Kernel Methods

Semyon Malamud

EPFL

Table of Contents

Kernel Regression i

- suppose we have a sample $(y_i, x_i),\ x_i \in \mathbb{R}^d$ of n observations $i = 1, \cdots, n$
- A kernel is a distance function $K(x_i, x_j)$ measuring the distance between items in the feature space
- Examples:
  - Gaussian kernel $K(x_i, x_j) = e^{-\|x_i - x_j\|^2/\gamma}$ where γ is the *bandwidth*
  - Inner product kernels $K(x_i, x_j) = \phi(x_i' x_j)$
  - Normalized inner product kernels

$$K(x_i, x_j) \ = \ \psi(\|x_i\|)\, \psi(\|x_j\|)\, \phi\left(\frac{x_i' x_j}{\|x_i\|\ \|x_j\|}\right) \tag{1}$$

- **Kernel Regression**:
  - Build the matrix $\hat{K} \ = \ (K(x_i, x_j))_{i,j=1}^n$

Kernel Regression ii

- Build prediction

$$\hat{y}(x) \;=\; y'\hat{K}^{-1}k(x),\;\; k(x) \;=\; (K(x_i,x))_{i=1}^{n} \tag{2}$$

- **Kernel Regression is an interpolator**

$$\hat{y}(x_i) \;=\; y'\hat{K}^{-1}k(x_i) \;=\; y'(\delta_{i,j})_{j=1}^{n} \;=\; y_i \tag{3}$$

because

$$\hat{K}^{-1}\hat{K} \;=\; I \;\Leftrightarrow\; \hat{K}^{-1}k(x_i) \;=\; (\delta_{i,j})_{j=1}^{n} \tag{4}$$

- **Kernel Ridge Regression**:

$$\hat{y}(x) \;=\; y'(zI+\hat{K})^{-1}k(x),\;\; k(x) \;=\; (K(x_i,x))_{i=1}^{n} \tag{5}$$

Kernel Regression iii

- In matrix form, $X = (x_i)_{i=1}^n \in \mathbb{R}^{n \times d}$ and $\hat{K} = K(X, X) \in \mathbb{R}^{n \times n}$ and

$$\begin{aligned} \hat{y}(X; z) &= y'(zI + \hat{K})^{-1} \hat{K} = y'(zI + \hat{K})^{-1} (zI + \hat{K} - zI) \\ &= y' - \underbrace{zy'(zI + \hat{K})^{-1}}_{\textit{in sample bias}} \end{aligned} \tag{6}$$

Kernel Ridge Versus Linear Ridge Regression

Let $X \in \mathbb{R}^{n \times d}$ be all features stacked together. Using

$$(zI + XX')^{-1}X \;=\; X(zI + X'X)^{-1}\,, \tag{7}$$

we get

Theorem

When $K(x_i, x_j) = \phi(x_i' x_j)$ $\phi(x) \;=\; x$, we get a linear ridge regression:

$$\begin{aligned} \hat{K} \;&=\; XX',\; k(x) \;=\; Xx, \\ \hat{y}(x) \;&=\; y' \underbrace{(zI + XX')^{-1}X}_{=X(zI+X'X)^{-1}} x \;=\; \underbrace{y'X(zI + X'X)^{-1}}_{\text{ridge regression } \hat{\beta}(z)} x \end{aligned} \tag{8}$$

Inner Product Kernels in High Dimension: Linear Regression when $d \sim n$ i

Here, we discuss the remarkable results of El Karoui and the idea of Gaussian equivalence.

- Suppose $X_i \in \mathbb{R}^d$ are i.i.d., $i = 1, \cdots, n$, and define

$$M_{i,j} = f(X_i' X_j / d) \tag{9}$$

- n/d and d/n remain bounded, and both n, d go to ∞
- $X_i = \Sigma^{1/2} Y_i$ where Y_i are i.i.d.
- f is smooth

Inner Product Kernels in High Dimension: Linear Regression when $d \sim n$ ii

- Then, $\|M - K\| \to 0$ where

$$K = (f(0) + f''(0)\frac{\text{tr}(\Sigma^2)}{2d^2})\mathbf{1}\mathbf{1}' + f'(0)\frac{XX'}{d} + \underbrace{v_d I_n}_{\textit{implicit regularization}}$$

where

$$v_p = f(\text{tr}(\Sigma)/d) - f(0) - f'(0)\,\text{tr}(\Sigma)/d \underbrace{>}_{\textit{when } f \textit{ is convex}} 0$$

Distance Kernels in High Dimension: Linear Regression when $d \sim n$ i

A similar result holds for distance kernels such as, e.g., the Gaussian kernel $K(x_i, x_j) = e^{-\|x_i - x_j\|^2/L^2}$:

- Suppose $X_i \in \mathbb{R}^d$ are i.i.d., $i = 1, \cdots, n$, and define

$$M_{i,j} \;=\; f(\|X_i - X_j\|^2/d) \tag{10}$$

- define

$$\tau \;=\; 2\,\mathrm{tr}(\Sigma)/d$$

and

$$\psi \;=\; (\|X_i\|^2/d - \mathrm{tr}(\Sigma)/d)_{i=1}^n$$

- n/d and d/n remain bounded, and both n, d go to ∞
- $X_i = \Sigma^{1/2} Y_i$ where Y_i are i.i.d.

Distance Kernels in High Dimension: Linear Regression when $d \sim n$ ii

- f is smooth
- Then, $\|M - K\| \to 0$ where

$$K = f(\tau)\mathbf{1}\mathbf{1}' + f'(\tau)[\mathbf{1}\psi' + \psi\mathbf{1}' - 2XX'/d] + 0.5f''(\tau)A + \underbrace{v_d I_n}_{\substack{\text{implicit} \\ \text{regularization}}}$$

where

$$v_d = f(0) + \tau f'(\tau) - f(\tau)$$

and

$$A = \mathbf{1}(\psi \circ \psi)' + (\psi \circ \psi)\mathbf{1}' + 2\psi\psi' + 4\,\mathrm{tr}(\Sigma^2)d^{-2}\mathbf{1}\mathbf{1}'$$

Intuition

Concentration of quadratic forms:

$$X_i' X_j / d \;=\; Y_i' \Sigma^{1/2} \Sigma^{1/2} Y_j / d \;\approx\; d^{-1} \operatorname{tr}(\Sigma) \delta_{i,j} \tag{11}$$

Thus,

$$\|X_i - X_j\|^2 / d \;=\; (1 - \delta_{i,j}) 2 d^{-1} \operatorname{tr}(\Sigma)\,. \tag{12}$$

What About Out-Of-Sample Performance?

