Chapter 5
Kernel Methods I

Ernest K. Ryu
Seoul National University

Mathematical Machine Learning Theory
Spring 2024

"In mathematics, a kernel is an object to which the author assigns the name K." — Jan 6, 2022, Sam Power (@sp_monte_carlo)[1]

[1] https://twitter.com/sp_monte_carlo/status/1478783658714673159

Outline

Prologue: Linear learning with finite nonlinear features

Kernels

Reproducing kernel Hilbert space (RKHS)

Shift invariant kernels

Representer theorem and kernel trick

Linear learning with nonlinear features

Consider the setup with $\phi\colon \mathcal{X} \to \mathbb{R}^d$, where d may be smaller or larger than the "dimension" of $\mathcal{X}$. (We later consider infinite d.)

Consider

$$\underset{\theta}{\text{minimize}} \quad \mathop{\mathbb{E}}_{(X,Y)\sim P}[\ell(f_\theta(X), Y)],$$

where f_θ is a *linear*[2] *prediction function*

$$f_\theta(\cdot) = \langle \theta, \phi(\cdot) \rangle = \sum_{i=1}^{d} \theta_i \phi_i(\cdot)$$

and $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^d$.

Equivalently, consider the dataset

$$(\breve{X}_1, Y_1), \ldots, (\breve{X}_N, Y_N),$$

with $\breve{X}_i = \phi(X_i)$, and $f_\theta(X_i) = \langle \theta, \breve{X}_i \rangle$.

[2] Linear in the parameters θ, but nonlinear in the input X.

Absorbing bias into linear weights

What if we want a bias? So, what if we want to learn

$$f_{\theta,b}(\cdot) = \langle \theta, \phi(\cdot) \rangle + b.$$

Define

$$\tilde{\phi}(\cdot) = \begin{bmatrix} \phi(\cdot) \\ 1 \end{bmatrix} \in \mathbb{R}^{d+1}, \qquad \tilde{\theta} = \begin{bmatrix} \theta \\ b \end{bmatrix} \in \mathbb{R}^{d+1}$$

and note

$$\tilde{f}_{\tilde{\theta}}(\cdot) = \langle \tilde{\theta}, \tilde{\phi}(\cdot) \rangle = f_{\theta,b}(\cdot).$$

Trick: Absorb bias into linear weights.
WLOG, consider $f_\theta(\cdot) = \langle \theta, \phi(\cdot) \rangle$ without biases.

Kernel SGD

Training with linear f_θ:

$$\underset{\theta}{\text{minimize}} \quad \mathbb{E}_{(X,Y)\sim P}[\ell(\theta \cdot \phi(X), Y)].$$

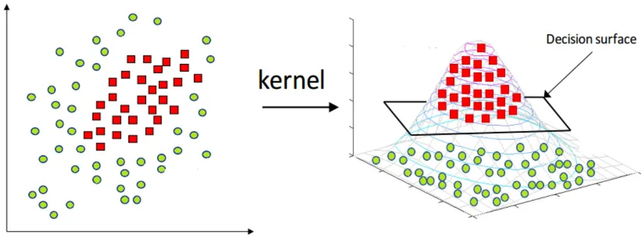
In ML and DL applications, $\ell(\cdot, y)$ is often convex in its first input (for fixed y). E.g., MSE and cross-entropy losses. Therefore, training is a *convex* optimization problem.

Therefore we can establish global convergence guarantees for SGD:

$$\begin{aligned} i(k) &\sim \text{Uniform}\{1, \ldots, N\} \\ \theta^{k+1} &= \theta^k - \alpha_k \ell'(\theta^k \cdot \phi(X_{i(k)}), Y_{i(k)})\phi(X_{i(k)}). \end{aligned}$$

Decision boundaries linear in ϕ, nonlinear in X

Linear classifiers yield decision boundaries that are linear *in the features*.



Most ML tasks are nonlinear in X, and features nonlinear in X are needed to perform classification well.

Feature map (ϕ) $\to$ Kernel (K) and RKHS ($\mathcal{H}$)

Consider $\phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_d(x))$ Assume $\phi_1, \ldots, \phi_d$ are linearly independent as functions. Consider $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$K(x', x) = \langle \phi(x), \phi(x') \rangle_{\mathbb{R}^d}.$$

Let

$$\mathcal{H} = \operatorname{span}\{\phi_k\}_{k=1}^d.$$

For any

$$f = \sum_{k=1}^d \alpha_k \phi_k \in \mathcal{H}, \qquad g = \sum_{k=1}^d \beta_k \phi_k \in \mathcal{H},$$

define the inner product

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{k=1}^d \alpha_k \beta_k.$$

Then, $\mathcal{H}$ is a finite-dimensional Hilbert space.

Reproducing property

Our $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ makes $\{\phi_1, \ldots, \phi_d\}$ an orthonormal basis of $\mathcal{H}$.
If $f(\cdot) = \sum_{k=1}^{d} \alpha_k \phi_k(\cdot)$, then

$$\langle f(\cdot), \phi_k(\cdot) \rangle_{\mathcal{H}} = \alpha_k$$

for $k = 1, \ldots, d$.

Note that

$$K(\cdot, x) = \sum_{k=1}^{d} \phi_k(x) \phi_k(\cdot) \in \mathcal{H}$$

for all $x \in \mathcal{X}$.

Reproducing property

K has the *reproducing property* with respect to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$:
If $f(\cdot) = \sum_{k=1}^{d} \alpha_k \phi_k(\cdot)$, then

$$\begin{aligned}
\langle f, K(\cdot,x)\rangle_{\mathcal{H}} &= \left\langle f(\cdot), \sum_{k=1}^{d} \phi_k(x)\phi_k(\cdot) \right\rangle_{\mathcal{H}} \\
&= \sum_{k=1}^{d} \phi_k(x) \left\langle f(\cdot), \phi_k(\cdot)\right\rangle_{\mathcal{H}} \\
&= \sum_{k=1}^{d} \alpha_k \phi_k(x) = f(x),
\end{aligned}$$

i.e., inner product with $K(\cdot,x)$ is evaluation at x. To put it differently yet,

$$\langle\cdot, K(\cdot,x)\rangle_{\mathcal{H}} \colon \mathcal{H} \to \mathbb{R}$$

is the point evaluation (linear) operator at point x.

We say K is a *reproducing kernel* of $\mathcal{H}$.

Example: Polynomial space

Let $\mathcal{X} = \mathbb{R}$. Let

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \\ \vdots \\ x^{d-1} \end{bmatrix}.$$

Then,

$$\mathcal{H} = \{\text{Polynomials of degree} < d\}$$

and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the $\mathbb{R}^d$-inner product of the monomial coefficients.

The kernel is

$$K(x', x) = \langle \phi(x), \phi(x') \rangle_{\mathbb{R}^d} = \sum_{i=1}^{d} (x)^{i-1} (x')^{i-1}.$$

Connection to 2-layer neural networks

Let $\mathcal{X} = \mathbb{R}^n$. Let $\phi_1, \dots, \phi_d$ be defined as

$$\phi_k(x) = \sigma(a_k^\intercal x + b_k)$$

for $k = 1, \dots, d$. Then

$$\mathcal{H} = \left\{ \sum_{k=1}^{d} u_k \sigma(a_k^\intercal x + b_k) \,|\, u_1, \dots, u_d \in \mathbb{R} \right\},$$

i.e., $\mathcal{H}$ is the set of 2-layer nerual networks with hidden layer weights and biases fixed to $a_1, \dots, a_d$ and $b_1, \dots, b_d$.

Performing kernel SGD corresponds to training the output layer weights of a 2-layer neural network with the hidden layer weights and biases fixed (and not trained).

Feature engineering

Feature engineering is the task of choosing (often hand-crafting) ϕ for a given ML task.

There was a time when ML was primarily about feature engineering.[3]
In modern deep learning, features are learned.

[3] One can argue that in modern machine learning *practice*, feature engineering is still the main engineering challenge.

