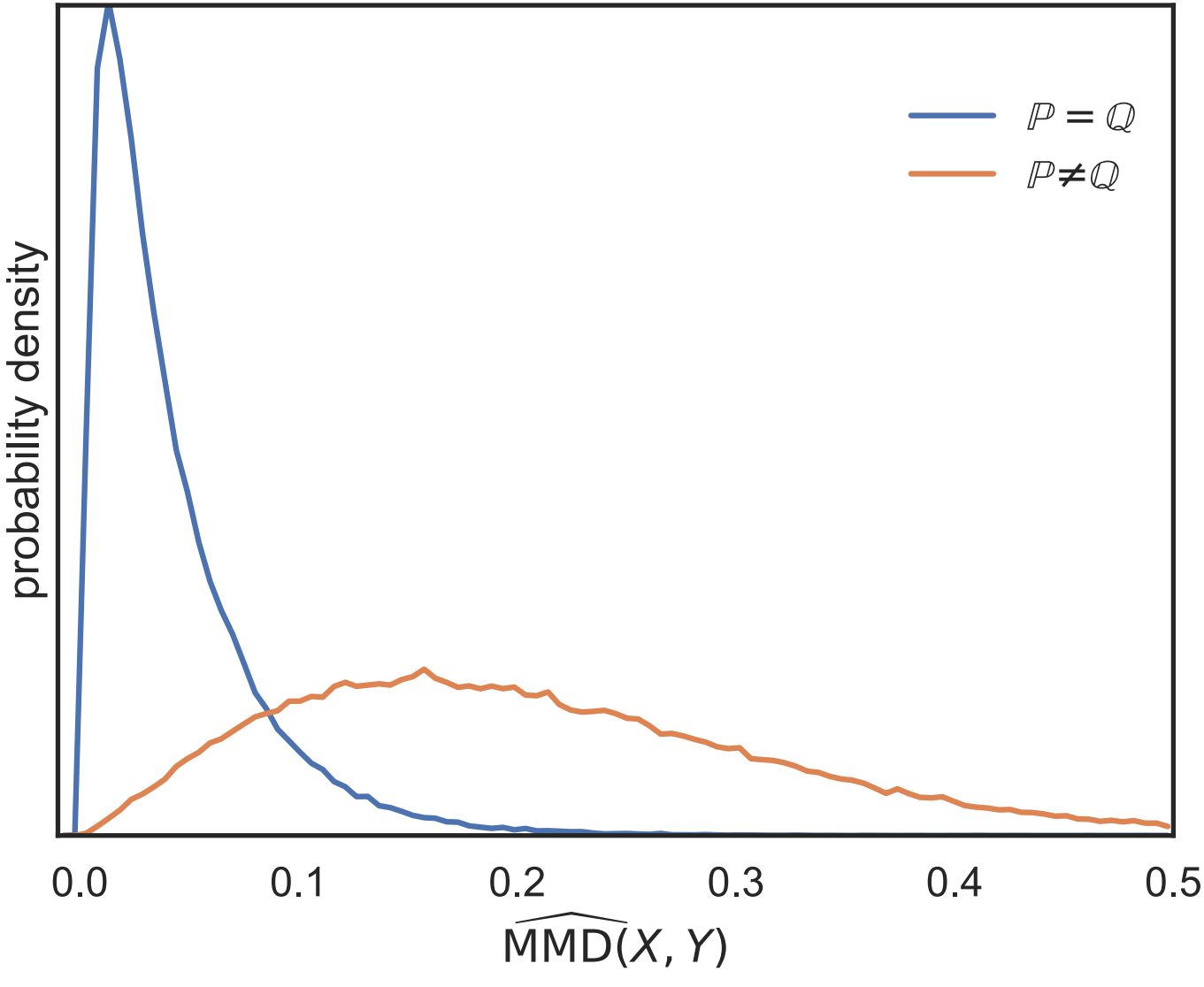
$$H_0 : \mathbb{P} = \mathbb{Q} \qquad H_1 : \mathbb{P} \neq \mathbb{Q}$$

- Reject H_0 if $\widehat{\mathrm{MMD}}(X, Y) > c_\alpha$

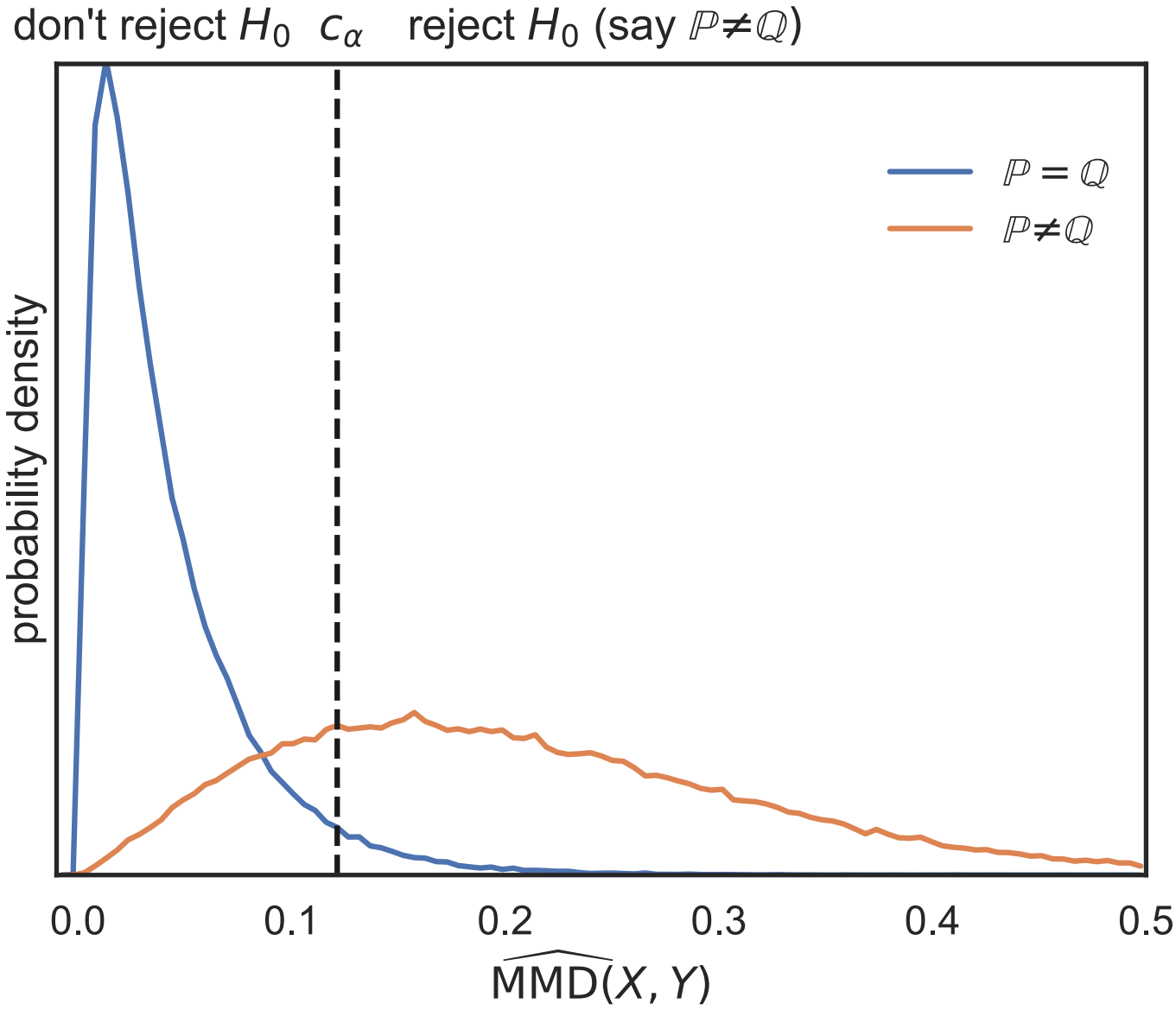
What's a hypothesis test again?



