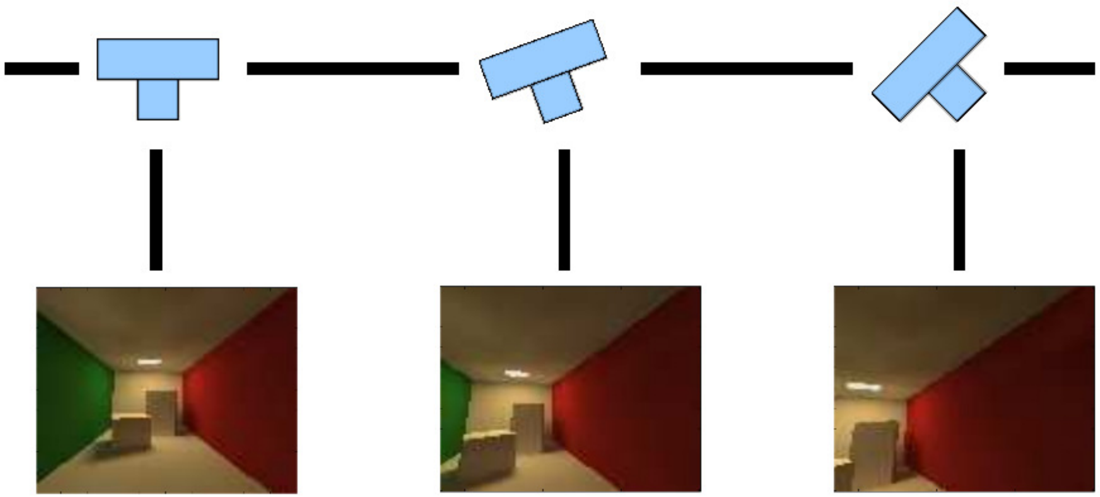
Kernel methods for Bayesian inference

Arthur Gretton

Gatsby Computational Neuroscience Unit

Lancaster, Nov. 2014

Motivating Example: Bayesian inference without a model



- 3600 downsampled frames of 20×20 RGB pixels ($Y_t \in [0, 1]^{1200}$)
- 1800 training frames, remaining for test.
- Gaussian noise added to Y_t.

Challenges:

- No parametric model of camera dynamics (only samples)
- No parametric model of map from camera angle to image (only samples)
- Want to do filtering: Bayesian inference

ABC: an approach to Bayesian inference without a model

Bayes rule:

$$\mathbf{P}(y|x) = \frac{\mathbf{P}(x|y)\pi(y)}{\int \mathbf{P}(x|y)\pi(y)dy}$$

- $\mathbf{P}(x|y)$ is likelihood
- $\pi(y)$ is prior

ABC: an approach to Bayesian inference without a model

Bayes rule:

$$\mathbf{P}(y|x) = \frac{\mathbf{P}(x|y)\pi(y)}{\int \mathbf{P}(x|y)\pi(y)dy}$$

- $\mathbf{P}(x|y)$ is likelihood
- $\pi(y)$ is prior

One approach: Approximate Bayesian Computation (ABC)

ABC: an approach to Bayesian inference without a model

Bayes rule:

$$\mathbf{P}(y|x) = \frac{\mathbf{P}(x|y)\pi(y)}{\int \mathbf{P}(x|y)\pi(y)dy}$$

- $\mathbf{P}(x|y)$ is likelihood
- $\pi(y)$ is prior

One approach: Approximate Bayesian Computation (ABC)

ABC generates a sample from $p(Y|x^*)$ as follows:

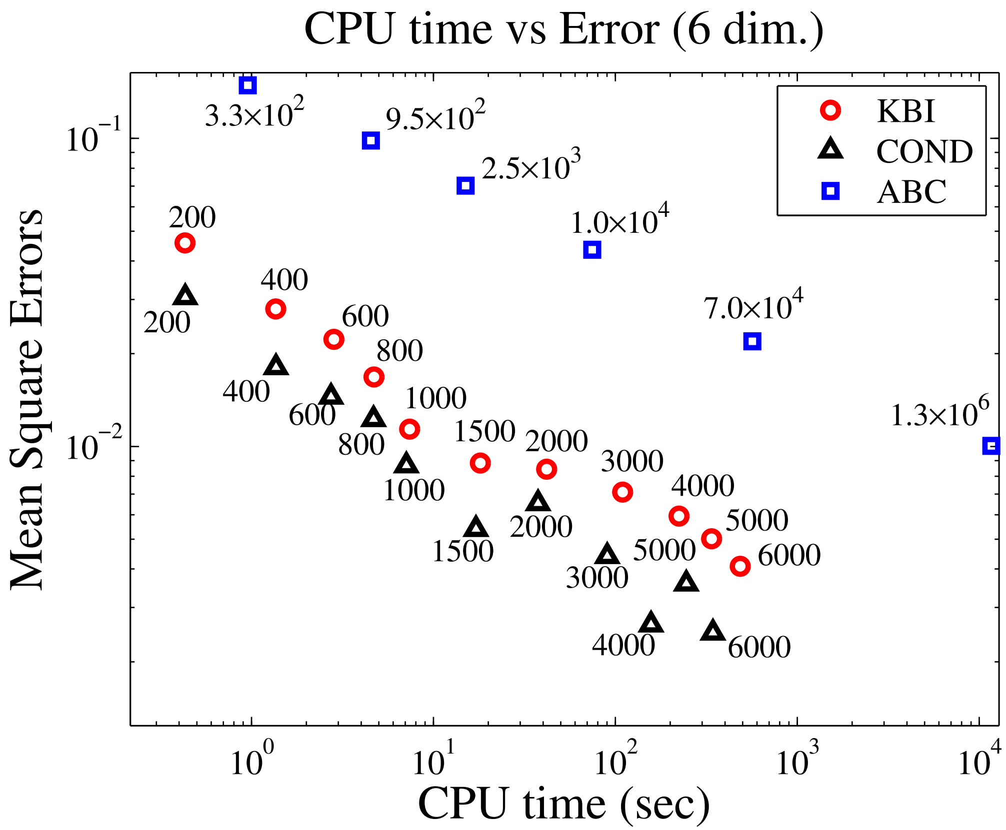
1. generate a sample y_t from the prior π,
2. generate a sample x_t from $\mathbf{P}(X|y_t)$,
3. if $D(x^*, x_t) < \tau$, accept $y = y_t$; otherwise reject,
4. go to (i).

In step (3), D is a distance measure, and τ is a tolerance parameter.