Learning features with deep neural networks

Let $\theta = (\theta^{(1)}, \theta^{(2)})$ and let

$$f_\theta(x) = \langle \theta^{(1)}, \phi_{\theta^{(2)}}(x) \rangle.$$

In other words, f_θ is a deep neural network, $\theta^{(1)}$ is the trainable parameters for the output linear layer (FC1), and $\theta^{(2)}$ is the trainable parameters for the earlier layers.

A deep neural network uses a prediction function non-linear in its parameter θ. Most modern deep neural networks have this form.

However, if $\theta^{(2)}$ is fixed, then f_θ *is* linear in $\theta^{(1)}$. Deep learning can be interpreted as a process in which the feature mapping $\phi_{\theta^{(2)}}$ is learned along with its linear weights $\theta^{(1)}$.

Outline

Prologue: Linear learning with finite nonlinear features

Kernels

Reproducing kernel Hilbert space (RKHS)

Shift invariant kernels

Representer theorem and kernel trick

Kernel: Definition

Let $\mathcal{X}$ be a nonempty set. Let $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We say K is symmetric if $K(x', x) = K(x, x')$ for all $x, x' \in \mathcal{X}$. Given $N \in \mathbb{N}$ and $x_1, \dots, x_N \in \mathcal{X}$, let $G \in \mathbb{R}^{N \times N}$ be

$$G_{ij} = K(x_i, x_j), \qquad i, j \in \{1, \dots, N\}.$$

We call G the *kernel matrix* or the *Gramian matrix* of K. Then K is a *positive definite kernel* (PDK) if G is symmetric positive semidefinite for any $N \in \mathbb{N}$ and $x_1, \dots, x_N \in \mathcal{X}$. Equivalently, K is positive definite if it is symmetric and

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j K(x_i, x_j) \geq 0$$

for all $N \in \mathbb{N}, x_1, \dots, x_N \in \mathcal{X}$ and $c \in \mathbb{R}^N$.

We discuss the building blocks of PDKs. This machinery will allow us to construct PDKs and identify PDKs.

Strictly positive definite kernels

The inconsistent naming warrants some clarification. A matrix $G \in \mathbb{R}^{N \times N}$ is symmetric positive definite if all eigenvalues are strictly positive ($>$) and symmetric positive **semi**definite if all eigenvalues are nonnegative ($\geq$). In contrast, a **strictly** positive definite kernel, as defined below, refers to the strict notion ($>$) while positive definite kernels correspond to the non-strict notion ($\geq$).

We say $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a strictly positive definite kernel if for any $N \in \mathbb{N}$ and distinct $x_1, \ldots, x_N \in \mathcal{X}$, the corresponding Gramian matrix G is symmetric (strictly) positive definite. Equivalently, K is strictly positive definite if it is symmetric and

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j K(x_i, x_j) > 0$$

for all $N \in \mathbb{N}$, distinct $x_1, \ldots, x_N \in \mathcal{X}$, and nonzero $c \in \mathbb{R}^N$.

Inner products of feature maps

Let $\phi\colon \mathcal{X} \to \mathcal{H}$ for some Hilbert space $\mathcal{H}$ (not necessarily an RKHS) equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and induced norm $\| \cdot \|_{\mathcal{H}}$. Then, $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$K(x', x) = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$$

is a PDK, since, for all $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}$, and $c \in \mathbb{R}^N$,

$$\begin{aligned}
\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j K(x_i, x_j) &= \sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{N} c_i \phi(x_i), \sum_{j=1}^{N} c_j \phi(x_j) \right\rangle_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^{N} c_i \phi(x_i) \right\|_{\mathcal{H}}^2 \\
&\geq 0.
\end{aligned}$$

Example: Linear kernel

The simplest instance is $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \mathbb{R}^d$, $\phi(x) = x$, and

$$K(x', x) = \langle x, x' \rangle_{\mathbb{R}^d}.$$

Example: Tensor product

Let $f_1, \ldots, f_d$ be functions from $\mathcal{X}$ to $\mathbb{R}$. Then, $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$K(x', x) = \sum_{i=1}^{d} f_i(x) f_i(x')$$

is a PDK. Using the notation of tensor products, which we further discuss later, we can equivalently write

$$K = \sum_{i=1}^{d} f_i \otimes f_i.$$

(Analogous to expressing a matrix as a sum of d rank-1 outer products.)

Proof. The sum of d tensor products is an instance of a PDK defined through the feature map

$$\phi(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_d(x) \end{bmatrix} \in \mathbb{R}^d.$$

$\square$

Example: Min kernel

Let $\mathcal{X} = [0, \infty)$. Then, $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$K(x', x) = \min(x, x')$$

is a PDK.

Proof. For $L^2(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} \,|\, (\int |f(x)|^2 dx)^{1/2} < \infty\}$, let $\phi \colon \mathcal{X} \to L^2(\mathbb{R})$ be defined by $\phi(x) = \mathbf{1}_{[0,x]}$. Then

$$K(x', x) = \langle \phi(x), \phi(x') \rangle_{L^2(\mathbb{R})} = \langle \mathbf{1}_{[0,x]}, \mathbf{1}_{[0,x']} \rangle_{L^2(\mathbb{R})} = \min(x, x').$$

$\square$

Operations preserving PDKs

Given simple PDKs, we can construct more complex PDKs through operations preserving positive definiteness.

Let K_1 and K_2 be PDKs mapping $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. Then

- αK_1 for any $\alpha \geq 0$
- $K_1 + K_2$
- $K_1 K_2$

are PDKs. The first two claims are clear. The third claim means $K_3 \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$K_3(x', x) = K_1(x', x) K_2(x', x), \qquad \forall\, x, x \in \mathcal{X}$$

is PDK, and it follows from the Schur product theorem.

Schur product theorem

Theorem

Let $A \in \mathbb{R}^{N\times N}$ and $B \in \mathbb{R}^{N\times N}$ be symmetric positive semidefinite. Then the Hadamard product $C = A \odot B$, defined by $C_{ij} = A_{ij}B_{ij}$ for $i, j \in \{1, \ldots, N\}$, is symmetric positive semidefinite.

Proof. Let

$$A = \sum_{i=1}^{N} \lambda_i u_i u_i^\mathsf{T}, \qquad B = \sum_{i=1}^{N} \nu_i v_i v_i^\mathsf{T}$$

be the eigenvalue decompositions of A and B with respective orthonormal eigenvectors $u_1, \ldots, u_N$ and $v_1, \ldots, v_N$. Since $\odot$ is bilinear,

$$\begin{aligned} C = A \odot B &= \left(\sum_{i=1}^{N} \lambda_i u_i u_i^\mathsf{T}\right) \odot \left(\sum_{j=1}^{N} \nu_j v_j v_j^\mathsf{T}\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \nu_j \left(u_i u_i^\mathsf{T}\right) \odot \left(v_j v_j^\mathsf{T}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \nu_j (u_i \odot v_j)(u_i \odot v_j)^\mathsf{T} \end{aligned}$$

is a sum of N^2 (rank-0 or rank-1) symmetric positive semidefinite matrices and therefore is symmetric positive semidefinite. $\square$

Sums and integrals of PDKs

Let $\{K_i\}_{i \in \mathbb{N}}$ be a sequence of PDKs mapping $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. If

$$K_\infty(x', x) = \sum_{i=1}^{\infty} K_i(x', x)$$

finitely exists for all $x, x' \in \mathcal{X}$, then K_∞ is a PDK. Let $\{K_w\}_{w \in \mathcal{W}}$ be a family of PDKs mapping $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. Let μ be a nonnegative measure on $\mathcal{W}$. If

$$K(x', x) = \int_{\mathcal{W}} K_w(x', x) \, d\mu(w)$$

is well-defined (measurable and finitely integrable) for all $x, x' \in \mathcal{X}$, then K is a PDK.

Proof.

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j K(x_i, x_j) = \int_{\mathcal{W}} \sum_{i=1}^{N} \sum_{j=1}^{N} \underbrace{c_i c_j K_w(x_i, x_j)}_{\geq 0} \, d\mu(w) \geq 0.$$

□

Example: Polynomial kernel

Let $\mathcal{X} = \mathbb{R}^d$ and $p \in \mathbb{N}$. Then, $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined as

$$K(x', x) = (\langle x, x' \rangle + 1)^p$$

is a PDK.