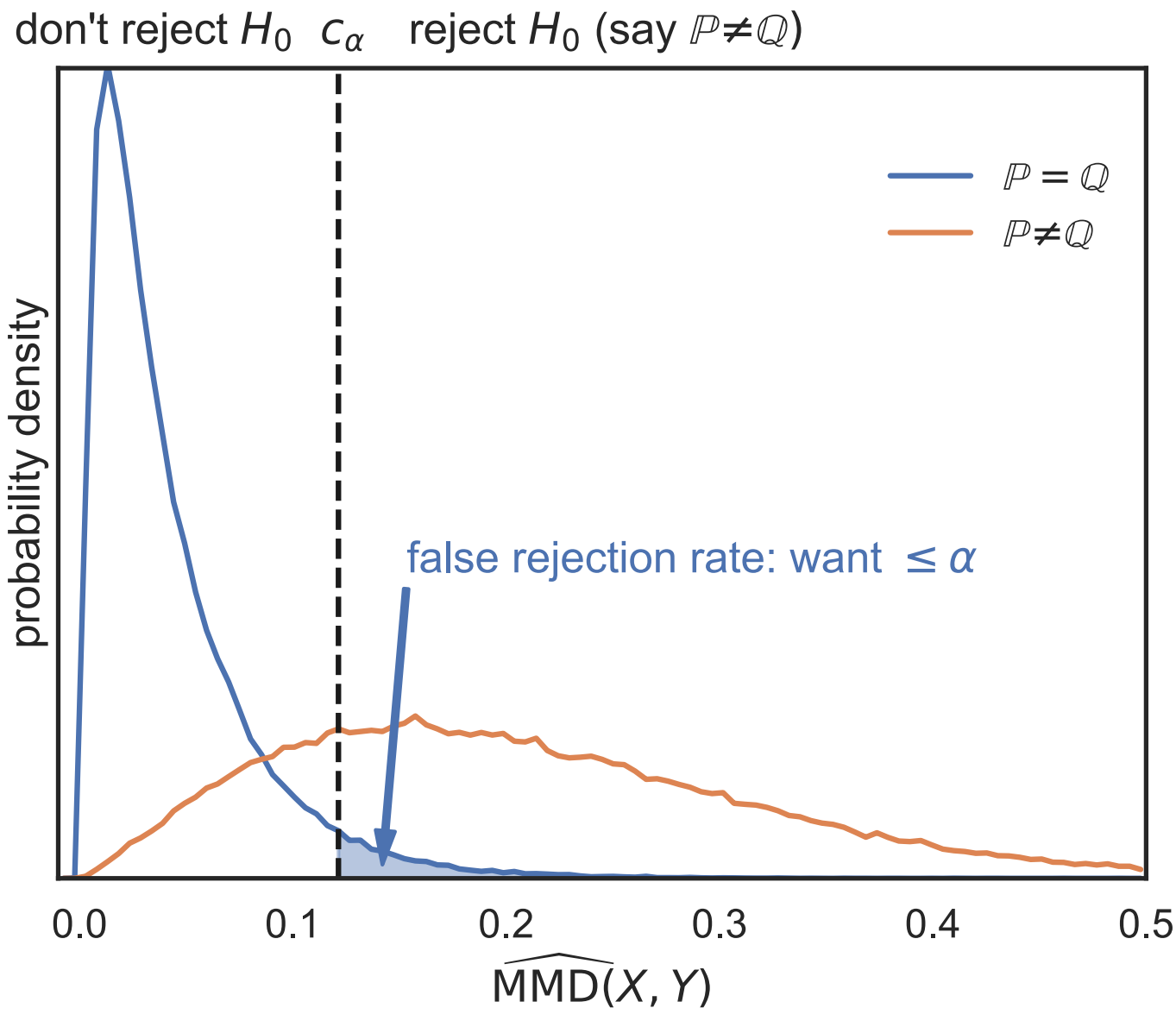
What's a hypothesis test again?



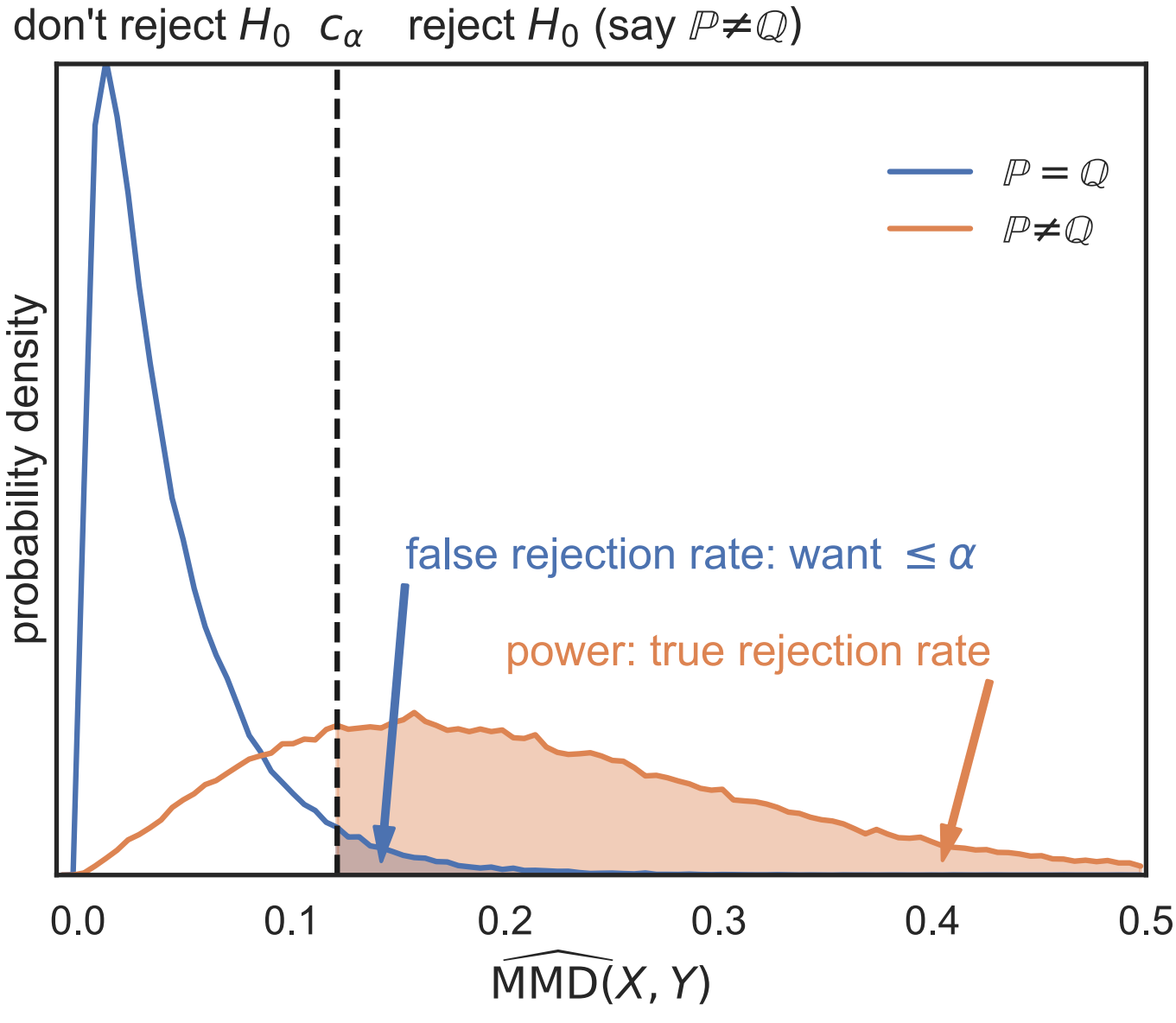
What's a hypothesis test again?



What's a hypothesis test again?



What's a hypothesis test again?