Motivating example 2: simple Gaussian case

- $p(x, y)$ is $\mathcal{N}((0, \mathbf{1}_d^T)^T, V)$ with V a randomly generated covariance

Posterior mean on x: ABC vs kernel approach



Overview

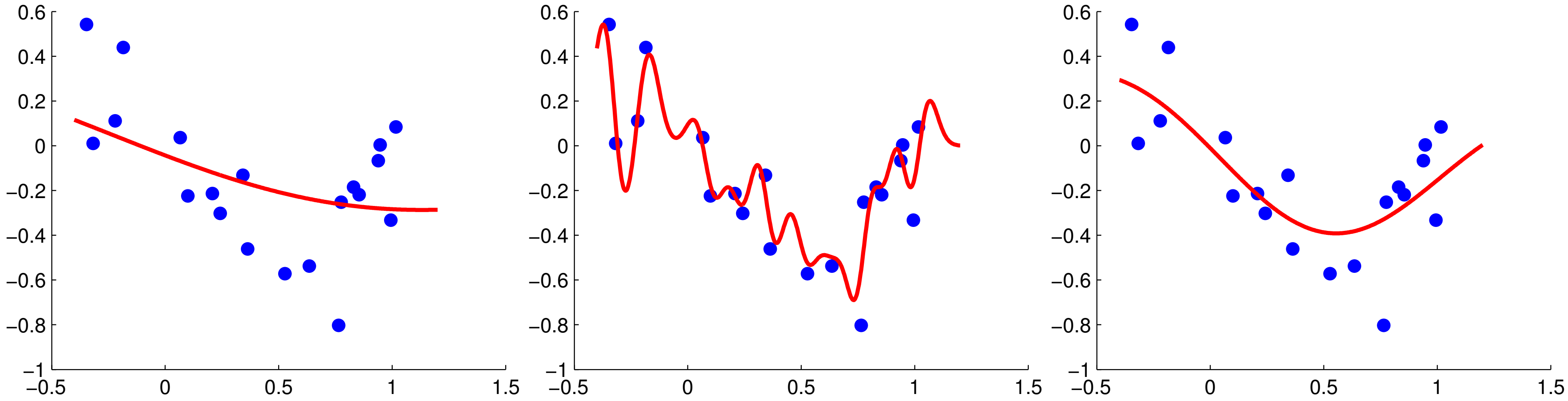
- Introduction to reproducing kernel Hilbert spaces
  - Hilbert space
  - Kernels and feature spaces
  - Reproducing property
  - Mapping probabilities to feature space

Overview

- Introduction to reproducing kernel Hilbert spaces
  - Hilbert space
  - Kernels and feature spaces
  - Reproducing property
  - Mapping probabilities to feature space

- Nonparametric Bayesian inference
  - Learning conditional probabilities: smooth regression to an RKHS
  - Kernelized Bayesian inference

Functions in a reproducing kernel Hilbert space



Kernel methods can control **smoothness** and **avoid overfitting/underfitting**.

Hilbert space

Inner product

Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an inner product on $\mathcal{H}$ if

1. Linear: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
2. Symmetric: $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

Hilbert space

Inner product

Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an inner product on $\mathcal{H}$ if

1. Linear: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
2. Symmetric: $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

Norm induced by the inner product: $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$

Hilbert space

Inner product

Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an inner product on $\mathcal{H}$ if

1. Linear: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
2. Symmetric: $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

Norm induced by the inner product: $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$

Hilbert space: Inner product space containing Cauchy sequence limits.

Kernel

Kernel: Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if there exists an $\mathbb{R}$-Hilbert space and a map $\varphi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') := \langle \varphi_x, \varphi_{x'} \rangle_{\mathcal{H}} \,.$$

- Almost no conditions on $\mathcal{X}$ (eg, $\mathcal{X}$ itself doesn't need an inner product, eg. documents).
- A single kernel can correspond to several possible feature vectors. A trivial example for $\mathcal{X} := \mathbb{R}$:

$$\varphi_x^{(1)} = x \qquad \text{and} \qquad \varphi_x^{(2)} = \begin{bmatrix} x/\sqrt{2} \\ x/\sqrt{2} \end{bmatrix}$$

Finite dim. RKHS with polynomial features

Example: A three dimensional space of features of points in $\mathbb{R}^2$:

$$\begin{aligned}\varphi : \mathbb{R}^2 &\to \mathbb{R}^3 \\ x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &\mapsto \varphi_x = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix},\end{aligned}$$

with kernel

$$k(x, y) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}^\top \begin{bmatrix} y_1 \\ y_2 \\ y_1 y_2 \end{bmatrix}$$

(the standard inner product in $\mathbb{R}^3$ between features). Denote this feature space by $\mathcal{H}$.

Finite dim. RKHS with polynomial features

Define a linear function of the inputs x_1, x_2, and their product $x_1 x_2$,

$$f(x) = f_1 x_1 + f_2 x_2 + f_3 x_1 x_2.$$

f in a space of functions mapping from $\mathcal{X} = \mathbb{R}^2$ to $\mathbb{R}$. Equivalent representation for f,

$$f(\cdot) = \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix}^\top.$$

$f(\cdot)$ refers to the function as an object (here as a vector in $\mathbb{R}^3$)
$f(x) \in \mathbb{R}$ is function evaluated at a point (a real number).

Finite dim. RKHS with polynomial features

Define a linear function of the inputs x_1, x_2, and their product $x_1 x_2$,

$$f(x) = f_1 x_1 + f_2 x_2 + f_3 x_1 x_2.$$

f in a space of functions mapping from $\mathcal{X} = \mathbb{R}^2$ to $\mathbb{R}$. Equivalent representation for f,

$$f(\cdot) = \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix}^\top.$$

$f(\cdot)$ refers to the function as an object (here as a vector in $\mathbb{R}^3$)
$f(x) \in \mathbb{R}$ is function evaluated at a point (a real number).

$$f(x) = f(\cdot)^\top \varphi_x = \langle f(\cdot), \varphi_x \rangle_{\mathcal{H}}$$

Evaluation of f at x is an **inner product in feature space** (here standard inner product in $\mathbb{R}^3$)
$\mathcal{H}$ is a space of functions mapping $\mathbb{R}^2$ to $\mathbb{R}$.

Finite dim. RKHS with polynomial features

φ_y is a mapping from $\mathbb{R}^2$ to $\mathbb{R}^3$...

...which also parametrizes a function mapping $\mathbb{R}^2$ to $\mathbb{R}$.

$$k(\cdot, y) := \begin{bmatrix} y_1 & y_2 & y_1 y_2 \end{bmatrix}^\top = \varphi_y,$$

Given y, there is a vector $k(\cdot, y)$ in $\mathcal{H}$ such that

$$\langle k(\cdot, y), \varphi_x \rangle_{\mathcal{H}} = a x_1 + b x_2 + c x_1 x_2,$$

where $a = y_1$, $b = y_2$, and $c = y_1 y_2$

Finite dim. RKHS with polynomial features

φ_y is a mapping from $\mathbb{R}^2$ to $\mathbb{R}^3$...

...which also parametrizes a function mapping $\mathbb{R}^2$ to $\mathbb{R}$.

$$k(\cdot, y) := \begin{bmatrix} y_1 & y_2 & y_1 y_2 \end{bmatrix}^\top = \varphi_y,$$

Given y, there is a vector $k(\cdot, y)$ in $\mathcal{H}$ such that

$$\langle k(\cdot, y), \varphi_x \rangle_{\mathcal{H}} = a x_1 + b x_2 + c x_1 x_2,$$

where $a = y_1$, $b = y_2$, and $c = y_1 y_2$

Due to symmetry,

$$\begin{aligned} \langle k(\cdot, x), \varphi_y \rangle &= u y_1 + v y_2 + w y_1 y_2 \\ &= k(x, y). \end{aligned}$$

We can write $\varphi_x = k(\cdot, x)$ and $\varphi_y = k(\cdot, y)$ without ambiguity: canonical feature map

The reproducing property

This example illustrates the two defining features of an RKHS:

- **The reproducing property**:
  $\forall x \in \mathcal{X}, \forall f(\cdot) \in \mathcal{H}, \;\; \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$
  ...or use shorter notation $\langle f, \varphi_x \rangle_{\mathcal{H}}$.
- In particular, for any $x, y \in \mathcal{X}$,

