Chapter 4, Part 4: Constructing Kernels

Advanced Topics in Statistical Machine Learning

Eugenio Clerico

Hilary 2026

eugenio.clerico@stats.ox.ac.uk

Constructing Kernels

There are three equivalent ways of constructing a kernel:

- Defining a feature map $\phi(x)$ and then taking the inner product: $K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$
- As a positive definite function: $\sum_i \sum_j a_i a_j K(x_i, x_j) \geq 0$
- By choosing an RKHS $\mathcal{H}$ and then considering its reproducing kernel $\mathcal{H}$

Yet in practice one often starts from some *simple* kernels, and use them to construct new kernels! We now see a few ways to do so.

Some simple bricks to construct kernels

The following functions are kernels:

- $K(x, y) = c$, for some constant $c \geq 0$.
- $K(x, y) = x \cdot y$, for $x, y \in \mathbb{R}^d$.
- $K(x, y) = \phi(x) \cdot \phi(y)$, for $\phi : \mathcal{X} \to \mathbb{R}^d$.
- $K(x, y) = \delta_{x,y}$ (1 if $x = y$ and 0 ow), for any $\mathcal{X}$.
- $K(x, y) = \min(x, y)$, for $x \geq 0$, $y \geq 0$.[1]
- $K(x, y) = \cos(x - y)$, for $x, y \in \mathbb{R}$.[2]

[1] $\min(x, y) = \int_{[0,\infty)} \mathbf{1}_{[0,x]}(u)\mathbf{1}_{[0,y]}(u)\mathrm{d}u$, which is an inner product in $L^2([0, \infty))$.

[2] $\cos(x - y) = \cos x cos y + \sin x \sin y$, which is an inner product in $\mathbb{R}^2$.

Some examples of non-kernels

The following functions are **not** kernels:

- $K(x, y) = c$, for $c < 0$.
  (The Gram matrix for a single point is (c), which is not PSD.)
- $K(x, y) = x + y$, for $x, y \in \mathbb{R}$.
  (Cauchy-Schwarz in $\mathcal{H}$ reads $|K(x, y)|^2 \leq K(x, x)K(y, y)$.
  This fails to be true for $x = 0$, $y = 1$.)
- $K(x, y) = \max(x, y)$, for $x, y \in \mathbb{R}$.
  (Again Cauchy-Schwarz fails for $x = 0$, $y = 1$.)
- $K(x, y) = \sin(x - y)$, for $x, y \in \mathbb{R}$.
  (This is not even symmetric!)

Linear combination of kernels

Lemma 1

Given kernels $K_1, \ldots, K_d$ on $\mathcal{X}$ and **non-negative** *coefficients $\alpha_1, \ldots \alpha_d > 0$, then $K = \sum_{i=1}^{d} \alpha_i K_i$ is also a kernel on $\mathcal{X}$.*

Symmetry is trivial. We need to check positive definiteness. Pick $(x_1, \ldots, x_N) \in \mathcal{X}^N$, and $v \in \mathbb{R}^N$. We have

$$v^\top \hat{K} v = \sum_{i=1}^{d} \alpha_i v^\top \hat{K}_i v \geq 0\,,$$

where the hat denotes the Gram matrix of the respective kernel.

Note:

For negative coefficients the property might fail!

$K_1 - K_2$ is not a kernel in general!

Product of kernels

Lemma 2

Let K' and K'' be kernels on $\mathcal{X}$. Then K defined as $K(x,y) = K'(x,y)K''(x,y)$ is a kernel.

- Symmetry is trivial. We need to check positive definiteness.
- **Fact:** If A is a PSD $n \times n$ matrix, there is a $n \times \mathrm{rank}(A)$ matrix B such that $A = BB^\top$.[3]
- Fix $(x_1, \ldots, x_N) \in \mathcal{X}^N$, and $v \in \mathbb{R}^N$: $\hat{K}' = CC^\top$, $\hat{K}'' = DD^\top$.
- We have $\hat{K}'_{ij} = \sum_p C_{ip}C_{jp}$, $\hat{K}''_{ij} = \sum_q D_{iq}D_{jq}$.
- $\hat{K}_{ij} = \hat{K}'_{ij}\hat{K}''_{ij} = \sum_{p,q}(C_{ip}D_{iq})(C_{jp}D_{jq})$.
- $v^T\hat{K}v = \sum_{i,j,p,q} v_i v_j C_{ip}D_{iq}C_{jp}D_{jq} = \sum_{p,q}\left(\sum_i v_i C_{ip}D_{iq}\right)^2 \geq 0$.

