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

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

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:

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

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

$$K(x_i, x_j) \;=\; f(x_i' x_j / d) \tag{9}$$

- 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, $\|\hat{K} - K_*\| \to 0$ where

$$K_* = \left(f(0) + f''(0)\frac{\mathrm{tr}(\Sigma^2)}{2d^2}\right)\mathbf{1}\mathbf{1}' + f'(0)\frac{XX'}{d} + \underbrace{v_d I_n}_{\textit{implicit regularization}}$$

where

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

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

$$K(x_i, x_j) \;=\; f(\|x_i - x_j\|^2/d) \tag{10}$$

- 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, $\|\hat{K} - K_*\| \to 0$ where

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

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:

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

Thus,

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

What About Out-Of-Sample Performance?

- Let us augment

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

- The result still applies:

$$\tilde{K} \approx c_1 \tilde{X}\tilde{X}' + c_2 I + rank\_three - perturbation \tag{14}$$

and, hence, kernel ridge prediction is

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

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

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

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

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 $\big(K(x_i, x_j)\big)_{i,j=1}^{n}$, is symmetric positive semi-definite.

Theorem

A kernel is positive definite if and only if

$$K(x_i, x_j) = \int_\Theta f(x_i; \theta) f(x_j; \theta) p(\theta) d\theta \tag{16}$$

for some f.

General Random Features and Kernel Ridge Regression i

- using a discrete approximation

$$K(x_i, x_j) \; = \; \int f(x_i;\theta)f(x_j;\theta)p(\theta)d\theta \; \approx \; P^{-1}\sum_k f(x_i;\theta_k)f(x_j;\theta_k) \tag{17}$$

implies that we can:

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

$$S_i \; = \; P^{-1/2}(f(x_i;\theta_k))_{k=1}^{P} \; = \; P^{-1/2}f(x_i) \tag{18}$$

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

- indeed, (17) implies with $S = (P^{-1/2}(f(x_i;\theta_k))) \in \mathbb{R}^{n\times P}$ that

$$\hat{K} \;=\; (K(x_i,x_j))_{i,j=1}^n \;\approx\; \sum_k (P^{-1} f(x_i;\theta_k) f(x_j;\theta_k))_{i,j=1}^n \;=\; SS' \quad (19)$$

and

$$k(x) \;=\; (K(x;x_i))_{i=1}^n \;\approx\; \sum_k (P^{-1} f(x;\theta_k) f(x_i;\theta_k))_{i=1}^n = \; P^{-1/2} S f(x) \quad (20)$$

so that using

$$S'(zI+SS')^{-1} \;=\; (zI+S'S)^{-1}S', \quad (21)$$

we get

$$\hat{y}(x) = k(x)'(zI + \hat{K})^{-1}y \approx P^{-1/2}f(x)'S'(zI + SS')^{-1}y$$
$$= P^{-1/2}f(x)'\underbrace{(zI + S'S)^{-1}S'y}_{\hat{\beta}} \tag{22}$$

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

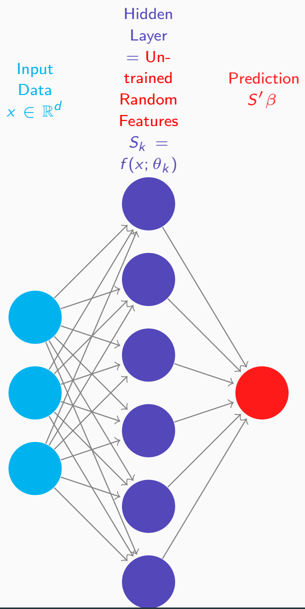
$$S_i = P^{-1/2}(\sigma(\theta_k' x_i + b_k))_{k=1}^{P} \tag{23}$$

where θ_k are weights and b_k are biases sampled from a distibution $p(\theta, b)$ of our choice.