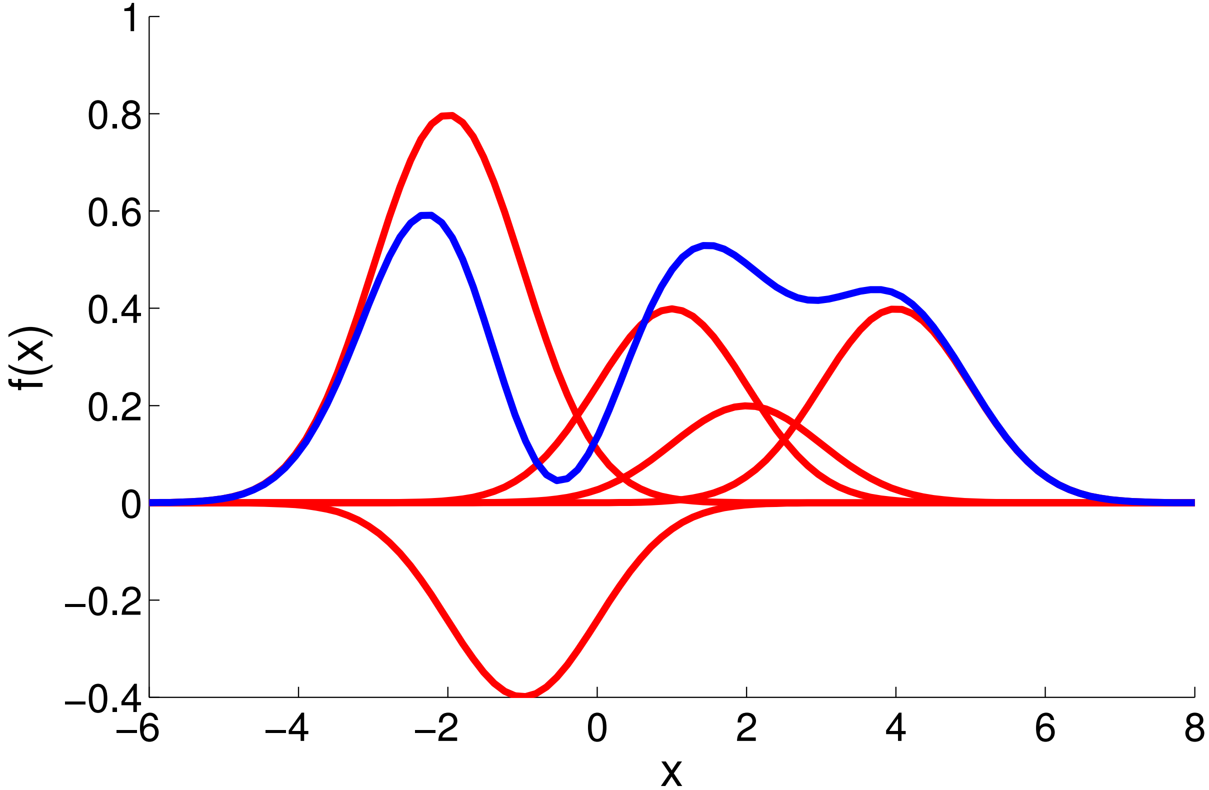
$$k(x, y) = \langle k\left(\cdot, x\right), k\left(\cdot, y\right) \rangle_{\mathcal{H}}.$$

Note: the feature map of every point is in the feature space:
$\forall x \in \mathcal{X}, \;\; k(\cdot, x) = \varphi_x \in \mathcal{H},$

Infinite dimensional feature space

Reproducing property for function with Gaussian kernel:

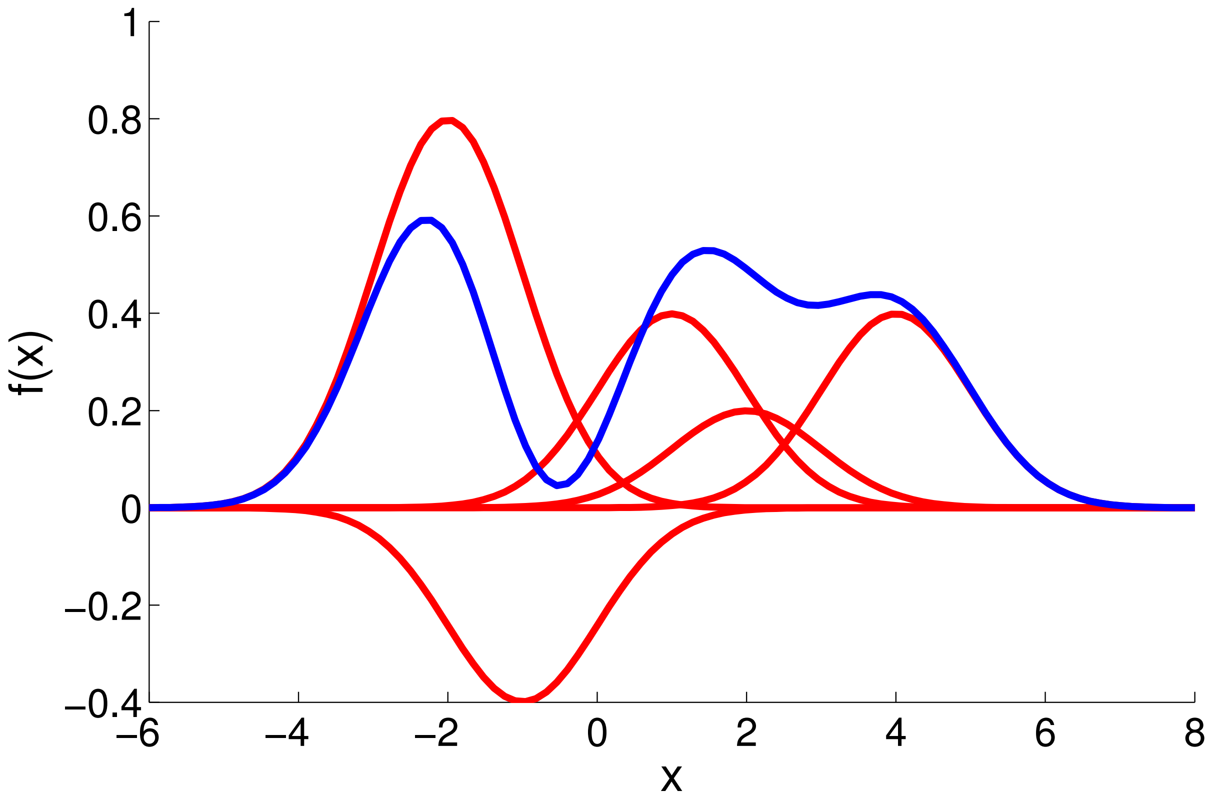
$f(x) := \sum_{i=1}^{m} \alpha_i k(x_i, x) = \langle \sum_{i=1}^{m} \alpha_i \varphi_{x_i}, \varphi_x \rangle_{\mathcal{H}}$.



Infinite dimensional feature space

Reproducing property for function with Gaussian kernel:

$f(x) := \sum_{i=1}^{m} \alpha_i k(x_i, x) = \langle \sum_{i=1}^{m} \alpha_i \varphi_{x_i}, \varphi_x \rangle_{\mathcal{H}}$.