Example: Exponential kernel

Let $\mathcal{X} = \mathbb{R}^d$. Then, $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined as

$$K(x', x) = \exp(\langle x, x' \rangle) = \sum_{p=0}^{\infty} \frac{1}{p!} (\langle x, x' \rangle)^p.$$

is a PDK.

Example: Cosine kernel

Let $\mathcal{X} = \mathbb{R}$. Then, $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$K(x', x) = \cos(x - x') = \cos(x)\cos(x') + \sin(x)\sin(x')$$

is a PDK.

Example: Kernels with integers

Let $\mathcal{X} = \mathbb{N}$. Then, $K\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ defined as

$$K(x', x) = 2^{xx'} = e^{(\log 2)xx'}$$

is PDK.

(The point is that the theory of kernels are applicable to non-vector data types $\mathcal{X}$. There are also kernels for strings of variable lengths for NLP applications.)

Outline

Prologue: Linear learning with finite nonlinear features

Kernels

Reproducing kernel Hilbert space (RKHS)

Shift invariant kernels

Representer theorem and kernel trick

Reproducing kernel Hilbert space (RKHS)

Let $\mathcal{X}$ be a nonempty set (No further assumption on $\mathcal{X}$ yet). Let $\mathcal{H}$ be a (real) Hilbert space of functions $f\colon \mathcal{X} \to \mathbb{R}$ equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and induced norm $\| \cdot \|_{\mathcal{H}}$. By definition, $\|f\|_{\mathcal{H}} = 0$ if and only if $f(x) = 0$ for all $x \in \mathcal{X}$.[4]

$K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel (RK) of $\mathcal{H}$ if

$$K(x, \cdot) \in \mathcal{H}, \qquad \forall\, x \in \mathcal{H},$$

and K has the *reproducing property*

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

If $\mathcal{H}$ has an RK, it is a *reproducing kernel Hilbert space* (RKHS).

[4] Clarification on next slide.

RKHS vs. L^p spaces

To clarify, $\mathcal{H}$ is a space *functions*, not a space of equivalence classes of functions. Therefore, the point evaluation $f(x)$ is well defined for any $x \in \mathcal{X}$.

If $f \in L^2$, then f is not a single function, but rather a set of functions that differ only on a set of measure 0, and the point evaluation $f(x)$ is undefined. ($\int_{B(x,\varepsilon)} f(x)\,dx$ is well defined, but $f(x)$ is undefined.)

RKHS vs. L^p spaces

RKHS function spaces (rather than L^p spaces) are how people think about functions in machine learning theory.

The output of an ML algorithm is a prediction function $\hat{f}$ (and we use $\hat{f}$ for point evaluations, not integrals). RKHS is the class of Hilbert spaces on which point evaluation is continuous.

Therefore, the requirements of RKHSs that the evaluation functional is continuous is a natural requirement, provided that you insist on working with Hilbert spaces. (Some recent research tries to understand deep learning as finding $\hat{f}$ within Banach spaces.)

Example: Band-limited L^2 functions

Let $B > 0$ and $\mathcal{X} = \mathbb{R}$. Let

$$\mathcal{H} = \left\{ f : \mathbb{R} \to \mathbb{R} \,\middle|\, \int_{\mathbb{R}\setminus[-B,B]} |\hat{f}(\omega)|^2 \, d\omega = 0, \ \|f\|_{\mathcal{H}} < \infty \right\}$$

$$\langle f, g \rangle_{\mathcal{H}} = \int_{\mathbb{R}} f(x)g(x) \, dx = \frac{1}{2\pi} \int_{-B}^{B} \hat{f}(\omega)\overline{\hat{g}}(\omega) \, d\omega$$

be the Hilbert space of band-limited L^2 functions.

Then, $\mathcal{H}$ is an RKHS with RK

$$K(x', x) = 2B\mathrm{sinc}(B(x - x')) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\omega x'} e^{i\omega x} \mathbf{1}_{[-B,B]}(\omega) \, d\omega.$$

To see why, note that $\widehat{K(x,\cdot)}(\omega) = e^{i\omega x}\mathbf{1}_{[-B,B]}(\omega)$, so $K(x,\cdot) \in \mathcal{H}$ for all $x \in \mathbb{R}$, and

$$\langle f, K(x,\cdot) \rangle_{\mathcal{H}} = \frac{1}{2\pi} \int_{-B}^{B} \hat{f}(\omega) e^{-i\omega x} \, d\omega = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\omega) e^{-i\omega x} \, d\omega = f(x),$$

so K has the reproducing property.

Continuity of point evaluation

RKHSs can be equivalently defined by continuity of point evaluation.

Theorem

Let $\mathcal{X}$ be a nonempty set. Let $\mathcal{H}$ be a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$. $\mathcal{H}$ is an RKHS if and only if the evaluation functional L_x, defined as $L_x[f] = f(x)$, is bounded (continuous) for all $x \in \mathcal{X}$.

Proof. Assume $\mathcal{H}$ is an RKHS. For any $x \in \mathcal{X}$,

$$|L_x[f]| = |\langle f, K(x, \cdot)\rangle_{\mathcal{H}}| \le \|f\|_{\mathcal{H}} \|K(x, \cdot)\|_{\mathcal{H}}, \qquad \forall\, f \in \mathcal{H}$$

and $\|K(x, \cdot)\|_{\mathcal{H}} < \infty$ since $K(x, \cdot) \in \mathcal{H}$. So L_x is bounded.[5]

Next, assume $L_x \colon \mathcal{H} \to \mathbb{R}$ is bounded in $\mathcal{H}$. By the Riesz representation theorem, there exists a $h_x \in \mathcal{H}$ such that

$$L_x[f] = \langle h_x, f\rangle_{\mathcal{H}}, \qquad \forall\, f \in \mathcal{H}.$$

Let $K(x', x) = h_x(x')$ for all $x, x' \in \mathcal{X}$. □

[5] "Bounded" means bounded/continuous linear operator.

Kernel (K) $\Leftrightarrow$ RKHS ($\mathcal{H}$)

There is a one-to-one correspondence between PDKs and RKHS.

First, establish uniqueness: If a $\mathcal{H}$ exists for a K, then it is unique.
and if a K exists for a $\mathcal{H}$, then it is unique.

Theorem

If $\mathcal{H}$ is an RKHS, its reproducing kernel $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is unique.

Proof. Let K and K' be two RK of an RKHS $\mathcal{H}$. Then for any $x \in \mathcal{X}$,

$$\begin{aligned}
&\|K(x,\cdot) - K'(x,\cdot)\|_{\mathcal{H}}^2 \\
&\qquad = \langle K(x,\cdot) - K'(x,\cdot), K(x,\cdot) - K'(x,\cdot)\rangle_{\mathcal{H}} \\
&\qquad = \langle K(x,\cdot), K(x,\cdot) - K'(x,\cdot)\rangle_{\mathcal{H}} - \langle K'(x,\cdot), K(x,\cdot) - K'(x,\cdot)\rangle_{\mathcal{H}} \\
&\qquad = K(x,x) - K'(x,x) - K(x,x) + K'(x,x) \\
&\qquad = 0.
\end{aligned}$$

Therefore, $K = K'$. $\square$

Kernel (K) $\Leftrightarrow$ RKHS ($\mathcal{H}$)

First, establish uniqueness: If a K exists for a $\mathcal{H}$, then it is unique.

Theorem

If $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel, its Hilbert space $\mathcal{H}$ is unique.

Proof. Let $\mathcal{H}$ be an RKHS of a reproducing kernel K. Since $K(x,\cdot) \in \mathcal{H}$ for all $x \in \mathcal{X}$, we have

$$\mathcal{S} = \operatorname{span}\{K(x,\cdot)\,|\,x \in \mathcal{X}\} \subseteq \mathcal{H}$$

and $\overline{\mathcal{S}} \subseteq \mathcal{H}$. We claim $\overline{\mathcal{S}} = \mathcal{H}$, which holds if and only if 0 is the only element in $\mathcal{H}$ orthogonal to all vectors in $\mathcal{S}$. Indeed, if $h \in \mathcal{H}$ satisfies

$$\langle h, K(x,\cdot)\rangle = 0, \qquad \forall\, x \in \mathcal{X},$$

then $h(x) = 0$ for all $x \in \mathcal{X}$, by the reproducing property, and $h = 0$. Since, any RKHS of K is precisely characterized by $\overline{\mathcal{S}} = \mathcal{H}$, it is unique. $\square$

Kernel (K) $\Leftrightarrow$ RKHS ($\mathcal{H}$)

We now complete the proof of the one-to-one correspondence by showing existence: There exists a $\mathcal{H}$ exists for a K; and there exists a K for a $\mathcal{H}$.

