Kernel Method

Wei Fan, Jiaqiang Li

University of Science and Technology of China

December 7, 2022

Outline

1. Some facts on Hilbert Space
2. Positive Semidefinite Kernel
   - Reproducing Kernel Hilbert Space
   - Mercer's Theorem, Feature Map and Kernel Trick
   - Operations on RKHS
   - Kernel Ridge Regression
3. Random Features for Large-Scale Kernel Machines
   - Random Fourier Features
   - Random Binning Features
4. Kernel Embedding for Probability Measure
   - From Data Point to Probability Measure
   - Maximum Mean Discrepancy
5. Coupled Kernel Embedding

Hilbert space

- An **inner product** on a vector space $\mathbb{V}$ is a mapping $\langle \cdot, \cdot \rangle_{\mathbb{V}} : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ such that

$$\begin{aligned} \langle f, g \rangle_{\mathbb{V}} &= \langle g, f \rangle_{\mathbb{V}} && \forall f, g \in \mathbb{V}, \\ \langle f, f \rangle_{\mathbb{V}} &\geq 0 && \forall f \in \mathbb{V}, \text{ with equality iff } f = 0 \\ \langle f + \alpha g, h \rangle_{\mathbb{V}} &= \langle f, h \rangle_{\mathbb{V}} + \alpha \langle g, h \rangle_{\mathbb{V}} && \forall f, g, h \in \mathbb{V} \text{ and } \alpha \in \mathbb{R}. \end{aligned}$$

- A vector space equipped with an inner product is an inner product space, equipped with a norm via $\|f\|_{\mathbb{V}} := \sqrt{\langle f, f \rangle_{\mathbb{V}}}$.
- A sequence $(f_n)_{n=1}^{\infty}$ with elements in $\mathbb{V}$ is Cauchy if, for all $\epsilon > 0$, there exists some integer $N(\epsilon)$ such that

$$\|f_n - f_m\|_{\mathbb{V}} < \epsilon \quad \forall n, m \geq N(\epsilon).$$

Hilbert Space

- A Hilbert space $\mathbb{H}$ is an inner product space $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ in which every Cauchy sequence $(f_n)_{n=1}^{\infty}$ in $\mathbb{H}$ converges to some element $f^* \in \mathbb{H}$, e.g.
  - $l^2(\mathbb{N}) := \{(\theta_j)_{j=1}^{\infty} | \sum_{j=1}^{\infty} \theta_j^2 < \infty\}$;
  - $L^2[0,1]$, the space of functions $f : [0,1] \to \mathbb{R}$ that is Lebesgue-integrable.

  Both of them are **separable** Hilbert spaces for which there is a countable dense subset.
- A **linear functional** on a Hilbert space $\mathbb{H}$ is a mapping $L : \mathbb{H} \to \mathbb{R}$ that is linear, meaning that $L(f + \alpha g) = L(f) + \alpha L(g)$ for all $f, g \in \mathbb{H}, \alpha \in \mathbb{R}$.
- A linear functional is said to be **bounded** if there exists some $M < \infty$ such that $|L(f)| \leq M \|f\|_{\mathbb{H}}$ for all $f \in \mathbb{H}$. **A bounded linear functional must be continuous, and vice versa.**

Riesz representation

Theorem 1

Let L be a bounded linear functional on a Hilbert space. Then there exists a unique $g \in \mathbb{H}$ such that $L(f) = \langle f, g \rangle_{\mathbb{H}}$ for all $f \in \mathbb{H}$. (We refer to g as the representer of the functional L.)

- Example: A given vector $\alpha \in \mathbb{R}^n$ acts as a bounded liner functional on $\mathbb{R}^n$ by the inner product, i.e.

$$\alpha : \mathbb{R}^n \to \mathbb{R}, \quad \beta \mapsto \langle \alpha, \beta \rangle.$$

Positive semidefinite kernel functions

Definition 2 (Positive semidefinite kernel function)

A symmetric bivariate function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semidefinite (PSD) if for all integers $n \geq 1$ and elements $\{x_i\}_{i=1}^n \subset \mathcal{X}$, the $n \times n$ matrix with elements $K_{ij} := \mathcal{K}(x_i, x_j)$ is positive semidefinite.

- Non-negative linear combination of PSD kernels is a PSD kernel.
- Any finite product and sum of PSD kernels is a PSD kernel.
- A pointwise limit of PSD kernels is also a PSD kernel.
- $f : \mathcal{X} \to \mathbb{R}$, $\tilde{\mathcal{K}}(x, z) = f(x)\mathcal{K}(x, z)f(z)$ is a PSD kernel if $\mathcal{K}$ is.

Examples of PSD kernels

- Linear kernels: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{K}(x, x') := \langle x, x' \rangle$
- Polynomial kernels: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{K}(x, z) = (\langle x, z \rangle)^m$ (homogeneous), $\mathcal{K}(x, z) = (1 + \langle x, z \rangle)^m$ (inhomogeneous).
- Finite rank kernel: $K(t, s) = \sum_{k=1}^{K} \eta_k \phi_k(t) \phi_k(s)$, with $\eta_k \geq 0$.
- Covariance function of Gaussian process, $K(s, t) = \mathrm{Cov}(X_s, X_t)$, where $(X_t)_{t \in \mathcal{T}}$ is a Gaussian process. For example, the covariance function of Brownian motion: $\min(s, t)$.

Examples of PSD kernels

- Gaussian kernels: Given some compact subset $\mathcal{X} \subseteq \mathbb{R}^d$,

$$\mathcal{K}(x, z) = \exp\big(-\frac{1}{2\sigma^2}\|x - z\|_2^2\big).$$

- The kernel $\mathcal{K}(s, t) = \int G(s, z)G(t, z)\,\mathrm{d}z$. For example,

$$\mathcal{K}(x, y) = \int_0^1 \frac{(x - z)_+^{\alpha-1}}{(\alpha - 1)!}\frac{(y - z)_+^{\alpha-1}}{(\alpha - 1)!}dz,$$

  where $\alpha \geq 2$.

Reproducing kernel Hilbert spaces

Now we focus on the notion of a reproducing kernel Hilbert space (RKHS). For short:

- Focus on positive semidefinite kernel and use it to construct a Hilbert space in explicit way.
- Hilbert spaces in which the evaluation functionals (the linear mappings from $\mathbb{H}$ to $\mathbb{R}$ obtained by evaluating each function at a given point) are bounded.

Constructing an RKHS by a kernel

Theorem 3

Given any PSD kernel function $\mathcal{K}$, there is an unique ***Hilbert space of function*** $\mathbb{H}$ *on $\mathcal{X}$ in which the kernel satisfies the reproducing property [1]:*

- $\mathcal{K}(\cdot, x) \in \mathbb{H} \quad \forall x \in \mathcal{X}$,
- $\langle f, \mathcal{K}(\cdot, x) \rangle_{\mathbb{H}} = f(x) \quad \forall f \in \mathbb{H}$.

It is known as the reproducing Kernel Hilbert space associated with $\mathcal{K}$.

We prove this theorem in three steps:

- Construct an inner product space $\tilde{\mathbb{H}}$.
- Complete $\tilde{\mathbb{H}}$ to $\mathbb{H}$.
- $\mathbb{H}$ is the unique Hilbert space satisfies conditions.
- Importantly, given a RKHS, the kernel is uniquely determined and must be symmetric and positive.

Proof of Theorem 1: construct $\tilde{\mathbb{H}}$ I

Let $\tilde{\mathbb{H}} = \{f | f(\cdot) = \sum_{j=1}^{n} \alpha_j \mathcal{K}(\cdot, x_j), n \geq 1, \alpha_j \in \mathbb{R}, x_j \in \mathcal{X}\}$. It's easy to check $\tilde{\mathbb{H}}$ is a vector space. Note that $\mathcal{K}(\cdot, x) \in \tilde{\mathbb{H}}$, we define

$$\langle \mathcal{K}(\cdot, x), \mathcal{K}(\cdot, z) \rangle_{\mathbb{H}} = \mathcal{K}(x, z) \quad \forall x, z \in \mathcal{X}.$$

Then for $f(\cdot) = \sum_{j=1}^{n} \alpha_j \mathcal{K}(\cdot, x_j)$, $\tilde{f}(\cdot) = \sum_{j=1}^{n} \tilde{\alpha}_j \mathcal{K}(\cdot, x_j)$, we have

$$\langle f, \tilde{f} \rangle_{\tilde{\mathbb{H}}} := \sum_{j=1}^{n} \sum_{k=1}^{\tilde{n}} \alpha_j \tilde{\alpha}_k \mathcal{K}(x_j, \tilde{x}_k)$$

$$\langle f, \mathcal{K}(\cdot, x) \rangle_{\tilde{\mathbb{H}}} = \sum_{j=1}^{n} \alpha_j \mathcal{K}(x_j, x) = f(x)$$

Moreover, $\langle \cdot, \cdot \rangle_{\tilde{\mathbb{H}}}$ is indeed an inner product.