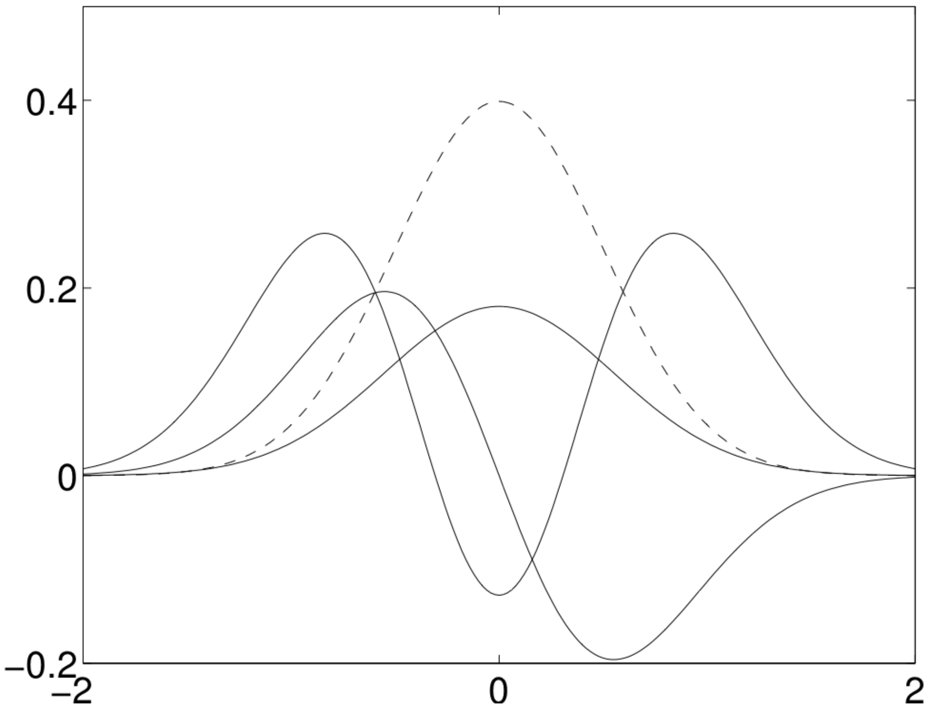
- What do the features φ_x look like (there are infinitely many of them, and they are not unique!)
- What do these features have to do with smoothness?

Infinite dimensional feature space

Gaussian kernel, $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$,

$$\lambda_k \propto b^k \qquad b < 1$$
$$e_k(x) \propto \exp(-(c-a)x^2)H_k(x\sqrt{2c}),$$

a, b, c are functions of σ, and H_k is kth order Hermite polynomial.



$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x') = \sum_{i=1}^{\infty} \left(\sqrt{\lambda_i} e_i(x)\right)\left(\sqrt{\lambda_i} e_i(x')\right) = \sum_{i=1}^{\infty} \varphi_x \varphi_{x'}$$

Infinite dimensional feature space

(Mercer) Let $\mathcal{X}$ be a compact metric space, k be a continous kernel, and μ be a finite Borel measure with $\text{supp}\{\mu\} = \mathcal{X}$. Then the convergence of

$$k(x,y) = \sum_j \lambda_j e_j(x) e_j(y)$$

is absolute and uniform (e_j is the continuous element of the L^2 equivalence class $\mathbf{e}_j$.).

The feature map is $\varphi_x = \begin{bmatrix} \ldots & \sqrt{\lambda_i} e_i(x) & \ldots \end{bmatrix}$

Infinite dimensional feature space

(Mercer RKHS) (Steinwart and Christmann, Theorem 4.51) Under the assumptions of Mercer's theorem,

$$\mathcal{H} := \left\{ \sum_i a_i \sqrt{\lambda_i} e_i \ : \ a_i \in \ell_2 \right\} \tag{1}$$

is an RKHS with kernel k.

Given two functions in the RKHS

$$f(x) := \sum_i a_i \sqrt{\lambda_i} e_i(x), \qquad g(x) := \sum_i b_i \sqrt{\lambda_i} e_i(x),$$

the inner product is $\langle f, g \rangle_{\mathcal{H}} = \sum_i a_i b_i$

Infinite dimensional feature space

Proof: There are two aspects requiring care:

1. Is $k(x, \cdot) \in \mathcal{H} \qquad \forall x \in \mathcal{X}$? **Requires Mercer's theorem**
2. Does the reproducing property hold? $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

Infinite dimensional feature space

Proof: There are two aspects requiring care:

1. Is $k(x, \cdot) \in \mathcal{H} \qquad \forall x \in \mathcal{X}$? **Requires Mercer's theorem**
2. Does the reproducing property hold? $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

First part:

By the definition of $\mathcal{H}$, the function in $\mathcal{H}$ indexed by x is

$$k(x, \cdot) = \sum_i \left(\sqrt{\lambda_i} e_i(x) \right) \left(\sqrt{\lambda_i} e_i(\cdot) \right).$$

Is this function in the RKHS? Yes, if the ℓ_2 norm of $\left(\sqrt{\lambda_i} e_i(x) \right)$ is bounded. This is due to Mercer: $\forall x \in \mathcal{X}$,

$$\|k(x, \cdot)\|_{\mathcal{H}}^2 = \left\| \left(\sqrt{\lambda_i} e_i(x) \right) \right\|_{\ell_2}^2 = k(x, x) < \infty.$$

Infinite dimensional feature space

Proof (continued):

Second part:

The reproducing property holds: using the inner product definition,

$$\langle f, k(x,\cdot)\rangle_{\mathcal{H}} = \sum_i f_i \left(\sqrt{\lambda_i} e_i(x)\right) = f(x),$$

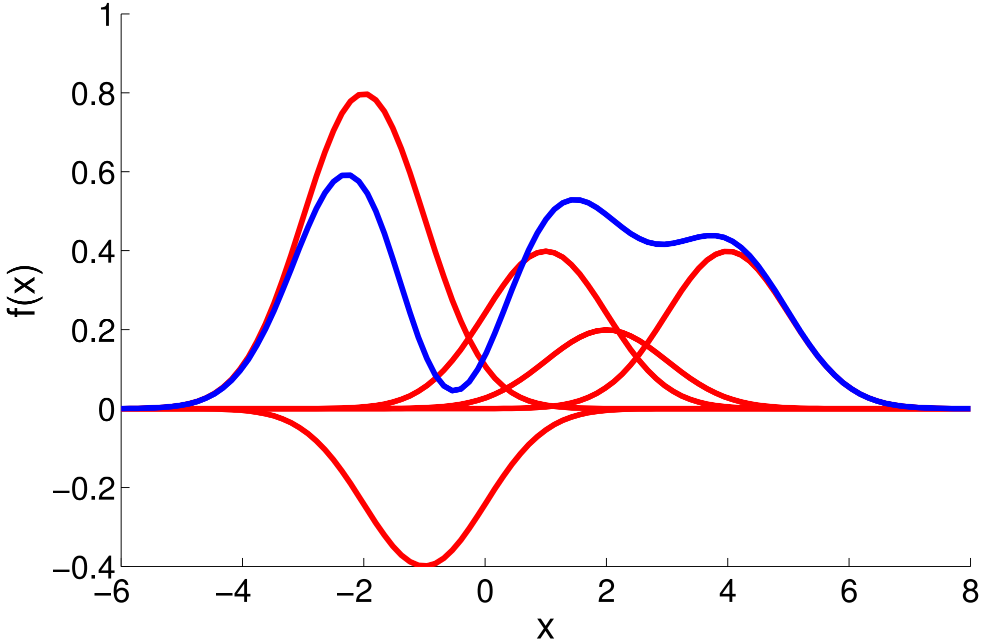
which is always well defined since both $f \in \ell_2$ and $k(x,\cdot) \in \ell_2$.

Infinite dimensional feature space

Example RKHS function, Gaussian kernel:

$$f(x) := \sum_{i=1}^{m} \alpha_i k(x_i, x) = \sum_{i=1}^{m} \alpha_i \left[\sum_{j=1}^{\infty} \lambda_j e_j(x_i) e_j(x) \right] = \sum_{j=1}^{\infty} f_j \left[\sqrt{\lambda_j} e_j(x) \right]$$

where $f_j = \sum_{i=1}^{m} \alpha_i \sqrt{\lambda_j} e_j(x_i)$.