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

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

$$T_K(f) \ = \ E_{\tilde{x}}[K(x,\tilde{x})f(\tilde{x})] \ = \ \int_\Omega K(x,\tilde{x})f(\tilde{x})p(d\tilde{x}) \ \approx \ \sum_i K(x,\tilde{x}_i)f(\tilde{x}_i) \tag{24}$$

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

$$\int_\Omega \int_\Omega K(x,\tilde{x})^2 p(d\tilde{x})p(dx) \ < \ \infty \, . \tag{25}$$

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

$$T_K(\phi_j) \;=\; \lambda_j \phi_j \;\Leftrightarrow\; \int_\Omega K(x, \tilde{x}) \phi_j(\tilde{x}) p(d\tilde{x}) \;=\; \lambda_j \phi_j(x) \tag{26}$$

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

$$f(x) \;=\; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \phi_j(x) \tag{27}$$

and

$$\begin{aligned} T_K(f) \;&=\; T_K(\sum_{j=1}^{\infty} \langle f, \phi_j \rangle \phi_j(x)) \;=\; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, T_K(\phi_j(x)) \\ &=\; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \lambda_j \, \phi_j(x) \,. \end{aligned} \tag{28}$$

Mercer Theorem iii

Furthermore,

$$\int_\Omega \int_\Omega K(x,\tilde{x})^2 p(d\tilde{x}) p(dx) \; = \; \sum_{j=1}^{\infty} \lambda_j^2 \; < \; \infty \,. \tag{29}$$

and, hence,

$$\lim_{j\to\infty} \lambda_j \; = \; 0\,. \tag{30}$$

Reproducing Kernel Hilbert Space i

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

$$T_K^{-1} f \;=\; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \lambda_j^{-1} \, \phi_j(x)\,. \tag{31}$$

By definition,

$$\begin{aligned} T_K(T_K^{-1} f) \;&=\; T_K \left(\sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \lambda_j^{-1} \, \phi_j(x) \right) \;=\; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \lambda_j^{-1} \, T_K \phi_j(x) \\ &=\; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \lambda_j^{-1} \, \lambda_j \phi_j(x) \;=\; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \phi_j(x) \;=\; f(x)\,, \end{aligned} \tag{32}$$

so this is indeed the inverse.

Reproducing 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 (30)). **Is a pure infinite-dimensional phenomenon.**

Reproducing Kernel Hilbert Space iii

- Define

$$\mathcal{H}_K = \left\{ f \in L_2(\Omega) : \sum_{j=1}^{\infty} \langle f, \phi_j \rangle^2 \lambda_j^{-1} < \infty \right\}, \tag{33}$$

and equip it with the inner product

$$\langle f, g \rangle_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \langle g, \phi_j \rangle \lambda_j^{-1} . \tag{34}$$

- Note that

$$\langle f, g\rangle_{\mathcal{H}_K} \;=\; \langle T_K^{-1} f, g\rangle_{L_2(\Omega)} \;=\; \int_\Omega f(x)(T_K^{-1} g)(x) p(dx) \tag{35}$$

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

$$\begin{aligned} g(x) \;=\; T_K(T_K^{-1} g)(x) \;&=\; \int K(x, \tilde{x})(T_K^{-1} g)(\tilde{x}) p(d\tilde{x}) \\ =\; \int_\Omega K_x(\tilde{x})(T_K^{-1} g)(x) p(dx) \;&=\; \langle K_x, g\rangle_{\mathcal{H}_K} . \end{aligned} \tag{36}$$

This remarkable identity is known as the reproducing kernel property.

- Hence, $\mathcal{H}_K$ is called a reproducing kernel Hilbert space.

OK, So what is RKHS?

Theorem

RKHS = set of functions for which (33) holds:

$$\mathcal{H}_K \;=\; \left\{ f \in L_2(\Omega) :\; \sum_{j=1}^{\infty} \langle f, \phi_j \rangle^2 \, \lambda_j^{-1} \;<\; \infty \right\} \tag{37}$$

- 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

$$\begin{aligned} \hat{f}(x) &= K(x,X)(zI+K(X,X))^{-1}y = \sum_{i=1}^{n} K(x,x_i)\xi_i, \\ \xi &= (zI+K(X,X))^{-1}y \in \mathbb{R}^n. \end{aligned} \tag{38}$$

Theorem

$K_{x_i}(x) = K(x,x_i) \in \mathcal{H}_K$ *always!*

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

- **Proof.** It follows from the definition of RKHS, but let us do a direct derivation; it is instructive. We have

$$K_{x_i}(x) \;=\; K(x, x_i) \;=\; \sum_{j=1}^{\infty} \underbrace{\lambda_j \phi_j(x_i)}_{basis\ \ coefficients} \quad \phi_j(x)\,, \tag{39}$$

implying that the basis coefficients are $\lambda_j \phi_j(x_i)$. Then, we need to check that they satisfy (33):

$$\sum_{j=1}^{\infty} (\lambda_j \phi_j(x_i))^2 \, \lambda_j^{-1} \;=\; \sum_{j=1}^{\infty} \lambda_j (\phi_j(x_i))^2 \;=\; K(x_i, x_i) \;<\; \infty\,. \tag{40}$$

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.