Proof of Theorem 1: construct $\tilde{\mathbb{H}}$ II

Here we only check that $\langle f, f \rangle_{\tilde{\mathbb{H}}} \geq 0$ iff $f = 0$, or equivalently that $f(x) = \sum_{i=1}^{n} \alpha_i \mathcal{K}(x, x_i) = 0, \forall x \in \mathcal{X}$. Note that $\forall (a, x) \in \mathbb{R} \times \mathcal{X}$, we have

$$0 \leq \left\| a\mathcal{K}(\cdot, x) + \sum_{i=1}^{n} \alpha_i \mathcal{K}(\cdot, x_i) \right\|_{\tilde{\mathbb{H}}}^2 = a^2 \mathcal{K}(x, x) + 2a \sum_{i=1}^{n} \alpha_i \mathcal{K}(x, x_i)$$

Since $\mathcal{K}(x, x) \geq 0$ and the scalar $a \in \mathbb{R}$ is arbitrary, this inequality can hold only if $\sum_{i=1}^{n} \alpha_i \mathcal{K}(x, x_i) = 0$. i.e. $f = 0$.

Proof of Theorem 1: complete the space

- $\mathbb{H} = \left\{ f | f(x) := \lim_{n\to\infty} f_n(x), \left(f_n\right)_{n=1}^{\infty} \text{is a Cauchy sequence in } \tilde{\mathbb{H}} \right\}$.
- Define $\|f\|_{\mathbb{H}} := \lim_{n\to\infty} \|f_n\|_{\tilde{\mathbb{H}}}$. Any Cauchy sequence $(g_n)_{n=1}^{\infty}$ in $\tilde{\mathbb{H}}$ such that $\lim_{n\to\infty} g_n(x) = 0$ for all $x \in \mathcal{X}$, we also have $\lim_{n\to\infty} \|g_n\|_{\tilde{\mathbb{H}}} = 0$. So that the definition is sensible.
- The norm can be used to define an inner product on $\mathbb{H}$ via the polarization identity

$$\langle f, g \rangle_{\mathbb{H}} := \frac{1}{2} \left\{ \|f+g\|_{\mathbb{H}}^2 - \|f\|_{\mathbb{H}}^2 + \|g\|_{\mathbb{H}}^2 \right\}.$$

With this definition, it can be shown that $\langle \mathcal{K}(\cdot, x), f \rangle_{\mathbb{H}} = f(x)$ for all $f \in \mathbb{H}$, so that $\mathcal{K}(\cdot, x)$ is again reproducing over $\mathbb{H}$.

Proof of Theorem 1: Uniqueness

- Suppose that $\mathbb{G}$ is some other Hilbert space with $\mathcal{K}$ as its reproducing kernel, so that $\mathcal{K}(\cdot, x) \in \mathbb{G}$ for all $x \in \mathcal{X}$.
- Since $\mathbb{G}$ is complete and closed under linear operations, we must have $\mathbb{H} \subseteq \mathbb{G}$. Consequently, $\mathbb{H}$ is a closed linear subspace of $\mathbb{G}$, so that we can write $\mathbb{G} = \mathbb{H} \oplus \mathbb{H}^{\perp}$. Let $g \in \mathbb{H}^{\perp}$ be arbitrary, and note that $\mathcal{K}(\cdot, x) \in \mathbb{H}$.
- By orthogonality, we must have $0 = \langle \mathcal{K}(\cdot, x), g \rangle_{\mathbb{G}} = g(x)$, from which we conclude that $\mathbb{H}^{\perp} = \{0\}$, and hence that $\mathbb{H} = \mathbb{G}$ as claimed.

Kernel and RKHS

- From Theorem 3, we know that given a kernel we can construct a unique RKHS. Moreover, given a RKHS $\mathbb{H}$, $\exists!$ kernel function $\mathcal{K}$ satisfying the reproducing property.
- By $\mathcal{K}(s,t) = \langle \mathcal{K}(\cdot,s), \mathcal{K}(\cdot,t) \rangle_{\mathbb{H}}$, it's easy to check $\mathcal{K}$ is PSD.
- If $\exists\ \tilde{\mathcal{K}}$ satisfying the reproducing property too, then $\forall t, t' \in \mathcal{X}$, we have

$$\tilde{\mathcal{K}}(t,t') = \langle \tilde{\mathcal{K}}(\cdot,t'), \mathcal{K}(\cdot,t) \rangle_{\mathbb{H}} = \langle \mathcal{K}(\cdot,t), \tilde{\mathcal{K}}(\cdot,t') \rangle_{\mathbb{H}} = \mathcal{K}(t,t')$$

So the kernel is unique.

Hilbert space with bounded evaluation functionals

- By Riesz representation theorem, the reproducing property is equivalent to asserting that the function $\mathcal{K}(\cdot, x)$ acts as the representer for the evaluation functional at x, namely, the linear functional $L_x : \mathbb{H} \to \mathbb{R}$ that performs the operation $f \mapsto f(x)$.
- In any reproducing kernel Hilbert space, the evaluation functionals are all bounded.

Definition 4

A reproducing kernel Hilbert space $\mathbb{H}$ is a Hilbert space of functions on $\mathcal{X}$ s.t. for each $x \in \mathcal{X}$, the evaluation functional $L_x : H \to \mathbb{R}$ is bounded (i.e., there exists some $M < \infty$ such that $|L_x(f)| \leq M\|f\|_H$ for all $f \in \mathbb{H}$).

Hilbert space with bounded evaluation functionals

- L_x is bounded, then $\exists R_x \in \mathbb{H}$ such that

$$f(x) = L_x(f) = \langle f, R_x \rangle_{\mathbb{H}} \quad \forall f \in \mathbb{H}.$$

- Define a real-valued function $\mathcal{K}$ on the Cartesian product space $\mathcal{X} \times \mathcal{X}$ via $\mathcal{K}(x, z) := R_x(z)$. Then $\mathcal{K}$ is the reproducing kernel of H, hence H is the RKHS.
- The reproducing kernel of an RKHS is unique.

Example of RKHS

- Finite dimensional function space with orthogonal basis function $\psi_k(x)$ with inner product $\langle \sum_k a_k \psi_k, \sum_k b_k \psi_k \rangle = \sum_k \frac{a_k b_k}{\mu_k}$, the reproducing kernel of which is $\mathcal{K}(x, y) = \sum_k \mu_k \psi_k(t) \psi_k(s)$.
- Sobolev space $\mathbb{H}^\alpha[0, 1]$, which is the absolutely continuous function on $[0, 1]$ $|f^{(k)}(0) = 0,\ k < \alpha,\ f^{(\alpha)} \in L^2[0, 1]$, with the inner product $\langle f, g \rangle = \int f^{(\alpha)} g^{(\alpha)}$. The reproducing kernel of which is

$$\mathcal{K}(x, y) = \int_0^1 \frac{(x - z)_+^{\alpha-1}}{(\alpha - 1)!} \frac{(y - z)_+^{\alpha-1}}{(\alpha - 1)!} dz$$

if $\alpha > 1$, and $\mathcal{K}(x, y) = \min(x, y)$ if $\alpha = 1$.

Mercer's theorem

- We can represent kernel functions in terms of their eigenfunctions.
- A symmetric positive semi-definite matrix A can be seen as a kernel function on $\mathbb{R}^n \times \mathbb{R}^n$. Then we can find orthogonormal eigenvectors $\{\beta_1, \ldots, \beta_n\}$ of A and eigenvalues $\mu_1, \ldots, u_n$ such that

$$A = \sum_{i=1}^{n} \mu_i \beta_i \beta_i^T.$$

- Mercer's theorem can be seen as the generation of spectral representation of a matrix.

Theorem 5 (Mercer's theorem)

Suppose that $\mathcal{X}$ is compact, the kernel function $\mathcal{K}$ is continuous and PSD, and satisfies the Hilbert-Schmidt condition. Then there exist a sequence of eigenfunctions $(\phi_j)_{j=1}^{\infty}$ that form an orthonormal basis of $L^2(X;\mathbb{P})$, and non-negative eigenvalues $(\mu_j)_{j=1}^{\infty}$ such that

$$T_{\mathcal{K}}(\phi_j) = \mu_j \phi_j \quad \text{for } j = 1, 2, \ldots.$$

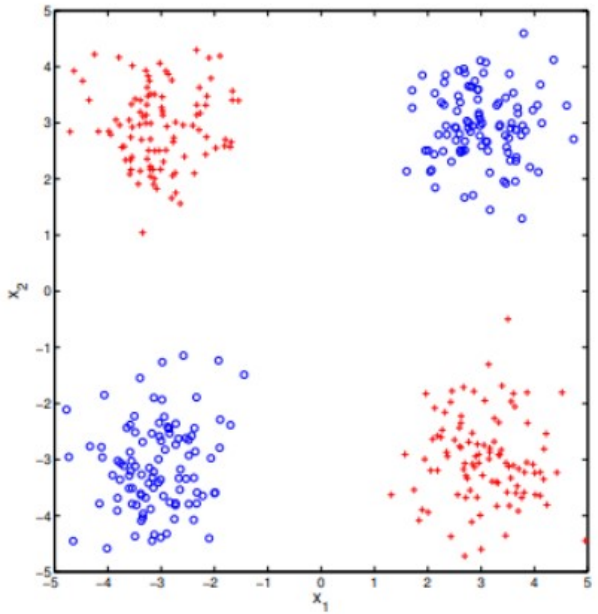
Moreover, the kernel function has the expansion

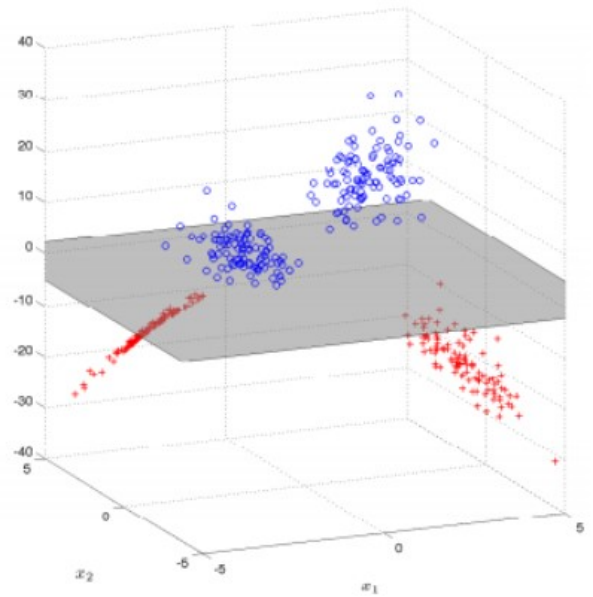
$$\mathcal{K}(x, z) = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(z)$$

where the convergence of the infinite series holds absolutely and uniformly.