Theorem (Moore–Aronszajn Theorem)

Let $\mathcal{X}$ be a nonempty set. Then $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a PDK if and only if it is an RK of an RKHS $\mathcal{H}$.

Proof. ($\Leftarrow$) Assume K is an RK of an RKHS $\mathcal{H}$. Then K is symmetric, since $K(x', x) = \langle K(x, \cdot), K(x', \cdot)\rangle_{\mathcal{H}} = \langle K(x', \cdot), K(x, \cdot)\rangle_{\mathcal{H}} = K(x, x')$ for all $x, x' \in \mathcal{X}$. Moreover, for any $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}$, and $c \in \mathbb{R}^N$, we have

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j K(x_i, x_j) &= \sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j \langle K(x_i, \cdot), K(x_j, \cdot)\rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{N} c_i K(x_i, \cdot), \sum_{j=1}^{N} c_j K(x_j, \cdot) \right\rangle_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^{N} c_i K(x_i, \cdot) \right\|_{\mathcal{H}}^2 \\
&\geq 0.
\end{aligned}$$

So K is a PDK.

($\Rightarrow$) Let $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDK. Define $\mathcal{H}_0$ to be the (not necessarily complete) vector space

$$\begin{aligned}\mathcal{H}_0 &= \operatorname{span}\{K(x,\cdot) \,|\, x \in \mathcal{X}\} \\ &= \left\{ \sum_{i=1}^{N} \alpha_i K(x_i,\cdot) \,\middle|\, N \in \mathbb{N},\, x_1,\ldots,x_N \in \mathcal{X},\, \alpha_1,\ldots,\alpha_N \in \mathbb{R} \right\}.\end{aligned}$$

For

$$f(\cdot) = \sum_{i=1}^{N} \alpha_i K(x_i,\cdot), \qquad g(\cdot) = \sum_{i=1}^{N'} \beta_i K(x_i',\cdot),$$

define

$$\begin{aligned}\langle f,g\rangle_{\mathcal{H}_0} &= \sum_{i=1}^{N}\sum_{j=1}^{N'} \alpha_i\beta_j K(x_i,x_j') \\ &= \sum_{i=1}^{N} \alpha_i \underbrace{\sum_{j=1}^{N'} \beta_j K(x_i,x_j')}_{=g(x_i)} = \sum_{j=1}^{N'} \beta_j \underbrace{\sum_{i=1}^{N} \alpha_i K(x_i,x_j')}_{=f(x_j')}\end{aligned}$$

Clearly, $\langle\cdot,\cdot\rangle_{\mathcal{H}_0}\colon \mathcal{H}_0 \times \mathcal{H}_0 \to \mathbb{R}$ is symmetric and bilinear. The value of $\langle\cdot,\cdot\rangle_{\mathcal{H}_0}$ is independent of the representation of f via $x_1,\dots,x_N,\alpha_1,\dots,\alpha_N$ and g via $x'_1,\dots,x'_{N'},\beta_1,\dots,\beta_{N'}$, since

$$\langle f,g\rangle_{\mathcal{H}_0} = \sum_{i=1}^{N} \alpha_i g(x_i) = \sum_{j=1}^{N'} \beta_j f(x'_j).$$

(So $\langle\cdot,\cdot\rangle_{\mathcal{H}_0}$ is well-defined.) Since K is a PDK, we have $\langle f,f\rangle_{\mathcal{H}_0} = \alpha^\intercal G\alpha \geq 0$, where $\alpha = (\alpha_1,\dots,\alpha_N)$ and $G \in \mathbb{R}^{N\times N}$ is the kernel matrix for $x_1,\dots,x_N$. So $\langle\cdot,\cdot\rangle_{\mathcal{H}_0}$ is a semi-inner product (it is an inner product, but we have so far shown that it is a semi-inner product.) so Cauchy–Schwartz inequality holds. We have the reproducing property

$$\langle f,K(x,\cdot)\rangle_{\mathcal{H}_0} = \sum_{i=1}^{N} \alpha_i K(x_i,x) = f(x), \qquad \forall x\in\mathcal{X},\ f\in\mathcal{H}_0.$$

Therefore,

$$|f(x)| \leq |\langle f,K(x,\cdot)\rangle_{\mathcal{H}_0}| \leq \|f\|_{\mathcal{H}_0}\|K(x,\cdot)\|_{\mathcal{H}_0} \leq \|f\|_{\mathcal{H}_0}\sqrt{K(x,x)},$$

and $\|f\|_{\mathcal{H}_0} = 0$ implies $f(x) = 0$ for all $x\in\mathcal{X}$, i.e., $f = 0$. Therefore, $\mathcal{H}_0$ is a pre-Hilbert space (a vector space equipped with an inner product).

Pointwise convergence and definition of $\mathcal{H}$

We complete $\mathcal{H}_0$ to get $\mathcal{H}$ by considering Cauchy sequences in $\mathcal{H}_0$.

Let $\{f_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ be a Cauchy sequence with respect to the $\|\cdot\|_{\mathcal{H}_0}$-norm. For any $x \in \mathcal{X}$,

$$\begin{aligned}
|f_m(x) - f_n(x)| &= |\langle f_m - f_n, K(x,\cdot)\rangle_{\mathcal{H}_0}| \\
&\leq \|f_m - f_n\|_{\mathcal{H}_0} \|K(x,\cdot)\|_{\mathcal{H}_0} \\
&= \|f_m - f_n\|_{\mathcal{H}_0} \sqrt{K(x,x)} \\
&\to 0
\end{aligned}$$

as $\min\{m,n\} \to \infty$. So, for all $x \in \mathcal{X}$, $\{f_k(x)\}_{k\in\mathbb{N}} \subset \mathbb{R}$ is a Cauchy sequence and converges to a limit. We define $f_\infty \colon \mathcal{X} \to \mathbb{R}$ to be the pointwise limit of $\{f_k\}_{k\in\mathbb{N}}$, i.e.,

$$f_\infty(x) = \lim_{k\to\infty} f_k(x).$$

We define $\mathcal{H}$ as the space of all pointwise limits of Cauchy sequences in $\mathcal{H}_0$. Clearly, $\mathcal{H}$ is a vector space. Moreover, $\mathcal{H}_0 \subseteq \mathcal{H}$, since for any $f \in \mathcal{H}_0$, the Cauchy sequence $f_k = f$ for all k has the limit f.

Definition of $\langle \cdot, \cdot \rangle_{\mathcal{H}}$

Let $f_\infty, g_\infty \in \mathcal{H}$ with Cauchy sequences $\{f_k\}_{k \in \mathbb{N}} \subset \mathcal{H}_0$ and $\{g_k\}_{k \in \mathbb{N}} \subset \mathcal{H}_0$ respectively converging to them. Define

$$\langle f_\infty, g_\infty \rangle_{\mathcal{H}} = \lim_{k \to \infty} \langle f_k, g_k \rangle_{\mathcal{H}_0}.$$