[3] You might notice that this can be seen as a consequence of Moore-Aronszajn thm, as A is a kernel on $\mathbb{R}^n$... Althought this is a bit like using a sledgehammer to crack a nut.

Pullback of a kernel

Lemma 3

Let K be a kernel on $\mathcal{X}$ and $\psi : \mathcal{Z} \to \mathcal{X}$. Then K_f, defined as $K_\psi(z, z') = K(\psi(z), \psi(z'))$ is a kernel on $\mathcal{Z}$.

K is clearly symmetric. Let $(\mathcal{W}, \phi)$ be a Hilbert space and feature map such that $K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{W}}$. Then $K_\psi(z, z') = \langle \phi \circ \psi(z), \phi \circ \psi(z') \rangle_{\mathcal{W}}$, so $\phi \circ \psi : \mathcal{Z} \to \mathcal{W}$ is a valid feature map for K_ψ.

Tensor product of kernels

Lemma 4

Let K' be a kernel on $\mathcal{X}$ and K'' be a kernel on $\mathcal{Y}$. Then $K = K' \otimes K''$ defined as $K((x,y),(x',y')) = K'(x,x')K''(y,y')$ is a kernel on $\mathcal{X} \times \mathcal{Y}$.

Let $\psi_1 : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ be given by $\psi_1(x,y) = x$, and $\psi_2 : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ by $\psi_2(x,y) = y$. Then K'_{ψ_1} and K'_{ψ_2} are the pulled back kernels on $\mathcal{X} \times \mathcal{Y}$. K is the component-wise product of these two kernels, hence a kernel.

Limit of kernels

Lemma 5

Let $(K_i)_{i\geq 1}$ be a sequence of kernels on $\mathcal{X}$. If $K_i \to K$ point-wise, then K is a kernel.

Symmetry is trivial. For positive definiteness, point-wise convergence implies convergence of the Gram matrices, and a sequence of PSD matrices converges to a PSD matrix (you can check it as an exercise...).

Power series of kernels

Lemma 6

Let K be a kernel on $\mathcal{X}$. Let $f : (-a, a) \to \mathbb{R}$ (with $a > 0$, possibly $a = \infty$) be a function that can be expressed as a power series:

$$f(u) = \sum_{t=0}^{\infty} c_t u^t \,, \quad \forall u \in (-a, a) \,.$$

Assume that $c_t \geq 0$ for all t and $|K(x, y)| < a$ for all x, y. Then $f \circ K$ is a kernel on $\mathcal{X}$.

For each finite t we have that K^t is a kernel, as a product of kernels. In particular, for each T we have that $S_T = \sum_{t=0}^{T} c_t K^t$ is a kernel as a linear combination of kernels, with non-negative coefficients. Since $f \circ K = \lim_{T \to \infty} S_T$ (point-wise), it is a kernel.

Application: RBF is a kernel

We can prove (without using the feature expansion) that RBF is a kernel from the previous results.

Recall that the RBF kernel is

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\gamma^2}\right) .$$

We can write

$$K(x, y) = \exp\left(-\frac{\|x\|^2}{2\gamma^2}\right) \exp\left(-\frac{\|y\|^2}{2\gamma^2}\right) \exp\left(\frac{x \cdot y}{\gamma^2}\right) .$$

Note that $(x, y) \mapsto x \cdot y$ is a kernel, so $e^{x \cdot y/\gamma^2}$ is a kernel as the Taylor expansion of $\exp$ has all non-negative coefficients!

Note that $(x, y) \mapsto \phi(x)\phi(y)$ ia a kernel for any real-valued function ϕ (which is a feature map on $\mathbb{R}$).

So $e^{-\|x\|^2/(2\gamma^2)} e^{-\|y\|^2/(2\gamma^2)}$ is a kernel (pick $\phi(x) = e^{-\|x\|^2/(2\gamma^2)}$).

We conclude that K is a kernel as the product of two kernels.