Feature maps

Sometime we want to map objects from low dimensions to higher dimensions so that we can divide two different kinds of data. To achieve this, we should find the feature map Φ which embeds the original data into a higher-dimensional space.





Feature maps

- By Mercer's theorem, we can associate an RKHS (with PSD $\mathcal{K}$) with the high dimensional space (isomorphism preserving the inner product), hence $\langle \Phi(x), \Phi(z) \rangle_{l_2(\mathbb{N})}$ could be calculated as the inner product of the corresponding elements in RKHS.
- (**Kernel trick**) We can compute inner products between the embedded data pairs $(\Phi(x), \Phi(z))$ directly by $\mathcal{K}$, rather than specifying the feature map.

A feature map specifying by a kernel

- Given a PSD $\mathcal{K} = \mathcal{K}(x, z) = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(z)$, we define the feature map as

$$\Phi : \mathcal{X} \to \mathbb{R}^{\infty}, \quad x \mapsto (\mu_1 \phi_1(x), \ldots, \mu_{\infty} \phi_{\infty}(x)),$$

  and the inner product of $\mathbb{R}^{\infty}$ is $\langle x, y \rangle = \sum_{j=1}^{\infty} \frac{x_j y_j}{\mu_j}$.
- Define isomorphism preserving the inner product:

$$\mathbb{R}^{\infty} \leftrightarrow \mathbb{H} : \Phi(x) \leftrightarrow \sum_{j=1}^{\infty} \mu_i \phi_j(x) \phi_j(\cdot) = \mathcal{K}(x, \cdot).$$

- $\langle \mathcal{K}(x, \cdot), \mathcal{K}(y, \cdot) \rangle_{\mathbb{H}} := \langle \Phi(x), \Phi(y) \rangle = \mathcal{K}(x, y)$.
- $\mathbb{H}$ must be a Hilbert space, hence be the RKHS. Therefore, given $\mathcal{K}$, the $\mathbb{R}^{\infty}$ (and feature map) could be uniquely determined, which is the kernel trick.

Sums of RKHS

Given two Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$ of functions defined on domains $\mathcal{X}_1, \mathcal{X}_2$ respectively, consider the space

$$\mathbb{H}_1 + \mathbb{H}_2 := \{f_1 + f_2 | f_j \in \mathbb{H}_j, j = 1, 2\},$$

Proposition 1

Suppose that $\mathbb{H}_1$ and $\mathbb{H}_2$ are both RKHSs with kernels $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively. Then the space $\mathbb{H} = \mathbb{H}_1 + \mathbb{H}_2$ with norm

$$\|f\|_{\mathbb{H}}^2 := \min_{\substack{f=f_1+f_2 \\ f_1 \in \mathbb{H}_1, f_2 \in \mathbb{H}_2}} \left\{ \|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2 \right\}$$

is an RKHS with kernel $\mathcal{K} = \mathcal{K}_1 + \mathcal{K}_2$.

- Example: Additive models.

Generalized Sobolev Space

- We remove the condition for $H^\alpha[0,1]$: $f^{(k)}(0) = 0,\ k < \alpha$ on the definition of Sobolev space $\mathbb{H}^\alpha[0,1]$, hence we add a null space of function H_0 with basis function $\psi_k(x) = \frac{x^k}{k!}, k = 0, \cdots, \alpha - 1$ with normal inner product, hence, we define

$$W^\alpha[0,1] = H^\alpha[0,1] + H_0.$$

The reproducing kernel of which is

$$\mathcal{K}(x,y) = \int_0^1 \frac{(x-z)_+^{\alpha-1}}{(\alpha-1)!}\frac{(y-z)_+^{\alpha-1}}{(\alpha-1)!}dz + \sum_k \psi_k(x)\psi_k(y)$$

if $\alpha > 1$. The associated RKHS norm is $\int (f^{(\alpha)})^2 + \sum_{k=0}^{\alpha-1}(f^{(k)}(0))^2$.

Proof of proposition 1

- Consider the direct sum $\mathbb{F} = \{(f_1, f_2) | f_j \in \mathbb{H}_j, j = 1, 2\}$ with the norm $\|(f_1, f_2)\|_{\mathbb{F}}^2 := \|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2$
- Consider $L : \mathbb{F} \to \mathbb{H}, (f_1, f_2) \mapsto f_1 + f_2$. N(L) is its nullspace.
- Let $N^{\perp}$ be the orthogonal complement of N(L) in F, and let $L_{\perp}$ be the restriction of L to $N^{\perp}$. Since $L_{\perp}$ is a bijection between $N^{\perp}$ and $\mathbb{H}$, we define an inner product on $\mathbb{H}$ via $\langle f, g \rangle_{\mathbb{H}} := \langle L_{\perp}^{-1}(f), L_{\perp}^{-1}(g) \rangle_{\mathbb{F}}$ $\mathbb{H}$ with this inner product is a Hilbert space.

Proof of proposition 1

- $\mathbb{H}$ is an RKHS with kernel $\mathcal{K} = \mathcal{K}_1 + \mathcal{K}_2$
  Since $\mathcal{K}_1(\cdot, x) \in \mathbb{H}_1, \mathcal{K}_2(\cdot, x) \in \mathbb{H}_2$, then we know $\mathcal{K}(\cdot, x) = \mathcal{K}_1(\cdot, x) + \mathcal{K}_2(\cdot, x) \in \mathbb{H}$.
  For a fixed $f \in \mathbb{F}$, let $(f_1, f_2) = L_{\perp}^{-1}(f) \in \mathbb{F}$, and for a fixed $x \in \mathcal{X}$, let $(g_1, g_2) = L_{\perp}^{-1}(\mathcal{K}(\cdot, x)) \in \mathbb{F}$.
  Since $(g_1 - \mathcal{K}_1(\cdot, x), g_2 - \mathcal{K}_2(\cdot, x)) \in \mathbb{N}(L)$, it must be orthogonal (in $\mathbb{F}$) to the element $(f_1, f_2) \in \mathbb{N}^{\perp}$.

$$\langle (g_1 - \mathcal{K}_1(\cdot, x), g_2 - \mathcal{K}_2(\cdot, x)), (f_1, f_2) \rangle_F = 0$$

$$\begin{aligned} f(x) &= f_1(x) + f_2(x) \\ &= \langle f_1, \mathcal{K}_1(\cdot, x) \rangle_{\mathbb{H}_1} + \langle f_2, \mathcal{K}_2(\cdot, x) \rangle_{\mathbb{H}_2} \\ &= \langle f_1, g_1 \rangle_{\mathbb{H}_1} + \langle f_2, g_2 \rangle_{\mathbb{H}_2} \\ &= \langle f, \mathcal{K}(\cdot, x) \rangle_{\mathbb{H}}. \end{aligned}$$

Proof of proposition 1

- For a given $f \in \mathbb{H}$, consider some pair $(f_1, f_2) \in \mathbb{F}$ such that $f = f_1 + f_2$, and define $(v_1, v_2) = (f_1, f_2) - L_\perp^{-1}(f)$. We have

$$\|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2 \overset{\text{(i)}}{=} \|(f_1, f_2)\|_{\mathbb{F}}^2 \overset{\text{(ii)}}{=} \|(v_1, v_2)\|_{\mathbb{F}}^2 + \left\|L_\perp^{-1}(f)\right\|_{\mathbb{F}}^2 \overset{\text{(iii)}}{=} \|(v_1, v_2)\|_{\mathbb{F}}^2 + \|f\|_{\mathbb{H}}^2,$$

  Hence we have

$$\|f\|_{\mathbb{H}}^2 \leq \|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2 ,$$

  with equality holding if and only if $(v_1, v_2) = (0, 0)$
  So $\|f\|_{\mathbb{H}}^2 = \min_{\substack{f=f_1+f_2 \\ f_1 \in \mathbb{H}_1, f_2 \in \mathbb{H}_2}} \left\{ \|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2 \right\}$.