For $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ to be well defined, the limit must exist and the limit must not depend on the Cauchy sequence converging to $f_\infty, g_\infty \in \mathcal{H}$. First,

$$\begin{aligned} |\langle f_m, g_m \rangle_{\mathcal{H}_0} - \langle f_n, g_n \rangle_{\mathcal{H}_0}| &= |\langle f_m - f_n, g_m \rangle_{\mathcal{H}_0} - \langle f_n, g_n - g_m \rangle_{\mathcal{H}_0}| \\ &\leq |\langle f_m - f_n, g_m \rangle_{\mathcal{H}_0}| + |\langle f_n, g_n - g_m \rangle_{\mathcal{H}_0}| \\ &\leq \|f_m - f_n\|_{\mathcal{H}_0} \|g_m\|_{\mathcal{H}_0} + \|f_n\|_{\mathcal{H}_0} \|g_n - g_m\|_{\mathcal{H}_0} \to 0 \end{aligned}$$

as $\min\{m, n\} \to \infty$. (Note $\{f_k\}_{k \in \mathbb{N}} \subset \mathcal{H}_0$ and $\{g_k\}_{k \in \mathbb{N}} \subset \mathcal{H}_0$ are bounded since Cauchy.) Next, let $\{f'_k\}_{k \in \mathbb{N}} \subset \mathcal{H}_0$ and $\{g'_k\}_{k \in \mathbb{N}} \subset \mathcal{H}_0$ also be Cauchy sequences respectively converging to f_∞ and g_∞. Then

$$\begin{aligned} |\langle f_n, g_n \rangle_{\mathcal{H}_0} - \langle f'_n, g'_n \rangle_{\mathcal{H}_0}| &= |\langle f_n - f'_n, g_n \rangle_{\mathcal{H}_0} - \langle f'_n, g'_n - g_n \rangle_{\mathcal{H}_0}| \\ &\leq |\langle f_n - f'_n, g_n \rangle_{\mathcal{H}_0}| + |\langle f'_n, g'_n - g_n \rangle_{\mathcal{H}_0}| \\ &\leq \|f_n - f'_n\|_{\mathcal{H}_0} \|g_n\|_{\mathcal{H}_0} + \|f'_n\|_{\mathcal{H}_0} \|g'_n - g_n\|_{\mathcal{H}_0} \to 0. \end{aligned}$$

$\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is an inner product

That $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is symmetric and bilinear is clear. Also, $\|\cdot\|_{\mathcal{H}}$ is nonnegative, since

$$\|f_\infty\|_{\mathcal{H}} = \lim_{k\to\infty} \|f_k\|_{\mathcal{H}_0} \geq 0$$

for $\{f_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ converging to f_∞. For $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ to be an inner product on $\mathcal{H}$, it remains to verify positive definiteness of $\|\cdot\|_{\mathcal{H}}$, i.e., that $\|f_\infty\|_{\mathcal{H}} = 0$ only if and only if $f_\infty(x) = 0$ for all $x \in \mathcal{X}$.
If $f_\infty = 0$, then $\|f_\infty\|_{\mathcal{H}} = 0$ since $0 \in \mathcal{H}_0$ and $\{f_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ with $f_k = 0$ converges to $f_\infty = 0$. Conversely, assume $\{f_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ converges to f_∞ and $\|f_\infty\|_{\mathcal{H}} = 0$. Then, for any $x \in \mathcal{X}$,

$$|f_\infty(x)| = \left|\lim_{k\to\infty} f_k(x)\right| = \left|\lim_{k\to\infty} \langle f_k, K(x,\cdot)\rangle_{\mathcal{H}_0}\right| \leq \lim_{k\to\infty} \|f_k\|_{\mathcal{H}_0} \|K(x,\cdot)\|_{\mathcal{H}_0}.$$

Since $\|f_k\|_{\mathcal{H}_0} \to \|f\|_{\mathcal{H}} = 0$, we conclude $f_\infty(x) = 0$ for all $x \in \mathcal{X}$.

$\mathcal{H}$ is complete

While Cauchy sequences in $\mathcal{H}_0$ have limits $\mathcal{H}$ by definition, it remains to establish that Cauchy sequences in $\mathcal{H}$ have a limit in $\mathcal{H}$.

Let $f_\infty^{(1)}, f_\infty^{(2)}, \ldots$ be a Cauchy sequence in $\mathcal{H}$, and let $\{f_k^{(1)}\}_{k\in\mathbb{N}}, \{f_k^{(2)}\}_{k\in\mathbb{N}}, \ldots$ be Cauchy sequences in $\mathcal{H}_0$ with respective limits $f_\infty^{(1)}, f_\infty^{(2)}, \ldots$. Let $\{k(j)\}_{j\in\mathbb{N}} \subseteq \mathbb{N}$ be a sequence such that $\|f_{k(j)}^{(j)} - f_\infty^{(j)}\| \to 0$ as $j \to \infty$. Then

$$\begin{aligned}
\|f_{k(i)}^{(i)} - f_{k(j)}^{(j)}\|_{\mathcal{H}_0} &= \|f_{k(i)}^{(i)} - f_{k(j)}^{(j)}\|_{\mathcal{H}} \\
&\le \|f_{k(i)}^{(i)} - f_\infty^{(i)}\|_{\mathcal{H}} + \|f_\infty^{(i)} - f_\infty^{(j)}\|_{\mathcal{H}} + \|f_\infty^{(j)} - f_{k(j)}^{(j)}\|_{\mathcal{H}} \\
&\to 0
\end{aligned}$$

as $\min\{i, j\} \to \infty$. Therefore, $\{f_{k(j)}^{(j)}\}_{j\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{H}_0$ and it has a limit $\mathbf{f} \in \mathcal{H}$. Finally,

$$\|\mathbf{f} - f_\infty^{(j)}\|_{\mathcal{H}} \le \|\mathbf{f} - f_{k(j)}^{(j)}\|_{\mathcal{H}} + \|f_{k(j)}^{(j)} - f_\infty^{(j)}\|_{\mathcal{H}} \to 0$$

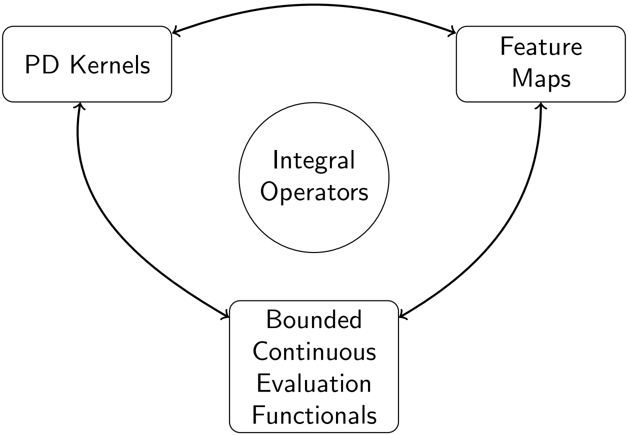
as $j \to \infty$. Since the Cauchy sequence $f_\infty^{(1)}, f_\infty^{(2)}, \ldots$ in $\mathcal{H}$ converges to a limit $\mathbf{f}$ in $\mathcal{H}$, we conclude $\mathcal{H}$ is complete.

K is an RK for $\mathcal{H}$

We have established that K has the reproducing property for $\mathcal{H}_0$ and that $K(x,\cdot) \in \mathcal{H}_0 \subseteq \mathcal{H}$ for all $x \in \mathcal{X}$. It remains to show that K has the reproducing property for all of $\mathcal{H}$. Let $f_\infty \in \mathcal{H}$ and let $\{f_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ be a Cauchy sequence converging to f_∞. Then

$$\underbrace{f_k(x)}_{\to f_\infty(x)} = \underbrace{\langle f_k, K(x,\cdot)\rangle_{\mathcal{H}_0}}_{\to \langle f_\infty, K(x,\cdot)\rangle_{\mathcal{H}}}.$$

□


PD Kernels
Feature Maps
Integral Operators
Bounded Continuous Evaluation Functionals

RKHS norm quantifies smoothness

The norm of a function in an RKHS controls how fast the function varies over $\mathcal{X}$ with respect to the (pseudo-)metric d_K, defined below.

Alternatively, one says, $\|f\|_{\mathcal{H}}$ quantifies the "smoothness" or "complexity" of f. In the context of machine learning and optimization, "smoothness" often refers to the variation of the function, and does not directly refer to (infinite) differentiability. Specifically, for $f \in \mathcal{H}$,

$$\begin{aligned} |f(x) - f(x')| &= |\langle f, K(x,\cdot) - K(x',\cdot)\rangle_{\mathcal{H}}| \\ &\leq \|f\|_{\mathcal{H}} \|K(x,\cdot) - K(x',\cdot)\|_{\mathcal{H}} \\ &= \|f\|_{\mathcal{H}} d_K(x', x), \end{aligned}$$

so f is $\|f\|_{\mathcal{H}}$-Lipschitz continuous as a map from $(\mathcal{X}, d_K)$ to $(\mathbb{R}, |\cdot|)$.