NOTE that this enforces smoothing:

λ_j decay as e_j become rougher,

f_j decay since $\|f\|_{\mathcal{H}}^2 = \sum_j f_j^2 < \infty$.

Reproducing kernel Hilbert space

$\mathcal{H}$ a Hilbert space of $\mathbb{R}$-valued functions on non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *reproducing kernel* of $\mathcal{H}$, and $\mathcal{H}$ is a *reproducing kernel Hilbert space*, if

- $\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (the reproducing property).

In particular, for any $x, y \in \mathcal{X}$,

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}. \qquad (2)$$

Original definition: kernel an inner product between feature maps. Then $\varphi_x = k(\cdot, x)$ a valid feature map.

Probabilities in feature space: the mean trick

The kernel trick

- Given $x \in \mathcal{X}$ for some set $\mathcal{X}$, define feature map $\varphi_x \in \mathcal{H}$,

$$\varphi_x = \left[\ldots \sqrt{\lambda_i} e_i(x) \ldots \right] \in \ell_2$$

- For positive definite $k(x, x')$,

$$k(x, x') = \langle \varphi_x, \varphi_{x'} \rangle_{\mathcal{H}}$$

- The kernel trick: $\forall f \in \mathcal{H}$,

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{H}}$$

Probabilities in feature space: the mean trick

The kernel trick

- Given $x \in \mathcal{X}$ for some set $\mathcal{X}$, define feature map $\varphi_x \in \mathcal{H}$,

$$\varphi_x = \left[\dots \sqrt{\lambda_i} e_i(x) \dots \right] \in \ell_2$$

- For positive definite $k(x, x')$,

$$k(x, x') = \langle \varphi_x, \varphi_{x'} \rangle_{\mathcal{H}}$$

- The kernel trick: $\forall f \in \mathcal{H}$,

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{H}}$$

The mean trick

- Given $\mathbf{P}$ a Borel probability measure on $\mathcal{X}$, define feature map $\mu_{\mathbf{P}} \in \mathcal{H}$

$$\mu_{\mathbf{P}} = \left[\dots \sqrt{\lambda_i} \mathbf{E}_{\mathbf{P}} \left[e_i(X) \right] \dots \right] \in \ell_2$$

- For positive definite $k(x, x')$,

$$\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(X, Y) = \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{H}}$$

for $X \sim \mathbf{P}$ and $Y \sim \mathbf{Q}$.

- The mean trick: (we call $\mu_{\mathbf{P}}$ a mean/distribution embedding)

$$\mathbf{E}_{\mathbf{P}}(f(X)) = \mathbf{E}_{\mathbf{P}} \left[\langle \varphi_X, f \rangle_{\mathcal{F}} \right]$$

Feature embeddings of probabilities

The kernel trick:

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{H}}$$

The mean trick:

$$\mathbf{E}_{\mathsf{P}}(f(X)) = \langle \mu_{\mathsf{P}}, f \rangle_{\mathcal{F}}$$

Empirical mean embedding:

$$\widehat{\mu}_{\mathsf{P}} = m^{-1} \sum_{i=1}^{m} \varphi_{x_i} \qquad x_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}$$

μ_{P} gives you expectations of all RKHS functions

When k characteristic, then μ_{P} unique, e.g. Gauss, Laplace, . . .

Nonparametric Bayesian inference using distribution embeddings

Bayes again

Bayes rule:

$$\mathbf{P}(y|x) = \frac{\mathbf{P}(x|y)\pi(y)}{\int \mathbf{P}(x|y)\pi(y)dy}$$

- $\mathbf{P}(x|y)$ is likelihood
- π is prior

How would this look with kernel embeddings?

Bayes again

Bayes rule:

$$\mathbf{P}(y|x) = \frac{\mathbf{P}(x|y)\pi(y)}{\int \mathbf{P}(x|y)\pi(y)dy}$$

- $\mathbf{P}(x|y)$ is likelihood
- π is prior

How would this look with kernel embeddings?

Define RKHS $\mathcal{G}$ on $\mathcal{Y}$ with feature map ψ_y and kernel $l(y, \cdot)$

We need a conditional mean embedding: for all $g \in \mathcal{G}$,

$$\mathbf{E}_{Y|x^*} g(Y) = \langle g, \mu_{\mathbf{P}(y|x^*)} \rangle_{\mathcal{G}}$$

This will be obtained by RKHS-valued ridge regression

Ridge regression and the conditional feature mean

Ridge regression from $\mathcal{X} := \mathbb{R}^d$ to a finite *vector* output $\mathcal{Y} := \mathbb{R}^{d'}$ (these could be d' nonlinear features of y):

Define training data

$$X = \begin{bmatrix} x_1 & \ldots & x_m \end{bmatrix} \in \mathbb{R}^{d \times m} \qquad Y = \begin{bmatrix} y_1 & \ldots & y_m \end{bmatrix} \in \mathbb{R}^{d' \times m}$$

Ridge regression and the conditional feature mean

Ridge regression from $\mathcal{X} := \mathbb{R}^d$ to a finite *vector* output $\mathcal{Y} := \mathbb{R}^{d'}$ (these could be d' nonlinear features of y):

Define training data

$$X = \begin{bmatrix} x_1 & \ldots & x_m \end{bmatrix} \in \mathbb{R}^{d \times m} \qquad Y = \begin{bmatrix} y_1 & \ldots & y_m \end{bmatrix} \in \mathbb{R}^{d' \times m}$$

Solve

$$\breve{A} = \arg\min_{A \in \mathbb{R}^{d' \times d}} \left(\|Y - AX\|^2 + \lambda \|A\|_{\mathrm{HS}}^2 \right),$$

where

$$\|A\|_{\mathrm{HS}}^2 = \mathrm{tr}(A^\top A) = \sum_{i=1}^{\min\{d,d'\}} \gamma_{A,i}^2$$

Ridge regression and the conditional feature mean

Ridge regression from $\mathcal{X} := \mathbb{R}^d$ to a finite *vector* output $\mathcal{Y} := \mathbb{R}^{d'}$ (these could be d' nonlinear features of y):

Define training data

$$X = \begin{bmatrix} x_1 & \ldots & x_m \end{bmatrix} \in \mathbb{R}^{d \times m} \qquad Y = \begin{bmatrix} y_1 & \ldots & y_m \end{bmatrix} \in \mathbb{R}^{d' \times m}$$

Solve

$$\breve{A} = \quad \arg\min_{A \in \mathbb{R}^{d' \times d}} \left(\|Y - AX\|^2 + \lambda \|A\|_{\mathrm{HS}}^2 \right),$$

where

$$\|A\|_{\mathrm{HS}}^2 = \mathrm{tr}(A^\top A) = \sum_{i=1}^{\min\{d,d'\}} \gamma_{A,i}^2$$

Solution: $\breve{A} = C_{YX} \left(C_{XX} + m\lambda I \right)^{-1}$

Ridge regression and the conditional feature mean

Prediction at new point x:

$$
\begin{aligned}
y^* &= \breve{A}x \\
&= C_{YX}\left(C_{XX} + m\lambda I\right)^{-1} x \\
&= \sum_{i=1}^{m} \beta_i(x) y_i
\end{aligned}
$$

where

$$\beta_i(x) = (K + \lambda m I)^{-1} \left[\begin{array}{ccc} k(x_1, x) & \ldots & k(x_m, x) \end{array} \right]^\top$$

and

$$K := X^\top X \qquad\qquad k(x_1, x) = x_1^\top x$$

Ridge regression and the conditional feature mean

Prediction at new point x:

$$\begin{aligned} y^* &= \breve{A}x \\ &= C_{YX}\left(C_{XX} + m\lambda I\right)^{-1} x \\ &= \sum_{i=1}^{m} \beta_i(x) y_i \end{aligned}$$

where

$$\beta_i(x) = (K + \lambda m I)^{-1} \left[\begin{array}{ccc} k(x_1, x) & \dots & k(x_m, x) \end{array} \right]^\top$$

and

$$K := X^\top X \qquad\qquad k(x_1, x) = x_1^\top x$$

What if we do everything in kernel space?