MMD-based testing

- H_0: $n\widehat{\mathrm{MMD}}^2$ converges in distribution to...something
  - Infinite mixture of χ^2s, params depend on $\mathbb{P}$ and k

MMD-based testing

- H_0: $n\widehat{\mathrm{MMD}}^2$ converges in distribution to...something
  - Infinite mixture of χ^2s, params depend on $\mathbb{P}$ and k
  - Can estimate threshold with *permutation testing*

MMD-based testing

- H_0: $n\widehat{\mathrm{MMD}}^2$ converges in distribution to...something
  - Infinite mixture of χ^2s, params depend on $\mathbb{P}$ and k
  - Can estimate threshold with *permutation testing*
- H_1: $\sqrt{n}(\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2) \xrightarrow{d}$ asymptotically normal

MMD-based testing

- H_0: $n\widehat{\mathrm{MMD}}^2$ converges in distribution to...something
  - Infinite mixture of χ^2s, params depend on $\mathbb{P}$ and k
  - Can estimate threshold with *permutation testing*
- H_1: $\sqrt{n}(\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2) \xrightarrow{d}$ asymptotically normal
- Any characteristic kernel gives consistent test

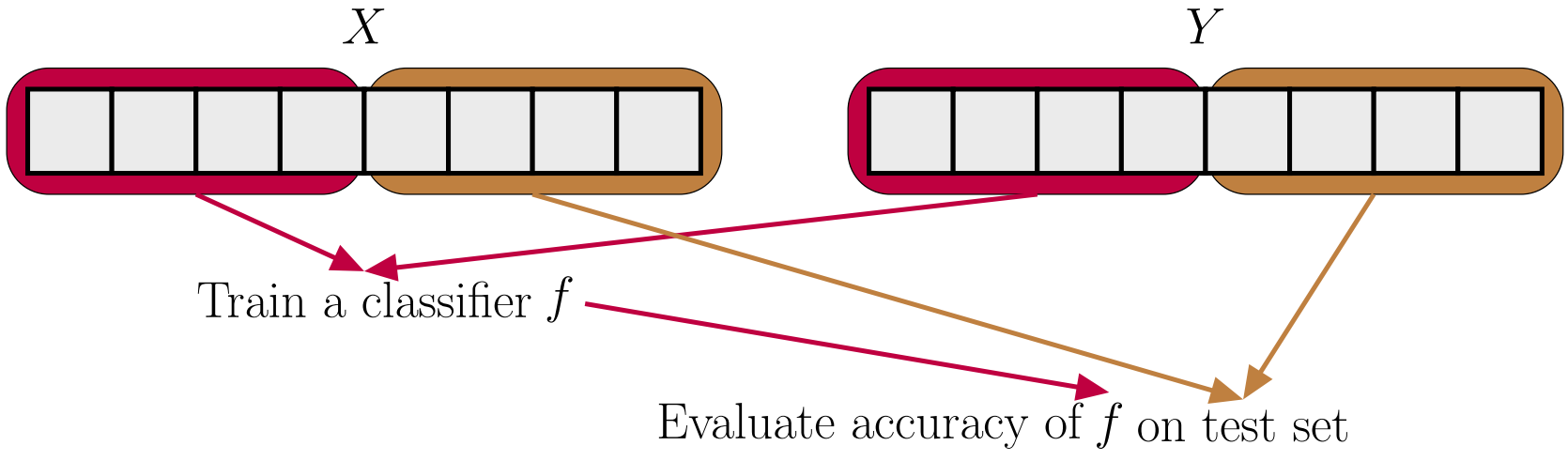
MMD-based testing

- H_0: $n\widehat{\mathrm{MMD}}^2$ converges in distribution to...something
  - Infinite mixture of χ^2s, params depend on $\mathbb{P}$ and k
  - Can estimate threshold with *permutation testing*
- H_1: $\sqrt{n}(\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2) \xrightarrow{d}$ asymptotically normal
- Any characteristic kernel gives consistent test...eventually

MMD-based testing

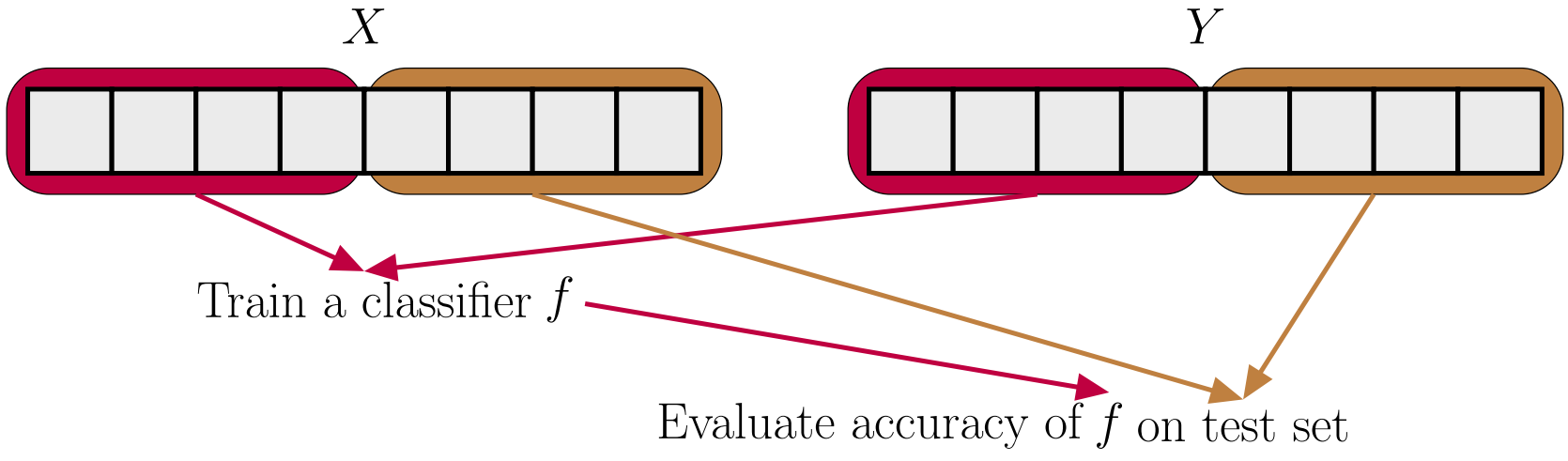
- H_0: $n\widehat{\mathrm{MMD}}^2$ converges in distribution to...something
  - Infinite mixture of χ^2s, params depend on $\mathbb{P}$ and k
  - Can estimate threshold with *permutation testing*
- H_1: $\sqrt{n}(\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2) \xrightarrow{d}$ asymptotically normal
- Any characteristic kernel gives consistent test...eventually
- Need enormous n if kernel is bad for problem

Classifier two-sample tests



- $\hat{T}(X, Y)$ is the accuracy of f on the test set
- Under H_0, classification impossible: $\hat{T} \sim \mathbf{Binomial}(n, \frac{1}{2})$

Classifier two-sample tests



- $\hat{T}(X, Y)$ is the accuracy of f on the test set
- Under H_0, classification impossible: $\hat{T} \sim \mathbf{Binomial}(n, \frac{1}{2})$
- With $k(x, y) = \frac{1}{4} f(x) f(y)$ where $f(x) \in \{-1, 1\}$, get $\widehat{\mathrm{MMD}}(X, Y) = \left| \hat{T}(X, Y) - \frac{1}{2} \right|$

Deep learning and deep kernels

- $k(x, y) = \frac{1}{4} f(x) f(y)$ is one form of *deep kernel*

Deep learning and deep kernels

- $k(x, y) = \frac{1}{4} f(x) f(y)$ is one form of *deep kernel*
- Deep models are usually of the form $f(x) = w^\mathsf{T} \phi_\psi(x)$
  - With a *learned* $\phi_\psi(x) : \mathcal{X} \to \mathbb{R}^D$

Deep learning and deep kernels

- $k(x, y) = \frac{1}{4} f(x) f(y)$ is one form of *deep kernel*
- Deep models are usually of the form $f(x) = w^\mathsf{T} \phi_\psi(x)$
  - With a *learned* $\phi_\psi(x) : \mathcal{X} \to \mathbb{R}^D$
- If we fix ψ, have $f \in \mathcal{H}_\psi$ with $k_\psi(x, y) = \phi_\psi(x)^\mathsf{T} \phi_\psi(y)$

Deep learning and deep kernels

- $k(x, y) = \frac{1}{4} f(x) f(y)$ is one form of *deep kernel*
- Deep models are usually of the form $f(x) = w^\mathsf{T} \phi_\psi(x)$
  - With a *learned* $\phi_\psi(x) : \mathcal{X} \to \mathbb{R}^D$
- If we fix ψ, have $f \in \mathcal{H}_\psi$ with $k_\psi(x, y) = \phi_\psi(x)^\mathsf{T} \phi_\psi(y)$
  - Same idea as NNGP approximation

