More Kernels

CPSC 532S: Modern Statistical Learning Theory
9 March 2022
cs.ubc.ca/~dsuth/532S/22/

Admin: Projects

- **Literature survey** option:
  - Read several related papers on a learning theory topic
  - Write a document that overviews the results + proof techniques, relates their assumptions, etc
- **Extension** option:
  - Extend/analyze 1-2 learning theory papers
  - Maybe do some experiments checking assumptions/conclusions/etc
  - Maybe weaken some assumptions in the paper, prove interesting corollary, etc
  - Write a document overviewing the paper + proof and describing new results
- **Novel analysis** option:
  - Analyze an algorithm/setting that hasn't been (satisfyingly) analyzed yet
  - Analysis should be nontrivial; can be based on class or related techniques
  - Failure okay if you show why it *should* have worked + why it didn't
  - But probably have a survey or extension "backup plan"

Admin: Projects

- Do in groups of 1-3; counts as one assignment but can’t be dropped
- Suggestions for topics will be up **soon**, but you can also pick your own
- 10 points: a **very short proposal** (~1 paragraph, including papers), by **Wed Mar 16**
  - Make a private Piazza post with me + your group
  - I’ll give you feedback ASAP
  - Can change topic afterwards if needed, but talk to me if significant
- 20 points: **in-class presentation**, on **Wed April 6**
  - Around 5-10 mins depending on # of groups
  - Come in person if you can, otherwise can do by Zoom – let me know if an issue
  - Explain the topic, new results if relevant, 1-2 papers inc. proof if survey
- 70 points: the **project report**, due on **Fri April 8**
  - NeurIPS format, 4-10 pages (plus appendices if necessary)

Reproducing kernel Hilbert space (RKHS)

- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive semidefinite **kernel**

Reproducing kernel Hilbert space (RKHS)

$$\begin{bmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) \\ \vdots & \ddots & \vdots \\ k(x_n,x_1) & \cdots & k(x_n,x_n) \end{bmatrix}$$

- $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is a positive semidefinite **kernel**
  - For all $n \geq 1, x_1, \ldots, x_n \in \mathscr{X}$, the matrix $\left[k(x_i, x_j)\right]_{ij}$ is psd

$a^T K a \geq 0$

positive semidefinite : $a^T K a \geq 0 \quad \forall a \in \mathbb{R}^n$; $\lambda_{min}(K) \geq 0$

"positive definite" $\Leftrightarrow$

strictly positive definite : $a^T K a > 0 \quad \forall a \neq 0$; $\lambda_{min}(K) > 0$

Reproducing kernel Hilbert space (RKHS)

- $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is a positive semidefinite **kernel**
  - For all $n \geq 1, x_1, \ldots, x_n \in \mathscr{X}$, the matrix $\left[k(x_i, x_j)\right]_{ij}$ is psd
  - Equivalent: there is some Hilbert space $\mathscr{H}'$ and $\phi' : \mathscr{X} \to \mathscr{H}'$ where $k(x, y) = \langle \phi'(x), \phi'(y) \rangle_{\mathscr{H}'}$

Reproducing kernel Hilbert space (RKHS)

- $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is a positive semidefinite **kernel**
  - For all $n \geq 1, x_1, \ldots, x_n \in \mathscr{X}$, the matrix $\left[k(x_i, x_j)\right]_{ij}$ is psd
  - Equivalent: there is some Hilbert space $\mathscr{H}'$ and $\phi' : \mathscr{X} \to \mathscr{H}'$ where $k(x, y) = \langle \phi'(x), \phi'(y) \rangle_{\mathscr{H}'}$

$\langle k(x,\cdot), k(y,\cdot) \rangle_{\mathcal{H}_k} = k(x,y)$

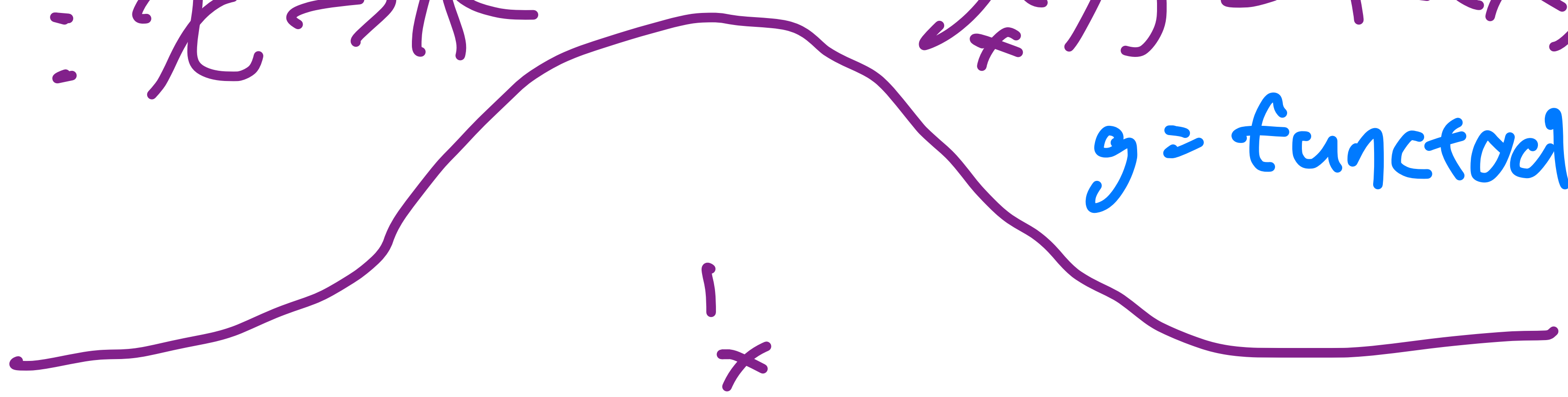
- An **RKHS** with kernel k, $\mathscr{H}_k$, is a Hilbert space of functions $f : \mathscr{X} \to \mathbb{R}$ with

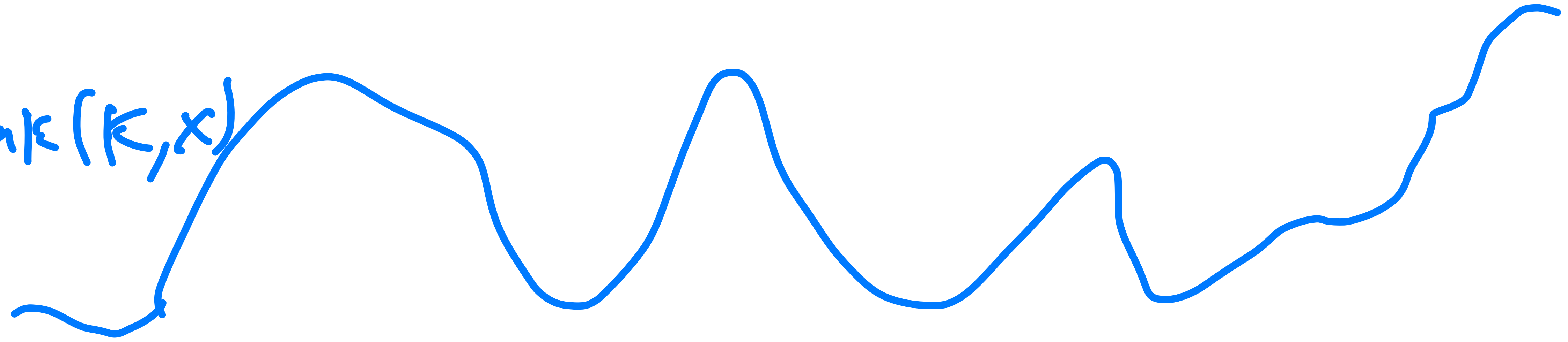
$\forall x,$ $\quad k(x, \cdot\,) = \left[y \mapsto k(x, y)\right] \in \mathscr{H}_k \quad$ and $\quad f(x) = \langle f, k(x, \cdot\,) \rangle_{\mathscr{H}_k}$ $\quad \forall f \in \mathcal{H}_k$