Tensor products I

The tensor product of $\mathbb{H}_1$ and $\mathbb{H}_2$ is defined as

$$\mathbb{H} = \mathbb{H}_1 \otimes \mathbb{H}_2 = \left\{ h : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R} | h = \sum_{j=1}^{n} f_j g_j, n \in \mathbb{N}, f_j \in \mathbb{H}_1, g_j \in \mathbb{H}_2 \right\}$$

If $h = \sum_{j=1}^{n} f_j g_j, \tilde{h} = \sum_{k=1}^{m} \tilde{f}_k \tilde{g}_k$, their inner product can be defined as

$$\langle h, \tilde{h} \rangle_{\mathbb{H}} := \sum_{j=1}^{n} \sum_{k=1}^{m} \langle f_j, \tilde{f}_k \rangle_{\mathbb{H}_1} \langle g_j, \tilde{g}_k \rangle_{\mathbb{H}_2}$$

Write it as $\langle h, \tilde{h} \rangle_{\mathbb{H}} = \sum_{k=1}^{m} \langle (h \odot \tilde{f}_k), \tilde{g}_k \rangle_{\mathbb{H}_2}$, $(h \odot \tilde{f}_k) \in \mathbb{H}_2$ is the function given by $x_2 \mapsto \langle h(\cdot, x_2), \tilde{f}_k \rangle_{\mathbb{H}_1}$. In this way we know that the inner product does not depend on the representation of $h, \tilde{h}$.

Tensor products II

Proposition 2

Suppose that $\mathbb{H}_1$ and $\mathbb{H}_2$ are reproducing kernel Hilbert spaces of real-valued functions with domains $\mathcal{X}_1$ and $\mathcal{X}_2$, and equipped with kernels $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively. Then the tensor product space $\mathbb{H} = \mathbb{H}_1 \otimes \mathbb{H}_2$ is an $RKHS$ of real-valued functions with domain $\mathcal{X}_1 \times \mathcal{X}_2$, and with kernel function

$$\mathcal{K}\left((x_1, x_2), (x_1', x_2')\right) = \mathcal{K}_1\left(x_1, x_1'\right)\mathcal{K}_2\left(x_2, x_2'\right).$$

Let $f = \sum_{j,k=1}^{n} \alpha_{j,k}\phi_j\psi_k$ be an arbitrary element of $\mathbb{H}$, we have

$$\begin{aligned}\left\langle f, \mathcal{K}\left((\cdot,\cdot),(x_1,x_2)\right)\right\rangle_{\mathbb{H}} &= \sum_{j,k=1}^{n} \alpha_{j,k}\left\langle \phi_j, \mathcal{K}_1\left(\cdot, x_1\right)\right\rangle_{\mathbb{H}_1}\left\langle \psi_k, \mathcal{K}_2\left(\cdot, x_2\right)\right\rangle_{\mathbb{H}_2} \\ &= \sum_{j,k=1}^{n} \alpha_{j,k}\phi_j\left(x_1\right)\psi_k\left(x_2\right) = f\left(x_1, x_2\right)\end{aligned}$$

Penalized Least Square

Now we focus on the penalized least square

$$\arg \min_{\beta \in \mathbb{R}^\infty} \sum_{i=1}^{n} L(y_i, \langle \beta, \Phi(x_i) \rangle) + \lambda ||\beta||_\infty^2.$$

By the isomorphism mentioned above, it's equivalent to

$$\arg \min_{f \in \mathbb{H}} \sum_{i=1}^{n} L(y_i, f(x_i)\rangle) + \lambda ||f||_{\mathbb{H}}^2,$$

where $\mathbb{H}$ is the RKHS with $\mathcal{K}$. It can be shown that

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i \mathcal{K}(x, x_i).$$

Finite dimensional representation

- Let $f_\alpha(\cdot) := \sum_{i=1}^n \alpha_i \mathcal{K}(\cdot, x_i)$, $\mathbb{L} := \{f_\alpha \mid \alpha \in \mathbb{R}^n\}$.
- Note that $\mathbb{L}$ is also the closed Hilbert space, we have $\mathbb{H} = \mathbb{L} \oplus \mathbb{L}^\perp$, then for any $f \in \mathbb{H}$,

$$f = f_\alpha + f_\perp.$$

  where $f_\alpha \in \mathbb{L}$ and $f_\perp$ is orthogonal to $\mathbb{L}$.
- We have

$$f(x_j) = \langle f, \mathcal{K}(\cdot, x_j) \rangle_{\mathbb{H}} = \langle f_\alpha + f_\perp, \mathcal{K}(\cdot, x_j) \rangle_{\mathbb{H}} = f_\alpha(x_j)$$

- $\|f_\alpha + f_\perp\|_H^2 = \|f_\alpha\|_{\mathbb{H}}^2 + \|f_\perp\|_{\mathbb{H}}^2 \geq \|f_\alpha\|_{\mathbb{H}}^2$.

Kernel ridge regression

Proposition 3

For all $\lambda_n > 0$, the kernel ridge regression estimate can be written as

$$\widehat{f}(\cdot) = \sum_{i=1}^{n} \widehat{\alpha}_i \mathcal{K}(\cdot, x_i),$$

where the optimal weight vector $\widehat{\alpha} \in \mathbb{R}^n$ is given by

$$\widehat{\alpha} = (K + \lambda_n \mathbf{I}_n)^{-1} y,$$

if $L(x, y) = (x - y)^2$.

This is also equivalent to asumming that f is a mean zero Gaussian process with covariance function $\mathcal{K}$, and $\hat{f}$ is the posterior mean of f given that raw data y.

KRR estimate

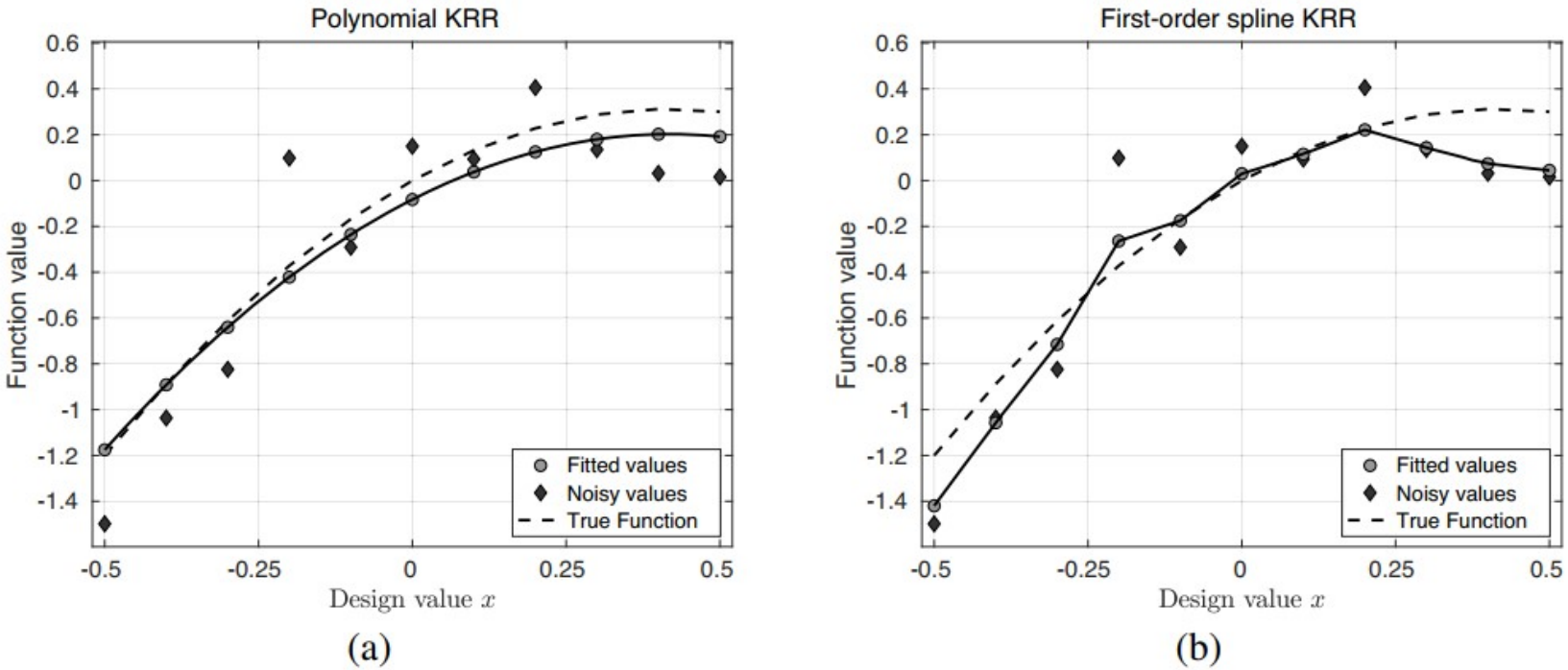


Figure 12.2 Illustration of kernel ridge regression estimates of function $f^*(x) = \frac{3x}{2} - \frac{9}{5}x^2$ based on $n = 11$ samples, located at design points $x_i = -0.5 + 0.10\,(i-1)$ over the interval $[-0.5, 0.5]$. (a) Kernel ridge regression estimate using the second-order polynomial kernel $\mathcal{K}(x, z) = (1 + xz)^2$ and regularization parameter $\lambda_n = 0.10$. (b) Kernel ridge regression estimate using the first-order Sobolev kernel $\mathcal{K}(x, z) = 1 + \min\{x, z\}$ and regularization parameter $\lambda_n = 0.10$.

Smoothing Spline

Now we assume $\mathbb{H} = W^2[0,1] = H^2[0,1] + H_0$, with penalty $||\mathcal{P}f||_H$. Among $\mathcal{P}$ is the projection operator onto $H^2[0,1]$. Hence, $||\mathcal{P}f||_H = \int (f^{(2)})^2$, which is the smoothing penalty. Similarly,

$$\arg\min_{f\in\mathbb{H}} \sum_{i=1}^{n} L(y_i, f(x_i)\rangle) + \lambda||\mathcal{P}f||_{\mathbb{H}}^2.$$

It can be shown that

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i \mathcal{P}(\mathcal{K}(\cdot, x_i))|_{\cdot=x} + \sum_{k=0}^{1} \beta_k \frac{x^k}{k!},$$

which is the natural cubic spline.