Theorem

$$\hat{f}(x) \;=\; K(x,X)(zI + K(X,X))^{-1} y\,. \tag{41}$$

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)

$$\hat{f} = \arg\min_{f \in \mathcal{H}_K} \left\{ \sum_{i=1}^{n} (y_i - f(x_i))^2 + z\|f\|^2_{\mathcal{H}_K} \right\}. \tag{42}$$

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

$$\hat{f} = \arg\min_{f \in \mathcal{H}_K} \{\|f\|^2_{\mathcal{H}_K} : (y_i - f(x_i)) = 0 \ \forall \ i\}. \tag{43}$$

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

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

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

$$f(x_i) \ = \ \langle f, K(\cdot, x_i)\rangle \ = \ \langle f_\parallel, K(\cdot, x_i)\rangle \ = \ f_\parallel(x_i) \tag{44}$$

so that

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

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:

$$\begin{aligned} H_0(x) \;&=\; 1, \quad H_1(x) \;=\; -\,e^{x^2}\left[-\,2x\,e^{-x^2}\right] \;=\; 2x, \\ H_2(x) \;&=\; e^{x^2}\left[(-2x)e^{-x^2} \;-\; 2\,e^{-x^2}\right] \;=\; 4x^2 \;-\; 2, \quad \dots \end{aligned} \tag{45}$$

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 $1/2$.

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 iii

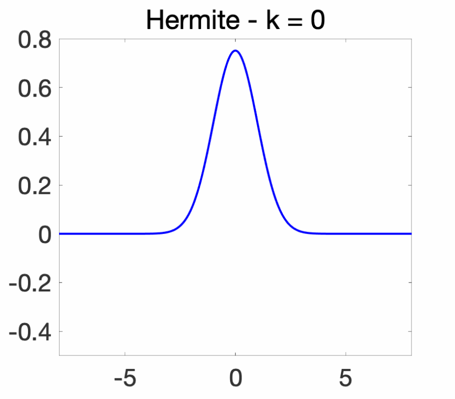


Figure: Hermite polynomial animation.

Define

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

Theorem

The Gaussian kernel admits the decomposition

$$K(x_1, x_2) = \exp\Big(-\tfrac{\rho}{1-\rho^2} (x_1 - x_2)^2\Big) = \sum_{k=0}^{\infty} (1-\rho)\,\rho^k\, f_k(x_1)\, f_k(x_2).$$

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

$$n^{-1}K(X,X) \;=\; \frac{1}{n}(K(x_i,x_j))_{i,j=1}^n \tag{46}$$

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

$$\phi_k(x) \;\approx\; \sum_{i=1}^n K(x,x_i)q_k(i)\,. \tag{47}$$

Heuristic Proof:

$$\begin{aligned} \lambda_k \phi_k(x) &= \int K(x, \tilde{x}) \phi_k(\tilde{x}) d\mu(\tilde{x}) \approx \sum_{i=1}^{n} K(x, \tilde{x}_i) \phi_k(\tilde{x}_i) \\ \hat{\lambda}_k q_k(i) &= \sum_{i=1}^{n} K(x_i, x_j) q_k(j) \end{aligned} \tag{48}$$

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

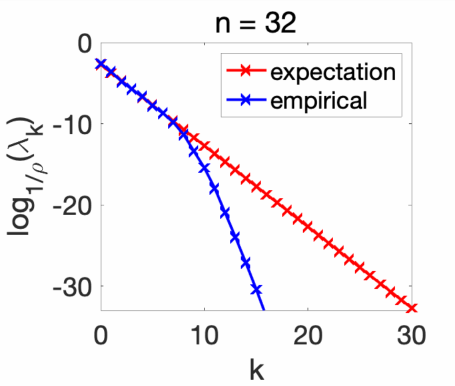


Figure: Convergence of Estimated Eigenvalues to $\lambda_k = (1-\rho)\,\rho^k$.