Deep learning and deep kernels

- $k(x, y) = \frac{1}{4} f(x) f(y)$ is one form of *deep kernel*
- Deep models are usually of the form $f(x) = w^\mathsf{T} \phi_\psi(x)$
  - With a *learned* $\phi_\psi(x) : \mathcal{X} \to \mathbb{R}^D$
- If we fix ψ, have $f \in \mathcal{H}_\psi$ with $k_\psi(x, y) = \phi_\psi(x)^\mathsf{T} \phi_\psi(y)$
  - Same idea as NNGP approximation
- Generalize to a **deep kernel**:

$$k_\psi(x, y) = \kappa\left(\phi_\psi(x), \phi_\psi(y)\right)$$

Normal deep learning $\subset$ deep kernels

- Take $k_\psi(x, y) = \frac{1}{4} f_\psi(x) f_\psi(y)$
- Final function in $\mathcal{H}_\psi$ will be $a f_\psi(x)$

Normal deep learning $\subset$ deep kernels

- Take $k_\psi(x, y) = \frac{1}{4} f_\psi(x) f_\psi(y) + 1$
- Final function in $\mathcal{H}_\psi$ will be $a f_\psi(x) + b$

Normal deep learning $\subset$ deep kernels

- Take $k_\psi(x, y) = \frac{1}{4} f_\psi(x) f_\psi(y) + 1$
- Final function in $\mathcal{H}_\psi$ will be $a f_\psi(x) + b$
- With logistic loss: this is Platt scaling

Normal deep learning $\subset$ deep kernels

- Take $k_\psi(x, y) = \frac{1}{4} f_\psi(x) f_\psi(y) + 1$
- Final function in $\mathcal{H}_\psi$ will be $a f_\psi(x) + b$
- With logistic loss: this is Platt scaling

On Calibration of Modern Neural Networks

Chuan Guo [*1] **Geoff Pleiss** [*1] **Yu Sun** [*1] **Kilian Q. Weinberger** [1]

“Normal deep learning $\subset$ deep kernels” – so?

- This definitely does *not* say that deep learning is (even approximately) a kernel method

“Normal deep learning ⊂ deep kernels” – so?

- This definitely does *not* say that deep learning is (even approximately) a kernel method
- …despite what some people might want you to think

Computer Science > Machine Learning

[Submitted on 30 Nov 2020]

Every Model Learned by Gradient Descent Is Approximately a Kernel Machine

Pedro Domingos

“Normal deep learning ⊂ deep kernels” – so?

- This definitely does *not* say that deep learning is (even approximately) a kernel method
- …despite what some people might want you to think

- We know theoretically deep learning can learn some things faster than any kernel method [see Malach+ ICML-21 + refs]

“Normal deep learning ⊂ deep kernels” – so?

- This definitely does *not* say that deep learning is (even approximately) a kernel method
- …despite what some people might want you to think

- We know theoretically deep learning can learn some things faster than any kernel method [see Malach+ ICML-21 + refs]
- But deep kernel learning ≠ traditional kernel models
  - exactly like how usual deep learning ≠ linear models

Optimizing power of MMD tests

- Asymptotics of $\widehat{\mathrm{MMD}}^2$ give us immediately that

$$\Pr_{H_1}\left(n\widehat{\mathrm{MMD}}^2 > c_\alpha\right) \approx \Phi\left(\frac{\sqrt{n}\,\mathrm{MMD}^2}{\sigma_{H_1}} - \frac{c_\alpha}{\sqrt{n}\sigma_{H_1}}\right)$$

MMD, σ_{H_1}, c_α are constants: first term usually dominates

Optimizing power of MMD tests

- Asymptotics of $\widehat{\mathrm{MMD}}^2$ give us immediately that

$$\Pr_{H_1}\left(n\widehat{\mathrm{MMD}}^2 > c_\alpha\right) \approx \Phi\left(\frac{\sqrt{n}\,\mathrm{MMD}^2}{\sigma_{H_1}} - \frac{c_\alpha}{\sqrt{n}\sigma_{H_1}}\right)$$

  MMD, σ_{H_1}, c_α are constants: first term usually dominates

- Pick k to maximize an estimate of $\mathrm{MMD}^2\,/\sigma_{H_1}$

Optimizing power of MMD tests

- Asymptotics of $\widehat{\mathrm{MMD}}^2$ give us immediately that

$$\Pr_{H_1}\left(n\widehat{\mathrm{MMD}}^2 > c_\alpha\right) \approx \Phi\left(\frac{\sqrt{n}\,\mathrm{MMD}^2}{\sigma_{H_1}} - \frac{c_\alpha}{\sqrt{n}\sigma_{H_1}}\right)$$

  MMD, σ_{H_1}, c_α are constants: first term usually dominates

- Pick k to maximize an estimate of $\mathrm{MMD}^2\,/\sigma_{H_1}$

- Use $\widehat{\mathrm{MMD}}$ from before, get $\hat{\sigma}_{H_1}$ from U-statistic theory

Optimizing power of MMD tests

- Asymptotics of $\widehat{\mathrm{MMD}}^2$ give us immediately that

$$\Pr_{H_1}\left(n\widehat{\mathrm{MMD}}^2 > c_\alpha\right) \approx \Phi\left(\frac{\sqrt{n}\,\mathrm{MMD}^2}{\sigma_{H_1}} - \frac{c_\alpha}{\sqrt{n}\sigma_{H_1}}\right)$$

  MMD, σ_{H_1}, c_α are constants: first term usually dominates

- Pick k to maximize an estimate of $\mathrm{MMD}^2 \,/\sigma_{H_1}$
- Use $\widehat{\mathrm{MMD}}$ from before, get $\hat{\sigma}_{H_1}$ from U-statistic theory
- Can show uniform $\mathcal{O}_P\left(n^{-\frac{1}{3}}\right)$ convergence of estimator

Optimizing power of MMD tests

- Asymptotics of $\widehat{\mathrm{MMD}}^2$ give us immediately that

$$\Pr_{H_1}\left(n\widehat{\mathrm{MMD}}^2 > c_\alpha\right) \approx \Phi\left(\frac{\sqrt{n}\,\mathrm{MMD}^2}{\sigma_{H_1}} - \frac{c_\alpha}{\sqrt{n}\sigma_{H_1}}\right)$$

  MMD, σ_{H_1}, c_α are constants: first term usually dominates

- Pick k to maximize an estimate of $\mathrm{MMD}^2 \,/\sigma_{H_1}$
- Use $\widehat{\mathrm{MMD}}$ from before, get $\hat{\sigma}_{H_1}$ from U-statistic theory
- Can show uniform $\mathcal{O}_P\left(n^{-\frac{1}{3}}\right)$ convergence of estimator
- Get better tests (even after data splitting)

Application: (S)MMD GANs

- An implicit generative model:
  - A generator net outputs samples from $\mathbb{Q}_\theta$

Application: (S)MMD GANs

- An implicit generative model:
  - A generator net outputs samples from $\mathbb{Q}_\theta$
  - Minimize estimate of $\mathrm{MMD}_\psi(\mathbb{P}^m, \mathbb{Q}_\theta^n)$ on a minibatch

Application: (S)MMD GANs

- An implicit generative model:
  - A generator net outputs samples from $\mathbb{Q}_\theta$
  - Minimize estimate of $\mathrm{MMD}_\psi(\mathbb{P}^m, \mathbb{Q}_\theta^n)$ on a minibatch
- MMD GAN: $\min_\theta \left[\max_\psi \mathrm{MMD}_\psi(\mathbb{P}, \mathbb{Q}_\theta)\right]$

Application: (S)MMD GANs

- An implicit generative model:
  - A generator net outputs samples from $\mathbb{Q}_\theta$
  - Minimize estimate of $\mathrm{MMD}_\psi(\mathbb{P}^m, \mathbb{Q}_\theta^n)$ on a minibatch
- MMD GAN: $\min_\theta \left[\max_\psi \mathrm{MMD}_\psi(\mathbb{P}, \mathbb{Q}_\theta)\right]$
- SMMD GAN: $\min_\theta \left[\max_\psi \mathrm{SMMD}_\psi(\mathbb{P}, \mathbb{Q}_\theta)\right]$
  - Scaled MMD uses kernel properties to ensure smooth loss for θ by making witness function smooth [Arbel+ NeurIPS-18]