Ridge regression and the conditional feature mean

Recall our setup:

- Given training *pairs:*

$$(x_i, y_i) \sim \mathbf{P}_{XY}$$

- $\mathcal{F}$ on $\mathcal{X}$ with feature map φ_x and kernel $k(x, \cdot)$
- $\mathcal{G}$ on $\mathcal{Y}$ with feature map ψ_y and kernel $l(y, \cdot)$

We define the covariance between feature maps:

$$C_{XX} = \mathbf{E}_X \left(\varphi_X \otimes \varphi_X \right) \qquad C_{XY} = \mathbf{E}_{XY} \left(\varphi_X \otimes \psi_Y \right)$$

and matrices of feature mapped training data

$$X = \left[\begin{array}{ccc} \varphi_{x_1} & \ldots & \varphi_{x_m} \end{array} \right] \quad Y := \left[\begin{array}{ccc} \psi_{y_1} & \ldots & \psi_{y_m} \end{array} \right]$$

Ridge regression and the conditional feature mean

Objective: [Weston et al. (2003), Micchelli and Pontil (2005), Caponnetto and De Vito (2007), Grunewalder et al. (2012, 2013)]

$$\breve{A} = \arg \min_{A \in \mathrm{HS}(\mathcal{F},\mathcal{G})} \left(\mathbf{E}_{XY} \|Y - AX\|^2_{\mathcal{G}} + \lambda \|A\|^2_{\mathrm{H}S} \right), \qquad \|A\|^2_{\mathrm{H}S} = \sum_{i=1}^{\infty} \gamma^2_{A,i}$$

Solution same as vector case:

$$\breve{A} = C_{YX} \left(C_{XX} + m\lambda I \right)^{-1},$$

Prediction at new x using kernels:

$$\begin{aligned} \breve{A}\varphi_x &= \begin{bmatrix} \psi_{y_1} & \ldots & \psi_{y_m} \end{bmatrix} (K + \lambda m I)^{-1} \begin{bmatrix} k(x_1, x) & \ldots & k(x_m, x) \end{bmatrix} \\ &= \sum_{i=1}^{m} \beta_i(x) \psi_{y_i} \end{aligned}$$

where $K_{ij} = k(x_i, x_j)$

Ridge regression and the conditional feature mean

How is loss $\|Y - AX\|^2_{\mathcal{G}}$ relevant to conditional expectation of some $\mathbf{E}_{Y|x} g(Y)$? Define: [Song et al. (2009), Grunewalder et al. (2013)]

$$\mu_{Y|x} := A\varphi_x$$

Ridge regression and the conditional feature mean

How is loss $\|Y - AX\|_{\mathcal{G}}^2$ relevant to conditional expectation of some $\mathbf{E}_{Y|x} g(Y)$? Define: [Song et al. (2009), Grunewalder et al. (2013)]

$$\mu_{Y|x} := A\varphi_x$$

We need A to have the property

$$\begin{aligned}
\mathbf{E}_{Y|x} g(Y) &\approx \langle g, \mu_{Y|x} \rangle_{\mathcal{G}} \\
&= \langle g, A\varphi_x \rangle_{\mathcal{G}} \\
&= \langle A^* g, \varphi_x \rangle_{\mathcal{F}} = (A^* g)(x)
\end{aligned}$$

Ridge regression and the conditional feature mean

How is loss $\|Y - AX\|^2_{\mathcal{G}}$ relevant to conditional expectation of some $\mathbf{E}_{Y|x} g(Y)$? Define: [Song et al. (2009), Grunewalder et al. (2013)]

$$\mu_{Y|x} := A\varphi_x$$

We need A to have the property

$$\begin{aligned}\mathbf{E}_{Y|x} g(Y) &\approx \langle g, \mu_{Y|x} \rangle_{\mathcal{G}} \\ &= \langle g, A\varphi_x \rangle_{\mathcal{G}} \\ &= \langle A^* g, \varphi_x \rangle_{\mathcal{F}} = (A^* g)(x)\end{aligned}$$

Natural risk function for conditional mean

$$\mathcal{L}(A, \mathbf{P}_{XY}) := \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\underbrace{\left(\mathbf{E}_{Y|X} g(Y)\right)}_{\text{Target}}(X) - \underbrace{(A^* g)}_{\text{Estimator}}(X) \right]^2,$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathsf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathsf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}\mathcal{L}(A, \mathsf{P}_{XY}) &:= \sup_{\|g\|\leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - \left(A^* g \right)(X) \right]^2 \\ &\leq \mathbf{E}_{XY} \sup_{\|g\|\leq 1} \left[g(Y) - \left(A^* g \right)(X) \right]^2\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - (A^* g)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - (A^* g)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[\langle g, \psi_Y \rangle_{\mathcal{G}} - \langle A^* g, \varphi_X \rangle_{\mathcal{F}} \right]^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \|\psi_Y - A\varphi_X\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\|\leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - (A^* g)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\|\leq 1} \left[g(Y) - (A^* g)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\|\leq 1} \left[\langle g, \psi_Y \rangle_{\mathcal{G}} - \langle g, A\varphi_X \rangle_{\mathcal{G}} \right]^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - (A^* g)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - (A^* g)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \langle g, \psi_Y - A\varphi_X \rangle_{\mathcal{G}}^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathsf{P}_{XY}) \leq \mathbf{E}_{XY} \|\psi_Y - A\varphi_X\|^2_{\mathcal{G}}$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathsf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - (A^* g)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - (A^* g)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \langle g, \psi_Y - A\varphi_X \rangle^2_{\mathcal{G}} \\
&\leq \mathbf{E}_{XY} \|\psi_Y - A\varphi_X\|^2_{\mathcal{G}}
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathsf{P}_{XY}) \leq \mathbf{E}_{XY} \left\|\psi_Y - A\varphi_X\right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathsf{P}_{XY}) &:= \sup_{\|g\|\leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y)\right)(X) - \left(A^* g\right)(X)\right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\|\leq 1} \left[g(Y) - \left(A^* g\right)(X)\right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\|\leq 1} \langle g, \psi_Y - A\varphi_X \rangle_{\mathcal{G}}^2 \\
&\leq \mathbf{E}_{XY} \left\|\psi_Y - A\varphi_X\right\|_{\mathcal{G}}^2
\end{aligned}$$

If we assume $\mathbf{E}_Y[g(Y)|X = x] \in \mathcal{F}$ then upper bound tight (next slide).