Ideal Kernel i

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

$$K_{ideal}(x, \tilde{x}) \;=\; f(x)\, f(\tilde{x}) \tag{49}$$

- Eigenfunction equation

$$\int K_{ideal}(x, \tilde{x})\psi(\tilde{x})\sigma(d\tilde{x}) \;=\; \lambda\, \psi(x) \tag{50}$$

takes the form

$$f(x) \int f(\tilde{x})\psi(\tilde{x})\sigma(d\tilde{x}) \;=\; \lambda\, \psi(x) \tag{51}$$

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

$$\lambda \;=\; \int f(\tilde{x})^2 \sigma(d\tilde{x}) \;=\; \left(\int\int K^2(x, \tilde{x})\sigma(dx)\sigma(d\tilde{x})\right)^{1/2}, \tag{52}$$

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

- with an ideal kernel,

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

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

$$\begin{aligned} K(x,X)(zI+K(X,X))^{-1} \;&=\; f(x)f(X)^\top(zI+f(X)f(X)^\top)^{-1} \\ =\; f(x)f(X)^\top\frac{z^{-1}}{1+z^{-1}\|f(X)\|^2} \;&=\; f(x)f(X)^\top\frac{1}{z+\|f(X)\|^2} \end{aligned} \tag{54}$$

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

$$\begin{aligned} \hat{f}(x) \;&=\; K(x,X)(zI+K(X,X))^{-1}y \\ &=\; f(x)f(X)^\top\frac{1}{z+\|f(X)\|^2}(f(X)+\varepsilon) \;=\; c\,f(x)\,, \end{aligned} \tag{55}$$

Ideal Kernel iii

where

$$c \;=\; \frac{\|f(X)\|^2 + f(X)^\top \varepsilon}{z + \|f(X)\|^2} \;=\; \frac{\frac{1}{n}\sum_i f(x_i)^2 + \frac{1}{n}\sum_i f(x_i)\varepsilon_i}{\frac{1}{n}z \;+\; \frac{1}{n}\sum_i f(x_i)^2} \approx\; 1 \quad (56)$$

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, $S_{k,t} = \frac{1}{P^{1/2}} f(X_t; \theta_k),\ k = 1, \cdots, P$. They have the true covariance matrix

$$E[S_t S_t'] \ = \ \Psi \tag{57}$$

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

$$\frac{1}{P} E[f(X; \theta_{j_1}) f(X; \theta_{j_2})] \ = \ \Psi_{j_1, j_2}\,. \tag{58}$$

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

$$\Psi h_j(j_1) \ = \ \hat{\lambda}_j\, h_j(j_1)\,. \tag{59}$$

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

$$\hat{\psi}_j(x) \ = \ \sum_{j=1}^{P} h_j(j_1)\, f(x; \theta_{j_1})\,. \tag{60}$$

Theorem

For a finite-dimensional kernel

$$K(x,\tilde{x}) \;=\; \frac{1}{P}\sum_{j=1}^{P} f(x;\theta_j)f(\tilde{x};\theta_j)\,, \tag{61}$$

the integral operator T_K *only has* P *non-zero eigenvalues* λ_j *coinciding with the eigenvalues of the matrix* Ψ,

$$\frac{1}{P}E[f(X;\theta_{j_1})f(X;\theta_{j_2})] \;=\; \Psi_{j_1,j_2}\,. \tag{62}$$

Furthermore, the eigenfunctions are given by

$$\psi_j(x) \;=\; \sum_{j_1=1}^{P} h_j(j_1)f(x;\theta_{j_1}) \tag{63}$$

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

$$f(x) \;=\; \sum_{j=1}^{\infty} \psi_j(x) c_j\,,\;\; where\;\; c_j \;=\; \langle \psi_j(x), f(x) \rangle \;=\; E[\psi_j(x) f(x)] \qquad (64)$$

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

$$\hat{f}_n(x) \;=\; K(x, X_n)^\top (zI + K(X_n, X_n))^{-1} y_n \qquad (65)$$

converges to the projection

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

Implicit Regularization i

Define the **implicit shrinkage**

$$Z_*(z;c) \;=\; \frac{z}{1 \;-\; c \;+\; c\,z\,m(-z;c)} \tag{66}$$

Recall to this end that

$$\begin{aligned} \tilde{m}(-z;c) \;&=\; z^{-1}(1-c+cz\hat{m}(-z)) \\ &=\; T^{-1}(T-P)z^{-1} + T^{-1}\,\mathrm{tr}((zI+S'S/T)^{-1})\,. \end{aligned} \tag{67}$$