Random Features for Large-Scale Kernel Machines

- Kernel Trick: For any positive definite function $k(\mathbf{x}, \mathbf{y})$ with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it defines an inner product and a lifting ϕ such that $\langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle = k(\mathbf{x}, \mathbf{y})$.
- Problem: Algorithm accesses data only through kernel evaluations $k(\mathbf{x}, \mathbf{y})$, so large training sets incur large computational and storage costs.
- Solution: An explicit mapping of the data to a low-dimensional Euclidean inner product space using a randomized feature map $\mathbf{z} : \mathbb{R}^d \to \mathbb{R}^D$, so that the inner product between a pair of transformed points approximates their kernel evaluation, as shown in [2]:

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle \approx \mathbf{z}(\mathbf{x})' \mathbf{z}(\mathbf{y}). \tag{1}$$

  Here, $\mathbf{z}$ is low-dimensional, that is $D < d$.
- In what follows, we show how to construct feature spaces that uniformly approximate popular shift-invariant kernels $k(\mathbf{x} - \mathbf{y})$.

Random Fourier Features

- In this section, our first set of random features project data points onto a randomly chosen line, and then pass the resulting scalar through a sinusoid.
- The following classical theorem from harmonic analysis provides the key insight behind this transformation:

Theorem 6 (Bochner)

A continuous kernel $k(x, y) = k(x - y)$ on $\mathbb{R}^d$ is positive definite if and only if $k(\delta)$ is the Fourier transform of a non-negative measure.

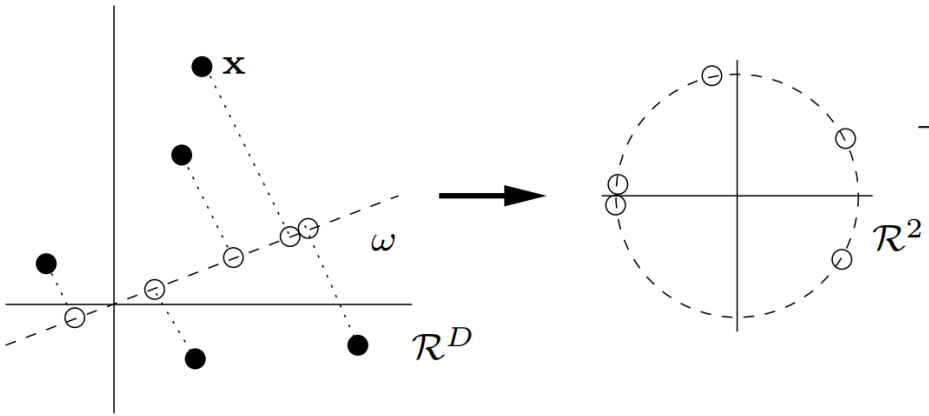
Random Fourier Features

For properly scaled kernel $k(\delta)$, Bochner's theorem guarantees that its Fourier transform $p(\omega)$ is a proper probability distribution. Define $\zeta_\omega(x) = e^{j\omega' \mathbf{x}}$, we have

$$k(\mathbf{x}-\mathbf{y}) = \int_{\mathcal{R}^d} p(\omega) e^{j\omega'(\mathbf{x}-\mathbf{y})} d\omega = E_\omega \left[\zeta_\omega(\mathbf{x})\zeta_\omega(\mathbf{y})^*\right]. \tag{2}$$

so $\zeta_\omega(\mathbf{x})\zeta_\omega(\mathbf{y})^*$ is an unbiased estimate of $k(\mathbf{x},\mathbf{y})$ when ω is drawn from p.

Random Fourier Features



Kernel Name	$k(\Delta)$	$p(\omega)$
Gaussian	$e^{-\frac{\|\Delta\|_2^2}{2}}$	$(2\pi)^{-\frac{D}{2}} e^{-\frac{\|\omega\|_2^2}{2}}$
Laplacian	$e^{-\|\Delta\|_1}$	$\prod_d \frac{1}{\pi(1+\omega_d^2)}$
Cauchy	$\prod_d \frac{2}{1+\Delta_d^2}$	$e^{-\|\Delta\|_1}$

Figure: Random Fourier Features. Each component of the feature map $z(x)$ projects x onto a random direction ω drawn from the Fourier transform $p(\omega)$ of $k(\Delta)$, and wraps this line onto the unit circle in $\mathbb{R}^2$. After transforming two points x and y in this way, their inner product is an unbiased estimator of $k(x, y)$. The table lists some popular shift-invariant kernels and their Fourier transforms. To deal with non-isotropic kernels, the data may be whitened before applying one of these kernels.

Random Fourier Features

- We suppose that both $p(\omega)$ and $k(\Delta)$ are real, so the integrand $e^{j\omega'(\mathbf{x}-\mathbf{y})}$ can be replaced by $\cos \omega'(\mathbf{x}-\mathbf{y})$.
- Define $z_\omega(\mathbf{x}) = [\cos(\omega'\mathbf{x}), \sin(\omega'\mathbf{x})]'$ gives a real-valued mapping that satisfies the condition

$$E\left[z_\omega(\mathbf{x})' z_\omega(\mathbf{y})\right] = k(\mathbf{x}, \mathbf{y}). \tag{3}$$

- Other possible mappings: $z_\omega(\mathbf{x}) = \sqrt{2}\cos(\omega'\mathbf{x} + b)$, where ω is drawn from $p(\omega)$ and b is drawn uniformly from $[0, 2\pi]$, also satisfy (3).

Random Fourier Features

- We only need to consider the approximation of $E\left[z_\omega(\mathbf{x})' z_\omega(\mathbf{y})\right]$, a common choice is

$$\mathbf{z}(\mathbf{x})'\mathbf{z}(\mathbf{y}) = \frac{1}{D}\sum_{j=1}^{D} z_{\omega_j}(\mathbf{x}) z_{\omega_j}(\mathbf{y}) \tag{4}$$

where $\mathbf{z}(\mathbf{x}) = \sqrt{\frac{1}{D}}\left[\cos\left(\omega_1'\mathbf{x}\right)\cdots\cos\left(\omega_D'\mathbf{x}\right)\sin\left(\omega_1'\mathbf{x}\right)\cdots\sin\left(\omega_D'\mathbf{x}\right)\right]'$ and $\omega_1, \cdots, \omega_D \in \mathcal{R}^d$ are D iid samples from p.

Random Fourier Features

- Applying Hoeffding's Inequality directly, we have

$$\Pr\left[\left|\mathbf{z}(\mathbf{x})'\mathbf{z}(\mathbf{y}) - k(\mathbf{x},\mathbf{y})\right| \geq \epsilon\right] \leq 2\exp\left(-D\epsilon^2/2\right). \tag{5}$$

- A stronger assertion guarantees the uniform concentration

Proposition 4 (Uniform convergence of Fourier features)

Let $\mathcal{M}$ be a compact subset of $\mathcal{R}^d$ with diameter $\operatorname{diam}(\mathcal{M})$. Then, for the mapping $\mathbf{z}$ defined in Algorithm 1, we have

$$\Pr\left[\sup_{x,y\in\mathcal{M}}\left|\mathbf{z}(\mathbf{x})'\mathbf{z}(\mathbf{y}) - k(\mathbf{x},\mathbf{y})\right| \geq \epsilon\right] \leq 2^8\left(\frac{\sigma_p \operatorname{diam}(\mathcal{M})}{\epsilon}\right)^2 e^{-\frac{D\epsilon^2}{4(d+2)}}.$$

where $\sigma_p^2 \equiv E_p\left[\omega'\omega\right]$ is the second moment of the Fourier transform of k. Further, $\sup_{x,y\in\mathcal{M}}\left|\mathbf{z}(\mathbf{x})'\mathbf{z}(\mathbf{y}) - k(\mathbf{y},\mathbf{x})\right| \leq \epsilon$ with any constant probability when $D = \Omega\left(\frac{d}{\epsilon^2}\log\frac{\sigma_p \operatorname{diam}(\mathcal{M})}{\epsilon}\right)$.

Random Fourier Features

Algorithm Random Fourier Features

Require: A positive definite shift-invariant kernel $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$.

Ensure: A randomized feature map $\mathbf{z}(\mathbf{x}) : \mathcal{R}^d \to \mathcal{R}^{2D}$ so that

$$\mathbf{z}(\mathbf{x})'\mathbf{z}(\mathbf{y}) \approx k(\mathbf{x} - \mathbf{y}).$$

Compute the Fourier transform p of the kernel k:

$$p(\omega) = \frac{1}{2\pi} \int e^{-j\omega'\Delta} k(\Delta) d\Delta.$$

Draw D iid samples $\omega_1, \cdots, \omega_D \in \mathcal{R}^d$ from p.

Let $\mathbf{z}(\mathbf{x}) \equiv \sqrt{\frac{1}{D}} \left[\cos\left(\omega_1'\mathbf{x}\right), \cdots, \cos\left(\omega_D'\mathbf{x}\right), \sin\left(\omega_1'\mathbf{x}\right), \cdots, \sin\left(\omega_D'\mathbf{x}\right)\right]'$.