- Let us augment

$$\tilde{K} \;=\; \begin{pmatrix} K(x,x) & K(x,X) \\ K(X,x) & K(X,X) \end{pmatrix} \;=\; K(\tilde{X},\tilde{X}) \;\in\; \mathbb{R}^{(n+1)\times(n+1)} \qquad (13)$$

- The result still applies:

$$\tilde{K} \;\approx\; c_1\tilde{X}\tilde{X}' \;+\; c_2 I \;+\; \mathit{rank\_three} - \mathit{perturbation} \qquad (14)$$

and, hence, kernel ridge prediction is

$$\hat{y}(x) \;=\; K(x,X)(zI + K(X,X))^{-1} \qquad (15)$$

is approximately linear in X (up to a rank_three-perturbation)

Picking Non-Linearities with Extreme Non-Smoothness of Kernels i

- So, the kernel is a sort of linear regression when $d \sim n$
- At least, when f is smooth, the CLT guarantees that the perturbation is small.
- What about non-smooth f? But does it even matter? The answer is yes!!
- One of the most powerful kernels, $K(x_1, x_2) = e^{-\|x_1 - x_2\|}$ is not smooth at the origin: $f(x) = e^{-x^{1/2}}$.
- It turns out one can do it, but it is tough Kernel Matrix with Non-Smooth Kernels

Picking Non-Linearities with Extreme Non-Smoothness of Kernels ii

- Suppose $X_i \sim N(0, I/P)$ are i.i.d. and

$$M_{i,j} = f(X_i' X_j)(1 - \delta_{i,j})$$

(killed diagonal). Then, we get a form of the Marcenko-Pastur equation that is cubic in m

- The trick is to replace Taylor expansion with an expansion in special orthogonal polynomials

Summary

- Curse of Dimensionality: When $d \sim n$, we cannot learn non-linearities with simple kernel methods. Kernels collapse to a linear ridge
  - Why linear? It is linked to eigenfunctions !
- non-smooth kernels can help (formulas involve $f'(0)$!). E.g., $f(\|x_i - x_j\|^2)$ where $f'(0) = \infty$, such as $f(x) = e^{-x^{1/2}}$.
- Need ways to escape the curse of dimensionality: Feature Learning

Table of Contents

Definition of a Positive-Definite Kernel

A positive-definite kernel is a function

$$K : \Omega \times \Omega \rightarrow \mathbb{R}$$

such that, for any finite set $\{x_1, \ldots, x_n\} \subset \Omega$, the *kernel matrix* in $\mathbb{R}^{n \times n}$, defined by pairwise evaluations $\left(K(x_i, x_j)\right)_{i,j=1}^{n}$, is symmetric positive semi-definite.

Hilbert Spaces i

- A (real) Hilbert Space is a set H such that we can take linear combinations: $x, y \in H,\ ax + by \in H$
- And we have an **inner product**

$$\langle x, y \rangle \tag{16}$$

satisfying standard linearity properties:

$$\langle ax + by, z \rangle \ = a\langle x, z \rangle \ + \ b\langle y, z \rangle \tag{17}$$

- It is complete under the norm

$$\|x\| \ = \ \langle x, x \rangle^{1/2}$$

Hilbert Spaces ii

- In asset pricing, $H = L_2(\Omega)$, and

$$\langle X, Y \rangle \;=\; E[XY]\,.$$

In particular, for an SDF M, we can write prices as

$$P(X) \;=\; E[MX] \;=\; \langle M, X \rangle$$

Kernels and Feature Maps

Theorem

A kernel is positive definite if and only if

$$K(x_i, x_j) = \int_{\Theta} f(x_i;\theta) f(x_j;\theta) p(\theta) d\theta \tag{18}$$

for some f. Let $H = L_2(\Theta; p(\theta)d\theta)$. Let also
$f : x \to f(x) = (f(x;\theta))_{\theta\in\Theta}$ be the feature map. Then,

$$K(x_i; x_j) = E_\theta[f(x_i) f(x_j)] \tag{19}$$

Proof of Sufficiency:

$$\begin{aligned} a'\hat{K}a &= \sum_{i,j} K(x_i, x_j) a_i a_j = \sum_{i,j} E[f(x_i) f(x_j)] a_i a_j \\ &= E[(\sum_i a_i f(x_i))^2] \geq 0 . \end{aligned} \tag{20}$$

Examples of Feature Maps i

Theorem 1 (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.

$$k(x-y) = \int_{\mathbb{R}^d} p(\omega) e^{i\omega^\top (x-y)} d\omega = \mathbb{E}_\omega[\zeta_\omega(x)\bar{\zeta}_\omega(y)], \zeta_\omega(x) = e^{i\omega^\top x} \quad (21)$$

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

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

Table: Popular shift-invariant kernels and their corresponding Fourier transforms.

Examples of Feature Maps ii

To obtain a real-valued random feature for $k(x, y) = e^{-\|x-y\|^2/2}$, note that both the probability distribution $p(\omega)$ and the kernel $k(\Delta)$ are real, so the integrand $e^{i\omega^\top(x-y)}$ may be replaced with $\cos \omega^\top(x - y)$. Defining

$$z_\omega(x) = \begin{bmatrix} \cos(\omega^\top x) \\ \sin(\omega^\top x) \end{bmatrix}$$

gives a real-valued mapping that satisfies the condition $\mathbb{E}[z_\omega(x)^\top z_\omega(y)] = k(x, y)$, since $z_\omega(x)^\top z_\omega(y) = \cos \omega^\top(x - y)$. Other mappings such as $z_\omega(x) = \sqrt{2}\cos(\omega^\top x + b)$, where ω is drawn from $p(\omega)$ and b is drawn uniformly from $[0, 2\pi]$, also satisfy the condition $\mathbb{E}[z_\omega(x)^\top z_\omega(y)] = k(x, y)$.