Application: (S)MMD GANs

- An implicit generative model:
  - A generator net outputs samples from $\mathbb{Q}_\theta$
  - Minimize estimate of $\mathbf{MMD}\ \psi(\mathbb{P}^m, \mathbb{Q}_\theta^n)$ on a minibatch
- MMD GAN: $\mathbf{min}_\theta \left[\mathbf{max}_\psi\ \mathbf{MMD}_\psi(\mathbb{P}, \mathbb{Q}_\theta)\right]$
- SMMD GAN: $\mathbf{min}_\theta \left[\mathbf{max}_\psi\ \mathbf{SMMD}_\psi(\mathbb{P}, \mathbb{Q}_\theta)\right]$
  - Scaled MMD uses kernel properties to ensure smooth loss for θ by making witness function smooth [Arbel+ NeurIPS-18]
  - Uses $\langle f, \partial_{x_1} k(x, \cdot)\rangle_{\mathcal{H}} = \partial_{x_1} f(x)$

Application: (S)MMD GANs

- An implicit generative model:
  - A generator net outputs samples from $\mathbb{Q}_\theta$
  - Minimize estimate of $\mathbf{MMD}_\psi(\mathbb{P}^m, \mathbb{Q}_\theta^n)$ on a minibatch
- MMD GAN: $\mathbf{min}_\theta \left[\mathbf{max}_\psi \mathbf{MMD}_\psi(\mathbb{P}, \mathbb{Q}_\theta)\right]$
- SMMD GAN: $\mathbf{min}_\theta \left[\mathbf{max}_\psi \mathbf{SMMD}_\psi(\mathbb{P}, \mathbb{Q}_\theta)\right]$
  - Scaled MMD uses kernel properties to ensure smooth loss for θ by making witness function smooth [Arbel+ NeurIPS-18]
  - Uses $\langle f, \partial_{x_1} k(x, \cdot)\rangle_{\mathcal{H}} = \partial_{x_1} f(x)$
  - Standard WGAN-GP better thought of in kernel framework

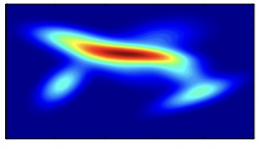
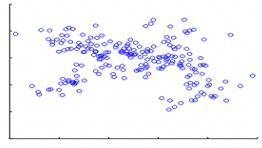
Application: distribution regression/classification/...

- We can define a kernel on distributions by, e.g.,

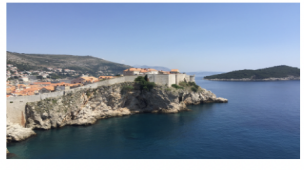
$$k(\mathbb{P}, \mathbb{Q}) = \exp\left(-\frac{1}{2\sigma^2}\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})\right)$$

- Some pointers:
  [Muandet+ NeurIPS-12] [Sutherland 2016] [Szabó+ JMLR-16]

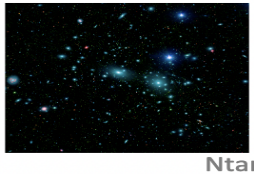
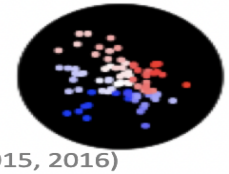
distribution observed sample label

9 components

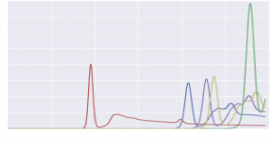



“seaside city”

Mass $7 \times 10^{14}\ M_{\odot}$ and more...

Ntampaka et al. (ApJ 2015, 2016)

Jin et al. (NSS 2016)

no Cs137 present

	AGEP	SEX	...	RACSOR	RACWHT
0	75	1	...	0	1
1	25	0	...	0	1
...	...	...	...	...	...

Flaxman et al. (KDD 2015)

county voted
54% for Obama

Example: age from face images [Law+ AISTATS-18]

Bayesian distribution regression: incorporate $\mu_{\mathbb{P}}$ uncertainty

$$\left\{ \quad \text{  }, \text{  }, \text{  } \quad \right\} \to 35$$

Example: age from face images [Law+ AISTATS-18]

Bayesian distribution regression: incorporate $\mu_{\mathbb{P}}$ uncertainty

$$\left\{ \quad \text{  }, \text{  }, \text{  } \quad \right\} \to 35$$

IMDb database [Rothe+ 2015]: 400k images of 20k celebrities

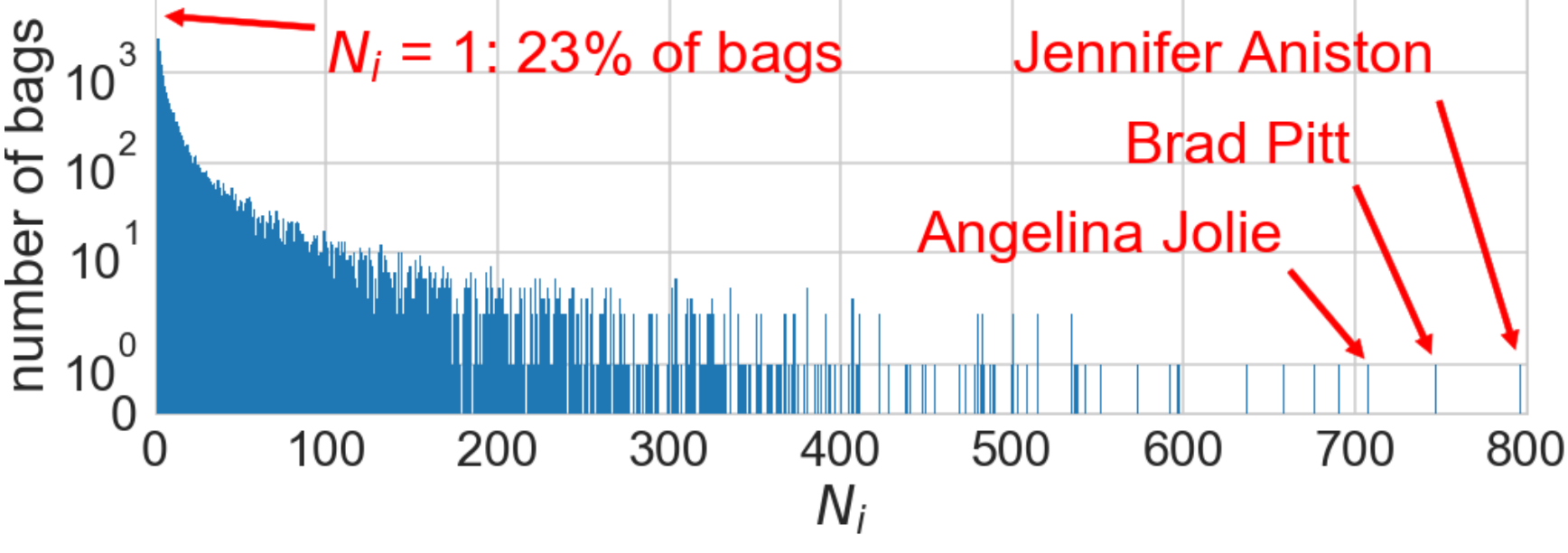
Example: age from face images [Law+ AISTATS-18]