Outline

Prologue: Linear learning with finite nonlinear features

Kernels

Reproducing kernel Hilbert space (RKHS)

Shift invariant kernels

Representer theorem and kernel trick

Bochner's theorem

Let $\mathcal{X} = \mathbb{R}^d$. We say $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is *shift-invariant* if there exists a function $\kappa\colon \mathbb{R}^d \to \mathbb{R}$ such that

$$K(x', x) = \kappa(x - x').$$

Theorem (Bochner)

Let $x \in \mathcal{X} = \mathbb{R}^d$. Then, $K(x', x) = \kappa(x - x')$ is a PDK if and only if

$$\kappa(t) = \int_{\mathbb{R}^d} e^{-i\omega^\intercal t} d\mu(\omega)$$

for some (real) nonnegative finite measure $\mu \in \mathcal{M}_+(\mathbb{R}^d)$.

Proof of ($\Leftarrow$).

$$K(x', x) = \int_{\mathbb{R}^d} e^{-i\omega^\intercal (x-x')}\, d\mu(\omega) = \Re \int_{\mathbb{R}^d} e^{-i\omega^\intercal (x-x')}\, d\mu(\omega)$$

$$= \int_{\mathbb{R}^d} \cos(\omega^\intercal (x - x')) d\mu(\omega) = \int_{\mathbb{R}^d} (\cos(\omega^\intercal x)\cos(\omega^\intercal x') + \sin(\omega^\intercal x)\sin(\omega^\intercal x'))\, d\mu(\omega).$$

We omit ($\Rightarrow$) since it requires more work and we do not use it. $\square$

Example: Sinc kernel

Let $B > 0$ and $\mathcal{X} = \mathbb{R}$. Then, $K\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined as

$$K(x', x) = 2B\mathrm{sinc}(B(x - x')) = \begin{cases} \frac{2\sin(B(x-x'))}{(x-x')} & \text{if } x \neq x' \\ 0 & \text{if } x = x' \end{cases}$$

is a PDK, since

$$2B\mathrm{sinc}(B(t)) = \int_{\mathbb{R}} e^{-i\omega t} \mathbf{1}_{[-B,B]}(\omega)\, d\omega.$$

Example: 1-D Gaussian kernel

Let $\sigma > 0$ and $\mathcal{X} = \mathbb{R}$. Then, $K\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined as

$$K(x', x) = e^{-\frac{(x-x')^2}{2\sigma^2}}$$

is a PDK, since

$$K(x', x) = \underbrace{e^{\frac{xx'}{\sigma^2}}}_{\text{exponential kernel}} \underbrace{e^{-\frac{(x)^2}{2\sigma^2}} e^{-\frac{(x')^2}{2\sigma^2}}}_{\text{tensor product}}.$$

Alternatively, we can conclude K is PDK through

$$e^{-\frac{t^2}{2\sigma^2}} = \frac{\sigma}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-i\omega t} e^{-\frac{\sigma^2 \omega^2}{2}} \, d\omega.$$

Example: Gaussian kernel with covariance matrix

Let $\Sigma \in \mathbb{R}^{n\times n}$ be symmetric positive definite. Then

$$K(x', x) = \exp\Big(-\frac{(x-x')^\intercal \Sigma^{-1}(x-x')}{2}\Big)$$

is a PDK.

Justification in homework.

Example: 1-D Laplace kernel

Let $\gamma > 0$ and $\mathcal{X} = \mathbb{R}$. Then, $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined as

$$K(x', x) = \frac{1}{2} e^{-\gamma |x - x'|}$$

is a PDK, since

$$\frac{1}{2} e^{-\gamma |t|} = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\omega t} \frac{\gamma}{\gamma^2 + \omega^2} \, d\omega.$$

(Integral can be evaluated via contour integration.)

Example: Laplace kernel

Let

$$\kappa(x) = \exp(-\frac{\|x - x'\|_2}{r})$$

Then, one can show that

$$\hat{\kappa}(\omega) = 2^d \pi^{d-1} \Gamma((d+1)/2) \frac{r^d}{(1 + r^2\|\omega\|_2^2)^{(d+1)/2}}$$

Note that

$$\frac{1}{\hat{\kappa}(\omega)} \propto (1 + r^2\|\omega\|_2^2)^{(d+1)/2}$$

Explicit construction of norm for shift-invariant RKHS

We now informally derive the RKHS norm and inner product corresponding to shift-invariant PDKs.

Let

$$K(x', x) = \kappa(x - x')$$

be PDK. (So $\hat{\kappa}$ is nonnegative.) Assume $\kappa \in L^1$, which implies that $\hat{\kappa}$ exists. Then

$$K(x', x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \sqrt{\hat{\kappa}(\omega)} e^{-i\omega^\intercal x} \overline{\sqrt{\hat{\kappa}(\omega)} e^{-i\omega^\intercal x'}} \, d\omega = \int_{\mathbb{R}^d} \phi_\omega(x) \overline{\phi_\omega(x')} \, d\omega$$

Now we have an explicit feature map $\phi_\cdot(x) \in L^2$. Then,

$$f(x) = \langle \phi_\cdot(x), \theta(\cdot) \rangle_{L^2} = \int \theta(\omega) \phi_\omega(x) \, d\omega$$

means

$$\theta(\omega) = \frac{1}{(2\pi)^{d/2}} \frac{\hat{f}(\omega)}{\sqrt{\hat{\kappa}(\omega)}}.$$

So $\{\phi_\omega(\cdot)\}_{\omega \in \mathbb{R}^d}$ serves as a linearly independent basis of the x-space, and $\theta(\omega)$ serves as the coefficient for each $\phi_\omega(\cdot)$ when representing f.

Explicit construction of norm for shift-invariant RKHS

With analogous steps as in the finite-dimensional case, we expect

$$\|f\|_{\mathcal{H}}^2 = \|\theta\|_{L^2}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\kappa}(\omega)} \, d\omega$$

and

$$\mathcal{H} = \{f : \|f\|_{\mathcal{H}} < \infty\}, \qquad \langle f, g \rangle_{\mathcal{H}} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\hat{\kappa}(\omega)} \, d\omega.$$

To verify that this $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the RKHS with RK K, we need to check

- $\mathcal{H}$ is a Hilbert space.
- $K(x', \cdot) = \kappa(\cdot - x') \in \mathcal{H}$ for all $x \in \mathbb{R}^d$.
- $\langle f, K(x, \cdot) \rangle_{\mathcal{H}} = f(x)$ for all $x \in \mathbb{R}^d$.

The first point is an exercise in analysis, so we skip it.

Explicit construction of norm for shift-invariant RKHS

We now show $K(x', \cdot) = \kappa(\cdot - x') \in \mathcal{H}$ for all $x \in \mathbb{R}^d$. Since

$$\widehat{K(x', \cdot)}(\omega) = \hat{\kappa}(\omega) e^{i\omega^\intercal x'},$$

we have

$$\|K(x', \cdot)\|_{\mathcal{H}}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{\kappa}(\omega)\, d\omega = \kappa(0) = K(x, x) < \infty.$$

Next,

$$\langle f, K(\cdot, x) \rangle_{\mathcal{H}} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{f(\hat{\omega}) \hat{\kappa}(\omega) e^{-i\omega^\intercal x'}}{\hat{\kappa}(\omega)}\, d\omega = f(x).$$

So our guess is verified.

Norm of shift-invariant RKHS

Theorem

Let $\kappa\colon \mathbb{R}^d \to \mathbb{R}$ *be such that* $\kappa \in L^1$*. Let*

$$K(x', x) = \kappa(x - x')$$

be a PDK.[6] *Then, the RKHS* $\mathcal{H}$ *corresponding to* K *has inner product and norm*

$$\langle f, g\rangle_{\mathcal{H}} = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\hat{\kappa}(\omega)}\, d\omega, \qquad \|f\|_{\mathcal{H}}^2 = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\kappa}(\omega)}\, d\omega.$$

[6] So $\hat{\kappa} \geq 0$ exists as a function rather than as a measure.