To express the kernel $k(x - y)$ in terms of real-valued random features, we can use the following integral representation:

Examples of Feature Maps iii

$$k(x-y) = \mathbb{E}_{\omega,b}\left[\sqrt{2}\cos(\omega^\top x + b)\cdot\sqrt{2}\cos(\omega^\top y + b)\right],$$

where $b \sim \text{Uniform}[0, 2\pi]$, and ω is drawn from the Fourier transform $p(\omega)$ of the kernel $k(\Delta)$. Expanding the cosine terms, we have:

$$\cos(\omega^\top x+b)\cdot\cos(\omega^\top y+b) = \frac{1}{2}\left[\cos(\omega^\top(x-y)) + \cos(\omega^\top(x+y)+2b)\right].$$

Taking the expectation over b, the second term averages out to zero since b is uniformly distributed, leaving:

$$k(x-y) = \int_{\mathbb{R}^d} p(\omega)\cos(\omega^\top(x-y))\,d\omega.$$

Alternatively, using the 2-dimensional real-valued mapping:

Examples of Feature Maps iv

$$z_\omega(x) = \begin{bmatrix} \cos(\omega^\top x) \\ \sin(\omega^\top x) \end{bmatrix},$$

we obtain:

$$k(x-y) = \mathbb{E}_\omega \left[z_\omega(x)^\top z_\omega(y) \right],$$

where the inner product $z_\omega(x)^\top z_\omega(y)$ simplifies to:

$$\cos(\omega^\top x)\cos(\omega^\top y) + \sin(\omega^\top x)\sin(\omega^\top y) = \cos(\omega^\top (x-y)).$$

Thus, the integral representation becomes:

$$k(x-y) = \int_{\mathbb{R}^d} p(\omega)\cos(\omega^\top (x-y))\, d\omega.$$

General Random Features and Kernel Ridge Regression i

- using a discrete approximation

$$K(x_i, x_j) \;=\; \int f(x_i;\theta) f(x_j;\theta) p(\theta) d\theta \;\approx\; P^{-1} \sum_k f(x_i;\theta_k) f(x_j;\theta_k) \tag{22}$$

  implies that we can:

- Generate random features: Sample $\theta_k, k = 1, \cdots, P$ from $p(\theta)$
- compute random features

$$S_i \;=\; P^{-1/2} (f(x_i;\theta_k))_{k=1}^{P} \;=\; P^{-1/2} f(x_i) \tag{23}$$

- run a ridge regression of y_i on $S_i \in \mathbb{R}^P$

General Random Features and Kernel Ridge Regression ii

- indeed, (22) implies with $S = (P^{-1/2}(f(x_i;\theta_k))) \in \mathbb{R}^{n\times P}$ that

$$\hat{K} \;=\; (K(x_i,x_j))_{i,j=1}^n \;\approx\; \sum_k (P^{-1} f(x_i;\theta_k) f(x_j;\theta_k))_{i,j=1}^n \;=\; SS' \quad (24)$$

and

$$k(x) \;=\; (K(x;x_i))_{i=1}^n \;\approx\; \sum_k (P^{-1} f(x;\theta_k) f(x_i;\theta_k))_{i=1}^n = \; P^{-1/2} S f(x) \quad (25)$$

so that using

$$S'(zI + SS')^{-1} \;=\; (zI + S'S)^{-1} S', \quad (26)$$

General Random Features and Kernel Ridge Regression iii

we get

$$\begin{aligned}\hat{y}(x) &= k(x)'(zI + \hat{K})^{-1}y \approx P^{-1/2}f(x)'S'(zI + SS')^{-1}y \\ &= P^{-1/2}f(x)'\underbrace{(zI + S'S)^{-1}S'y}_{\hat{\beta}}\end{aligned} \tag{27}$$

- most common choice: choose an activation function $\sigma(x)$ and define random features as

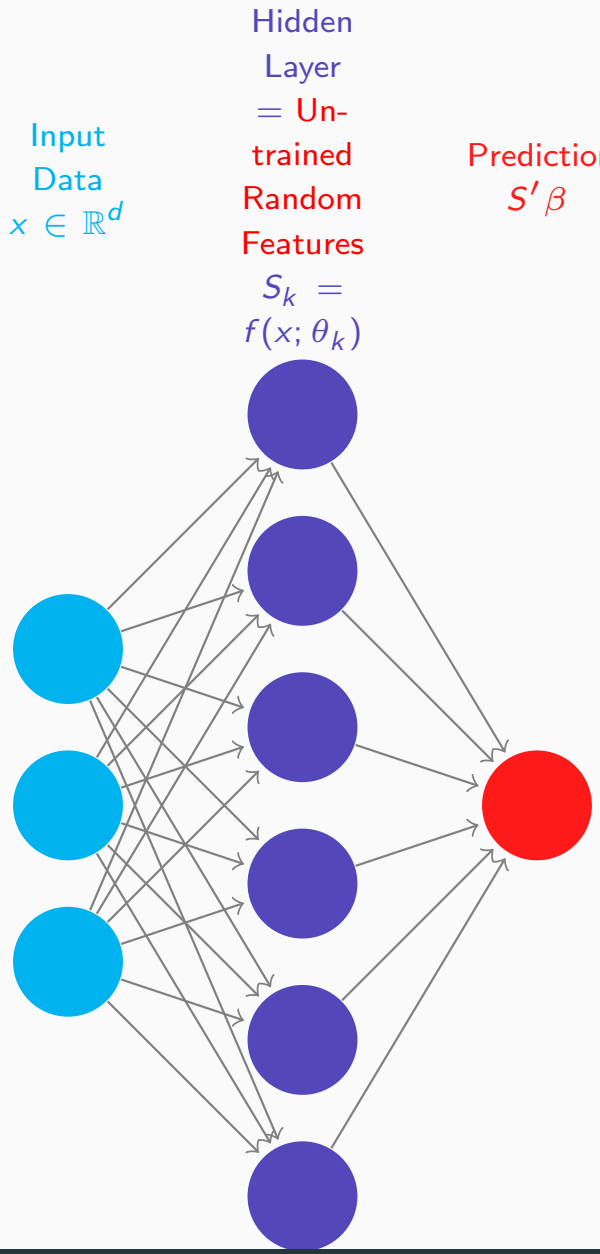
$$S_i = P^{-1/2}(\sigma(\theta_k' x_i + b_k))_{k=1}^{P} \tag{28}$$

where θ_i are weights and b_i are biases.

Theorem

Kernel Ridge Regression = Neural Network with one hidden layer and a linear output neuron where the hidden layer weights are not trained

Kernel=Shallow Neural Network



Link

Understanding Kernels

Table of Contents

Kernels in Plato's Cave

- An ML model is a map

$$x \to \hat{f}(\underbrace{x}_{\substack{\text{test}\\ \text{features}}} ; \underbrace{X, y}_{\substack{\text{train}\\ \text{data}}}) \tag{29}$$

- In the standard parametric world, in the interpolation regime, we solve

$$f(X_i; \theta) = y_i \,, \ \theta \in \mathbb{R}^P, \ i = 1, \cdots, n \,. \tag{30}$$

- This defines a foliation: over each (y, X), we have a $(P - n)$-dimensional manifold attached.
- Gradient descent picks particular points on this manifold: Is is a form of projection
- Kernels also project ground truth on something
- How can we describe this something?