Bayesian distribution regression: incorporate $\mu_{\mathbb{P}}$ uncertainty

$$\left\{ \quad , \quad , \quad \right\} \to 35$$



IMDb database [Rothe+ 2015]: 400k images of 20k celebrities

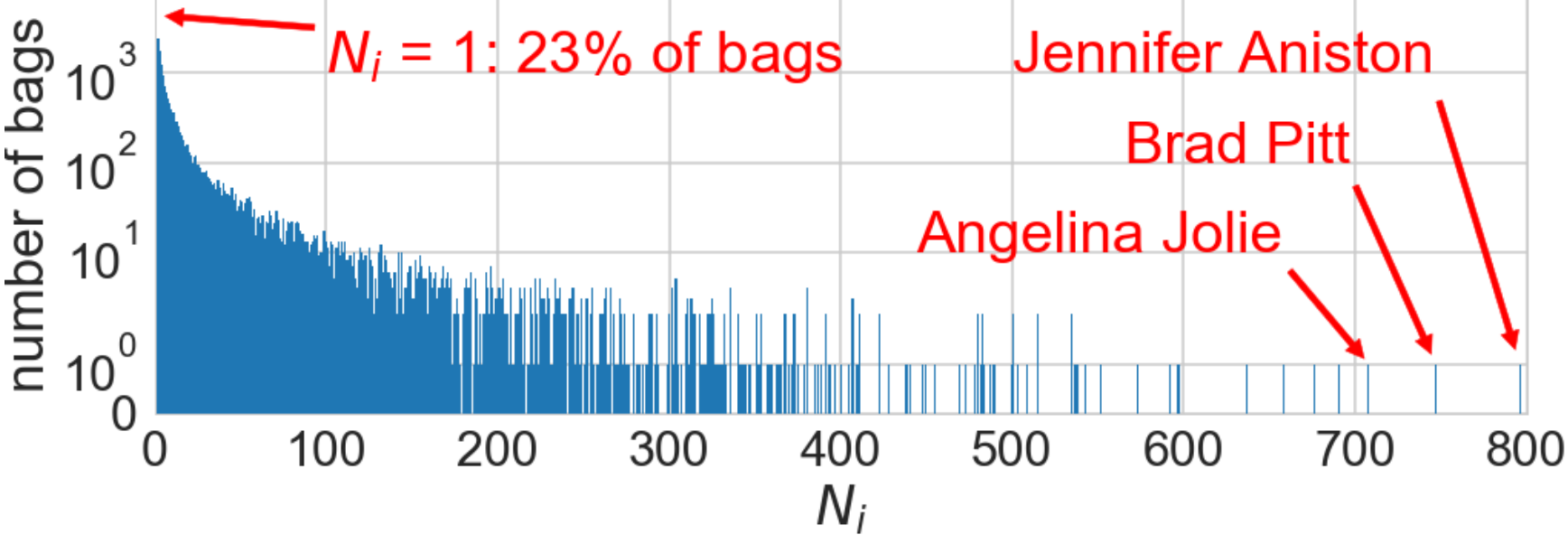


Example: age from face images [Law+ AISTATS-18]

Bayesian distribution regression: incorporate $\mu_{\mathbb{P}}$ uncertainty

$$\left\{ \text{  }, \text{  }, \text{  }, \text{  }, \text{  } \right\} \to 35$$

IMDb database [Rothe+ 2015]: 400k images of 20k celebrities

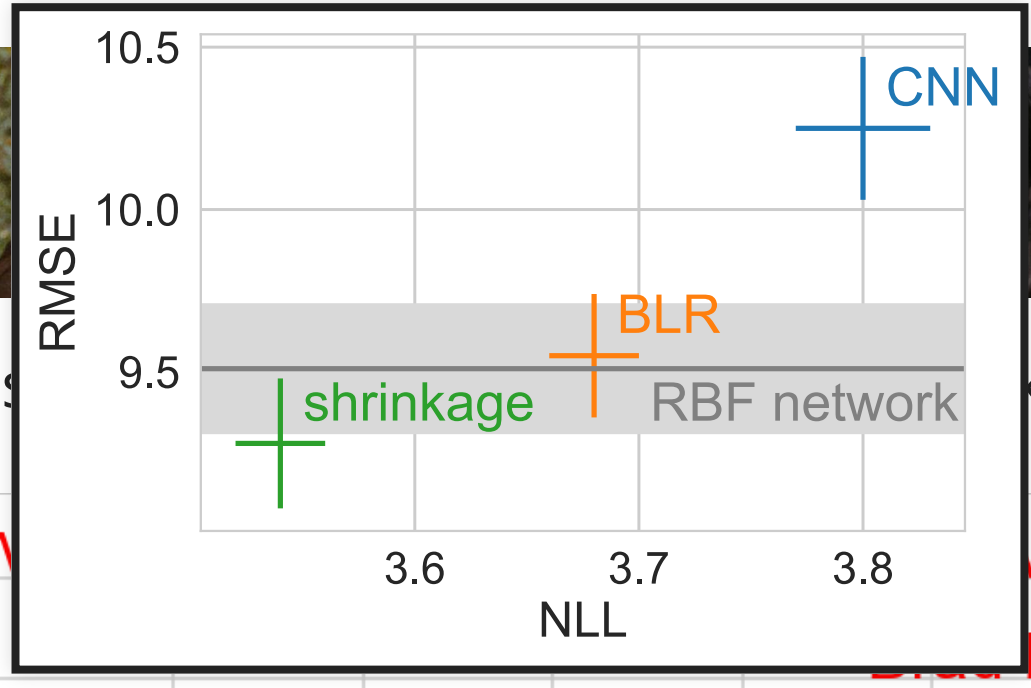


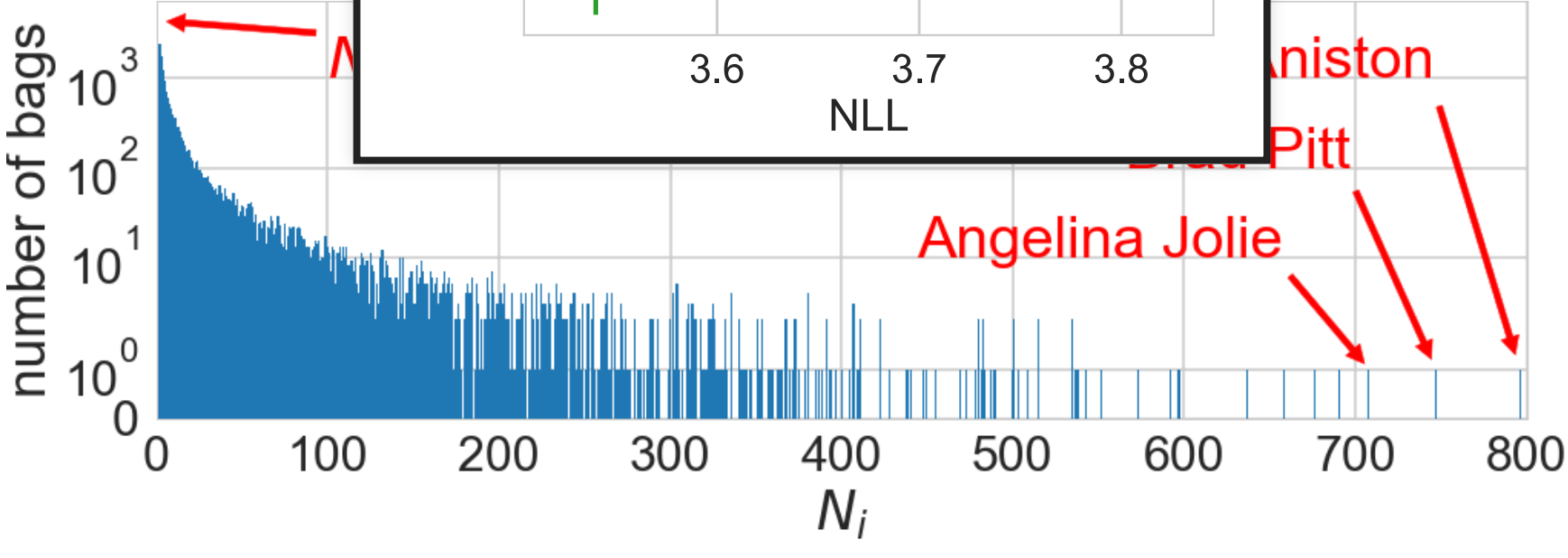
Example: age from face images [Law+ AISTATS-18]

Bayesian distribution regression: incorporate $\mu_{\mathbb{P}}$ uncertainty

$\rightarrow 35$

IMDb databas k celebrities





Independence

- $X \perp\!\!\!\perp Y$ iff $\mathbf{Cov}(f(X), g(Y)) = 0$ for all measurable f, g