Norm of shift-invariant RKHS

Algorithmically speaking, we will later see that we need to efficiently evaluate κ, not $\hat{\kappa}$. In particular, the formula

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\kappa}(\omega)} \, d\omega$$

will not be used in our computation.

However, the norm in the Fourier domain will allow us to think about the RKHS induced by K theoretically. Specifically, we will be able to identify the $\mathcal{H}$ with appropriate Sobolev spaces.

Example: Laplace kernel

Let

$$\kappa(x) = \exp(-\frac{\|x - x'\|_2}{r})$$

Then,

$$\frac{1}{\hat{\kappa}(\omega)} \propto (1 + r^2\|\omega\|_2^2)^{(d+1)/2}$$

and the RKHS norm is

$$\|f\|_{\mathcal{H}}^2 \propto \int_{\mathbb{R}^d} (1 + r^2\|\omega\|_2^2)^{(d+1)/2} |\hat{f}(\omega)|^2 \, d\omega.$$

For d odd, this is the $H^{(d+1)/2}$ Sobolev norm. (For even d, this still is the $H^{(d+1)/2}$ Sobolev norm if we appropriately define fractional derivatives.)

Note that the "bandwidth" r determines the relative amount we penalize the derivative. When r is large, kernel methods prefer smoother (smaller derivative) functions. When r is small, kernel methods are more favorable towards less smooth (larger derivatives) functions.

Example: 1-D Laplace kernel

When $d = 1$, then,

$$\hat{\kappa}(\omega) = \frac{2r}{1 + r^2\omega^2}$$

and

$$\begin{aligned}
\|f\|_{\mathcal{H}}^2 &= \frac{1}{2\pi}\int_{\mathbb{R}} \frac{|\hat{f}(\omega)|^2}{\hat{\kappa}(\omega)}\, d\omega = \frac{1}{2r}\frac{1}{2\pi}\int_{\mathbb{R}} |\hat{f}(\omega)|^2\, d\omega + \frac{r}{2}\frac{1}{2\pi}\int_{\mathbb{R}} |\omega\hat{f}(\omega)|^2\, d\omega \\
&= \frac{1}{2r}\int_{\mathbb{R}} |f(x)|^2\, dx + \frac{r}{2}\int_{\mathbb{R}} |f'(x)|^2\, dx \\
&= \frac{1}{2r}\|f\|_{L^2}^2 + \frac{r}{2}\|f'\|_{L^2}^2
\end{aligned}$$

Example: Gaussian kernel

Let $\sigma > 0$ and $\mathcal{X} = \mathbb{R}$. Then, $K\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined as

$$K(x', x) = e^{-\frac{(x-x')^2}{2\sigma^2}} = \kappa(x - x')$$

With some calculations, we get

$$\frac{1}{\hat{\kappa}(\omega)} = \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{s=0}^{\infty} \frac{\sigma^{2s}\omega^{2s}}{2^s s!}$$

and

$$\|f\|_{\mathcal{H}}^2 = \frac{\sigma}{\sqrt{2\pi}} \sum_{s=0}^{\infty} \frac{\sigma^{2s}}{2^s s!} \|f^{(2s)}\|_{L^2}^2$$

Intuitively speaking, $\|\cdot\|_{\mathcal{H}}$ penalizes all even derivatives. (One can show that elements of $\mathcal{H}$ are infinitely differentiable.)

Matérn kernel

Let κ be such that

$$\frac{1}{\hat{\kappa}(\omega)} = \frac{\Gamma(s-d/2)}{2^d \pi^{d/2} \Gamma(s) (2(s-d/2))^{s-d/2} r^d} (2(s-d/2) + r^2 \|\omega\|_2^2)^s$$

for $s > d/2$, where s is an integer. (If $s \leq d/2$, then $K(x,x') = \kappa(x-x')$ is invalid as $\hat{\kappa} \notin L^1$ and $K(x,x) = \infty$.) The Matérn kernel generalizes the Laplace kernel, which has $s = (d+1)/2$.

Then, we have

$$K(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\frac{\sqrt{2\nu}\|x-x'\|_2}{r} \Big)^\nu K_\nu \Big(\frac{\sqrt{2\nu}\|x-x'\|_2}{r} \Big),$$

where Γ is the gamma function and K_ν is the modified Bessel function, and $\nu + d/2 = s$.

Higher values of s imply the RKHS is more restricted; higher orders of differentiability are required, and the number of differentiability requirements grows with d. (This indicates a limitation of translation-invariant kernel methods. Since d is large, the RKHSs using Matérn kernels look for very smooth functions.)

Matérn kernel $K \Leftrightarrow$ Sobolev space $\mathcal{H}$

The RKHS corresponding to the Matérn kernel is the Sobolev space H^s.

If $s \leq d/2$, then the corresponding space

$$\{f \in L^2 \mid \int_{\mathbb{R}^d} (1 + r^2\|\omega\|_2^2)^s |\hat{f}(\omega)|^2 \, d\omega < \infty\}$$

This is a Hilbert space H^s, but it is not an RKHS.

Proof. If H^s with $s \leq d/2$ were an RKHS, it would have kernel K, which would have the reproducing property

$$\langle K(x,\cdot), f(\cdot)\rangle_{\mathcal{H}} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} (1+r^2\|\omega\|_2^2)^s \hat{f}(\omega) \frac{e^{-i\omega^\intercal x}}{(1 + r^2\|\omega\|_2^2)^s} \, d\omega = f(x).$$

This allows us to identify $\widehat{K(x,\cdot)}$, but this would lead to the conclusion that $K(x,x) = \infty$. $\square$

Outline

Prologue: Linear learning with finite nonlinear features

Kernels

Reproducing kernel Hilbert space (RKHS)

Shift invariant kernels

Representer theorem and kernel trick

Learning in RKHS

We now consider

$$\underset{f\in\mathcal{H}}{\text{minimize}} \quad \mathop{\mathbb{E}}_{(X,Y)\sim P}[\ell(f(X),Y)] + \lambda\|f\|_{\mathcal{H}}^2,$$

where $\mathcal{H}$ is an RKHS with kernel K and $\lambda \ge 0$. Since we don't have access to $\mathbb{E}_P$, we use N training datapoints $(X_1,Y_1),\dots,(X_N,Y_N) \sim P$ and solve

$$\underset{f\in\mathcal{H}}{\text{minimize}} \quad \frac{1}{N}\sum_{i=1}^{N}\ell(f(X_i),Y_i) + \lambda\|f\|_{\mathcal{H}}^2.$$

Infinite-dimensional problem if $\dim\mathcal{H} = \infty$. How to solve with finite computation?

Representer theorem

The representer theorem shows that the solution must lie in an N-dimensional subspace of $\mathcal{H}$.

Theorem

Let L be any function. Let $\mathcal{X}$ be a nonempty set, $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a PDK, $\mathcal{H}$ the corresponding RKHS, $X_1, \dots, X_N \in \mathcal{X}$, and $Y_1, \dots, Y_N \in \mathbb{R}$. Consider the optimization problem

$$\underset{f \in \mathcal{H}}{\text{minimize}} \quad L(\{(X_i, Y_i, f(X_i))\}_{i=1}^N) + Q(\|f\|_{\mathcal{H}}),$$

where $Q\colon \mathbb{R}_+ \to \mathbb{R}$ is a strictly increasing function. L is assumed to be any function (not necessarily convex). Then, if a minimizer exists, any minimizer must be in

$$\operatorname{span}(\{K(X_i, \cdot)\}_{i=1}^N).$$

Proof. Let

$$\mathcal{S} = \mathrm{span}(\{K(X_i, \cdot)\}_{i=1}^N) \subseteq \mathcal{H}.$$

In homework 3, you are to show that $f \in \mathcal{S}^\perp$ implies $f(X_i) = 0$ for all $i = 1, \dots, N$.