Spectral Theorem for Symmetric Matrices

Theorem

If $A = (A(i,j))_{i,j=1}^n$ *is symmetric, then* $A = U\,\mathrm{diag}(\lambda_j)U'$. *A defines an operator*

$$Ax \;=\; UDU'x,\; x \in \mathbb{R}^n\,.$$

If ϕ_j *are the eigenvectors, then*

$$Ax \;=\; \sum_{j=1}^{n} \lambda_j \phi_j \langle \phi_j, x \rangle\,.$$

Furthermore,

$$\sum_{i,j} A_{i,j}^2 \;=\; \sum_j \lambda_j^2\,.$$

Mercer Theorem i

Theorem Suppose that $K: \Omega \times \Omega \to \mathbb{R}$ is a positive definite kernel. Let us equip Ω with a probability measure $p(dx)$ and let $L_2(\Omega)$ be the set of square integrable functions. Define the Integral Operator

$$T_K(f) = E_{\tilde{x}}[K(x,\tilde{x})f(\tilde{x})] = \int_\Omega K(x,\tilde{x})f(\tilde{x})p(d\tilde{x}) \approx \sum_i K(x,\tilde{x}_i)f(\tilde{x}_i) \tag{31}$$

(like an infinite-dimensional matrix). Suppose that K is square integrable:

$$\int_\Omega \int_\Omega K(x,\tilde{x})^2 p(d\tilde{x})p(dx) < \infty. \tag{32}$$

Mercer Theorem ii

Then, T_K has an orthonormal basis of eigenfunctions $\phi_j(x)$, $j \geq 1$, that depend in a mysterious and complex way on both K and $p(dx)$, with the corresponding eigenvalues $\lambda_j \geq 0$, so that

$$T_K(\phi_j) = \lambda_j \phi_j \Leftrightarrow \int_\Omega K(x, \tilde{x}) \phi_j(\tilde{x}) p(d\tilde{x}) = \lambda_j \phi_j(x) \tag{33}$$

and ϕ_j form a basis of $L_2(\Omega)$ so that any function $f \in L_2$ can be written as

$$f(x) = \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \phi_j(x) \tag{34}$$

and

$$\begin{aligned} T_K(f) &= T_K(\sum_{j=1}^{\infty} \langle f, \phi_j \rangle \phi_j(x)) = \sum_{j=1}^{\infty} \langle f, \phi_j \rangle T_K(\phi_j(x)) \\ &= \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \lambda_j \phi_j(x) . \end{aligned} \tag{35}$$

Mercer Theorem iii

Furthermore,

$$\int_{\Omega}\int_{\Omega} K(x,\tilde{x})^2 p(d\tilde{x})p(dx) \;=\; \sum_{j=1}^{\infty}\lambda_j^2 \;<\; \infty\,. \tag{36}$$

and, hence,

$$\lim_{j\to\infty}\lambda_j \;=\; 0\,. \tag{37}$$

Reprodicing Kernel Hilbert Space i

- Let us define the inverse operator T_K^{-1} via

$$T_K^{-1} f \;=\; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \lambda_j^{-1} \, \phi_j(x) \,. \tag{38}$$

By definition,

$$\begin{aligned} T_K(T_K^{-1} f) \;&=\; T_K \left(\sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \lambda_j^{-1} \, \phi_j(x) \right) \;=\; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \lambda_j^{-1} \, T_K \phi_j(x) \\ &= \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \lambda_j^{-1} \, \lambda_j \phi_j(x) \;=\; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \phi_j(x) \;=\; f(x) \,, \end{aligned} \tag{39}$$

so this is indeed the inverse.

Reprodicing Kernel Hilbert Space ii

- In finite dimensions (**linear algebra!**), a symmetric positive definite matrix A is invertible if and only if all its eigenvalues are positive, $\lambda_j > 0$. In this case, the operator is **surjective**: A^{-1} is defined everywhere. I.e., for any vector y, there exists an x such that $Ax = y$. Equivalently, $x = A^{-1}y$ and A^{-1} is defined on all vectors.
- In infinite dimensions, this is not true anymore!
- T_K^{-1} exists but is not bounded (because, $\lambda_j \to 0$ by (37)). **Is a pure infinite-dimensional phenomenon.**

Reprodicing Kernel Hilbert Space iii

- Define

$$\mathcal{H}_K \;=\; \left\{ f(x) \in L_2(\Omega) : \; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle^2 \, \lambda_j^{-1} \;<\; \infty \right\}, \tag{40}$$

and equip it with the inner product

$$\langle f, g \rangle_{\mathcal{H}_K} \;=\; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \langle g, \phi_j \rangle \lambda_j^{-1} \, . \tag{41}$$

Reprodicing Kernel Hilbert Space iv

- Note that

$$\langle f, g \rangle_{\mathcal{H}_K} = \langle T_K^{-1} f, g \rangle_{L_2(\Omega)} = \int_\Omega f(x)(T_K^{-1} g)(x) p(dx) \tag{42}$$

By the definition of the T_K^{-1} operator,

$$\begin{aligned} g(x) &= T_K(T_K^{-1} g)(x) = \int K(x, \tilde{x})(T_K^{-1} g)(\tilde{x}) p(d\tilde{x}) \\ &= \int_\Omega K_x(\tilde{x})(T_K^{-1} g)(x) p(dx) = \langle K_x, g \rangle_{\mathcal{H}_K}, \end{aligned} \tag{43}$$

which is the reproducing kernel property!

OK, So what is RKHS?