Independence

- $X \perp\!\!\!\perp Y$ iff $\mathbf{Cov}(f(X), g(Y)) = 0$ for all measurable f, g
- Let's implement for RKHS functions $f \in \mathcal{H}_x, g \in \mathcal{H}_y$:

$$\mathbb{E}[f(X)]\,\mathbb{E}[g(Y)]$$

Independence

- $X \perp\!\!\!\perp Y$ iff $\mathbf{Cov}(f(X), g(Y)) = 0$ for all measurable f, g
- Let's implement for RKHS functions $f \in \mathcal{H}_x, g \in \mathcal{H}_y$:

$$\mathbb{E}[f(X)]\,\mathbb{E}[g(Y)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}_x} \langle \mu_{\mathbb{Q}}, g \rangle_{\mathcal{H}_y}$$

Independence

- $X \perp\!\!\!\perp Y$ iff $\mathbf{Cov}(f(X), g(Y)) = 0$ for all measurable f, g
- Let's implement for RKHS functions $f \in \mathcal{H}_x, g \in \mathcal{H}_y$:

$$\begin{aligned}\mathbb{E}[f(X)]\,\mathbb{E}[g(Y)] &= \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}_x} \langle \mu_{\mathbb{Q}}, g \rangle_{\mathcal{H}_y} \\ &= \langle f, (\mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}) g \rangle_{\mathcal{H}_x}\end{aligned}$$

Independence

- $X \perp\!\!\!\perp Y$ iff $\mathbf{Cov}(f(X), g(Y)) = 0$ for all measurable f, g
- Let's implement for RKHS functions $f \in \mathcal{H}_x, g \in \mathcal{H}_y$:

$$\begin{aligned}\mathbb{E}[f(X)]\,\mathbb{E}[g(Y)] &= \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}_x} \langle \mu_{\mathbb{Q}}, g \rangle_{\mathcal{H}_y} \\ &= \langle f, (\mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}) g \rangle_{\mathcal{H}_x}\end{aligned}$$

$$\mathbb{E}[f(X)g(Y)]$$

Independence

- $X \perp\!\!\!\perp Y$ iff $\mathbf{Cov}(f(X), g(Y)) = 0$ for all measurable f, g
- Let's implement for RKHS functions $f \in \mathcal{H}_x, g \in \mathcal{H}_y$:

$$\begin{aligned}
\mathbb{E}[f(X)]\,\mathbb{E}[g(Y)] &= \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}_x} \langle \mu_{\mathbb{Q}}, g \rangle_{\mathcal{H}_y} \\
&= \langle f, (\mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}) g \rangle_{\mathcal{H}_x} \\
\mathbb{E}[f(X)g(Y)] &= \mathbb{E}[\langle f, k_x(X, \cdot) \rangle_{\mathcal{H}_x} \langle k_y(Y, \cdot), g \rangle_{\mathcal{H}_y}]
\end{aligned}$$

Independence

- $X \perp\!\!\!\perp Y$ iff $\mathbf{Cov}(f(X), g(Y)) = 0$ for all measurable f, g
- Let's implement for RKHS functions $f \in \mathcal{H}_x, g \in \mathcal{H}_y$:

$$\begin{aligned}
\mathbb{E}[f(X)]\,\mathbb{E}[g(Y)] &= \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}_x} \langle \mu_{\mathbb{Q}}, g \rangle_{\mathcal{H}_y} \\
&= \langle f, (\mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}) g \rangle_{\mathcal{H}_x} \\
\mathbb{E}[f(X)g(Y)] &= \mathbb{E}[\langle f, k_x(X, \cdot) \rangle_{\mathcal{H}_x} \langle k_y(Y, \cdot), g \rangle_{\mathcal{H}_y}] \\
&= \langle f, \mathbb{E}[k_x(X, \cdot) \otimes k_y(Y, \cdot)]\, g \rangle_{\mathcal{H}_x}
\end{aligned}$$

Independence

- $X \perp\!\!\!\perp Y$ iff $\mathbf{Cov}(f(X), g(Y)) = 0$ for all measurable f, g
- Let's implement for RKHS functions $f \in \mathcal{H}_x, g \in \mathcal{H}_y$:

$$
\begin{aligned}
\mathbb{E}[f(X)]\,\mathbb{E}[g(Y)] &= \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}_x} \langle \mu_{\mathbb{Q}}, g \rangle_{\mathcal{H}_y} \\
&= \langle f, (\mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}) g \rangle_{\mathcal{H}_x} \\
\mathbb{E}[f(X)g(Y)] &= \mathbb{E}[\langle f, k_x(X, \cdot) \rangle_{\mathcal{H}_x} \langle k_y(Y, \cdot), g \rangle_{\mathcal{H}_y}] \\
&= \langle f, \mathbb{E}[k_x(X, \cdot) \otimes k_y(Y, \cdot)]\, g \rangle_{\mathcal{H}_x} \\
\mathbf{Cov}(f(X), g(Y)) &= \langle f, C_{XY}\, g \rangle_{\mathcal{H}_x}
\end{aligned}
$$

where $C_{XY} : \mathcal{H}_y \to \mathcal{H}_x$ is

$$
\mathbb{E}[k_x(X, \cdot) \otimes k_y(Y, \cdot)] - \mathbb{E}[k_x(X, \cdot)] \otimes \mathbb{E}[k_y(Y, \cdot)]
$$

Cross-covariance operator and independence

- $\mathrm{Cov}(f(X), g(Y)) = \langle f, C_{XY}\, g \rangle_{\mathcal{H}_x}$
- $C_{XY} = \mathbb{E}[k_x(X, \cdot) \otimes k_y(Y, \cdot)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}$

Cross-covariance operator and independence

- $\mathrm{Cov}(f(X), g(Y)) = \langle f, C_{XY}\, g \rangle_{\mathcal{H}_x}$
- $C_{XY} = \mathbb{E}[k_x(X, \cdot) \otimes k_y(Y, \cdot)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}$
- If $X \perp\!\!\!\perp Y$, then $C_{XY} = 0$

Cross-covariance operator and independence

- $\mathrm{Cov}(f(X), g(Y)) = \langle f, C_{XY}\, g \rangle_{\mathcal{H}_x}$
- $C_{XY} = \mathbb{E}[k_x(X, \cdot) \otimes k_y(Y, \cdot)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}$
- If $X \perp\!\!\!\perp Y$, then $C_{XY} = 0$
- If $C_{XY} = 0$, $\mathrm{Cov}(f(X), g(Y)) = 0 \quad \forall f \in \mathcal{H}_x, g \in \mathcal{H}_y$

Cross-covariance operator and independence

- $\mathrm{Cov}(f(X), g(Y)) = \langle f, C_{XY}\, g \rangle_{\mathcal{H}_x}$
- $C_{XY} = \mathbb{E}[k_x(X, \cdot) \otimes k_y(Y, \cdot)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}$
- If $X \perp\!\!\!\perp Y$, then $C_{XY} = 0$
- If $C_{XY} = 0$, $\mathrm{Cov}(f(X), g(Y)) = 0 \quad \forall f \in \mathcal{H}_x, g \in \mathcal{H}_y$
- If k_x, k_y are characteristic:
  - $C_{XY} = 0$ implies $X \perp\!\!\!\perp Y$ [Szabó/Sriperumbudur JMLR-18]

Cross-covariance operator and independence

- $\mathrm{Cov}(f(X), g(Y)) = \langle f, C_{XY}\, g \rangle_{\mathcal{H}_x}$
- $C_{XY} = \mathbb{E}[k_x(X, \cdot) \otimes k_y(Y, \cdot)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}$
- If $X \perp\!\!\!\perp Y$, then $C_{XY} = 0$
- If $C_{XY} = 0$, $\mathrm{Cov}(f(X), g(Y)) = 0 \quad \forall f \in \mathcal{H}_x, g \in \mathcal{H}_y$
- If k_x, k_y are characteristic:
  - $C_{XY} = 0$ implies $X \perp\!\!\!\perp Y$ [Szabó/Sriperumbudur JMLR-18]
  - $X \perp\!\!\!\perp Y$ iff $C_{XY} = 0$