Conditions for ridge regression = conditional mean

Proof: conditional mean obtained by ridge regression when

$\mathbf{E}_Y[g(Y)|X = x] \in \mathcal{F}$

Given a function $g \in \mathcal{G}$. Assume $E_{Y|X}[g(Y)|X = \cdot] \in \mathcal{F}$. Then

$$\boxed{C_{XX} E_{Y|X}[g(Y)|X = \cdot] = C_{XY} g.}$$

Conditions for ridge regression = conditional mean

Proof: conditional mean obtained by ridge regression when

$\mathbf{E}_Y[g(Y)|X = x] \in \mathcal{F}$

Given a function $g \in \mathcal{G}$. Assume $E_{Y|X}[g(Y)|X = \cdot] \in \mathcal{F}$. Then

$$\boxed{C_{XX} E_{Y|X}[g(Y)|X = \cdot] = C_{XY} g.}$$

Proof: [Fukumizu et al., 2004]

For all $f \in \mathcal{F}$, by definition of C_{XX},

$$\begin{aligned}
&\left\langle f, C_{XX} E_{Y|X}[g(Y)|X = \cdot] \right\rangle_{\mathcal{F}} \\
&= \operatorname{cov}\left(f, E_{Y|X}[g(Y)|X = \cdot]\right) \\
&= E_X\left(f(X)\, E_{Y|X}[g(Y)|X]\right) \\
&= E_{XY}(f(X)g(Y)) \\
&= \langle f, C_{XY} g \rangle ,
\end{aligned}$$

by definition of C_{XY}.

Kernel Bayes' law

- Prior: $Y \sim \pi(y)$
- Likelihood: $(X|y) \sim \mathbf{P}(x|y)$ with some joint $\mathbf{P}(x, y)$

Kernel Bayes' law

- Prior: $Y \sim \pi(y)$
- Likelihood: $(X|y) \sim \mathbf{P}(x|y)$ with some joint $\mathbf{P}(x, y)$
- Joint distribution: $\mathbf{Q}(x, y) = \mathbf{P}(x|y)\pi(y)$

Warning: $\mathbf{Q} \neq \mathbf{P}$, *change of measure* from $\mathbf{P}(y)$ to $\pi(y)$

- Marginal for x:

$$\mathbf{Q}(x) := \int \mathbf{P}(x|y)\pi(y)dy.$$

- Bayes' law:

$$\mathbf{Q}(y|x) = \frac{\mathbf{P}(x|y)\pi(y)}{\mathbf{Q}(x)}$$

Kernel Bayes' law

- Posterior embedding via the usual conditional update,

$$\mu_{\mathbf{Q}(y|x)} = C_{\mathbf{Q}(y,x)} C^{-1}_{\mathbf{Q}(x,x)} \phi_x.$$

Kernel Bayes' law

- Posterior embedding via the usual conditional update,

$$\mu_{\mathbf{Q}(y|x)} = C_{\mathbf{Q}(y,x)} C_{\mathbf{Q}(x,x)}^{-1} \phi_x.$$

- Given mean embedding of prior: $\mu_\pi(y)$
- Define marginal covariance:

$$C_{\mathbf{Q}(x,x)} = \int (\varphi_x \otimes \varphi_x) \; \mathbf{P}(x|y)\pi(y)dx = C_{(xx)y} C_{yy}^{-1} \mu_{\pi(y)}$$

- Define cross-covariance:

$$C_{\mathbf{Q}(y,x)} = \int (\phi_y \otimes \varphi_x) \; \mathbf{P}(x|y)\pi(y)dx = C_{(yx)y} C_{yy}^{-1} \mu_{\pi(y)}.$$

Kernel Bayes' law: consistency result

- How to compute posterior expectation from data?
- Given samples: $\{(x_i, y_i)\}_{i=1}^n$ from $\mathbf{P}_{xy}$, $\{(u_j)\}_{j=1}^n$ from prior π.
- Want to compute $\mathbf{E}[g(Y)|X=x]$ for g in $\mathcal{G}$
- For any $x \in \mathcal{X}$,

$$\left|\mathbf{g}_y^T R_{Y|X} \mathbf{k}_X(x) - \mathbf{E}[f(Y)|X=x]\right| = O_p(n^{-\frac{4}{27}}), \quad (n \to \infty),$$

where

- $\mathbf{g}_y = (g(y_1), \ldots, g(y_n))^T \in \mathbb{R}^n$.
- $\mathbf{k}_X(x) = (k(x_1, x), \ldots, k(x_n, x))^T \in \mathbb{R}^n$
- $R_{Y|X}$ learned from the samples, contains the u_j

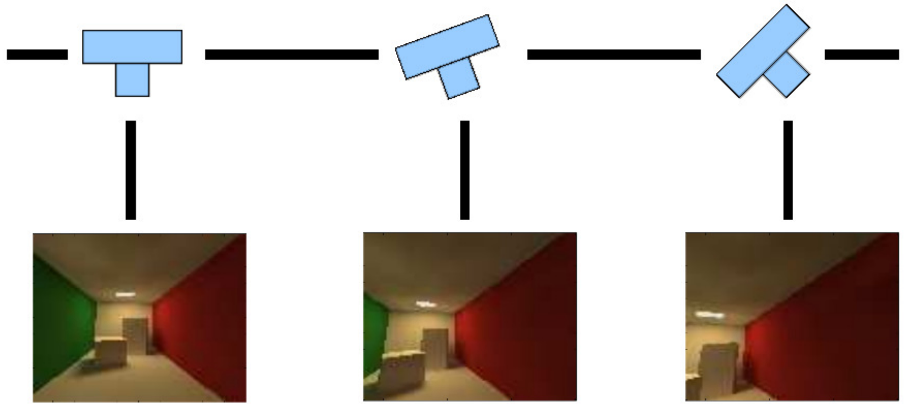
Smoothness assumptions:

- $\pi/p_Y \in \mathcal{R}(C_{YY}^{1/2})$, where p_Y p.d.f. of $\mathbf{P}_Y$,
- $E[g(Y)|X=\cdot] \in \mathcal{R}(C_{\mathbf{Q}(xx)}^2)$.

Experiment: Kernel Bayes' law vs EKF

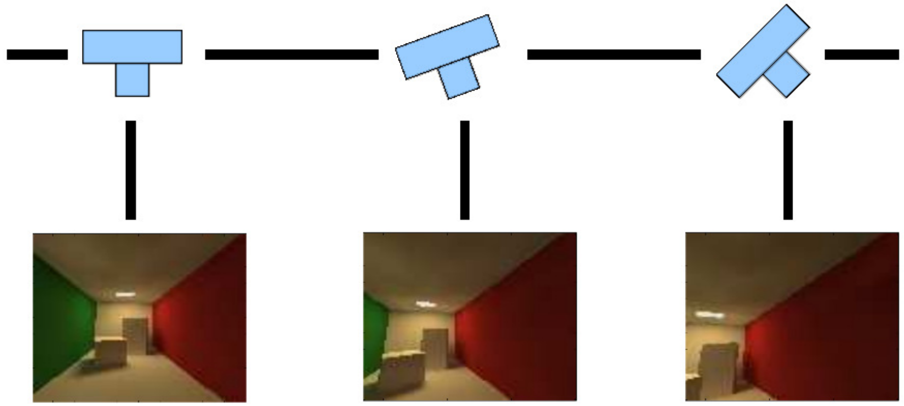
Experiment: Kernel Bayes' law vs EKF

- Compare with extended Kalman filter (EKF) on camera orientation task
- 3600 downsampled frames of 20×20 RGB pixels ($Y_t \in [0, 1]^{1200}$)
- 1800 training frames, remaining for test.
- Gaussian noise added to Y_t.



Experiment: Kernel Bayes' law vs EKF

- Compare with extended Kalman filter (EKF) on camera orientation task
- 3600 downsampled frames of 20×20 RGB pixels ($Y_t \in [0, 1]^{1200}$)
- 1800 training frames, remaining for test.
- Gaussian noise added to Y_t.