$[k(x,\cdot)] : \mathcal{X} \to \mathbb{R}$

$g_x(y) = k(x,y)$

g = functools.partial(k, x)





Moore-Aronszajn Theorem

- Building $\mathcal{H}$ for a given psd k:
  - Start with $\mathcal{H}_0 = \mathbf{span}(\{k(x, \cdot) : x \in \mathcal{X}\})$

Moore-Aronszajn Theorem

- Building $\mathcal{H}$ for a given psd k:
  - Start with $\mathcal{H}_0 = \mathbf{span}(\{k(x,\cdot) : x \in \mathcal{X}\})$
  - Define $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ from $\langle k(x,\cdot), k(y,\cdot) \rangle_{\mathcal{H}_0} = k(x,y)$

Moore-Aronszajn Theorem

- Building $\mathcal{H}$ for a given psd k:
  - Start with $\mathcal{H}_0 = \mathrm{span}(\{k(x,\cdot) : x \in \mathcal{X}\})$
  - Define $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ from $\langle k(x,\cdot), k(y,\cdot) \rangle_{\mathcal{H}_0} = k(x,y)$
  - Take $\mathcal{H}$ to be completion of $\mathcal{H}_0$ in the metric from $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$

$\sum \alpha_i \, k(x,\cdot)$

$\int \alpha(x) \, k(x,\cdot) \, dx$

Moore-Aronszajn Theorem

- Building $\mathcal{H}$ for a given psd k:
  - Start with $\mathcal{H}_0 = \mathbf{span}(\{k(x, \cdot) : x \in \mathcal{X}\})$
  - Define $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ from $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}_0} = k(x, y)$
  - Take $\mathcal{H}$ to be completion of $\mathcal{H}_0$ in the metric from $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$
  - Get that the reproducing property holds for $k(x, \cdot)$ in $\mathcal{H}$

Moore-Aronszajn Theorem

- Building $\mathcal{H}$ for a given psd k:
  - Start with $\mathcal{H}_0 = \mathbf{span}(\{k(x,\cdot) : x \in \mathcal{X}\})$
  - Define $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ from $\langle k(x,\cdot), k(y,\cdot) \rangle_{\mathcal{H}_0} = k(x,y)$
  - Take $\mathcal{H}$ to be completion of $\mathcal{H}_0$ in the metric from $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$
  - Get that the reproducing property holds for $k(x,\cdot)$ in $\mathcal{H}$
  - Can also show uniqueness

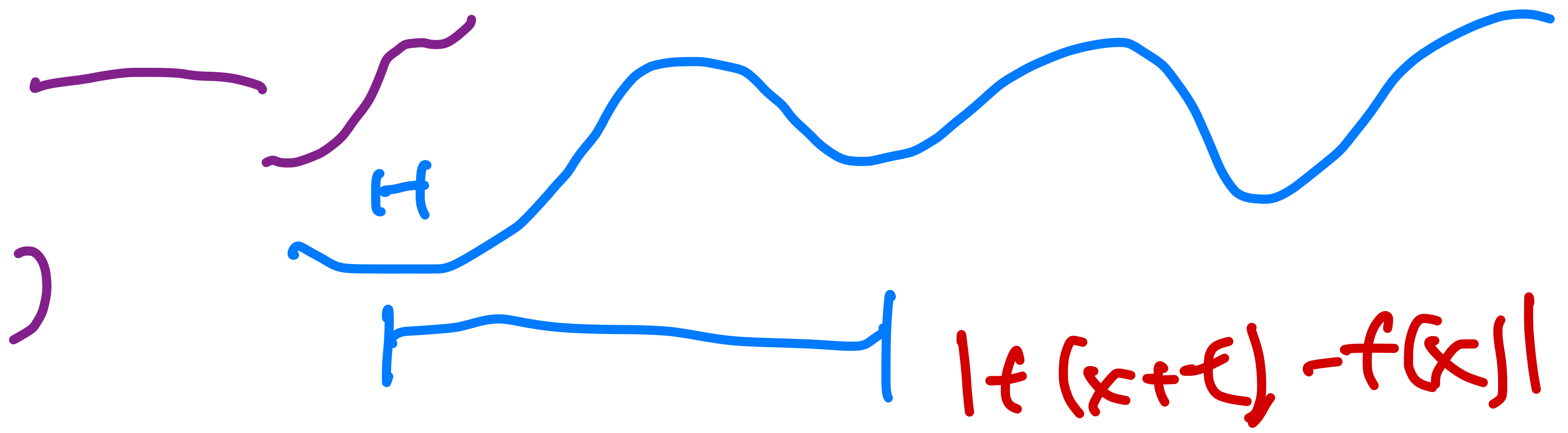
Moore-Aronszajn Theorem

- Building $\mathcal{H}$ for a given psd k:
  - Start with $\mathcal{H}_0 = \mathrm{span}(\{k(x, \cdot) : x \in \mathcal{X}\})$
  - Define $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ from $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}_0} = k(x, y)$
  - Take $\mathcal{H}$ to be completion of $\mathcal{H}_0$ in the metric from $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$
  - Get that the reproducing property holds for $k(x, \cdot)$ in $\mathcal{H}$
  - Can also show uniqueness
- Theorem: k is psd iff it's the reproducing kernel of an RKHS

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with $k(x, y) = k(y, x)$

$\|\sum_{i=1}^{m} a_i k(\tilde{x}_i, \cdot)\|^2_{\mathcal{H}_k} = a^T \tilde{K} a$

$\tilde{K}_{ij} = k(\tilde{x}_i, \tilde{x}_j)$



$|f(x+\epsilon) - f(x)| \le 2\|f\|_{\mathcal{H}} \left(1 - \exp\left(-\frac{\|\epsilon\|^2}{2\sigma^2}\right)\right)$

A quick check: linear kernels

- $k(x, y) = x^{\mathsf{T}} y$ on $\mathcal{X} = \mathbb{R}^d$

  $k(x, \cdot) = [y \mapsto x^T y]$ " = " x^T

- If $f(y) = \sum_{i=1}^{n} a_i k(x_i, y)$, then $f(y) = \left[\sum_{i=1}^{n} a_i x_i\right]^{\mathsf{T}} y$
- Closure doesn't add anything here, since $\mathbb{R}^d$ is closed
- So, linear kernel gives you RKHS of linear functions
- $\|f\|_{\mathcal{H}} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j)} = \|\sum_{i=1}^{n} a_i x_i\|$

  $= \sqrt{\langle f, f \rangle} = \sqrt{\langle \sum a_i x_i, \sum a_i x_i \rangle}$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2}_{L_S(f)} + \lambda \|f\|_{\mathcal{H}}^2$$

Kernel ridge regression

$$\hat{f} = \operatorname*{arg\,min}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

Linear kernel gives normal ridge regression:

$$\hat{f}(x) = \hat{w}^\mathsf{T} x; \quad \hat{w} = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (w^\mathsf{T} x_i - y_i)^2 + \lambda \|w\|^2$$

Nonlinear kernels will give nonlinear regression!

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$?