Random Binning Features

- In this section, our random map partitions the input space using randomly shifted grids at randomly chosen resolutions and assigns to an input point a binary bit string that corresponds to the bin in which it falls.
- We construct the bin in a way that the probability $\mathbf{x}$ and $\mathbf{y}$ falls in the same bin is proportional to $k(\mathbf{x}, \mathbf{y})$.
- The inner product is proportional to the times that two points are binned together.

Random Binning Features

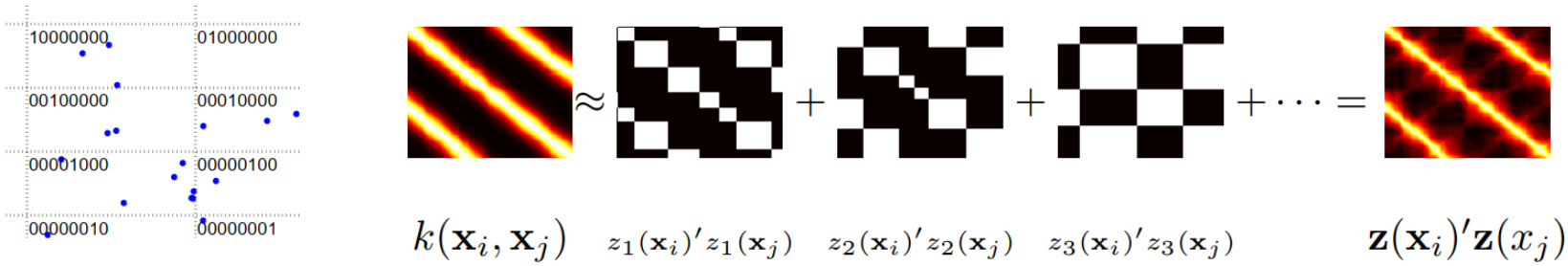


Figure: Random Binning Features. (left) The algorithm repeatedly partitions the input space using a randomly shifted grid at a randomly chosen resolution and assigns to each point $\mathbf{x}$ the bit string $z(\mathbf{x})$ associated with the bin to which it is assigned. (right) The binary adjacency matrix that describes this partitioning has $z\left(\mathbf{x}_i\right)' z\left(\mathbf{x}_j\right)$ in its ij th entry and is an unbiased estimate of kernel matrix.

Random Binning Features

- We start by considering the one-dimensional cases.
- We consider partition the real number line into intervals $[u + n\delta, u + (n + 1)\delta]$ for all integers n, where u is drawn uniformaly from the interval $[0, \delta]$.
- The probability that x and y falls in the same bin is $\max(0, \frac{|x-y|}{\delta})$, and we define $k_{hat}(x, y; \delta) = \max(0, \frac{|x-y|}{\delta})$.
- If we let $z(x)$ being a binary indicator vector over the bins, then

$$\Pr_u \left[z(x)'z(y) = 1 \mid \delta \right] = E_u \left[z(x)'z(y) \mid \delta \right] = k_{hat}(x, y; \delta). \quad (6)$$

Random Binning Features

- We need to find an appropriate probability distribution $p(\delta)$ that satisfies $k(x,y) = \int_o^\infty k(x,y;\delta)p(\delta)d\delta$, which guarantees as $\delta \sim p$,

$$E_{\delta,u}[z(x)'z(y)] = k(x,y). \tag{7}$$

- Such p is obtained follows from the Lemma below:

Lemma 7

Suppose a function $k(\Delta) : \mathcal{R} \to \mathcal{R}$ is twice differentiable and has the form $\int_0^\infty p(\delta) \max\left(0, 1 - \frac{\Delta}{\delta}\right) d\delta$. Then $p(\delta) = \delta\ddot{k}(\delta)$.

Random Binning Features

- For separable multivariate shift-invariant kernels of the form $k(\mathbf{x}-\mathbf{y}) = \prod_{m=1}^{d} k_m(|x^m - y^m|)$ can be constructed in a similar way if each k_m can be written as a convex combination of hat kernels. We apply the above binning process over each dimension of $\mathbb{R}^d$ independently.
- We can again reduce the variance by setting

$$\mathbf{z}(\mathbf{x})'\mathbf{z}(\mathbf{y}) = \frac{1}{P}\sum_{p=1}^{P} z_p(\mathbf{x}) z_p(\mathbf{y}). \tag{8}$$

Random Binning Features

- The following theorem shows again that the convergence holds simultaneously for all points.

Proposition 5

Let $\mathcal{M}$ be a compact subset of $\mathcal{R}^d$ with diameter $\operatorname{diam}(\mathcal{M})$. Let $\alpha = E[1/\delta]$ and let L_k denote the Lipschitz constant of k with respect to the L_1 norm. With $\mathbf{z}$ as above, we have

$$\Pr\left[\sup_{\mathbf{x},\mathbf{y}\in\mathcal{M}} \left|\mathbf{z}(\mathbf{x})'\mathbf{z}(\mathbf{y}) - k(\mathbf{x},\mathbf{y})\right| \leq \epsilon\right] \geq 1 - 36dP\alpha \operatorname{diam}(\mathcal{M}) e^{-\frac{\frac{P\epsilon^2}{8} + \ln\frac{\epsilon}{L_k}}{d+1}}.$$

Random Binning Features

Algorithm Random Binning Features

Require: A point $\mathbf{x} \in \mathbb{R}^d$. A kernel function $k(\mathbf{x}, \mathbf{y}) = \prod_{m=1}^{d} k_m \left(|x^m - y^m|\right)$, so that $p_m(\Delta) \equiv \Delta \ddot{k}_m(\Delta)$ is a probability distribution on $\Delta \geq 0$.

Ensure: A randomized feature map $\mathbf{z}(\mathbf{x})$ so that $\mathbf{z}(\mathbf{x})'\mathbf{z}(\mathbf{y}) \approx k(\mathbf{x} - \mathbf{y})$.

for $p = 1 \cdots P$ **do**

Draw grid parameters $\delta, \mathbf{u} \in \mathbb{R}^d$ with the pitch $\delta^m \sim p_m$, and shift u^m from the uniform distribution on $[0, \delta^m]$.

Let z return the coordinate of the bin containing $\mathbf{x}$ as a binary indicator vector $z_p(\mathbf{x}) \equiv \text{hash}\left(\left\lceil \frac{x^1 - u^1}{\delta^1} \right\rceil, \cdots, \left\lceil \frac{x^d - u^d}{\delta^d} \right\rceil\right)$.

end for

$\mathbf{z}(\mathbf{x}) \equiv \sqrt{\frac{1}{P}} \left[z_1(\mathbf{x}) \cdots z_P(\mathbf{x})\right]'$.

From Data Point to Probability Measure

Recall: Reproducing Kernel Hilbert Space

- Let $\mathcal{X}$ be a fixed non-empty set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a real-valued positive definite kernel function associated with the Hilbert space $\mathscr{H}$. For $f \in \mathscr{H}$, we have

$$\langle k(\mathbf{x}, \cdot), f \rangle_{\mathscr{H}} = f(\mathbf{x}). \tag{9}$$

- We may view the kernel evaluation as an inner product in $\mathscr{H}$ induced by a map from $\mathcal{X}$ into $\mathscr{H}$

$$\mathbf{x} \longmapsto k(\mathbf{x}, \cdot). \tag{10}$$

- In other words, $k(\mathbf{x}, \cdot)$ is a high dimensional representer of $\mathbf{x}$.

From Data Point to Probability Measure

We try to replace the data point $\mathbf{x}$ with Dirac measure $\delta_{\mathbf{x}}$.

- For any measurable function f on $\mathcal{X}$, we have

$$\int f(\mathbf{t})\delta_{\mathbf{x}}(\mathbf{t})\mathrm{d}t = f(\mathbf{x}). \tag{11}$$

- When f belongs to the Hilbert space $\mathscr{H}$ of functions on $\mathcal{X}$ with reproducing kernel k, we can rewrite (11) using the reproducing property of $\mathscr{H}$ as

$$\begin{aligned}\int f(\mathbf{t})\delta_{\mathbf{x}}(\mathbf{t})\mathrm{d}t &= \int \langle f, k(\mathbf{t},\cdot)\rangle_{\mathscr{H}}\delta_{\mathbf{x}}(\mathbf{t})\mathrm{d}t \\ &= \left\langle f, \int k(\mathbf{t},\cdot)\delta_{\mathbf{x}}(\mathbf{t})\mathrm{d}t \right\rangle_{\mathscr{H}} = \langle f, k(\mathbf{x},\cdot)\rangle_{\mathscr{H}}.\end{aligned} \tag{12}$$

From Data Point to Probability Measure

We may view Dirac measure $\delta_{\mathbf{x}}$ in the following perspectives

- It may be viewed as a representer of evaluation of the following functional

$$f \longmapsto \int f(\mathbf{t})\delta_{\mathbf{x}}(\mathbf{t})\mathrm{d}t, \tag{13}$$

  namely, the expectation of f w.r.t. the Dirac measure $\delta_{\mathbf{x}}$.