Functions in RKHSs

- For every RKHS $\mathcal{H}$, the linear span $\mathrm{Span}\{k_x : x \in \mathcal{X}\}$ is **dense** in $\mathcal{H}$ (wrt to the norm $\|\cdot\|_{\mathcal{H}}$).[4]
- This means that every $h \in \mathcal{H}$ can be written as
$$h(x) = \sum_{t=1}^{\infty} \alpha_t K(x_t, x)\,,$$
for some $\alpha_t \in \mathbb{R}$, $x_t \in \mathcal{X}$, where the series converges in $\mathcal{H}$.
- Note that convergence in $\mathcal{H}$ implies convergence point-wise, but the reverse is not true in general (a series might converge point-wise to a function which is not in the RKHS)!
- A necessary and sufficient condition for $h = \sum_{t=1}^{\infty} \alpha_t k_{x_t}$ to be in $\mathcal{H}$ is $S_\alpha = \sum_{s,t=1}^{\infty} \alpha_s \alpha_t K(x_s, x_t) < \infty$.
In such case $\|h\|_{\mathcal{H}} = \sqrt{S_\alpha}$.

[4] The simplest way to see this is to notice that the orthogonal to this span is $\{0\}$ by the reproducing property (it must be that $h(x) = \langle k_x, h\rangle_{\mathcal{H}} = 0\ \forall x$, if h is in the orthogonal).

Visualising the RBF's RKHS

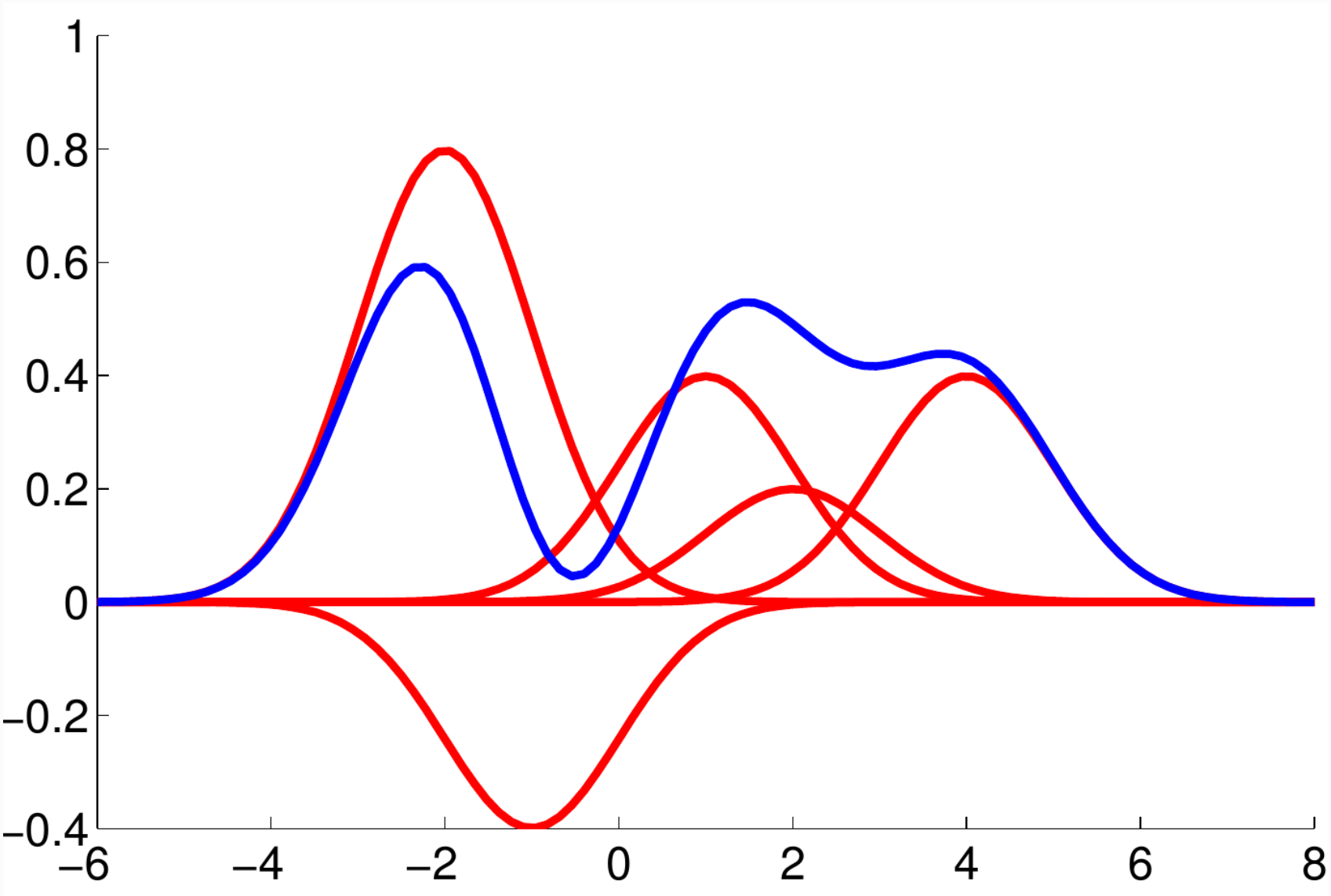

Figure 1: Visualizing $\mathcal{H}$ for $k(x, x') = \exp\left(-\frac{1}{2}(x - x')^2\right)$