Cross-covariance operator and independence

- $\mathrm{Cov}(f(X), g(Y)) = \langle f, C_{XY}\, g \rangle_{\mathcal{H}_x}$
- $C_{XY} = \mathbb{E}[k_x(X, \cdot) \otimes k_y(Y, \cdot)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}$
- If $X \perp\!\!\!\perp Y$, then $C_{XY} = 0$
- If $C_{XY} = 0$, $\mathrm{Cov}(f(X), g(Y)) = 0 \quad \forall f \in \mathcal{H}_x, g \in \mathcal{H}_y$
- If k_x, k_y are characteristic:
  - $C_{XY} = 0$ implies $X \perp\!\!\!\perp Y$ [Szabó/Sriperumbudur JMLR-18]
  - $X \perp\!\!\!\perp Y$ iff $C_{XY} = 0$
  - $X \perp\!\!\!\perp Y$ iff $0 = \|C_{XY}\|^2_{\mathrm{HS}}$ (sum squared singular values)
    - HSIC: "Hilbert-Schmidt Independence Criterion"

HSIC

$$C_{XY} = \mathbb{E}[k_x(X, \cdot) \otimes k_y(Y, \cdot)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}$$

$$\|C_{XY}\|^2_{\mathrm{HS}} = \|\mu_{\mathbb{P}_{XY}} - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_x \otimes \mathcal{H}_y}$$

HSIC

$$
\begin{aligned}
C_{XY} &= \mathbb{E}[k_x(X,\cdot) \otimes k_y(Y,\cdot)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}} \\
\|C_{XY}\|^2_{\mathrm{HS}} &= \|\mu_{\mathbb{P}_{XY}} - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_x \otimes \mathcal{H}_y} \\
&= \mathrm{MMD}(\mathbb{P}_{XY}, \mathbb{P} \times \mathbb{Q})^2
\end{aligned}
$$

HSIC

$$
\begin{aligned}
C_{XY} &= \mathbb{E}[k_x(X,\cdot) \otimes k_y(Y,\cdot)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}} \\
\|C_{XY}\|^2_{\mathrm{HS}} &= \|\mu_{\mathbb{P}_{XY}} - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_x \otimes \mathcal{H}_y} \\
&= \mathrm{MMD}(\mathbb{P}_{XY}, \mathbb{P} \times \mathbb{Q})^2 \\
&= \mathbb{E}[k_x(X,X')k_y(Y,Y')] \\
&\quad - 2\,\mathbb{E}[k_x(X,X')k_x(Y,Y'')] \\
&\quad + \mathbb{E}[k_x(X,X')]\,\mathbb{E}[k_y(Y,Y')]
\end{aligned}
$$

HSIC

$$
\begin{aligned}
C_{XY} &= \mathbb{E}[k_x(X,\cdot) \otimes k_y(Y,\cdot)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}} \\
\|C_{XY}\|^2_{\mathrm{HS}} &= \|\mu_{\mathbb{P}_{XY}} - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_x \otimes \mathcal{H}_y} \\
&= \mathrm{MMD}(\mathbb{P}_{XY}, \mathbb{P} \times \mathbb{Q})^2 \\
&= \mathbb{E}[k_x(X,X')k_y(Y,Y')] \\
&\quad - 2\,\mathbb{E}[k_x(X,X')k_x(Y,Y'')] \\
&\quad + \mathbb{E}[k_x(X,X')]\,\mathbb{E}[k_y(Y,Y')]
\end{aligned}
$$

- Linear case: C_{XY} is cross-covariance matrix, HSIC is squared Frobenius norm

HSIC

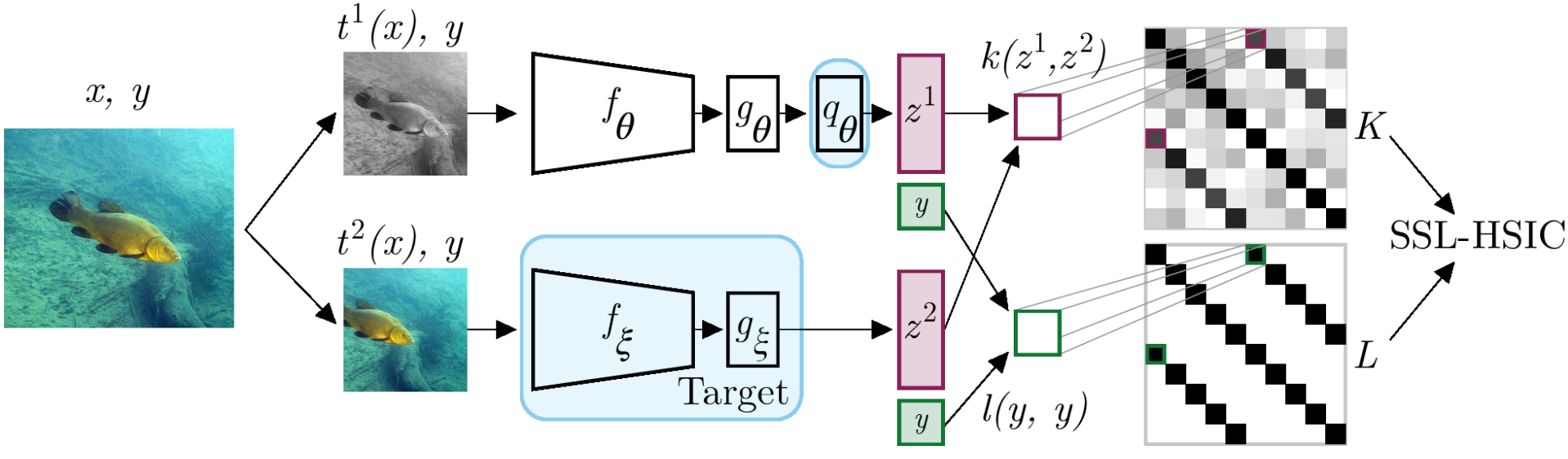
$$
\begin{aligned}
C_{XY} &= \mathbb{E}[k_x(X,\cdot) \otimes k_y(Y,\cdot)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}} \\
\|C_{XY}\|^2_{\mathbf{HS}} &= \|\mu_{\mathbb{P}_{XY}} - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_x \otimes \mathcal{H}_y} \\
&= \mathbf{MMD}(\mathbb{P}_{XY}, \mathbb{P} \times \mathbb{Q})^2 \\
&= \mathbb{E}[k_x(X,X')k_y(Y,Y')] \\
&\quad - 2\,\mathbb{E}[k_x(X,X')k_x(Y,Y'')] \\
&\quad + \mathbb{E}[k_x(X,X')]\,\mathbb{E}[k_y(Y,Y')]
\end{aligned}
$$

- Linear case: C_{XY} is cross-covariance matrix, HSIC is squared Frobenius norm
- Default estimator (biased, but simple): $\mathbf{Tr}(HK_XHK_Y)$ where $H = I - \mathbf{1}\mathbf{1}^\mathsf{T}$

HSIC applications

- Independence testing [Gretton+ NeurIPS-07]
- Clustering [Song+ ICML-07]
- Feature selection [Song+ JMLR-12]
- Self-supervised learning [Li+ NeurIPS-21]
- ⋮
- Broadly: easier-to-estimate, sometimes-nicer version of mutual information

Example: SSL-HSIC [Li+ NeurIPS-21]



- Maximizes dependence between image features f and its identity on a minibatch
- Using a learned deep kernel based on g

Recap

- Mean embedding $\mu_{\mathbb{P}} = \mathbb{E}\, k(X, \cdot)$
- $\mathbf{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$ is 0 iff $\mathbb{P} = \mathbb{Q}$ (for characteristic kernels)
- $\mathbf{HSIC}(X, Y) = \|C_{XY}\|_{\mathrm{HS}} = \mathbf{MMD}(\mathbb{P}_{XY}, \mathbb{P} \times \mathbb{Q})^2$ is 0 iff $X \perp\!\!\!\perp Y$ (for characteristic k_x, k_y)
- After break: last interactive session exploring testing
- More details:
  - Close connections to Gaussian processes [Kanagawa+ 'GPs and Kernel Methods' 2018]
  - Mean embeddings: survey [Muandet+ 'Kernel Mean Embedding of Distributions']