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**

- Let $\mathcal{H}_X = \mathrm{span}\{k(x_i, \cdot)\}_{i=1}^n$
  $\mathcal{H}_\perp$ its orthogonal complement in $\mathcal{H}$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**

- Let $\mathcal{H}_X = \text{span}\{k(x_i, \cdot)\}_{i=1}^{n}$
  $\mathcal{H}_\perp$ its orthogonal complement in $\mathcal{H}$
- Decompose $f = f_X + f_\perp$ with $f_{\mathcal{X}} \in \mathcal{H}_X$, $f_\perp \in \mathcal{H}_\perp$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**

- Let $\mathcal{H}_X = \text{span}\{k(x_i, \cdot)\}_{i=1}^{n}$
  $\mathcal{H}_\perp$ its orthogonal complement in $\mathcal{H}$
- Decompose $f = f_X + f_\perp$ with $f_X \in \mathcal{H}_X$, $f_\perp \in \mathcal{H}_\perp$
- $f(x_i) = \langle f_X + f_\perp, k(x_i, \cdot)\rangle_{\mathcal{H}} = \langle f_X, k(x_i, \cdot)\rangle_{\mathcal{H}}$

$$\langle f_\perp, \underbrace{k(x_i, \cdot)}_{\in \mathcal{H}_X} \rangle_{\mathcal{H}} = 0$$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**

- Let $\mathcal{H}_X = \mathrm{span}\{k(x_i, \cdot)\}_{i=1}^n$
  $\mathcal{H}_\perp$ its orthogonal complement in $\mathcal{H}$
- Decompose $f = f_X + f_\perp$ with $f_X \in \mathcal{H}_X$, $f_\perp \in \mathcal{H}_\perp$
- $f(x_i) = \langle f_X + f_\perp, k(x_i, \cdot) \rangle_{\mathcal{H}} = \langle f_X, k(x_i, \cdot) \rangle_{\mathcal{H}}$
- $\|f\|_{\mathcal{H}}^2 = \|f_X\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2$ $+ 2 \underbrace{\langle f_X, f_\perp \rangle_{\mathcal{H}}}_{0}$

$= \langle f_X + f_\perp, f_X + f_\perp \rangle_{\mathcal{H}}$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**

- Let $\mathcal{H}_X = \operatorname{span}\{k(x_i, \cdot)\}_{i=1}^{n}$
  $\mathcal{H}_\perp$ its orthogonal complement in $\mathcal{H}$
- Decompose $f = f_X + f_\perp$ with $f_X \in \mathcal{H}_X$, $f_\perp \in \mathcal{H}_\perp$
- $f(x_i) = \langle f_X + f_\perp, k(x_i, \cdot) \rangle_{\mathcal{H}} = \langle f_X, k(x_i, \cdot) \rangle_{\mathcal{H}}$
- $\|f\|_{\mathcal{H}}^2 = \|f_X\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2$
- Minimizer needs $f_\perp = 0$, and so $\hat{f} = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$

$\hat{f}(\tilde{x}) = \sum_{i=1}^{n} \alpha_i k(x_i, \tilde{x})$
$= \alpha^T K_S(\tilde{x})$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, \cdot)$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, \cdot)$

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{n} \alpha_j k(x_i, x_j) - y_i \right)^2 = \sum_{i=1}^{n} ([K\alpha]_i - y_i)^2$$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, \cdot)$

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{n} \alpha_j k(x_i, x_j) - y_i \right)^2 = \sum_{i=1}^{n} ([K\alpha]_i - y_i)^2 = \|K\alpha - y\|^2$$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, \cdot)$

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{n} \alpha_j k(x_i, x_j) - y_i \right)^2 = \sum_{i=1}^{n} ([K\alpha]_i - y_i)^2 = \|K\alpha - y\|^2$$

$$= \alpha^\top K^2 \alpha - 2y^\top K\alpha + y^\top y$$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, \cdot)$

$$\begin{aligned} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \alpha_j k(x_i, x_j) - y_i \right)^2 &= \sum_{i=1}^{n} ([K\alpha]_i - y_i)^2 = \|K\alpha - y\|^2 \\ &= \alpha^\mathsf{T} K^2 \alpha - 2y^\mathsf{T} K\alpha + y^\mathsf{T} y \end{aligned}$$

$$\left\| \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \right\|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i k(x_i, x_j) \alpha_j$$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, \cdot)$

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{n} \alpha_j k(x_i, x_j) - y_i \right)^2 = \sum_{i=1}^{n} ([K\alpha]_i - y_i)^2 = \|K\alpha - y\|^2$$

$$= \alpha^\mathsf{T} K^2 \alpha - 2y^\mathsf{T} K\alpha + y^\mathsf{T} y$$

$$\left\| \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \right\|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i k(x_i, x_j) \alpha_j = \alpha^\mathsf{T} K \alpha$$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, \cdot)$

$$\hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^n} \alpha^\top K^2 \alpha - 2y^\top K\alpha + y^\top y + n\lambda\alpha^\top K\alpha$$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, \cdot)$

$$\begin{aligned}\hat{\alpha} &= \arg\min_{\alpha \in \mathbb{R}^n} \alpha^\top K^2 \alpha - 2y^\top K\alpha + y^\top y + n\lambda\alpha^\top K\alpha \\ &= \arg\min_{\alpha \in \mathbb{R}^n} \alpha^\top K(K + n\lambda I)\alpha - 2y^\top K\alpha\end{aligned}$$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, \cdot)$

$$\begin{aligned} \hat{\alpha} &= \arg\min_{\alpha \in \mathbb{R}^n} \alpha^\top K^2 \alpha - 2y^\top K\alpha + y^\top y + n\lambda \alpha^\top K\alpha \\ &= \arg\min_{\alpha \in \mathbb{R}^n} \alpha^\top K(K + n\lambda I)\alpha - 2y^\top K\alpha \end{aligned}$$

Setting derivative to zero gives $K(K + n\lambda I)\hat{\alpha} = Ky$,
satisfied by $\hat{\alpha} = (K + n\lambda I)^{-1} y$

Kernel ridge regression

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How to find $\hat{f}$? **Representer Theorem**: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, \cdot)$

$$\begin{aligned}\hat{\alpha} &= \arg\min_{\alpha \in \mathbb{R}^n} \alpha^\top K^2 \alpha - 2y^\top K\alpha + y^\top y + n\lambda\alpha^\top K\alpha \\ &= \arg\min_{\alpha \in \mathbb{R}^n} \alpha^\top K(K + n\lambda I)\alpha - 2y^\top K\alpha\end{aligned}$$

Setting derivative to zero gives $K(K + n\lambda I)\hat{\alpha} = Ky$,

satisfied by $\hat{\alpha} = (K + n\lambda I)^{-1} y$

$$k_S(x) = \begin{bmatrix} k(x_1, x) \\ \vdots \\ k(x_n, x) \end{bmatrix}$$

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, x) = \hat{\alpha}^\top k_S(x) = y^\top (K + n\lambda I)^{-1} k_S(x)$$

Other kernel algorithms

- Representer theorem applies if R strictly increasing:

$$\min_{f \in \mathcal{H}} L(f(x_1), \cdots, f(x_n)) + R(\|f\|_{\mathcal{H}})$$

- Classification algorithms:
  - Support vector machines: L is hinge loss
  - Kernel logistic regression: L is logistic loss
- Principal component analysis, canonical correlation analysis
- Many, many more...