If $P > T$, then $S'S/T$ is degenerate and has the same eigenvalues as SS'/T, plus $(P-T)$ zero eigenvalues. Similarly, if $P < T$, then SS'/T will have $T-P$ zero eigenvalues, so that we always have that

$$\mathrm{tr}((zI+S'S/T)^{-1}) \;=\; z^{-1}(P-T) \;+\; \mathrm{tr}((zI+SS'/T)^{-1}) \tag{68}$$

Implicit Regularization ii

Thus,

$$Z_*(z;c) \;=\; \frac{z}{T^{-1}\,\mathrm{tr}((zI + SS'/T)^{-1})} \;\approx\; \frac{z}{T^{-1}\,\mathrm{tr}((zI + K(X;X))^{-1})} \tag{69}$$

Implicit Regularization iii

Theorem

We have

$$z\, m(-z;c) \;=\; Z_*(z;c)\, m(-Z_*(z;c)) \tag{70}$$

That is, $(zI+\hat{\Psi})^{-1}$ *behaves as if we are doing* $(Z_* I + \Psi)^{-1}$.

Furthermore,

$$Z_* \;=\; z \;+\; cZ_* \int \frac{xdH(x)}{x+Z_*} \tag{71}$$

so that

$$Z_* \;\in\; [z,\; z+c]\,. \tag{72}$$

Implicit Regularization iv

Formally, in finite samples,

$$\begin{aligned} Z_* \;&=\; z \;+\; cZ_* \int \frac{xdH(x)}{x+Z_*} \;\approx\; z \;+\; Z_* \frac{P}{T} P^{-1}\,\mathrm{tr}(\Psi(\Psi + Z_* I)^{-1}) \\ &=\; z \;+\; Z_* T^{-1}\,\mathrm{tr}(\Psi(\Psi + Z_* I)^{-1}) \;\approx\; z \;+\; Z_* T^{-1}\,\mathrm{tr}(T_K(T_K + Z_* I)^{-1}) \end{aligned} \tag{73}$$

where

$$T_K \tag{74}$$

is the infinite-dimensional integral operator and

$$T^{-1}\,\mathrm{tr}(T_K(T_K + Z_* I)^{-1}) \;=\; T^{-1} \sum_{i=1}^{\infty} \lambda_i (\lambda_i + Z_* I)^{-1}$$

In most real world examples, $\lambda_i \sim i^{-\alpha},\ \alpha > 1.$

Implicit Regularization v

Let $\lambda_i = i^{-\alpha}$ for some $\alpha > 0$, and define the sum:

$$S(\kappa) := \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \kappa} = \sum_{i=1}^{\infty} \frac{i^{-\alpha}}{i^{-\alpha} + \kappa}.$$

We want to approximate $S(\kappa)$ for small $\kappa > 0$.

First, note that the function $f(i) = \frac{i^{-\alpha}}{i^{-\alpha}+\kappa}$ is positive and decreasing. We approximate the sum by an integral:

$$S(\kappa) \approx \int_1^{\infty} \frac{x^{-\alpha}}{x^{-\alpha} + \kappa}\, dx.$$

Make the substitution $t = x^{-\alpha}$, which implies $x = t^{-1/\alpha}$ and

$$dx = -\frac{1}{\alpha} t^{-1/\alpha - 1}\, dt.$$

Implicit Regularization vi

Changing variables gives:

$$\int_1^\infty \frac{x^{-\alpha}}{x^{-\alpha}+\kappa}\,dx = \int_0^1 \frac{t}{t+\kappa}\cdot\frac{1}{\alpha}t^{-1/\alpha-1}\,dt = \frac{1}{\alpha}\int_0^1 \frac{t^{-\frac{1}{\alpha}}}{t+\kappa}\,dt.$$

Now let $u = t/\kappa$, so that $t = \kappa u$ and $dt = \kappa du$. Then:

$$\frac{1}{\alpha}\int_0^{1/\kappa} \frac{(\kappa u)^{-\frac{1}{\alpha}}}{\kappa u+\kappa}\cdot\kappa\,du = \frac{1}{\alpha}\kappa^{1-\frac{1}{\alpha}}\int_0^{1/\kappa}\frac{u^{-\frac{1}{\alpha}}}{u+1}\,du.$$

As $\kappa \to 0$, $1/\kappa \to \infty$, and we get:

$$S(\kappa) \sim \frac{1}{\alpha}\kappa^{1-\frac{1}{\alpha}}\int_0^\infty \frac{u^{-\frac{1}{\alpha}}}{u+1}\,du.$$