Theorem

RKHS = set of functions for which (40) holds:

$$\mathcal{H}_K = \left\{ f(x) \in L_2(\Omega) : \sum_{j=1}^{\infty} \langle f, \phi_j \rangle^2 \lambda_j^{-1} < \infty \right\}, \tag{44}$$

- Since $\lambda_j \to 0$, this is a non-trivial condition: it means that $\langle f, \phi_j \rangle$ go to zero fast as $j \to \infty$, faster than $\lambda_j^{1/2}$.
- ϕ_j tend to oscillate more when j ia large (a bit like $\sin(xj)$ waves)
- the smaller j, the smoother the function
- thus RKHS = functions that do not oscillate too much; or, equivalently, functions that are sufficiently smooth

Why Do We Care? Regression, Alignment, and the Inductive Bias i

- **Theorem** Ridge regression always generates prediction $\hat{f}(x) \in \mathcal{H}_K$. Thus, kernel ridge always predicts (=extrapolates!), assuming the function is smooth and does not oscillate too much. **If the function is not smooth and/or oscillates a lot, we are in trouble!**

  Proof. Kernel Ridge Prediction is

$$\begin{aligned} \hat{f}(x) &= K(x,X)(zI + K(X,X))^{-1}y = \sum_{i=1}^{n} K(x,x_i)\xi_i, \\ \xi &= (zI + K(X,X))^{-1}y \in \mathbb{R}^n. \end{aligned} \tag{45}$$

Why Do We Care? Regression, Alignment, and the Inductive Bias ii

Now, $K_{x_i}(x) = K(x, x_i) \in \mathcal{H}_K$ always! It follows from the definition of RKHS, but let us do a direct derivation; it is instructive. We have

$$K_{x_i}(x) \;=\; K(x, x_i) \;=\; \sum_{j=1}^{\infty} \underbrace{\lambda_j \phi_j(x_i)}_{\textit{basis coefficients}} \quad \phi_j(x)\,, \tag{46}$$

implying that the basis coefficients are $\lambda_j \phi_j(x_i)$. Then, we need to check that they satisfy (40):

$$\sum_{j=1}^{\infty} (\lambda_j \phi_j(x_i))^2 \, \lambda_j^{-1} \;=\; \sum_{j=1}^{\infty} \lambda_j (\phi_j(x_i))^2 \;=\; K(x_i, x_i) \;<\; \infty\,. \tag{47}$$

This is striking: while RKHS does depend on the underlying distribution $p(dx)$, the prediction of the ridge regression always belongs to the intersection of all possible RKHS generated by K.

Kernel Inductive Bias and Minimum Norm Interpolation i

Theorem

$$\hat{f}(x) \;=\; K(x,X)(zI + K(X,X))^{-1}y\,. \tag{48}$$

If $y_i = f^*(x_i) + \varepsilon_i,\ E[\varepsilon_i|x] = 0$, *and* x_i, ε_i *are i.i.d. and* $f^*(x) \in \mathcal{H}_K$, *then* $\hat{f}(x) \to f^*(x)$ *in* $\mathcal{H}_K$ *as* $n \to \infty$, *with probability one.*

So, what exactly does $\hat{f}$ pick in finite samples?

Kernel Inductive Bias and Minimum Norm Interpolation ii

Theorem (Kernel Ridge Inductive Bias=Small $\mathcal{H}_K$-norm)

$$\hat{f} = \arg\min_{f \in \mathcal{H}_K} \left\{ \sum_{i=1}^{n} (y_i - f(x_i))^2 + z\|f\|^2_{\mathcal{H}_K} \right\}. \tag{49}$$

When $z = 0$, we get the minimum $\mathcal{H}_K$-norm interpolator,

$$\hat{f} = \arg\min_{f \in \mathcal{H}_K} \{\|f\|^2_{\mathcal{H}_K} : (y_i - f(x_i)) = 0 \ \forall \ i\}. \tag{50}$$

Why is this striking? Well, $\mathcal{H}_K$ depends on the true probability distribution, which we do not know! And yet, mysteriously, Ridge regression finds it! Small $\mathcal{H}_K$ is the smoothness inductive bias of kernels!!

The proof is based on the Representer Theorem

Representer Theorem

Theorem: The solution to a regularized empirical risk minimization problem in RKHS has the form:

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

- Instead of searching over all functions in $\mathcal{H}_K$, we optimize over α.
- This allows efficient computation using kernel matrices.

Proof of Representer Theorem

Step 1: Function Decomposition

$$f = f_{\|} + f_{\perp}, \quad f_{\|} \in \mathcal{H}_n, \quad f_{\perp} \perp \mathcal{H}_n.$$

- $\mathcal{H}_n = \text{span}\{K(\cdot, x_i)\}_{i=1}^n$.
- $f_{\perp}$ is orthogonal and does not affect the empirical risk because

$$f(x_i) \;=\; \langle f, K(\cdot, x_i)\rangle \;=\; \langle f_{\|}, K(\cdot, x_i)\rangle \;=\; f_{\|}(x_i) \tag{51}$$

so that

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + z\|f\|_{\mathcal{H}_K}^2 \;=\; \sum_{i=1}^{n}(y_i - f_{\|}(x_i))^2 + z(\|f_{\|}\|_{\mathcal{H}_K}^2 + \|f_{\perp}\|_{\mathcal{H}_K}^2)$$

Conclusion of the Proof

Step 2: Regularization Effect

$$\|f\|_{\mathcal{H}_K}^2 = \|f_\parallel\|_{\mathcal{H}_K}^2 + \|f_\perp\|_{\mathcal{H}_K}^2.$$

Since $f_\perp$ only increases the regularization term, the optimal solution satisfies $f_\perp = 0$. Hence,

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

Conclusion: Every minimizer of the regularized problem is a linear combination of kernel evaluations.

Homework

Complete the proof: minimization over $f^*(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)$ (i.e., over α_i) gives the kernel ridge.

Also, prove the $z = 0$ case!

Table of Contents

Hermite Polynomials i

There are many equivalent ways to define the Hermite polynomials. A natural one is through the so-called *Rodrigues' formula*:

$$H_k(x) \;=\; (-1)^k\, e^{x^2}\, \frac{d^k}{dx^k}\left[e^{-x^2}\right].$$

From this definition, one can deduce:

$$\begin{aligned} H_0(x) \;&=\; 1, \quad H_1(x) \;=\; -\,e^{x^2}\left[-\,2x\,e^{-x^2}\right] \;=\; 2x, \\ H_2(x) \;&=\; e^{x^2}\left[(-2x)e^{-x^2} \;-\; 2\,e^{-x^2}\right] \;=\; 4x^2 \;-\; 2, \quad \ldots \end{aligned} \tag{52}$$

Other simple properties (provable by recursion) include:

- $H_k(x)$ is a polynomial of degree k.
- $H_k(x)$ has the same parity as k (even/odd).

Hermite Polynomials ii

- The leading coefficient of $H_k(x)$ is 2^k.

Using integration by parts, one shows that for $k \neq \ell$,

$$\int_{-\infty}^{+\infty} H_k(x)\, H_\ell(x)\, e^{-x^2}\, dx \;=\; 0,$$

and for $k = \ell$,

$$\int_{-\infty}^{+\infty} \big(H_k(x)\big)^2\, e^{-x^2}\, dx \;=\; \sqrt{\pi}\, 2^k\, k!.$$

Hence, the Hermite polynomials $\{H_k\}$ are orthogonal with respect to the Gaussian distribution of mean 0 and variance $1/2$.