Average MSE and standard errors (10 runs)

	KBR (Gauss)	KBR (Tr)	Kalman (9 dim.)	Kalman (Quat.)
$\sigma^2 = 10^{-4}$	0.210 ± 0.015	0.146 ± 0.003	1.980 ± 0.083	0.557 ± 0.023
$\sigma^2 = 10^{-3}$	0.222 ± 0.009	0.210 ± 0.008	1.935 ± 0.064	0.541 ± 0.022

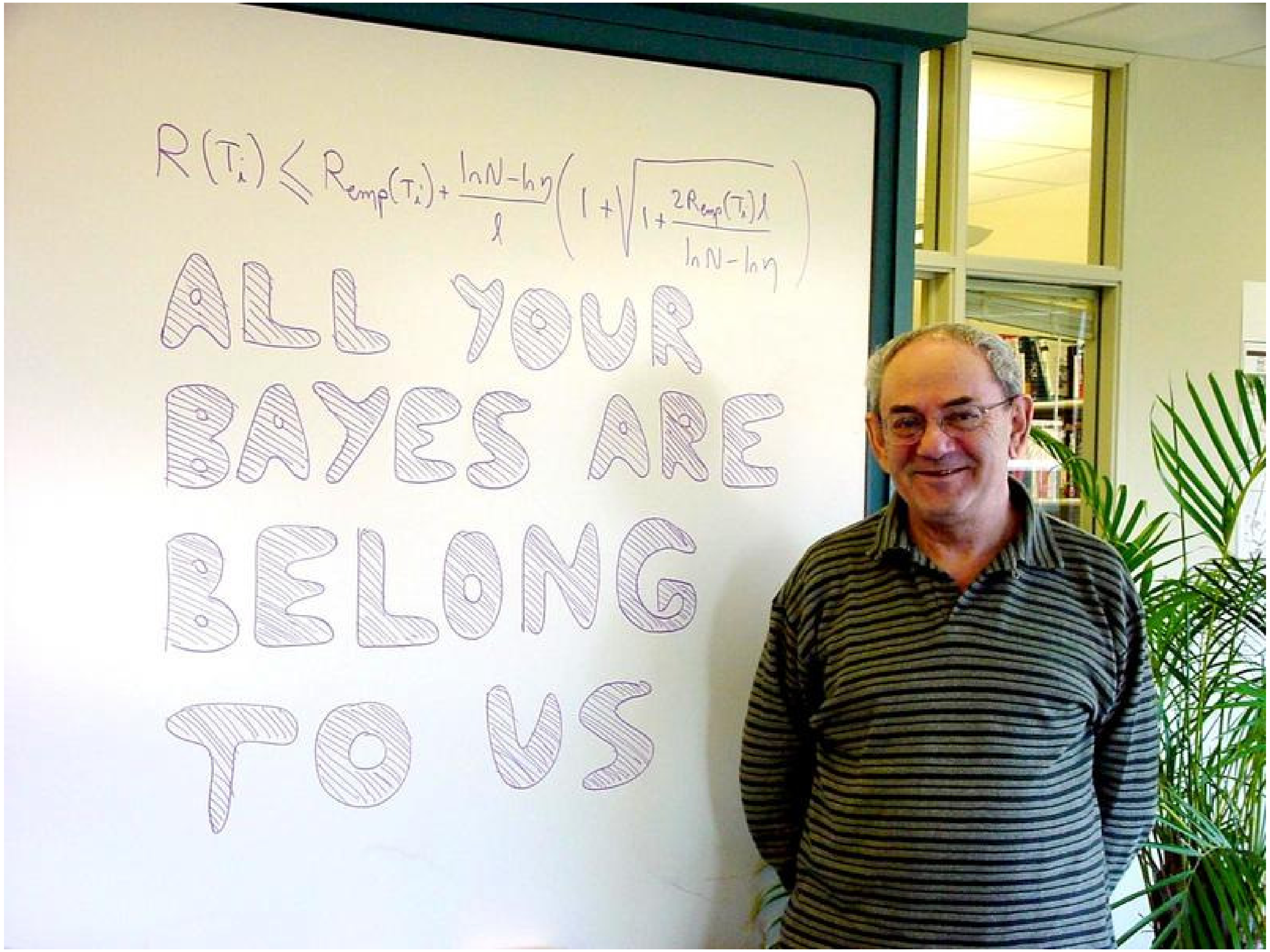
Overview

- Introduction to reproducing kernel Hilbert spaces
  - Hilbert space
  - Kernels and feature spaces
  - Reproducing property
- Nonparametric Bayesian inference
  - Mapping probabilities to feature space
  - Learning conditional probabilities: smooth regression to an RKHS
  - Kernelized Bayesian inference

Co-authors

- **From UCL:**
  - Luca Baldasssarre
  - Steffen Grunewalder
  - Guy Lever
  - Sam Patterson
  - Massimiliano Pontil
- **External:**
  - Kenji Fukumizu, ISM
  - Bernhard Schoelkopf, MPI
  - Alex Smola, Google/CMU
  - Le Song, Georgia Tech
  - Bharath Sriperumbudur, Penn. State

Questions?


$R(T_i) \leq R_{emp}(T_i) + \frac{\ln N - \ln \eta}{\ell}\left(1 + \sqrt{1 + \frac{2R_{emp}(T_i)\ell}{\ln N - \ln \eta}}\right)$
ALL YOUR BAYES ARE BELONG TO US

Selected references

Characteristic kernels and mean embeddings:

- Smola, A., Gretton, A., Song, L., Schoelkopf, B. (2007). A hilbert space embedding for distributions. ALT.
- Sriperumbudur, B., Gretton, A., Fukumizu, K., Schoelkopf, B., Lanckriet, G. (2010). Hilbert space embeddings and metrics on probability measures. JMLR.
- Gretton, A., Borgwardt, K., Rasch, M., Schoelkopf, B., Smola, A. (2012). A kernel two- sample test. JMLR.
- Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. Annals of Statistics.

Two-sample, independence, conditional independence tests:

- Gretton, A., Fukumizu, K., Teo, C., Song, L., Schoelkopf, B., Smola, A. (2008). A kernel statistical test of independence. NIPS
- Fukumizu, K., Gretton, A., Sun, X., Schoelkopf, B. (2008). Kernel measures of conditional dependence.
- Gretton, A., Fukumizu, K., Harchaoui, Z., Sriperumbudur, B. (2009). A fast, consistent kernel two-sample test. NIPS.
- Gretton, A., Borgwardt, K., Rasch, M., Schoelkopf, B., Smola, A. (2012). A kernel two- sample test. JMLR

Selected references (continued)

Conditional mean embedding, RKHS-valued regression:

- Weston, J., Chapelle, O., Elisseeff, A., Schölkopf, B., and Vapnik, V., (2003). Kernel Dependency Estimation, NIPS.
- Micchelli, C., and Pontil, M., (2005). On Learning Vector-Valued Functions. Neural Computation.
- Caponnetto, A., and De Vito, E. (2007). Optimal Rates for the Regularized Least-Squares Algorithm. Foundations of Computational Mathematics.
- Song, L., and Huang, J., and Smola, A., Fukumizu, K., (2009). Hilbert Space Embeddings of Conditional Distributions. ICML.
- Grunewalder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., Pontil, M. (2012). Conditional mean embeddings as regressors. ICML.
- Grunewalder, S., Gretton, A., Shawe-Taylor, J. (2013). Smooth operators. ICML.

Kernel Bayes rule:

- Song, L., Fukumizu, K., Gretton, A. (2013). Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. IEEE Signal Processing Magazine.
- Fukumizu, K., Song, L., Gretton, A. (2013). Kernel Bayes rule: Bayesian inference with positive definite kernels, JMLR

References

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

L. Song, J. Huang, A. J. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions. In *Proceedings of the International Conference on Machine Learning*, 2009.