Implicit Regularization vii

The integral is known:

$$\int_0^\infty \frac{u^{s-1}}{u+1}\, du = \pi / \sin(\pi s), \quad \text{for } 0 < s < 1.$$

Setting $s = 1 - \frac{1}{\alpha}$, we get:

$$\int_0^\infty \frac{u^{-\frac{1}{\alpha}}}{u+1}\, du = \pi / \sin(\pi/\alpha).$$

Putting it all together:

$$S(\kappa) = \frac{\pi}{\alpha \sin(\pi/\alpha)} \kappa^{-1/\alpha} + O(1).$$

Thus,

$$Z_* \;\approx\; z \;+\; Z_* T^{-1} \frac{\pi}{\alpha \sin(\pi/\alpha)} Z_*^{-1/\alpha}$$

Implicit Regularization viii

For $z = 0$, we get $Z_* \sim T^{-\alpha}$. Thus, we have about T eigenvalues $i^{-\alpha} > T^{-\alpha}$.

Setup: Mercer's Theorem

Let $K : \Omega \times \Omega \to \mathbb{R}$ be a continuous, symmetric, positive-definite kernel on a compact domain $\Omega \subset \mathbb{R}^d$.

Define the integral operator:

$$(T_K f)(x) := \int_\Omega K(x, x') f(x') \, dx'$$

Mercer's theorem gives:

$$K(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(x')$$

λ_i: eigenvalues, ϕ_i: orthonormal eigenfunctions.

Theorem 1: Sobolev Kernel Decay

Statement: If the RKHS $\mathcal{H}_K$ is norm-equivalent to the Sobolev space $H^s(\Omega)$, then:

$$\lambda_i \asymp i^{-2s/d}$$

Conditions:

- $\Omega \subset \mathbb{R}^d$ compact with Lipschitz boundary.
- K is sufficiently smooth.

Theorem 2: Analytic Kernels

Statement: If $K \in C^\infty(\Omega \times \Omega)$ is real-analytic, then:

$$\lambda_i \leq C \exp(-c\, i^{1/d})$$

Implication: Exponential or super-polynomial decay of eigenvalues.

- Suppose $y = f(x)$
- $$\frac{1}{n}K(X,X) \;=\; \sum_{j=1}^{\infty} \lambda_j \frac{1}{n}\psi_j(X)\,\psi_j(X)^\top \;\in\; \mathbb{R}^{n\times n} \tag{75}$$
- with many train observations, we have

$$\frac{1}{n}\psi_{j_1}(X)^\top\psi_{j_2}(X) \;=\; \frac{1}{n}\sum_i \psi_{j_1}(x_i)\psi_{j_2}(x_i) \underbrace{\approx}_{LLN} E[\psi_{j_1}(x)\psi_{j_2}(x)] \;=\; 0 \tag{76}$$

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

$$(zI + n^{-1}K(X,X))^{-1} \;\approx\; \sum_{j=1}^{\infty} \underbrace{(\lambda_j + Z_* I)^{-1}}_{implicit\;\;regularization} \frac{1}{n}\psi_j(X)\,\psi_j(X)^\top \tag{77}$$

and

$$
\begin{aligned}
(zI + n^{-1}K(X,X))^{-1}y \;&=\; \sum_{j=1}^{\infty} \underbrace{(\lambda_j + Z_* I)^{-1}}_{\textit{implicit regularization}} \;\frac{1}{n}\psi_j(X)\,\psi_j(X)^\top y \\
&=\; \sum_{j=1}^{\infty} \underbrace{(\lambda_j + Z_* I)^{-1}}_{\textit{implicit regularization}} \;\frac{1}{n}\psi_j(X)\,\psi_j(X)^\top f(X) \\
&\approx\; \sum_{j=1}^{\infty} \underbrace{(\lambda_j + Z_* I)^{-1}}_{\textit{implicit regularization}} \;\psi_j(X)\,c_j
\end{aligned}
\tag{78}
$$