- It may also be viewed as a representer of the measure $\delta_{\mathbf{x}}$ in the Hilbert space

$$\delta_{\mathbf{x}} \longmapsto k(\mathbf{x}, \cdot). \tag{14}$$

From Data Point to Probability Measure

We extend our idea to more general cases

- If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are n distinct points in $\mathcal{X}$ and $a_1, \ldots, a_n$ are n non-zero real numbers, we consider a linear combination $\mu = \sum_{i=1}^n a_i \delta_{\mathbf{x}_i}$.
- For any measurable functions f, we have

$$\int f \left(\sum_{i=1}^n a_i \delta_{\mathbf{x}_i} \right) = \sum_{i=1}^n a_i \int f \delta_{\mathbf{x}_i} = \sum_{i=1}^n a_i f\left(\mathbf{x}_i\right). \tag{15}$$

- We obtain similar results to the case of Dirac measure. That is, the mapping

$$\sum_{i=1}^n a_i \delta_{\mathbf{x}_i} \longmapsto \sum_{i=1}^n a_i k\left(\mathbf{x}_i, \cdot\right). \tag{16}$$

- Furthermore, for any $f \in \mathscr{H}$,

$$\int f \, \mathrm{d}\mu = \left\langle f, \sum_{i=1}^n a_i k\left(\mathbf{x}_i, \cdot\right) \right\rangle_{\mathscr{H}}. \tag{17}$$

From Data Point to Probability Measure

Furthermore, we may extend the embedding into any probability measure, as shown in [3]. In what follows, we use $M_+^1(\mathcal{X})$ to denote the space of probability measures over a measurable space $\mathcal{X}$.

Definition 8

The kernel mean embedding of probability measures in $M_+^1(\mathcal{X})$ into an RKHS $\mathscr{H}$ endowed with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by a mapping

$$\mu : M_+^1(\mathcal{X}) \longrightarrow \mathscr{H}, \quad \mathbb{P} \longmapsto \int k(\mathbf{x}, \cdot) \mathrm{d}\mathbb{P}(\mathbf{x}).$$

From Data Point to Probability Measure

The condition below guarantee the embedding $\mu_{\mathbb{P}}$ exists and belongs to $\mathscr{H}$.

Lemma 9

If $\mathbb{E}_{X\sim\mathbb{P}}[\sqrt{k(X,X)}] < \infty$, *then* $\mu_{\mathbb{P}} \in \mathscr{H}$ *and* $\mathbb{E}_{\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathscr{H}}$.

Proof.

Let $\mathbf{L}_{\mathbb{P}}$ be a linear operator defined as $\mathbf{L}_{\mathbb{P}} f := \mathbb{E}_{X\sim\mathbb{P}}[f(X)]$. We assume $\mathbf{L}_{\mathbb{P}}$ is bounded for all $f \in \mathscr{H}$, i.e.,

$$\begin{aligned} |\mathbf{L}_{\mathbb{P}} f| = |\mathbb{E}_{X\sim\mathbb{P}}[f(X)]| &\leq \mathbb{E}_{X\sim\mathbb{P}}[|f(X)|] \\ &= \mathbb{E}_{X\sim\mathbb{P}}\left[|\langle f, k(X,\cdot)\rangle_{\mathscr{H}}|\right] \\ &\leq \mathbb{E}_{X\sim\mathbb{P}}\left[\sqrt{k(X,X)}\|f\|_{\mathscr{H}}\right]. \end{aligned}$$

Hence, by the Riesz representation theorem, there exists $h \in \mathscr{H}$ such that $\mathbf{L}_{\mathbb{P}} f = \langle f, h \rangle_{\mathscr{H}}$, which means $h = \mu_{\mathbb{P}}$. □

Example 10 (Inhomogeneous polynomial kernel)

Consider the inhomogeneous polynomial kernel

$$k(\mathbf{x}, \mathbf{x}) = \left(\left\langle \mathbf{x}, \mathbf{x}' \right\rangle + 1\right)^p, \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$$

of degree p. Using

$$\begin{aligned}
(\langle \mathbf{x}, \mathbf{y} \rangle + 1)^p =& 1 + \binom{p}{1} \langle \mathbf{x}, \mathbf{y} \rangle + \binom{p}{2} \langle \mathbf{x}, \mathbf{y} \rangle^2 + \binom{p}{3} \langle \mathbf{x}, \mathbf{y} \rangle^3 + \cdots \\
=& 1 + \binom{p}{1} \langle \mathbf{x}, \mathbf{y} \rangle + \binom{p}{2} \left\langle \mathbf{x}^{(2)}, \mathbf{y}^{(2)} \right\rangle \\
&+ \binom{p}{3} \left\langle \mathbf{x}^{(3)}, \mathbf{y}^{(3)} \right\rangle + \cdots
\end{aligned}$$

where $\mathbf{x}^{(i)} := \bigotimes_{k=1}^{i} \mathbf{x}$ denotes the i th-order tensor product.

Example 11 (continued)

The kernel mean embedding can be explicitly written as

$$\begin{aligned}\mu_{\mathbb{P}}(\mathbf{t}) = & \int (\langle \mathbf{x}, \mathbf{t} \rangle + 1)^p \, \mathrm{d}\mathbb{P}(\mathbf{x}) \\ = & 1 + \begin{pmatrix} p \\ 1 \end{pmatrix} \langle \mathbf{m}_{\mathbb{P}}(1), \mathbf{t} \rangle + \begin{pmatrix} p \\ 2 \end{pmatrix} \left\langle \mathbf{m}_{\mathbb{P}}(2), \mathbf{t}^{(2)} \right\rangle \\ & + \begin{pmatrix} p \\ 3 \end{pmatrix} \left\langle \mathbf{m}_{\mathbb{P}}(3), \mathbf{t}^{(3)} \right\rangle + \cdots\end{aligned}$$

where $\mathbf{m}_{\mathbb{P}}(i)$ denotes the i th moment of the distribution $\mathbb{P}$.

Application: Maximum Mean Discrepancy

In this section, we introduce Maximum Mean Discrepancy, an effective way of analyzing and comparing distribution proposed in [4], which is also an application of embeddings of probability measure.

Problem 12

Let x and y be random variables defined on a topological space $\mathcal{X}$, with respective Borel probability measures $\mathbb{P}$ and $\mathbb{Q}$. Given observations $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_n\}$, independently and identically distributed (i.i.d.) from $\mathbb{P}$ and $\mathbb{Q}$, respectively, can we decide whether $\mathbb{P} \neq \mathbb{Q}$?

Application: Maximum Mean Discrepancy

- When there's no ambiguity, we write $\mathbb{E}_x[f(x)] = \mathbb{E}_{x\sim\mathbb{P}}[f(x)]$ and $\mathbb{E}_y[f(y)] = \mathbb{E}_{y\sim\mathbb{Q}}[f(y)]$.
- We wish to determine a criterion that, in the population setting, takes on a unique and distinctive value only when $\mathbb{P} = \mathbb{Q}$.
- The Lemma below shows that $C_b(\mathcal{X})$ is in principle allows us to identify $\mathbb{P} = \mathbb{Q}$ uniquely, but is not practical to work with such a rich function class in the finite sample setting.

Lemma 13

Let $(\mathcal{X}, d)$ be a metric space, and let p, q be two Borel probability measures defined on $\mathcal{X}$. Then $\mathbb{P} = \mathbb{Q}$ if and only if $\mathbf{E}_x(f(x)) = \mathbf{E}_y(f(y))$ for all $f \in C_b(\mathcal{X})$, where $C_b(\mathcal{X})$ is the space of bounded continuous functions on $\mathcal{X}$.

Application: Maximum Mean Discrepancy

Definition 14 (Maximum Mean Discrepancy)

Let $\mathcal{F}$ be a class of functions $f : X \to \mathbb{R}$ and let $\mathbb{P}, \mathbb{Q}, x, y, X, Y$ be defined as above. We define the maximum mean discrepancy(MMD) as

$$\mathrm{MMD}[\mathcal{F}, \mathbb{P}, \mathbb{Q}] := \sup_{f \in \mathcal{F}} \left(\mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)] \right). \tag{18}$$

A unbiased empirical estimate of the MMD is obtained by replacing the population expectations with empirical expectations computed on the samples X and Y,

$$\mathrm{MMD}_b[\mathcal{F}, X, Y] := \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(y_i) \right). \tag{19}$$

Application: Maximum Mean Discrepancy

Here, we let $\mathcal{F}$ being the unit ball of RKHS $\mathscr{H}$.

Lemma 15

Assume the conditions in Lemma 9 holds, then

$$\mathrm{MMD}^2[\mathcal{F}, \mathbb{P}, \mathbb{Q}] = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}}^2$$

Proof.

$$\begin{aligned}\mathrm{MMD}^2[\mathcal{F}, \mathbb{P}, \mathbb{Q}] &= \left[\sup_{\|f\|_{\mathscr{H}} \le 1} \left(\mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)]\right)\right]^2 \\ &= \left[\sup_{\|f\|_{\mathscr{H}} \le 1} \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, f \rangle_{\mathscr{H}}\right]^2 = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}}^2 .\end{aligned}$$

Application: Maximum Mean Discrepancy

Follows from Lemma 15, for any universal RKHS $\mathscr{H}$, we have

Theorem 16

Let $\mathcal{F}$ be a unit ball in a universal RKHS $\mathscr{H}$, defined on the compact metric space $\mathcal{X}$, with associated continuous kernel $k(\cdot,\cdot)$. Then $\mathrm{MMD}[\mathcal{F},\mathbb{P},\mathbb{Q}]=0$ if and only if $\mathbb{P}=\mathbb{Q}$.