Rademacher complexity

- Let $\mathscr{H}_{k,B} = \{f \in \mathscr{H}_k : \|f\|_{\mathscr{H}_k} \leq B\}$
- Let $S = (x_1, \ldots, x_n)$ have kernel matrix $K \in \mathbb{R}^{n\times n}$: $K_{ij} = k(x_i, x_j)$

$$\hat{R}_S(\mathcal{H}_{k,B}) = \frac{1}{n}\mathbb{E}_\sigma \sup_{f\in\mathcal{H}_{k,B}} \underbrace{\sum_{i=1}^n \sigma_i \underbrace{f(x_i)}_{\langle f, k(x_i,\cdot)\rangle_{\mathcal{H}}}}_{\langle f, \sum_{i=1}^n \sigma_i k(x_i,\cdot)\rangle_{\mathcal{H}}}$$

$$\leq \frac{B}{n}\mathbb{E}_\sigma \left\|\sum_{i=1}^n \sigma_i k(x_i,\cdot)\right\|_{\mathcal{H}}$$

$$\leq \frac{B}{n}\sqrt{\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i k(x_i,\cdot) \right\|_{\mathcal{H}}^2} = \frac{B}{n}\sqrt{\mathbb{E}_\sigma \sigma^T K \sigma} = \frac{B}{n}\sqrt{\mathrm{Tr}(K)} \leq \frac{BR}{\sqrt{n}}$$

if $k(x,x) \leq R^2$

$$\mathbb{E}_\sigma \sigma^T K \sigma = \mathbb{E}\sum_i \sigma_i^2 k(x_i,x_i) + \sum_{i\neq j} \cancel{\mathbb{E}\,\sigma_i\sigma_j}\, k(x_i,x_j)$$

$$= \sum k(x_i,x_i) = \mathrm{Tr}(K)$$

Estimation error bounds: SVMs

- Same ramp loss analysis as before: if $\mathbb{E}k(x,x) \leq R^2$,

$$\mathscr{L}_{\mathscr{D}}^{0-1}(\hat{f}) \leq \mathscr{L}_{\mathscr{D}}^{\text{ramp}}(\hat{f}) + \frac{2RB}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$ for $\mathscr{H}_{k,B} = \{f \in \mathscr{H}_k : \|f\|_{\mathscr{H}_k} \leq B\}$

Estimation error bounds: SVMs

- Same ramp loss analysis as before: if $\mathbb{E}k(x,x) \leq R^2$,

$$\mathcal{L}_{\mathcal{D}}^{0-1}(\hat{f}) \leq \mathcal{L}_{\mathcal{D}}^{\mathrm{ramp}}(\hat{f}) + \frac{2RB}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$ for $\mathcal{H}_{k,B} = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq B\}$

- (or the version with $B = 2\max\{\|\hat{f}\|_{\mathcal{H}_k}, 1\}$ and a $\sqrt{\frac{1}{n}\log\log_2\|\hat{f}\|_{\mathcal{H}_k}}$ penalty)

Estimation error bounds: SVMs

- Same ramp loss analysis as before: if $\mathbb{E}k(x, x) \leq R^2$,

$$\mathscr{L}_{\mathscr{D}}^{0-1}(\hat{f}) \leq \mathscr{L}_{\mathscr{D}}^{\text{ramp}}(\hat{f}) + \frac{2RB}{\sqrt{n}} + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} \text{ for } \mathscr{H}_{k,B} = \{f \in \mathscr{H}_k : \|f\|_{\mathscr{H}_k} \leq B\}$$

  - (or the version with $B = 2\max\{\|\hat{f}\|_{\mathscr{H}_k}, 1\}$ and a $\sqrt{\frac{1}{n} \log \log_2 \|\hat{f}\|_{\mathscr{H}_k}}$ penalty)

- Stability analysis also still works: if $\Pr(k(x, x) \leq R^2) = 1$,

Estimation error bounds: SVMs

- Same ramp loss analysis as before: if $\mathbb{E}k(x,x) \le R^2$,

  $\mathcal{L}_{\mathcal{D}}^{0-1}(\hat{f}) \le \mathcal{L}_{\mathcal{D}}^{\text{ramp}}(\hat{f}) + \dfrac{2RB}{\sqrt{n}} + \sqrt{\dfrac{1}{2n}\log\dfrac{1}{\delta}}$ for $\mathcal{H}_{k,B} = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \le B\}$

  - (or the version with $B = 2\max\{\|\hat{f}\|_{\mathcal{H}_k}, 1\}$ and a $\sqrt{\frac{1}{n}\log\log_2\|\hat{f}\|_{\mathcal{H}_k}}$ penalty)

- Stability analysis also still works: if $\Pr(k(x,x) \le R^2) = 1$,

  - $\mathbb{E}_S[L_{\mathcal{D}}^{0-1}(\hat{f})] \le \inf\limits_{\|f\|_{\mathcal{H}_k} \le B} L_{\mathcal{D}}^{\text{hinge}}(f) + 2RB\sqrt{\frac{2}{n}}$ for Soft-SVM with $\lambda = \frac{R}{B}\sqrt{\frac{2}{n}}$

Estimation error bounds: KRR

- Assume $\Pr(k(x,x) \leq R^2) = 1$

Estimation error bounds: KRR

- Assume $\Pr(k(x,x) \leq R^2) = 1$
- If targets y are bounded, say $|y| \leq BR$ for simplicity: analyzed way back in lecture 8

Estimation error bounds: KRR

- Assume $\Pr(k(x, x) \leq R^2) = 1$
- If targets y are bounded, say $|y| \leq BR$ for simplicity: analyzed way back in lecture 8
  - For $\mathscr{H}_{k,B} = \{f \in \mathscr{H}_k : \|f\|_{\mathscr{H}_k} \leq B\}$, have $|f(x)| \leq B\sqrt{k(x, x)} \leq BR$

Estimation error bounds: KRR

- Assume $\Pr(k(x, x) \leq R^2) = 1$
- If targets y are bounded, say $|y| \leq BR$ for simplicity: analyzed way back in lecture 8
    - For $\mathscr{H}_{k,B} = \{f \in \mathscr{H}_k : \|f\|_{\mathscr{H}_k} \leq B\}$, have $|f(x)| \leq B\sqrt{k(x, x)} \leq BR$
    - Makes square loss effectively $(4BR)$-Lipschitz and bounded in $[0,4B^2R^2]$:

Estimation error bounds: KRR

- Assume $\Pr(k(x,x) \leq R^2) = 1$
- If targets y are bounded, say $|y| \leq BR$ for simplicity: analyzed way back in lecture 8
  - For $\mathscr{H}_{k,B} = \{f \in \mathscr{H}_k : \|f\|_{\mathscr{H}_k} \leq B\}$, have $|f(x)| \leq B\sqrt{k(x,x)} \leq BR$
  - Makes square loss effectively $(4BR)$-Lipschitz and bounded in $[0,4B^2R^2]$:
  - Get that $\sup_{f \in \mathscr{H}_{k,B}} L_{\mathscr{D}}^{\mathrm{sq}}(f) - L_S^{\mathrm{sq}}(f) \leq \dfrac{4B^2R^2}{\sqrt{n}}\left(1 + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)$

Estimation error bounds: KRR

- Assume $\Pr(k(x,x) \le R^2) = 1$
- If targets y are bounded, say $|y| \le BR$ for simplicity: analyzed way back in lecture 8
  - For $\mathscr{H}_{k,B} = \{f \in \mathscr{H}_k : \|f\|_{\mathscr{H}_k} \le B\}$, have $|f(x)| \le B\sqrt{k(x,x)} \le BR$
  - Makes square loss effectively $(4BR)$-Lipschitz and bounded in $[0, 4B^2R^2]$:
  - Get that $\sup_{f \in \mathscr{H}_{k,B}} L_{\mathscr{D}}^{\mathrm{sq}}(f) - L_S^{\mathrm{sq}}(f) \le \frac{4B^2R^2}{\sqrt{n}}\left(1 + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)$
- Stability analysis also works: if $\Pr(k(x,x) \le R^2) = 1$,

Estimation error bounds: KRR

- Assume $\Pr(k(x,x) \le R^2) = 1$
- If targets y are bounded, say $|y| \le BR$ for simplicity: analyzed way back in lecture 8
  - For $\mathcal{H}_{k,B} = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \le B\}$, have $|f(x)| \le B\sqrt{k(x,x)} \le BR$
  - Makes square loss effectively $(4BR)$-Lipschitz and bounded in $[0,4B^2R^2]$:
  - Get that $\displaystyle\sup_{f\in\mathcal{H}_{k,B}} L_{\mathcal{D}}^{\text{sq}}(f) - L_S^{\text{sq}}(f) \le \frac{4B^2R^2}{\sqrt{n}}\left(1 + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)$

- Stability analysis also works: if $\Pr(k(x,x) \le R^2) = 1$,
  - $\displaystyle\mathbb{E}_S[L_{\mathcal{D}}^{\text{sq}}(\hat{f})] \le \inf_{\|f\|_{\mathcal{H}_k}\le B} L_{\mathcal{D}}^{\text{sq}}(f) + RB\sqrt{\frac{150}{n}}$ for KRR with $\displaystyle\lambda = \frac{R}{B}\sqrt{\frac{50}{3n}}$

Universal kernels

- What about that $L_S(f)$ or $\inf\limits_{\|f\|_{\mathcal{H}_k}} L_{\mathcal{D}}(f)$ term?