- and, hence,

$$
\begin{aligned}
\hat{f}(x) &= \frac{1}{n} K(x, X)(zI + n^{-1} K(X, X))^{-1} y \\
&= \sum_{j=1}^{\infty} \lambda_j \frac{1}{n} \psi_j(x)\, \psi_j(X)^\top \sum_{j_1=1}^{\infty} \underbrace{(\lambda_{j_1} + Z_* I)^{-1}}_{implicit\ \ regularization} \ \psi_{j_1}(X)\, c_{j_1} \\
&\underbrace{\approx}_{E[\psi_j \psi_{j_1}] = \delta_{j,j_1}} \sum_{j=1}^{\infty} \lambda_j \psi_j(x)\, c_j (\lambda_j + Z_* I)^{-1}
\end{aligned} \tag{79}
$$

so that

$$
\begin{aligned}
f(x) &= \sum_{j=1}^{\infty} c_j \psi_j(x) \\
\hat{f}(x) &\approx \sum_{j=1}^{\infty} c_j \frac{\lambda_j}{\lambda_j + Z_*} \psi_j(x) = \sum_{j=1}^{\infty} c_j \frac{1}{1 + Z_* / \lambda_j} \psi_j(x)
\end{aligned} \tag{80}
$$

where Z_* is the effective shrinkage (implicit regularization).

- It kills eigenvalues below Z_* and is irrelevant for eigenvalues above that threshold.
- It is all about alignment: $\hat{f}(x)$ is close to $f(x)$ if and only if f is aligned with top eigenfunctions of K

Multiple Descent

Multiple Descent

Kernels and Gaussian Processes i

- A Gaussian Process is a Gaussian distribution on **functions**.
- Basically, instead of randomly sampling a vector $f \in \mathbb{R}^P$, nature randomly samples a whole function $f(x) : \mathbb{R}^d \to \mathbb{R}$. Each value $x \in \mathbb{R}^d$ is a "coordinate" (**thus, there is a continuum of coordinates**)
- A Gaussian vector $f \in \mathbb{R}^P$ is defined by a μ, Σ : $\mu(i) = E[f(i)],\ \Sigma(i,j) \ = \ \mathrm{Cov}(f(i), f(j))$
- In a Gaussian process, we just replace coordinates $i, j = 1, \cdots, P$ with variables $x_1, x_2 \in \mathbb{R}^d$. Thus, a Gaussian process is defined by

$$\mu(x) \ = \ E[f(x)] \tag{81}$$

and

$$K(x, \tilde{x}) \ = \ E[f(x)\, f(\tilde{x})] \tag{82}$$

Here, K is the **covariance kernel of the Gaussian Process**

- Standard notation: $f(x) \sim \mathcal{GP}(\mu(x), \Sigma(x))$
- Suppose now we are trying to learn a function $f(x)$ from many observations. We observe $f(x_i)$, $i = 1, \cdots, n$ and we want to learn $f(x)$. Well, we can use the **Gaussian conditioning formula**:

$$\begin{aligned}
\begin{pmatrix} X \\ Y \end{pmatrix} &\sim N(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}) \\
E[X|Y] &= \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mu_Y) \\
\mathrm{Var}[X|Y] &= \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}
\end{aligned} \tag{83}$$

Similarly,

$$\begin{pmatrix} f(x) \\ f(X) \end{pmatrix} \sim N(\begin{pmatrix} \mu(x) \\ \mu(X) \end{pmatrix}, \begin{pmatrix} K(x,x) & K(x,X) \\ K(X,x) & K(X,X) \end{pmatrix}) \tag{84}$$

and, hence

$$E[f(x)|(f(x_1),\cdots,f(x_n))] = \underbrace{\mu(x)}_{=E[f(x)]}$$
$$+ \underbrace{K(x,X)}_{=\mathrm{Cov}(f(x),f(X))} \underbrace{K(X,X)^{-1}}_{\mathrm{Var}[f(X)]^{-1}}(f(X) - \mu(X)) \tag{85}$$
$$\mathrm{Var}[f(x)|(f(x_1),\cdots,f(x_n))]$$
$$= \underbrace{K(x,x)}_{=\mathrm{Var}[f(x)]} - K(x,X)\,K(X,X)^{-1}K(X,x)$$

- If we have multiple OOS points, we get that, from the **Bayesian point of view,** we have that

$$\mathrm{Var}[f(X_{OOS})|(f(x_1),\cdots,f(x_n))]$$
$$= K(X_{OOS},X_{OOS}) - K(X_{OOS},X)\,K(X,X)^{-1}K(X,X_{OOS}) \tag{86}$$

Kernels and Gaussian Processes iv

- Similarly, if we observe $f(x_i) + \varepsilon_i$, we replace $K(X, X)$ with $K(X, X) + \sigma_\varepsilon^2 I$.