Let $f^\star$ be a minimizer. Let

$$f_\star = s + t$$

such that $s \in \mathcal{S}$ and $t \in \mathcal{S}^\perp$. Then

$$L(\{(X_i, Y_i, f^\star(X_i))\}_{i=1}^N) = L(\{(X_i, Y_i, s(X_i))\}_{i=1}^N)$$

while

$$Q(\|f^\star\|_\mathcal{H}) = Q\left(\sqrt{\|s\|_\mathcal{H}^2 + \|t\|_\mathcal{H}^2}\right) \geq Q(\|s\|_\mathcal{H}),$$

where equality holds if and only if $t = 0$. Since $f^\star$ is assumed to be a minimizer, we conclude $t = 0$. □

Therefore, to solve

$$\underset{f\in\mathcal{H}}{\text{minimize}} \quad \sum_{i=1}^{N} \ell(f(X_i), Y_i) + \lambda \|f\|_{\mathcal{H}}^2,$$

it is enough to search in

$$\operatorname{span}(\{K(X_i,\cdot)\}_{i=1}^N).$$

Therefore, parameterize the solution into the form

$$f = \sum_{k=1}^{N} \beta_k K(X_k,\cdot)$$

and then optimize over $\beta_1, \ldots, \beta_N$.

Kernel ridge regression

Quickly establish the following identity.

Lemma (Push-through identity)

Let $\gamma > 0$, $U \in \mathbb{R}^{m\times n}$, and $V \in \mathbb{R}^{n\times m}$. Then

$$(\gamma I + UV)^{-1}U = U(\gamma I + VU)^{-1},$$

assuming $(\gamma I + UV)$ is invertible.

Proof. Clearly,

$$U(\gamma I + VU) = (\gamma I + UV)U.$$

Left-multiply $(\gamma + UV)^{-1}$ and right-multiply $(\gamma + VU)^{-1}$. $\square$

Kernel ridge regression: Finite-dimensional feature map

Let $\mathcal{X}$ be a nonempty set. Let $X_1, \dots, X_N \in \mathcal{X}$, $Y_1, \dots, Y_N \in \mathbb{R}$, $\phi \colon \mathcal{X} \to \mathbb{R}^d$, and $\lambda > 0$. Let

$$\Phi = \begin{bmatrix} \phi(X_1)^\intercal \\ \phi(X_2)^\intercal \\ \vdots \\ \phi(X_N)^\intercal \end{bmatrix} \in \mathbb{R}^{N \times d}, \qquad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} \in \mathbb{R}^N.$$

Consider the *ridge regression*[7] problem

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{N} \sum_{i=1}^{N} (h_\theta(X_i) - Y_i)^2 + \lambda \|\theta\|^2,$$

where $h_\theta \colon \mathcal{X} \to \mathbb{R}$ is defined as $h_\theta(x) = \phi(x)^\intercal \theta$. Equivalently,[8] we write

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{N} \|\Phi\theta - Y\|^2 + \lambda \|\theta\|^2.$$

---

[7] Regression with ℓ^2-regularization is referred to as ridge regression in classical statistics.

[8] Linear regression is an instance of the finite-sum formulation and its goal is to obtain a prediction function $h_{\theta^\star}$ (which is linear in θ but need not be linear in x) rather than to obtain the parameters θ.

Because the objective function is convex, the solution $\theta^\star$ is found by setting the gradient to 0

$$0 = \frac{2}{N}\Phi^\intercal(\Phi\theta^\star - Y) + 2\lambda\theta^\star,$$

which solves to

$$\begin{aligned}\theta^\star &= \underbrace{(\Phi^\intercal\Phi + \lambda N I)^{-1}}_{d\times d}\underbrace{\Phi^\intercal Y}_{d\times 1} = \underbrace{\Phi^\intercal}_{d\times N}\underbrace{(\Phi\Phi^\intercal + \lambda N I)^{-1}}_{N\times N}\underbrace{Y}_{N\times 1} \\ &= \Phi^\intercal\underbrace{(G + \lambda N I)^{-1}Y}_{=\varphi^\star\in\mathbb{R}^N},\end{aligned}$$

where we used the kernel matrix $G \in \mathbb{R}^{N\times N}$

$$G_{ij} = \phi(X_i)^\intercal\phi(X_j)$$

and the push-through identity. Once "training" is complete, i.e., $\theta^\star$ has been computed, we make predictions on new data $x \in \mathcal{X}$ with

$$\begin{aligned}h_{\theta^\star}(\cdot) &= \phi(\cdot)^\intercal\theta^\star \\ &= \sum_{i=1}^{N}\varphi_i^\star K(\cdot, X_i).\end{aligned}$$

Kernel ridge regression: RKHS

Next, consider the same linear regression setup with the prediction function in an RKHS as the explicit optimization variable. Let $X_1, \ldots, X_N \in \mathcal{X}$, $Y_1, \ldots, Y_N \in \mathbb{R}$, $\lambda > 0$, $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDK, and $\mathcal{H}$ the corresponding RKHS. Consider *kernel ridge regression* problem

$$\underset{f \in \mathcal{H}}{\text{minimize}} \quad \frac{1}{N} \sum_{i=1}^{N} (f(X_i) - Y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2.$$

By the representer theorem, a minimizer has the expression

$$f(x) = \sum_{j=1}^{N} \varphi_j K(x, X_j),$$

so we plug this form in to get a finite-dimensional optimization problem

$$\underset{\varphi \in \mathbb{R}^N}{\text{minimize}} \quad \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \varphi_j K(X_j, X_i) - Y_i \right)^2 + \lambda \left\| \sum_{j=1}^{N} \varphi_j K(X_j, \cdot) \right\|_{\mathcal{H}}^2.$$

Using the kernel matrix $G \in \mathbb{R}^{N \times N}$, we equivalently write

$$\underset{\varphi \in \mathbb{R}^N}{\text{minimize}} \quad \frac{1}{N} \|G\varphi - Y\|^2 + \lambda \varphi^\intercal G \varphi.$$

Kernelized implementation

To conclude, given $X_1, \dots, X_N \in \mathcal{X}$, $Y_1, \dots, Y_N \in \mathbb{R}$, $\lambda > 0$, and a PDK $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we can implement kernel ridge regression in a kernelized manner by forming the kernel matrix $G \in \mathbb{R}^{N \times N}$ (requires $N(N+1)/2$ evaluations of $K(\cdot, \cdot)$ but no need to explicitly form a feature vector) and perform linear algebra computations to solve

$$\varphi^\star = (G + \lambda N I)^{-1} Y.$$

Then, prediction on new data $x \in \mathcal{X}$ can be made with

$$f^\star(x) = \sum_{j=1}^{N} K(x, X_j)\varphi_j.$$

RKHS SGD

Let $\mathcal{X}$ be a nonempty set, $\mathcal{H}$ an RKHS on $\mathcal{X}$ with RK K, and $\mathcal{Y} = \mathbb{R}$. Consider the optimization problem

$$\underset{f \in \mathcal{H}}{\text{minimize}} \quad \mathbb{E}_{(X,Y)\sim P}\left[\ell(f(X); Y)\right].$$

SGD in the RKHS is

$$\begin{aligned} f^{k+1} &= f^k - \alpha_k \nabla_f \ell(f^k(X_{k+1}); Y_{k+1}) \\ &= f^k - \alpha_k \nabla_f \ell(\langle f^k, K(X_{k+1}, \cdot)\rangle_{\mathcal{H}}; Y_{k+1}) \\ &= f^k - \underbrace{\alpha_k \ell'(f^k(X_{k+1}); Y_{k+1})}_{=\beta_k} K(X_{k+1}, \cdot) \\ &= f^k - \beta_k K(X_{k+1}, \cdot), \end{aligned}$$

where we set $f^0 = 0$.

RKHS SGD

The RKHS SGD can be implemented with

$$
\begin{aligned}
f^k(X_{k+1}) &= -\sum_{i=1}^{k} \beta_i K(X_i, X_{k+1}) \\
\beta_{k+1} &= \alpha_{k+1} \ell'(f^k(X_{k+1}); Y_{k+1}) \\
\text{Storage} &\leftarrow (\beta_{k+1}, X_{k+1})
\end{aligned}
$$

and

$$
f^N(x) = -\sum_{k=1}^{N} \beta_k K(X_k, x).
$$