We stopped here in class;
will do (most of) the rest
on Monday

Universal kernels

- What about that $L_S(f)$ or $\inf_{\|f\|_{\mathcal{H}_k}} L_{\mathcal{D}}(f)$ term?
- A continuous kernel on a compact metric space $\mathcal{X}$ is **universal** if $\mathcal{H}_k$ is dense in $C(\mathcal{X})$:
  for every continuous $g : \mathcal{X} \to \mathbb{R}$, every $\varepsilon > 0$, there is an $f \in \mathcal{H}_k$
  with $\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)| \leq \varepsilon$

Universal kernels

- What about that $L_S(f)$ or $\inf_{\|f\|_{\mathcal{H}_k}} L_{\mathcal{D}}(f)$ term?
- A continuous kernel on a compact metric space $\mathcal{X}$ is **universal** if $\mathcal{H}_k$ is dense in $C(\mathcal{X})$: for every continuous $g : \mathcal{X} \to \mathbb{R}$, every $\varepsilon > 0$, there is an $f \in \mathcal{H}_k$ with $\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)| \leq \varepsilon$
  - If $\mathcal{X}$ is a topological space *not* generated by a metric, there is no universal kernel (Steinwart/Christmann exercise 4.13)

Universal kernels

- What about that $L_S(f)$ or $\inf_{\|f\|_{\mathcal{H}_k}} L_{\mathcal{D}}(f)$ term?
- A continuous kernel on a compact metric space $\mathcal{X}$ is **universal** if $\mathcal{H}_k$ is dense in $C(\mathcal{X})$: for every continuous $g : \mathcal{X} \to \mathbb{R}$, every $\varepsilon > 0$, there is an $f \in \mathcal{H}_k$ with $\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)| \leq \varepsilon$
    - If $\mathcal{X}$ is a topological space *not* generated by a metric, there is no universal kernel (Steinwart/Christmann exercise 4.13)
    - Separates compact sets: if $X_1 \cap X_2 = \emptyset$ are compact subsets of $\mathcal{X}$, is an $f \in \mathcal{H}_k$ with $f(x) > 0$ for $x \in X_1$, $f(x) < 0$ for $x \in X_2$ (so VCdim = ∞)

Universal kernels

- What about that $L_S(f)$ or $\inf_{\|f\|_{\mathcal{H}_k}} L_{\mathcal{D}}(f)$ term?
- A continuous kernel on a compact metric space $\mathcal{X}$ is **universal** if $\mathcal{H}_k$ is dense in $C(\mathcal{X})$: for every continuous $g : \mathcal{X} \to \mathbb{R}$, every $\varepsilon > 0$, there is an $f \in \mathcal{H}_k$ with $\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)| \leq \varepsilon$
  - If $\mathcal{X}$ is a topological space *not* generated by a metric, there is no universal kernel (Steinwart/Christmann exercise 4.13)
  - Separates compact sets: if $X_1 \cap X_2 = \emptyset$ are compact subsets of $\mathcal{X}$, is an $f \in \mathcal{H}_k$ with $f(x) > 0$ for $x \in X_1$, $f(x) < 0$ for $x \in X_2$ (so VCdim = ∞)
    - Implies that as $B \to \infty$, get $\inf_{\mathcal{H}_{k,B}} L_S(f) \to 0$, $\inf_{\mathcal{H}_{k,B}} L_{\mathcal{D}}(f) \to$ Bayes error if $\mathcal{D}$ has compact support

Universal kernels

- What about that $L_S(f)$ or $\inf_{\|f\|_{\mathcal{H}_k}} L_{\mathcal{D}}(f)$ term?
- A continuous kernel on a compact metric space $\mathcal{X}$ is **universal** if $\mathcal{H}_k$ is dense in $C(\mathcal{X})$:
  for every continuous $g : \mathcal{X} \to \mathbb{R}$, every $\varepsilon > 0$, there is an $f \in \mathcal{H}_k$
  with $\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)| \leq \varepsilon$
  - If $\mathcal{X}$ is a topological space *not* generated by a metric, there is no universal kernel (Steinwart/Christmann exercise 4.13)
  - Separates compact sets: if $X_1 \cap X_2 = \emptyset$ are compact subsets of $\mathcal{X}$,
    is an $f \in \mathcal{H}_k$ with $f(x) > 0$ for $x \in X_1$, $f(x) < 0$ for $x \in X_2$ (so VCdim = ∞)
    - Implies that as $B \to \infty$, get $\inf_{\mathcal{H}_{k,B}} L_S(f) \to 0$, $\inf_{\mathcal{H}_{k,B}} L_{\mathcal{D}}(f) \to$ Bayes error if $\mathcal{D}$ has compact support
  - Can show universality via Stone-Weierstrass, or Fourier properties

Universal kernels

- What about that $L_S(f)$ or $\inf_{\|f\|_{\mathcal{H}_k}} L_{\mathcal{D}}(f)$ term?
- A continuous kernel on a compact metric space $\mathcal{X}$ is **universal** if $\mathcal{H}_k$ is dense in $C(\mathcal{X})$:
  for every continuous $g : \mathcal{X} \to \mathbb{R}$, every $\varepsilon > 0$, there is an $f \in \mathcal{H}_k$
  with $\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)| \leq \varepsilon$
  - If $\mathcal{X}$ is a topological space *not* generated by a metric, there is no universal kernel (Steinwart/Christmann exercise 4.13)
  - Separates compact sets: if $X_1 \cap X_2 = \emptyset$ are compact subsets of $\mathcal{X}$,
    is an $f \in \mathcal{H}_k$ with $f(x) > 0$ for $x \in X_1$, $f(x) < 0$ for $x \in X_2$ (so VCdim = ∞)
    - Implies that as $B \to \infty$, get $\inf_{\mathcal{H}_{k,B}} L_S(f) \to 0$, $\inf_{\mathcal{H}_{k,B}} L_{\mathcal{D}}(f) \to$ Bayes error if $\mathcal{D}$ has compact support
  - Can show universality via Stone-Weierstrass, or Fourier properties
  - $\exp(x^\top y)$, $\exp(-\frac{1}{2\sigma^2}\|x - y\|^2)$, $\exp(-\frac{1}{\sigma}\|x - y\|)$ are universal on compact subsets of $\mathbb{R}^d$

Universal kernels

- What about that $L_S(f)$ or $\inf_{\|f\|_{\mathcal{H}_k}} L_{\mathcal{D}}(f)$ term?
- A continuous kernel on a compact metric space $\mathcal{X}$ is **universal** if $\mathcal{H}_k$ is dense in $C(\mathcal{X})$:
  for every continuous $g : \mathcal{X} \to \mathbb{R}$, every $\varepsilon > 0$, there is an $f \in \mathcal{H}_k$
  with $\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)| \leq \varepsilon$
  - If $\mathcal{X}$ is a topological space *not* generated by a metric, there is no universal kernel (Steinwart/Christmann exercise 4.13)
  - Separates compact sets: if $X_1 \cap X_2 = \emptyset$ are compact subsets of $\mathcal{X}$,
    is an $f \in \mathcal{H}_k$ with $f(x) > 0$ for $x \in X_1$, $f(x) < 0$ for $x \in X_2$ (so VCdim = ∞)
    - Implies that as $B \to \infty$, get $\inf_{\mathcal{H}_{k,B}} L_S(f) \to 0$, $\inf_{\mathcal{H}_{k,B}} L_{\mathcal{D}}(f) \to$ Bayes error if $\mathcal{D}$ has compact support
  - Can show universality via Stone-Weierstrass, or Fourier properties
  - $\exp(x^\top y)$, $\exp(-\frac{1}{2\sigma^2}\|x-y\|^2)$, $\exp(-\frac{1}{\sigma}\|x-y\|)$ are universal on compact subsets of $\mathbb{R}^d$
  - Never true for finite-dimensional kernels