Defining the *Hermite functions*

$$\psi_k(x) \;=\; \big(\sqrt{\pi}\, 2^k\, k!\big)^{-\frac{1}{2}}\, H_k(x)\, e^{-x^2/2},$$

Hermite Polynomials iii

we obtain an orthonormal basis of $L^2(\mathbb{R})$. As k increases, these functions have increasingly wide "effective support" (though they extend over the entire real line) and exhibit increasingly oscillatory behavior, much like sines and cosines in the Fourier basis.

Hermite Polynomials iv

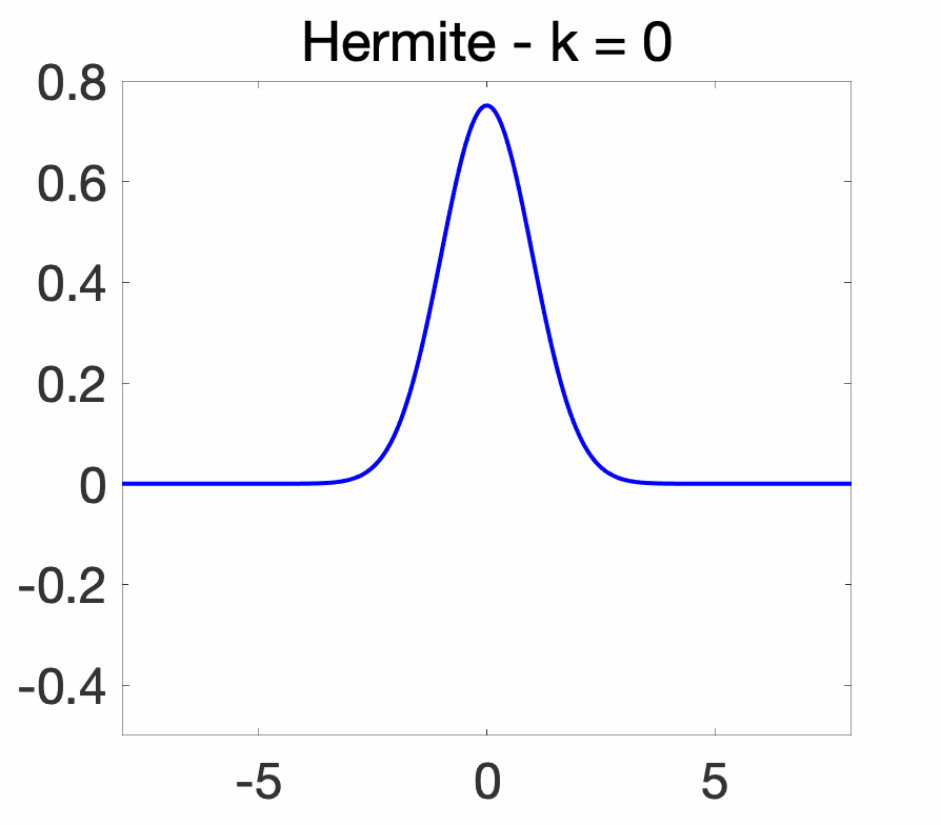


Figure: Hermite polynomial animation.

Hermite Polynomials v

Among such orthonormal bases, the Hermite functions happen to diagonalize the Fourier transform operator. In other words, the Fourier transform of ψ_k (for the definition making it an isometry of $L^2(dx)$) is equal to

$$(\mathcal{F}\psi_k)(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \psi_k(x)\, e^{-i\omega x}\, dx = (-i)^k\, \psi_k(\omega).$$

(Note that the eigenvalues are all of unit modulus, since we have an isometry.) I am not aware of any applications of this property in machine learning or statistics (though there probably are some).

In order to compute Hermite polynomials, the following recurrence relation is particularly useful:

$$H_{k+1}(x) \;=\; 2x\, H_k(x) \;-\; 2k\, H_{k-1}(x). \tag{1}$$

Hermite Polynomials vi

Such recurrences are always available for orthogonal polynomials (see **?**), but it takes here a particularly simple form

The following property is central in many proofs of the properties of Hermite polynomials: for all real t, we have

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} H_k(x) = e^{2xt-t^2}. \tag{2}$$

For the later developments, we need other properties which are less standard (there are many other interesting properties, which are not useful for this post, see *here*).

For $|\rho| < 1$, it states:

$$\exp\Big(-\tfrac{\rho}{1-\rho^2}(x-y)^2\Big) = \sqrt{1-\rho^2} \sum_{k=0}^{\infty} \frac{\rho^k}{2^k\, k!} H_k(x)\, H_k(y) \exp\Big(-\tfrac{\rho}{1+\rho}(x^2+y^2)\Big)$$

Hermite Polynomials vii

so that

$$\int_{-\infty}^{\infty} H_k(x)\, \exp\Big(-\frac{(x-\rho\, y)^2}{1-\rho^2}\Big)\, \mathrm{d}x \;=\; \sqrt{\pi}\, \rho^k\, \sqrt{1-\rho^2}\; H_k(y). \qquad (3)$$

Orthonormal Basis

Theorem

Define

$$f_k(x) \;=\; \frac{1}{\sqrt{N_k}}\, H_k(x) \exp\Big(-\frac{\rho}{1+\rho}\, x^2\Big),$$

where $N_k = 2^k\, k!\, \sqrt{\frac{1-\rho}{1+\rho}}$, *and* $H_k(x)$ *is the* k*th Hermite polynomial. Then,* $\{f_k\}_{k=0}^{\infty}$ *is an orthonormal basis for* $L^2(\mathrm{d}\mu)$ *when* $\mathrm{d}\mu$ *is the Gaussian distribution of mean 0 and variance* $\frac{1}{2}\,\frac{1+\rho}{1-\rho}$.

Homework: Prove this Theorem (this follows from the orthogonality property of Hermite polynomials).

Theorem

The Gaussian kernel admits the decomposition

$$K(x,y) \;=\; \exp\Big(-\tfrac{\rho}{1-\rho^2}\,(x-y)^2\Big) \;=\; \sum_{k=0}^{\infty}(1-\rho)\,\rho^k\,f_k(x)\,f_k(y).$$

Thus, the kernel operator in $L^2(\mathrm{d}\mu)$ *when* $\mathrm{d}\mu$ *is the Gaussian distribution of mean* 0 *and variance* $\frac{1}{2}\,\frac{1+\rho}{1-\rho}$ *has* f_k *as eigenfunctions and*

$$\lambda_k \;=\; (1-\rho)\,\rho^k.$$

as eigenvalues.