Regularity of functions

The regularity of K determines the regularity of the functions in $\mathcal{H}$:

- If K is continuous, every $h \in \mathcal{H}$ is continuous.
  $|h(x) - h(y)| \leq \|h\|_{\mathcal{H}} \|k_x - k_y\|_{\mathcal{H}}$ by Cauchy-Schwarz, and
  $\|k_x - k_y\|_{\mathcal{H}}^2 = K(x,x) + K(y,y) - 2K(x,y) \to 0$, as $y \to x$.
- If K is C^{2n} ($2n$ times differentiable with continuous derivatives), then every $h \in \mathcal{H}$ is C^n.
- If K is smooth (C^∞), then every $h \in \mathcal{H}$ is smooth.

In particular, every function in the RKHS generated by the RBF kernel is smooth!

Matérn Kernels

The smoothness of the RBF kernel is often overly restrictive. Matérn kernels K_ν, indexed by $\nu > 0$, allow for less regular functions.

K_ν is of class $C^{\lceil 2\nu \rceil - 1}$, so all functions in the associated RKHS are s-times differentiable if $\nu > s$.

Though we omit their full form here, we note they have simplified forms when $\nu = s + 1/2$ (here $\gamma > 0$ is a scaling hyperparameter):

- $\nu = \frac{1}{2}$: $K(x, x') = \exp\left(-\frac{1}{\gamma} \|x - x'\|_2\right)$,
- $\nu = \frac{3}{2}$: $K(x, x') = \left(1 + \frac{\sqrt{3}}{\gamma} \|x - x'\|_2\right) \exp\left(-\frac{\sqrt{3}}{\gamma} \|x - x'\|_2\right)$.

Exercise (non too trivial): prove explicitly that $h \in \mathcal{H}$ is differentiable for $\nu = \frac{3}{2}$.

Matérn Kernels

- As $\nu \to \infty$ the Matérn kernel converges to the RBF kernel.
- It can be shown $\|f\|^2_{\mathcal{H}}$ directly penalises the derivatives.

