

# CPSC 532D — 11. KERNELS

Danica J. Sutherland

University of British Columbia, Vancouver

Fall 2024

---

We've mentioned a couple times the idea of implementing a polynomial classifier as a special case of a linear one: in  $\mathbb{R}$ , a cubic classifier might look like

$$h(x) = w_0 + w_1x + w_2x^2 + w_3x^3$$

where we have four parameters in  $w$ . Notice that we can also write this as

$$h(x) = w \cdot \phi(x), \quad w \in \mathbb{R}^4, \quad \phi(x) = (1, x, x^2, x^3).$$

Now, consider the set of all cubic functions

$$\mathcal{F} = \{x \mapsto w \cdot \phi(x) = w_0 + w_1x + w_2x^2 + w_3x^3 : w \in \mathbb{R}^4\}.$$

We're going to introduce some machinery to think about  $\mathcal{F}$  as a function space, along the lines of the space  $C(\mathcal{X})$  from Definition 10.1. This will lead to *kernel methods* that allow us to optimize over  $\mathcal{F}$  using basically the same techniques as optimizing over linear spaces.

*“Kernel” is a super-overloaded word. This is not the same thing as in kernel density estimation, the kernel of a convolution, the kernel of a probability density, the kernel of a linear map, a CUDA kernel, an operating system kernel...*

## 11.1 DEFINING FUNCTION SPACES

To think of  $\mathcal{F}$  as a vector space of functions, let  $f, f' \in \mathcal{F}$  correspond to weight vectors  $w, w'$ . Then we can let  $f + f'$  be the function with weight vector  $w + w'$ , and  $af$  that with weight vector  $aw$ . This definition makes it a valid **vector space**:

**DEFINITION 11.1.** A real *vector space* is a non-empty set  $V$  along with the operations of *vector addition*, denoted  $v + w \in V$  for any  $v, w \in V$ , and *scalar multiplication*, denoted  $av \in V$  for any  $v \in V$  and  $a \in \mathbb{R}$ , satisfying the following requirements:

- Vector addition is associative: for all  $u, v, w \in V$ ,  $u + (v + w) = (u + v) + w$ .
- Vector addition is commutative: for all  $v, w \in V$ ,  $v + w = w + v$ .
- Vector addition has an identity: there is some *zero vector*  $0 \in V$  such that for all  $v \in V$ ,  $v + 0 = v$ .
- Vector addition has inverses: for each  $v \in V$ , there is some  $-v \in V$  such that  $v + (-v) = 0$ .
- Compatibility of scalar multiplication: for all  $a, b \in \mathbb{R}$  and  $v \in V$ ,  $a(bv) = (ab)v$ .
- Identity of scalar multiplication: for all  $v \in V$ ,  $(1)v = v$
- Distributive property I: for all  $a \in \mathbb{R}$  and  $v, w \in V$ ,  $a(v + w) = av + aw$ .
- Distributive property II: for all  $a, b \in \mathbb{R}$  and  $v \in V$ ,  $(a + b)v = av + bv$ .

---

For more, visit <https://cs.ubc.ca/~dsuth/532D/24w1/>.

---

A lot of the familiar linear algebra stuff you know and love from  $\mathbb{R}^d$  applies to any vector space as well.

**DEFINITION 11.2.** A real *normed vector space* is a real vector space  $V$  with a *norm*: a function  $V \rightarrow \mathbb{R}$ , written  $\|v\|$ , such that:

- Non-negativity: for all  $v \in V$ ,  $\|v\| \geq 0$ .
- Positive definiteness: for every  $v \in V$ ,  $\|v\| = 0$  if and only if  $v = 0$ .
- Absolute homogeneity: for every  $a \in \mathbb{R}$  and  $v \in V$ ,  $\|av\| = |a|\|v\|$ .
- Sub-additivity / triangle inequality: for every  $v, w \in V$ ,  $\|v + w\| \leq \|v\| + \|w\|$ .

The norm of a normed vector space induces the metric  $\rho(x, y) = \|x - y\|$ , which we can check satisfies the formal definition of a metric space:

**DEFINITION 11.3.** A *metric space* is a set  $\mathcal{X}$  along with a function  $\rho : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ , called the *metric*, satisfying the following properties:

- Non-negativity: for all  $x, y \in \mathcal{X}$ ,  $\rho(x, y) \geq 0$ .
- Positive definiteness for all  $x, y \in \mathcal{X}$ ,  $\rho(x, y) = 0$  if and only if  $x = y$ .
- Symmetry: for all  $x, y \in \mathcal{X}$ ,  $\rho(x, y) = \rho(y, x)$ .
- Triangle inequality: for all  $x, y, z \in \mathcal{X}$ ,  $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ .

**DEFINITION 11.4.** Consider a sequence  $x_1, x_2, \dots$  in a metric space  $\mathcal{X}$ .

This sequence has a *limit*  $x_\infty$  if for every  $\varepsilon > 0$ , there exists a positive integer  $N$  such that for all  $n > N$ ,  $\rho(x_n, x_\infty) < \varepsilon$ .

This sequence is called *Cauchy* if, for every  $\varepsilon > 0$ , there exists a positive integer  $N$  such that for all  $m, n > N$ ,  $\rho(x_m, x_n) < \varepsilon$ .

The metric space  $\mathcal{X}$  is called *complete* if all Cauchy sequences in  $\mathcal{X}$  have limits in  $\mathcal{X}$ .

**DEFINITION 11.5.** A real *Banach space* is a real normed vector space whose norm induces a complete vector space.

You can check that  $C(\mathcal{X})$  is a Banach space.