Empirical Eigenvalues and Empirical Eigenfunctions i

Theorem

The eigenvalues $\hat{\lambda}_k$ *of the kernel matrix*

$$n^{-1}K(X,X) \;=\; \frac{1}{n}(K(x_i,x_j))_{i,j=1}^{n} \tag{53}$$

converge to those of the kernel operator when x_i *are sampled i.i.d. from the* $\mu(dx)$*. Furthermore, given the eigenvectors* $q_k \;=\; (q_k(i))_{i=1}^{n} \in \mathbb{R}^n$ *of the kernel matrix can be used to construct Nystrom approximations to the true (unobserved!) eigenfunctions*

$$\phi_k(x) \;\approx\; \sum_{i=1}^{n} K(x,x_i)q_k(i)\,. \tag{54}$$

Empirical Eigenvalues and Empirical Eigenfunctions ii

Heuristic Proof:

$$\begin{aligned} \lambda_k \phi_k(x) &= \int K(x, \tilde{x}) \phi_k(\tilde{x}) d\mu(\tilde{x}) \approx \sum_{i=1}^{n} K(x, \tilde{x}_i) \phi_k(\tilde{x}_i) \\ \hat{\lambda}_k q_k(i) &= \sum_{i=1}^{n} K(x_i, x_j) q_k(j) \end{aligned} \tag{55}$$

Experiments. In order to showcase the exact eigenvalues of the expectation, we compare the eigenvalues with the ones of the empirical covariance operator for various values of the number of observations. We see that as n increases, the empirical eigenvalues match the exact ones for higher k: **smaller eigenvalues are harder to learn!**

Empirical Eigenvalues and Empirical Eigenfunctions iii

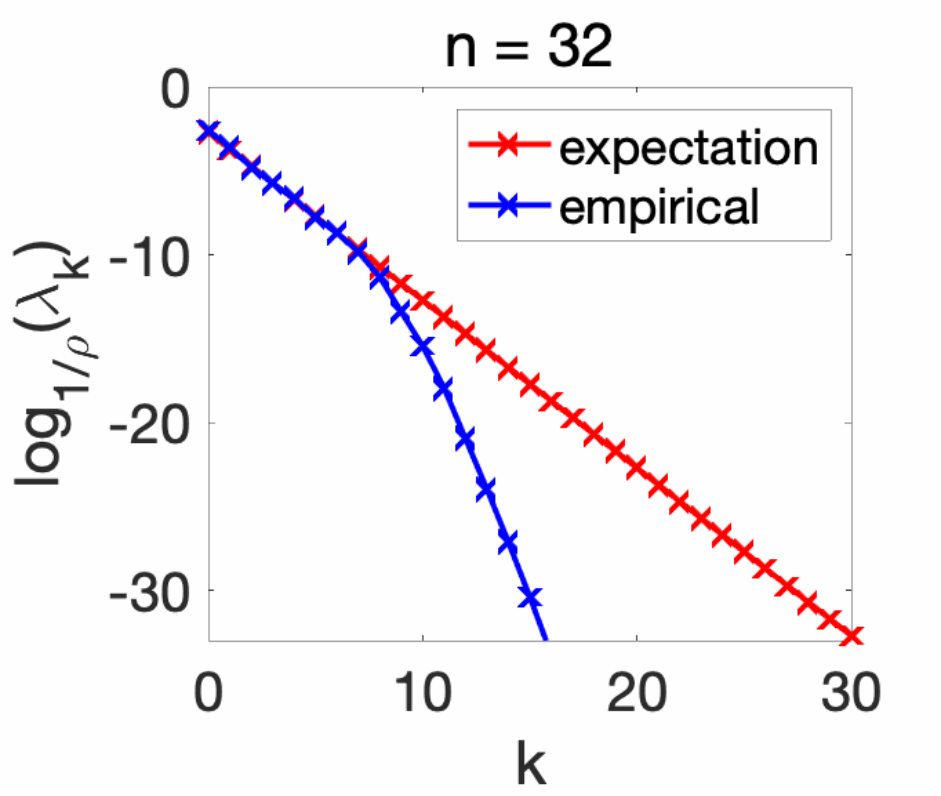


Figure: Convergence of Estimated Eigenvalues to $\lambda_k = (1-\rho)\,\rho^k$.

Ideal Kernel i

- The ideal kernel to learn $y = f(x)$ is

$$K_{ideal}(x, \tilde{x}) \;=\; f(x)\, f(\tilde{x}) \tag{56}$$

- Eigenfunction equation

$$\int K_{ideal}(x, \tilde{x}) \psi(\tilde{x}) \sigma(d\tilde{x}) \;=\; \lambda\, \psi(x) \tag{57}$$

takes the form

$$f(x) \int f(\tilde{x}) \psi(\tilde{x}) \sigma(d\tilde{x}) \;=\; \lambda\, \psi(x) \tag{58}$$

Thus, the only non-trivial eigenfunction is $\psi(x) = f(x)$ with the eigenvalue

$$\lambda \;=\; \int f(\tilde{x})^2 \sigma(d\tilde{x}) \;=\; \left(\int \int K^2(x, \tilde{x}) \sigma(dx) \sigma(d\tilde{x}) \right)^{1/2}, \tag{59}$$

Ideal Kernel ii

so that all other eigenvalues are identically zero: $\lambda = \lambda_1$, $\lambda_2 = \lambda_3 = \cdots = 0$.

- with an ideal kernel,

$$K(X,X) \;=\; (f(x_i)f(x_j))_{i,j=1}^{n} \;=\; f(X)f(X)^\top \tag{60}$$

has rank 1 and, hence, by the Sherman-Morrison formula,

$$\begin{aligned} K(x,X)(zI + K(X,X))^{-1} \;&=\; f(x)f(X)^\top (zI + f(X)f(X)^\top)^{-1} \\ &= f(x)f(X)^\top \frac{z^{-1}}{1 + z^{-1}\|f(X)\|^2} \;=\; f(x)f(X)^\top \frac{1}{z + \|f(X)\|^2} \end{aligned} \tag{61}$$

and, hence, if $y = f(X) + \varepsilon$, we get

$$\begin{aligned} \hat{f}(x) \;&=\; K(x,X)(zI + K(X,X))^{-1} y \\ &= f(x)f(X)^\top \frac{1}{z + \|f(X)\|^2}(f(X) + \varepsilon) \;=\; c\, f(x)\,, \end{aligned} \tag{62}$$

Ideal Kernel iii

where

$$c = \frac{\|f(X)\|^2 + f(X)^\top \varepsilon}{z + \|f(X)\|^2} = \frac{\frac{1}{n}\sum_i f(X_i)^2 + \frac{1}{n}\sum_i f(X_i)\varepsilon_i}{\frac{1}{n}z + \frac{1}{n}\sum_i f(X_i)^2} \approx 1 \quad (63)$$

by the law of large numbers when $E[f(X)\varepsilon] = 0$ when n is large.