Approximation error

- Know that as $B \to \infty$, get $\inf_{\mathcal{H}_{k,B}} L_S(f) \to 0$, $\inf_{\mathcal{H}_{k,B}} L_{\mathcal{D}}(f) \to$ Bayes error for compactly supported $\mathcal{D}$ (can use broader notion of universality in general)

Approximation error

- Know that as $B \to \infty$, get $\inf_{\mathscr{H}_{k,B}} L_S(f) \to 0$, $\inf_{\mathscr{H}_{k,B}} L_{\mathscr{D}}(f) \to$ Bayes error for compactly supported $\mathscr{D}$ (can use broader notion of universality in general)
  - But the rate at which this happens depends on $\mathscr{D}$

Approximation error

- Know that as $B \to \infty$, get $\inf_{\mathcal{H}_{k,B}} L_S(f) \to 0$, $\inf_{\mathcal{H}_{k,B}} L_{\mathcal{D}}(f) \to$ Bayes error for compactly supported $\mathcal{D}$ (can use broader notion of universality in general)
  - But the rate at which this happens depends on $\mathcal{D}$
- Usually compare to the **regression function** $f_{\mathcal{D}}(x) = \mathbb{E}[y \mid x]$

Approximation error

- Know that as $B \to \infty$, get $\inf_{\mathcal{H}_{k,B}} L_S(f) \to 0$, $\inf_{\mathcal{H}_{k,B}} L_{\mathcal{D}}(f) \to$ Bayes error for compactly supported $\mathcal{D}$ (can use broader notion of universality in general)
  - But the rate at which this happens depends on $\mathcal{D}$
- Usually compare to the **regression function** $f_{\mathcal{D}}(x) = \mathbb{E}[y \mid x]$
  - If $f_{\mathcal{D}} \in \mathcal{H}_k$, called **well-specified**:

Approximation error

- Know that as $B \to \infty$, get $\inf_{\mathscr{H}_{k,B}} L_S(f) \to 0$, $\inf_{\mathscr{H}_{k,B}} L_{\mathscr{D}}(f) \to$ Bayes error for compactly supported $\mathscr{D}$ (can use broader notion of universality in general)
  - But the rate at which this happens depends on $\mathscr{D}$
- Usually compare to the **regression function** $f_{\mathscr{D}}(x) = \mathbb{E}[y \mid x]$
  - If $f_{\mathscr{D}} \in \mathscr{H}_k$, called **well-specified**:
    - Stability for $B = \|f_{\mathscr{D}}\|_{\mathscr{H}_k}$: $\inf_{\|f\|_{\mathscr{H}_k} \leq B} L_{\mathscr{D}}(f)$ = Bayes error, excess error $\leq \mathscr{O}(1/\sqrt{n})$

Approximation error

- Know that as $B \to \infty$, get $\inf_{\mathscr{H}_{k,B}} L_S(f) \to 0$, $\inf_{\mathscr{H}_{k,B}} L_{\mathscr{D}}(f) \to$ Bayes error for compactly supported $\mathscr{D}$ (can use broader notion of universality in general)
    - But the rate at which this happens depends on $\mathscr{D}$
- Usually compare to the **regression function** $f_{\mathscr{D}}(x) = \mathbb{E}[y \mid x]$
    - If $f_{\mathscr{D}} \in \mathscr{H}_k$, called **well-specified**:
        - Stability for $B = \|f_{\mathscr{D}}\|_{\mathscr{H}_k}$: $\inf_{\|f\|_{\mathscr{H}_k} \leq B} L_{\mathscr{D}}(f)$ = Bayes error, excess error $\leq \mathcal{O}(1/\sqrt{n})$
        - Better rates (minimax-optimal) with "range-space condition" if $f_{\mathscr{D}}$ is "nice" in $\mathscr{H}_k$

Approximation error

- Know that as $B \to \infty$, get $\inf_{\mathcal{H}_{k,B}} L_S(f) \to 0$, $\inf_{\mathcal{H}_{k,B}} L_{\mathcal{D}}(f) \to$ Bayes error for compactly supported $\mathcal{D}$ (can use broader notion of universality in general)
  - But the rate at which this happens depends on $\mathcal{D}$
- Usually compare to the **regression function** $f_{\mathcal{D}}(x) = \mathbb{E}[y \mid x]$
  - If $f_{\mathcal{D}} \in \mathcal{H}_k$, called **well-specified**:
    - Stability for $B = \|f_{\mathcal{D}}\|_{\mathcal{H}_k}$: $\inf_{\|f\|_{\mathcal{H}_k} \leq B} L_{\mathcal{D}}(f)$ = Bayes error, excess error $\leq \mathcal{O}(1/\sqrt{n})$
    - Better rates (minimax-optimal) with "range-space condition" if $f_{\mathcal{D}}$ is "nice" in $\mathcal{H}_k$
      - Pretty different style of analysis, based on $\|\hat{f} - f_{\mathcal{D}}\|_{\mathcal{H}_k}$

Approximation error

- Know that as $B \to \infty$, get $\inf_{\mathcal{H}_{k,B}} L_S(f) \to 0$, $\inf_{\mathcal{H}_{k,B}} L_{\mathcal{D}}(f) \to$ Bayes error for compactly supported $\mathcal{D}$ (can use broader notion of universality in general)
  - But the rate at which this happens depends on $\mathcal{D}$
- Usually compare to the **regression function** $f_{\mathcal{D}}(x) = \mathbb{E}[y \mid x]$
  - If $f_{\mathcal{D}} \in \mathcal{H}_k$, called **well-specified**:
    - Stability for $B = \|f_{\mathcal{D}}\|_{\mathcal{H}_k}$: $\inf_{\|f\|_{\mathcal{H}_k} \leq B} L_{\mathcal{D}}(f)$ = Bayes error, excess error $\leq \mathcal{O}(1/\sqrt{n})$
    - Better rates (minimax-optimal) with "range-space condition" if $f_{\mathcal{D}}$ is "nice" in $\mathcal{H}_k$
      - Pretty different style of analysis, based on $\|\hat{f} - f_{\mathcal{D}}\|_{\mathcal{H}_k}$
- Misspecified case: more complicated analyses based on "approximation spaces"

Gaussian processes

- $f \sim \mathrm{GP}(m, k)$ is a **random function** $f : \mathcal{X} \to \mathbb{R}$ s.t., for any $x_1, \ldots, x_n$,

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \ldots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \ldots & k(x_n, x_n) \end{bmatrix} \right)$$

Gaussian processes

- $f \sim \mathrm{GP}(m, k)$ is a **random function** $f : \mathcal{X} \to \mathbb{R}$ s.t., for any $x_1, \ldots, x_n$,

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \ldots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \ldots & k(x_n, x_n) \end{bmatrix} \right)$$

- Mean function $m : \mathcal{X} \to \mathbb{R}$ can be any function; usually use 0

Gaussian processes

- $f \sim \mathrm{GP}(m, k)$ is a **random function** $f : \mathcal{X} \to \mathbb{R}$ s.t., for any $x_1, \ldots, x_n$,

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \ldots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \ldots & k(x_n, x_n) \end{bmatrix} \right)$$

- Mean function $m : \mathcal{X} \to \mathbb{R}$ can be any function; usually use 0
  - will see that we can just shift everything by m so that this is WLOG