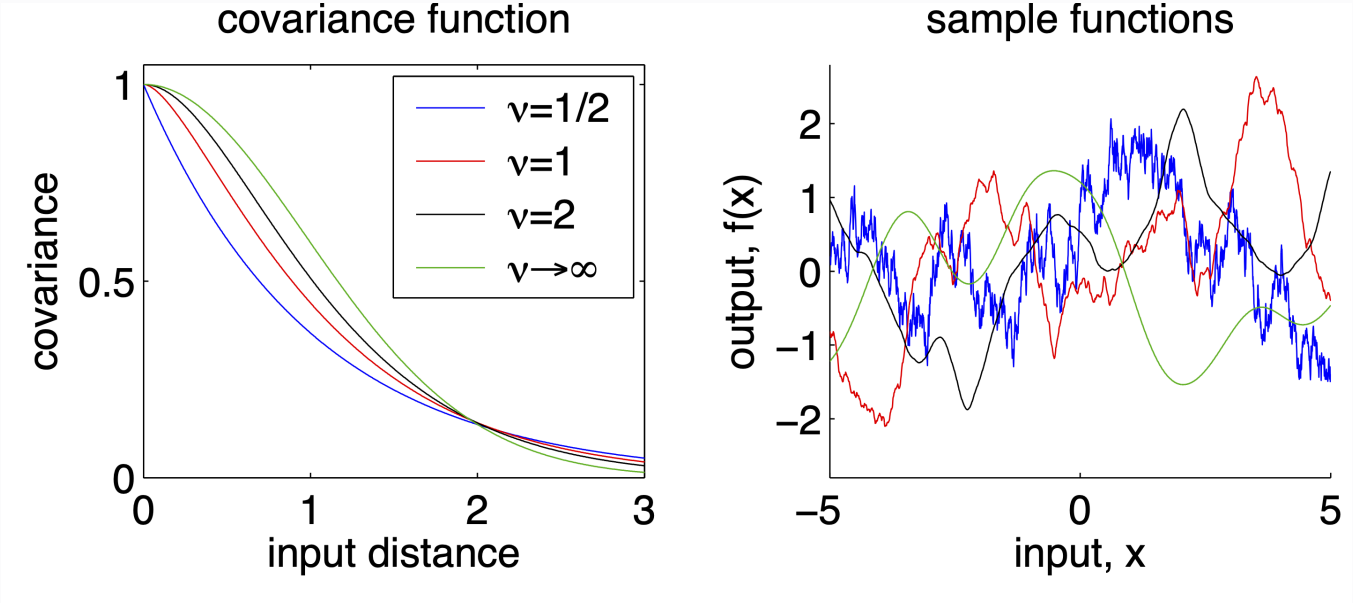

Figure 2: Characterisation of Matérn kernels. Source: Rasmussen and Williams, Gaussian Processes for Machine Learning, 2005

Some other commonly used kernels on $\mathbb{R}^n$

- **Polynomial**: $K(x, x') = (c + x^\top x')^m$, $c \in \mathbb{R}$, $m \in \mathbb{N}$ ($m = 1$ gives affine kernel).
- **Exponential**: $K(x, x') = \exp\left(\frac{x^\top x'}{\gamma}\right)$, $\gamma > 0$.
- **Laplace**: $K(x, x') = \exp\left(-\frac{1}{\gamma}\|x - x'\|_1\right)$, $\gamma > 0$ (Similar to Matérn $1/2$, but here we have the norm $\|u\|_1 = \sum_{t=1}^n |u_t|$).

Exercise: Use the rules we derived at the beginning of this lecture to show that these are kernels, starting from *simpler* kernels.

(Hint: For the Laplace kernel, first consider the case $n = 1$, then extend to $n > 1$ using tensor products of kernels. For $n = 1$, you might use that $K_m(x, y) = \max(0, 1 - |x - y|/m)$ is a kernel on $\mathbb{R}$, as it is can be written as an inner product in L^2:
$K_m(x, y) = \frac{1}{n}\int_{\mathbb{R}} \mathbf{1}_{[x,x+m]}(u)\mathbf{1}_{[y,y+m]}(u)\mathrm{d}u$.)

Kernel Ridge Regression

Kernel ridge regression is the kernelised version of regularised least squares linear regression

$$h^\star = \arg\min_{h \in \mathcal{H}} \left(\frac{1}{2} \sum_{i=1}^{n} (y_i - h(x_i))^2 + \lambda \|h\|_{\mathcal{H}}^2 \right) .$$

By the representer theorem, $h^\star = \sum_{i=1}^{n} \alpha_i^\star K(\cdot, x_i)$.
Replacing $h^\star$ with this expression and plugging it into the optimisation objective, we reduce the problem to a quadratic optimisation problem in $\mathbb{R}^n$:

$$\alpha^\star = \arg\min_{\alpha \in \mathbb{R}^n} \left(\frac{1}{2} \|Y - \hat{K}\alpha\|^2 + \lambda \alpha^\top \hat{K} \alpha \right) , \quad Y = (y_1 \dots y_n)^\top ,$$

whose solution has close form (see problem sheet).

Hyperparameters

- Even if the RKHS is very general, hyperparameters can still heavily influence what is learned in practice for **finite** data
- In particular, the common parameters of the length scale γ and regularisation strength λ can be particularly important.

[Coding examples]

Limitations of Kernels in High Dimensions

- Many common kernels are based only on Euclidean distances between points in the original space, e.g.
  $K(x, x') = \exp\left(-\frac{1}{2\gamma^2}\|x - x'\|_2^2\right)$
- This can lead to poor performance in high dimensions where such pairwise distances are not very informative: all points may be quite far away from each other
- This is not a limitation of kernel methods per se, put reflects the difficulty of constructing appropriate kernels for high dimensional problems
  - Here the machine learning challenge is typically more that of asserting which points are similar than it is of ensuring our predictor is sufficiently powerful; using the kernel trick is of limited help in this endeavor

Further Reading

- Go have a play: these things are super easy to code up and have a mess around with them is a good way to develop an understanding
- Chapter 4 of Carl Edward Rasmussen and Christopher Williams. **Gaussian Processes for Machine Learning.** The MIT Press, 2005, `http://www.gaussianprocess.org/gpml/chapters/` (will require some knowledge of Gaussian processes that will be covered later in the course)