Thus, ideal kernel has perfect alignment with the data, and hence, we can learn the true f easily.

Table of Contents

Suppose we have a bunch of random features or other signals, $S_{k,t} = \frac{1}{P^{1/2}} f(X_t; \omega_k),\ k = 1, \cdots, P$. They have the true covariance matrix

$$E[S_t S_t'] \ = \ \Psi \tag{64}$$

That is, assuming that that observations X_t across t are sampled i.i.d. from the same distribution $\sigma(dx)$, we get

$$\frac{1}{P} E[f(X; \omega_{j_1}) f(X; \omega_{j_2})] \ = \ \Psi_{j_1, j_2} \, . \tag{65}$$

Let now $h_j(j_1)$ be eigenvectors of Ψ :

$$\Psi h_j(j_1) \ = \ \hat{\lambda}_j \, h_j(j_1) \, . \tag{66}$$

We now show a surprising thing: There is a direct link between eigenvalues of Ψ and eigenvalues of the integral operator K. Namely, define

$$\hat{\psi}_j(x) \ = \ \sum_{j=1}^{P} h_j(j_1) \, f(x; \omega_{j_1}) \, . \tag{67}$$

Then,

$$\int K(x,\tilde{x})\hat{\psi}_j(\tilde{x})\sigma(d\tilde{x}) = \int K(x,\tilde{x})\sum_{j=1}^{P} h_j(j_1)\, f(x;\omega_j)\sigma(d\tilde{x})$$

$$= \sum_{j=1}^{P} h_j(j_1)\int K(x,\tilde{x})f(\tilde{x};\omega_j)\sigma(d\tilde{x})$$

$$= \sum_{j=1}^{P} h_j(j_1)\int \frac{1}{P}\sum_{j_1=1}^{P} f(x;\omega_{j_1})f(\tilde{x};\omega_{j_1})f(\tilde{x};\omega_j)\sigma(d\tilde{x})$$

$$= \sum_{j=1}^{P} h_j(j_1)\frac{1}{P}\sum_{j_1=1}^{P} f(x;\omega_{j_1})\int f(\tilde{x};\omega_{j_1})f(\tilde{x};\omega_j)\sigma(d\tilde{x})$$

$$= \sum_{j=1}^{P} h_j(j_1)\frac{1}{P}\sum_{j_1=1}^{P} f(x;\omega_{j_1})\int f(\tilde{x};\omega_{j_1})f(\tilde{x};\omega_j)\sigma(d\tilde{x}) \quad (68)$$

$$= \sum_{j=1}^{P} h_j(j_1)\sum_{j_1=1}^{P} f(x;\omega_{j_1})\frac{1}{P}E[f(X;\omega_{j_1})f(X;\omega_j)]$$

Thus, we have proved

Theorem

For a finite-dimensional kernel

$$K(x,\tilde{x}) \;=\; \frac{1}{P}\sum_{j=1}^{P} f(x;\omega_j)f(\tilde{x};\omega_j)\,, \tag{69}$$

the integral operator T_K only has P eigenvalues λ_j coinciding with the eigenvalues of the matrix Ψ,

$$\frac{1}{P}E[f(X;\omega_{j_1})f(X;\omega_{j_2})] \;=\; \Psi_{j_1,j_2}\,. \tag{70}$$

Furthermore, the eigenfunctions are given by

$$\psi_j(x) \;=\; \sum_{j_1=1}^{P} h_j(j_1)f(x;\omega_{j_1}) \tag{71}$$

Table of Contents

The following is a heuristic formulation of the (incredibly complex) result of Kernel Ridge in High Dimensions

Theorem

Suppose now we send $n, d \to \infty$. *Let*

$$f(x) \;=\; \sum_{j=1}^{\infty} \psi_j(x) c_j \,, \;\; \textit{where} \;\; c_j \;=\; \langle \psi_j(x), f(x) \rangle \;=\; E[\psi_j(x) f(x)] \quad (72)$$

Then, there exists an increasing function $n_*(d; L)$ *such that, for* $n \sim n_*(d; L)$ *and* z *sufficient small,*

$$\hat{f}_n(x) \;=\; K(x, X_n)^\top (zI + K(X_n, X_n))^{-1} y_n \quad (73)$$

converges to the projection

$$P_{\leq L} f(x) \;=\; \sum_{j=1}^{L} \psi_j(x) c_j$$

- Suppose $y = f(x)$

- $$\frac{1}{n}K(X,X) \;=\; \sum_{j=1}^{\infty} \lambda_j \frac{1}{n}\psi_j(X)\,\psi_j(X)^\top \;\in\; \mathbb{R}^{n\times n} \tag{74}$$

- with many train observations, we have

$$\frac{1}{n}\psi_{j_1}(X)^\top \psi_{j_2}(X) \;=\; \frac{1}{n}\sum_i \psi_{j_1}(x_i)\psi_{j_2}(x_i) \;\underbrace{\approx}_{LLN}\; E[\psi_{j_1}(x)\psi_{j_2}(x)] \;=\; 0 \tag{75}$$

Thus, $n^{-1/2}\psi_j(X)$ is approximately orthonormal basis of $\mathbb{R}^n$ and

$$(zI + n^{-1}K(X,X))^{-1} \;\approx\; \sum_{j=1}^{\infty} \underbrace{(\lambda_j + Z_*)^{-1}}_{implicit\ \ regularization} \frac{1}{n}\psi_j(X)\,\psi_j(X)^\top \tag{76}$$

and

$$(zI + n^{-1}K(X,X))^{-1}y = \sum_{j=1}^{\infty} \underbrace{(\lambda_j + Z_*)^{-1}}_{implicit\ \ regularization} \frac{1}{n}\psi_j(X)\,\psi_j(X)^\top y$$

$$= \sum_{j=1}^{\infty} \underbrace{(\lambda_j + Z_*)^{-1}}_{implicit\ \ regularization} \frac{1}{n}\psi_j(X)\,\psi_j(X)^\top f(X)$$

$$\approx \sum_{j=1}^{\infty} \underbrace{(\lambda_j + Z_*)^{-1}}_{implicit\ \ regularization} \psi_j(X)\,c_j \tag{77}$$