Gaussian processes

- $f \sim \mathrm{GP}(m, k)$ is a **random function** $f : \mathcal{X} \to \mathbb{R}$ s.t., for any $x_1, \ldots, x_n$,
$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \ldots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \ldots & k(x_n, x_n) \end{bmatrix} \right)$$
- Mean function $m : \mathcal{X} \to \mathbb{R}$ can be any function; usually use 0
  - will see that we can just shift everything by m so that this is WLOG
- Covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be any psd function, i.e. any kernel

Gaussian process regression

Gaussian process regression

- Assume a **prior** $f \sim \mathrm{GP}(m, k)$

Gaussian process regression

- Assume a **prior** $f \sim \mathrm{GP}(m, k)$
- Assume **likelihood** of observations by $y_i \sim \mathcal{N}(f(x_i), \sigma^2)$

Gaussian process regression

- Assume a **prior** $f \sim \mathrm{GP}(m, k)$
- Assume **likelihood** of observations by $y_i \sim \mathcal{N}(f(x_i), \sigma^2)$
  - $\mathbb{E}[y_i] = \mathbb{E}[f(x_i)], \quad \mathrm{Cov}(y_i, y_j) = \mathrm{Cov}(f(x_i), f(x_j)) + \sigma^2\delta_{ij}$

Gaussian process regression

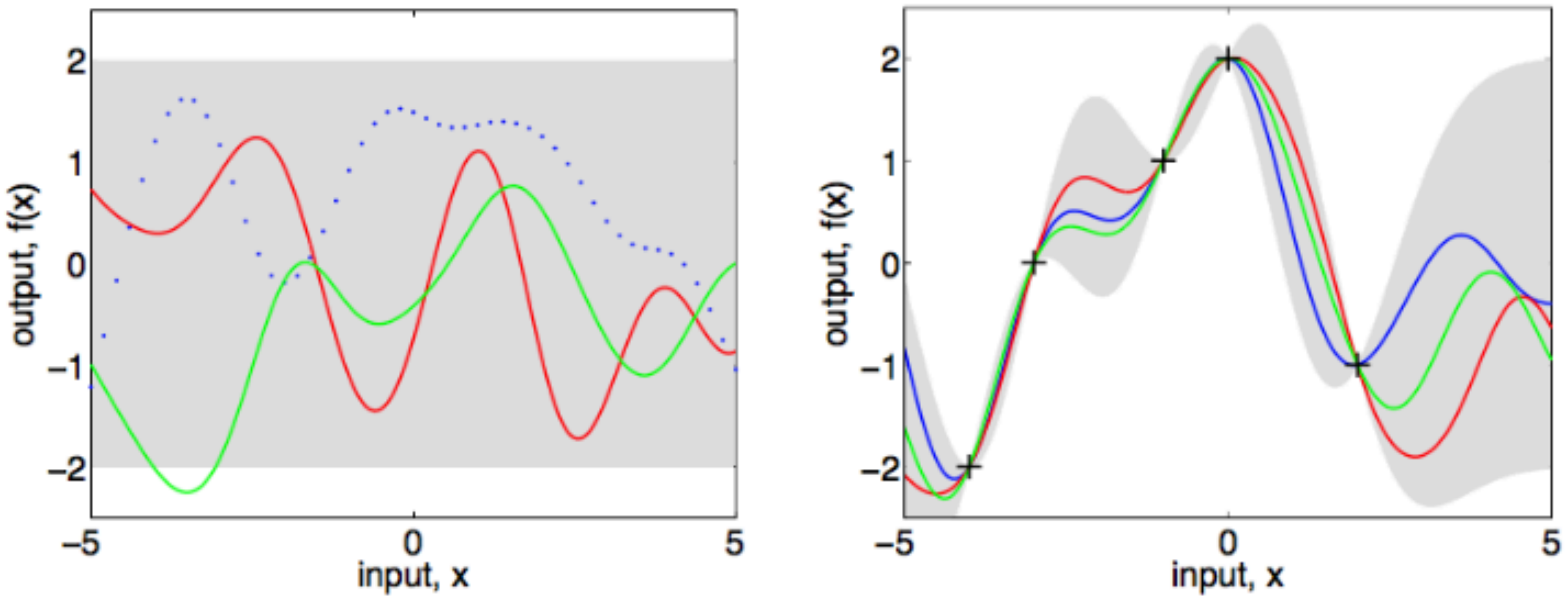
- Assume a **prior** $f \sim \mathrm{GP}(m, k)$
- Assume **likelihood** of observations by $y_i \sim \mathcal{N}(f(x_i), \sigma^2)$
  - $\mathbb{E}[y_i] = \mathbb{E}[f(x_i)], \quad \mathrm{Cov}(y_i, y_j) = \mathrm{Cov}(f(x_i), f(x_j)) + \sigma^2 \delta_{ij}$
- The **posterior** works out to be (via Kolmogorov Extension Theorem)

$$f \mid S \sim \mathrm{GP}\left(\left[x \mapsto y^\top (K_S + \sigma^2 I)^{-1} k_S(x)\right], \left[(x, x') \mapsto k(x, x') - k_S(x)^\top (K_S + \sigma^2 I)^{-1} k_S(x')\right]\right)$$

Gaussian process regression

- Assume a **prior** $f \sim \mathrm{GP}(m, k)$
- Assume **likelihood** of observations by $y_i \sim \mathcal{N}(f(x_i), \sigma^2)$
  - $\mathbb{E}[y_i] = \mathbb{E}[f(x_i)], \quad \mathrm{Cov}(y_i, y_j) = \mathrm{Cov}(f(x_i), f(x_j)) + \sigma^2 \delta_{ij}$
- The **posterior** works out to be (via Kolmogorov Extension Theorem)

$$f \mid S \sim \mathrm{GP}\left(\left[x \mapsto y^\top (K_S + \sigma^2 I)^{-1} k_S(x)\right], \left[(x, x') \mapsto k(x, x') - k_S(x)^\top (K_S + \sigma^2 I)^{-1} k_S(x')\right] \right)$$



(a), prior (b), posterior

More Gaussian Processes

- GP regression: can get **posterior contraction rates**
  - Look like KRR analysis for the mean, plus posterior variance decreasing
- Understanding posterior variance can be very useful!
  - e.g. Bayesian optimization / active learning / bandits / …
- GP classifiers: usual choice corresponds to kernel logistic regression

More resources

- Foundations: Berlinet and Thomas-Agnan, RKHSes in Probability and Stats (2004)
- Including more hardcore details: Steinwart and Christmann, SVMs (2008)
- Ridge regression analyses:
  - Smale and Zhou (2007) – fairly readable
  - Caponnetto and de Vito (2007) – minimax rate for “mostly”-well-specified, harder
  - Steinwart et al. (2009) – minimax in Sobolev spaces

- Rasmussen and Williams, Gaussian Processes for Machine Learning (2006)

- Connections between kernels and GPs: Kanagawa et al. (2018)
- Mean embeddings (slides after this, if we get there): Muandet et al. (2016)

Mean embeddings of distributions

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

This is "bonus material" we probably won't cover

Mean embeddings of distributions

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

Mean embeddings of distributions

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

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

Mean embeddings of distributions

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

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

Mean embeddings of distributions

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

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

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

Mean embeddings of distributions

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

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

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

- Okay. Why?