Application: Maximum Mean Discrepancy

- We can also express the MMD in terms of the associated kernel function k, here x' is an independent copy of x with the same distribution, and y' is an independent copy of y.

$$\mathrm{MMD}^2[\mathcal{F}, \mathbb{P}, \mathbb{Q}] = \mathbb{E}_{x,x'}\left[k\left(x, x'\right)\right] - 2\mathbb{E}_{x,y}[k(x, y)] + \mathbb{E}_{y,y'}\left[k\left(y, y'\right)\right] \tag{20}$$

- Based on that, we can derive some finite sample estimations of $\mathrm{MMD}^2[\mathcal{F}, \mathbb{P}, \mathbb{Q}]$.
- An unbiased empirical estimate is a sum of two U-statistics and a sample average

$$\begin{aligned}\mathrm{MMD}_u^2[\mathcal{F}, X, Y] = &\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k\left(x_i, x_j\right) \\ &+ \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k\left(y_i, y_j\right) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k\left(x_i, y_j\right).\end{aligned} \tag{21}$$

Application: Maximum Mean Discrepancy

- When $m = n$, a slightly simpler empirical estimate may be used. Let $Z := (z_1, \ldots, z_m)$ be m i.i.d. random variables, where $z := (x, y) \sim \mathbb{P} \times \mathbb{Q}$ (i.e., x and y are independent). An unbiased estimate of MMD^2 is

$$\mathrm{MMD}_u^2[\mathcal{F}, X, Y] = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} h\left(z_i, z_j\right), \tag{22}$$

which is a one-sample U-statistics with

$$h(z_i, z_j) := k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i).$$

Application: Maximum Mean Discrepancy

- MMD_b (19) can also be rewritten in the form of kernel, but is a V-statistics and also biased.

$$\begin{aligned}\mathrm{MMD}_b^2[\mathcal{F}, X, Y] = &\frac{1}{m^2}\sum_{i,j=1}^{m} k\left(x_i, x_j\right) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k\left(x_i, y_j\right) \\ &+ \frac{1}{n^2}\sum_{i,j=1}^{n} k\left(y_i, y_j\right).\end{aligned} \tag{23}$$

Coupled Kernel Embedding

Problem: LR (low-resolution) face recognition.

Figure: Demonstration of the main problem for LR face recognition, where one key question to be answered is related to the suitable definition of the metric, “How to compute the similarity (distance)?

Coupled Kernel Embedding

- Main question: How to measure the similarity (distance) between LR face images $l_i \in \mathbb{R}^m$ and HR (high-resolution) images $h_j \in \mathbb{R}^M$?
- Traditional approach: Let $y = g_{sr}(x) \in \mathbb{R}^M$ denotes the HR estimate of the LR image $x \in \mathbb{R}^m$, then calculate the distance in HR space as

$$d_{ij} := d_{sr}(h_i, l_j) = \text{dist}(h_i, g_{sr}(l_j)).$$

- Drawbacks: Performance are limited because the target of SR (super resolution) may not be consistent with that of classification, and time-consuming SR algorithms are not suitable for real-time applications.

Coupled Kernel Embedding

- We expect to project the data points in the original individual input spaces onto different reproducible kernel Hilbert spaces by coupled nonlinear functions (f_{sr} and g_{sr}), and calculate the distance by

$$d_{ij} = \text{dist}(f_{sr}(h_i), g_{sr}(l_j)). \tag{24}$$

- Based on the idea, we introduce Coupled Kernel Embedding (CKE) algorithm in the following section, which is proposed in [5].

Coupled Kernel Embedding

- Let Φ be a nonlinear mapping, and the HR input data space $\mathbb{R}^M$ is mapped onto a (potentially much higher dimensional) feature vector in the feature space $\mathcal{V}$

$$\Phi : \mathbb{R}^M \to \mathcal{V}, \quad x \mapsto \Phi(x).$$

- In the same way, let Ψ be another nonlinear mapping corresponding to LR images, and the input data $\mathbb{R}^m$ space with the other mode $(\mathrm{m} \ll \mathrm{M})$ can be mapped onto the feature space $\mathcal{W}$

$$\Psi : \mathbb{R}^m \to \mathcal{W}, \quad x \mapsto \Psi(x).$$

Coupled Kernel Embedding

- We project the kernel images $\Phi\left(\mathbb{R}^M\right)$ and $\Psi\left(\mathbb{R}^m\right)$ onto a common lower dimensional embedding feature space $\mathbb{R}^d$ as $P_H^T\Phi\left(h_i\right)$ and $P_L^T\Psi\left(l_j\right)$, respectively, the criterion for discriminant features extraction is formulated as

$$J\left(P_L, P_H\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left\|P_L^T\Psi\left(l_i\right) - P_H^T\Phi\left(h_j\right)\right\|_2^2 W_{ij}. \tag{25}$$

- Here, P_H and P_L are optimal projection matrices to be computed, N is the number of training images, and W_{ij} is the affinity weight between neighbors h_i and h_j, which serve to preserve the local relationship between data points in the original input spaces.

Coupled Kernel Embedding

- Let

$$P = \begin{pmatrix} P_L \\ P_H \end{pmatrix}, Z = \begin{pmatrix} \Psi(L) & 0 \\ 0 & \Phi(H) \end{pmatrix}, G = \begin{pmatrix} D_L & -W \\ -W^T & D_H \end{pmatrix}$$

where D_L and D_H are diagonal matrices defined on the basis of the weight matrix W as $D_L(i,i) = \sum_j W(i,j)$ and $D_H(i,i) = \sum_i W(i,j)$. We obtain

$$J(P) \triangleq J\left(P_L, P_H\right) = \mathrm{Tr}\left(P^T Z G Z^T P\right). \tag{26}$$

- Furthermore, we write the projection matrix as the linear combinations of input features Z, i.e., $P = ZU$.
- Let $\mathcal{K}_H(i,j) = \langle \Phi\left(h_i\right), \Phi\left(h_j\right) \rangle, \mathcal{K}_L(i,j) = \langle \Psi\left(l_i\right), \Psi\left(l_j\right) \rangle$ and

$$\mathcal{Z} = \begin{pmatrix} \mathcal{K}_L & 0 \\ 0 & \mathcal{K}_H \end{pmatrix}.$$

Coupled Kernel Embedding

- We obtain

$$J(U) = \mathrm{Tr}\left(U^T \mathcal{Z} G \mathcal{Z}^T U\right) \tag{27}$$

and the optimal projection U^* of CKE could be obtained by solving the following optimization problem:

$$U^* = \arg\min_U \mathrm{Tr}\left(U^T \mathcal{Z} G \mathcal{Z}^T U\right) \text{ s.t. } U^T \mathcal{Z} \mathcal{Z}^T U = I. \tag{28}$$

- The constraint is added to achieve scaling invariance in the embedding subspace as the matching errors for these coordinates should be measured on the same scale.

Coupled Kernel Embedding

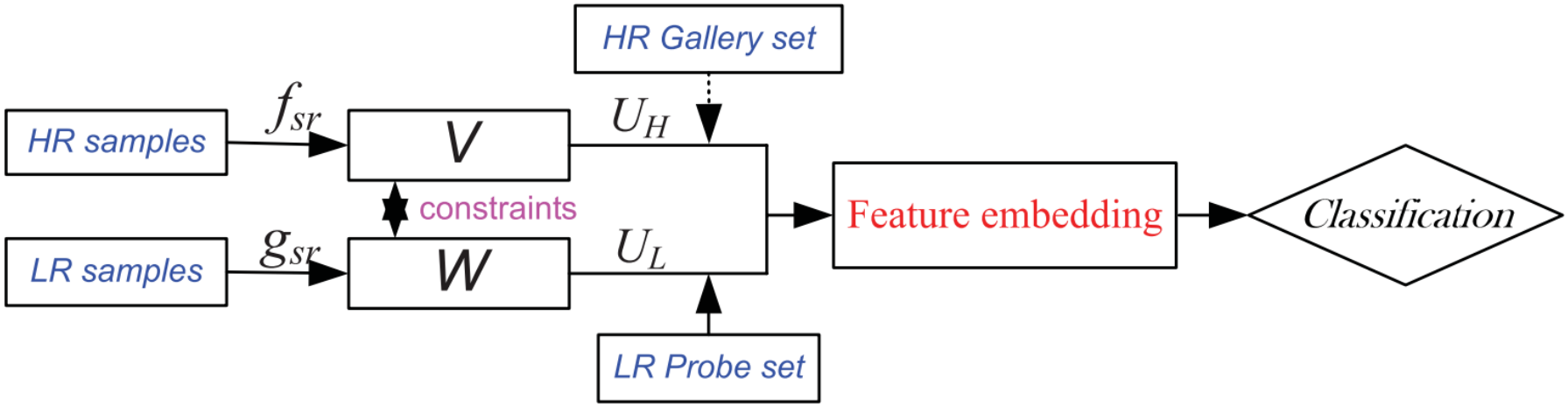


Figure: Flowchart of CKE algorithm applied to the double-modes classification problem, where the samples with different modes are first mapped onto different Hilbert spaces through nonlinear functions f_{sr} and g_{sr}, and then projected onto the learned subspace by linear transforms U_H and U_L, respectively. As a result, features-matching and classification can be efficiently implemented in this learned subspace.

References I

- M. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*.
  Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019.
- A. Rahimi, B. Recht, *et al.*, "Random features for large-scale kernel machines.," in *NIPS*, vol. 3, p. 5, Citeseer, 2007.
- K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," *arXiv preprint arXiv:1605.09522*, 2016.
- A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

References II

C.-X. Ren, D.-Q. Dai, and H. Yan, "Coupled kernel embedding for low-resolution face image recognition," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3770–3783, 2012.