Mean embeddings of distributions

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

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

  - Last step assumed e.g. $\mathbb{E}\sqrt{k(X, X)} < \infty$
- Okay. Why?
  - One reason: ML on distributions [Szabó+ JMLR-16]

Mean embeddings of distributions

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

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

  - Last step assumed e.g. $\mathbb{E} \sqrt{k(X, X)} < \infty$
- Okay. Why?
  - One reason: ML on distributions [Szabó+ JMLR-16]
  - More common reason: comparing distributions

Maximum Mean Discrepancy

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

Maximum Mean Discrepancy

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

- Last line is Integral Probability Metric (IPM) form

Maximum Mean Discrepancy

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

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

Maximum Mean Discrepancy

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

- 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

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

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

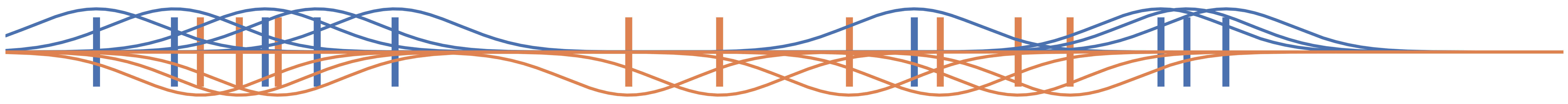
$$f^*(t) \propto \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, k(t,\cdot)\rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}\, k(t,X) - \mathbb{E}_{\mathbb{Q}}\, k(t,Y)$$

Maximum Mean Discrepancy

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

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

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

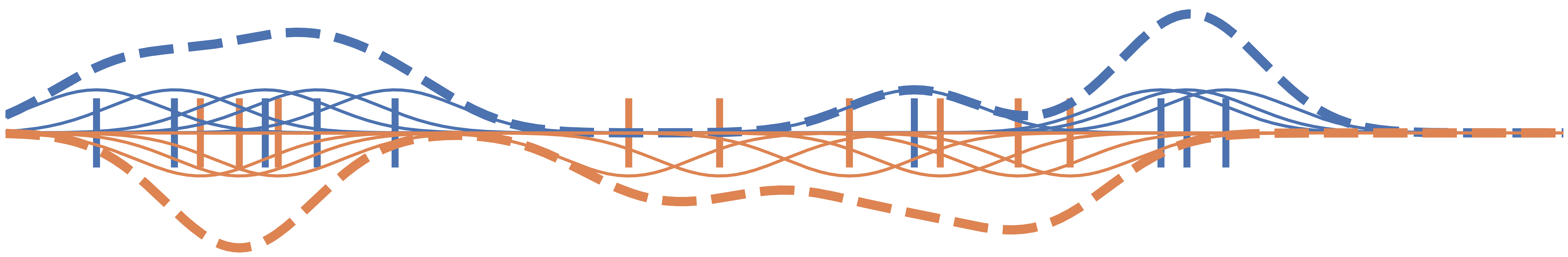


Maximum Mean Discrepancy

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

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

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

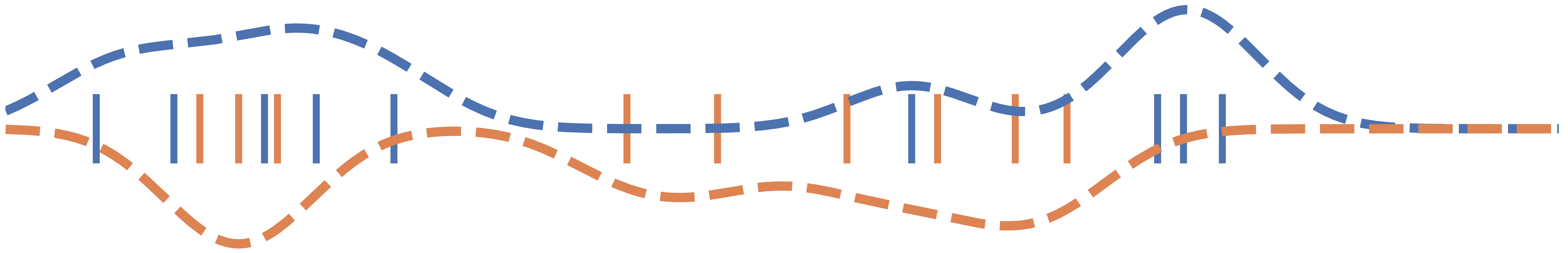


Maximum Mean Discrepancy

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

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

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

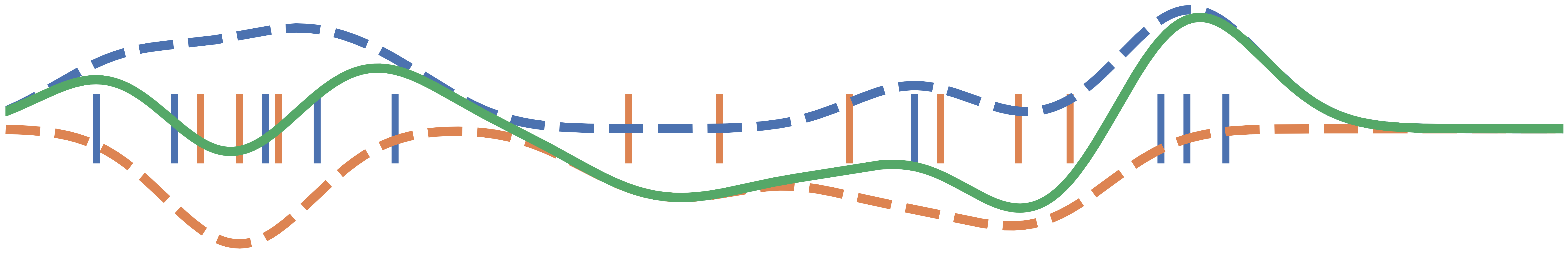


Maximum Mean Discrepancy

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

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

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

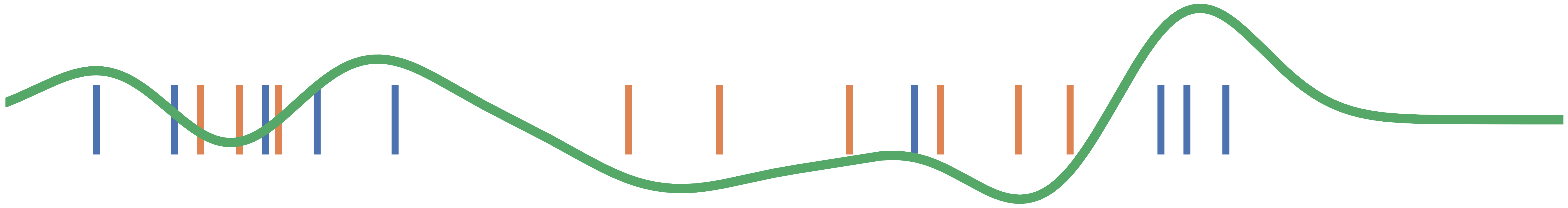


Maximum Mean Discrepancy

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

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

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

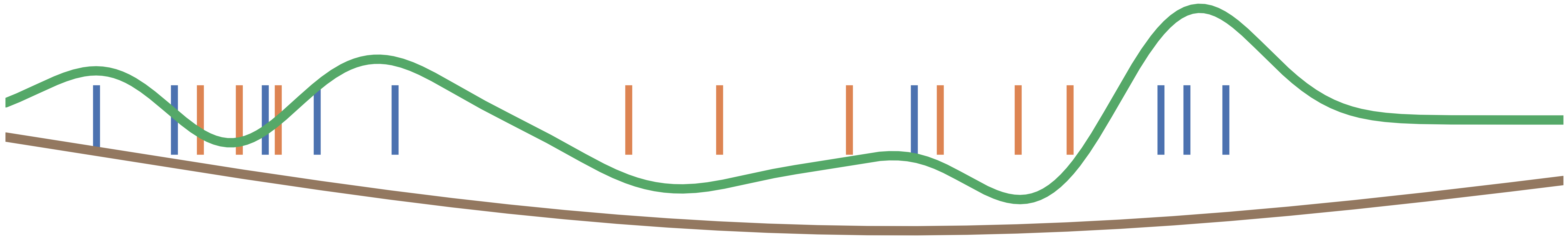


Maximum Mean Discrepancy

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

- 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

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

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

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