There's one other major structure in  $\mathbb{R}^d$  that we don't have yet: dot products.

**DEFINITION 11.6.** A real *inner product space* is a real vector space  $V$  together with an inner product, a function  $V \times V \rightarrow \mathbb{R}$  written  $\langle v, w \rangle$  satisfying

- Symmetry: for all  $v, w \in V$ ,  $\langle v, w \rangle = \langle w, v \rangle$ .
- Linearity: for all  $u, v, w \in V$  and  $a, b \in \mathbb{R}$ ,  $\langle au + bv, w \rangle = a\langle u, w \rangle + b\langle v, w \rangle$ .
- Positive-definiteness: if  $v \neq 0$ , then  $\langle v, v \rangle > 0$ .

An inner product space is also a normed vector space with  $\|v\| = \sqrt{\langle v, v \rangle}$ , and hence a metric space with  $\rho(v, w) = \|v - w\| = \sqrt{\langle v - w, v - w \rangle}$ .

**DEFINITION 11.7.** A real *Hilbert space* is a real inner product space whose induced metric space is complete.

---

## 11.2 POLYNOMIAL FUNCTIONS

Now, recall our function space

$$\mathcal{F} = \{x \mapsto w \cdot \phi(x) = w_0 + w_1x + w_2x^2 + w_3x^3 : w \in \mathbb{R}^4\}$$

with addition defined by adding weight vectors, and scalar multiplication by scaling the weight vectors. We can also define an inner product  $\langle f, f' \rangle_{\mathcal{F}}$  by  $w \cdot w'$ , also giving the norm  $\|f\|_{\mathcal{F}} = \|w\|$ . We can check that this satisfies all the conditions we need, including completeness, for  $\mathcal{F}$  to define a Hilbert space.

Now, let's think about a different function class. Choose any  $c > 0$  and define

$$\mathcal{F}_c = \{x \mapsto w \cdot \phi(x) = w_0\sqrt{c^3} + w_1\sqrt{3c^2}x + w_2\sqrt{3c}x^2 + w_3x^3 : w \in \mathbb{R}^4\},$$

then again define addition / scalar multiplication / inner products in terms of *these* weight vectors  $w$ . The reason for this reparameterization is that we get

$$\phi(x) \cdot \phi(x') = c^3 + 3c^2xx' + 3c(xx')^2 + (xx')^3 = (xx' + c)^3,$$

which makes  $\phi(x) \cdot \phi(x')$  much easier to compute. The same thing happens in higher dimensions or with higher polynomial degrees; for degree- $\ell$  polynomials in  $d$  dimensions, there are  $\mathcal{O}(d^\ell)$  parameters, but we can compute this inner product  $\phi(x) \cdot \phi(x')$  still in  $\mathcal{O}(d)$  time.

We call this function  $\phi(x) \cdot \phi(x')$  the *kernel function*:

$$k(x, x') = \phi(x) \cdot \phi(x').$$

We'll see soon that it's a very fundamental object.

The set of functions in  $\mathcal{F}$  and  $\mathcal{F}_c$  for any  $c$  are the same, as functions; addition and scalar multiplication also agree between all of them. But the inner product doesn't! So  $\|w\|$ , and hence  $\|f\|_{\mathcal{F}_c}$ , is different depending on your choice of  $c$ . (Larger  $c$  will mean the lower-order coefficients can be smaller in order to express the same function, and so means that  $\|f\|_{\mathcal{F}}$  is more determined by the coefficient on  $x^3$ .) This will be important when we use algorithms that depend on  $\|f\|_{\mathcal{F}}$ .

Now, let's do something slightly weird. Recall that

$$\phi(x) = (\sqrt{c^3}, \sqrt{c^2}x, \sqrt{c}x^2, x^3) \in \mathbb{R}^4.$$

Elements of  $\mathcal{F}_c$  are functions corresponding to any  $w \in \mathbb{R}^4$ . So what happens if we think of the element of  $\phi(x)$  as a weight vector for an element in  $\mathcal{F}_c$ ? This would give us a function of the form

$$\begin{aligned} x' \mapsto & \sqrt{c^3}\sqrt{c^3} + \sqrt{3c^2}x\sqrt{3c^2}x' + \sqrt{3c}x\sqrt{3c}(x')^2 + x^3(x')^3 \\ & = c^3 + 3c^2xx' + 3c(xx')^2 + (xx')^3 \\ & = (xx' + c)^3 = \phi(x) \cdot \phi(x'). \end{aligned}$$

That is, if we evaluate the function with weights  $\phi(x)$  at a point  $x'$ , we just get the kernel back. There isn't any magic here; we defined  $\mathcal{F}$  that way in the first place! Letting  $f_w \in \mathcal{F}$  denote the function with weight vector  $w$ , this means that

$$\langle f_{\phi(x)}, f_{\phi(x')} \rangle_{\mathcal{F}} = k(x, x').$$

Now, because it's a vector space, we know that  $\sum_{i=1}^n \alpha_i f_{\phi(x_i)} \in \mathcal{F}$  for any  $n$ ,  $\alpha_i \in \mathbb{R}$ , and choice of  $x_i$ . By the linearity properties of inner product spaces,

$$\left\langle \sum_{i=1}^n \alpha_i f_{\phi(x_i)}, f_{\phi(x)} \right\rangle_{\mathcal{F}} = \sum_{i=1}^n \alpha_i \langle f_{\phi(x_i)}, f_{\phi(x)} \rangle_{\mathcal{F}} = \sum_{i=1}^n \alpha_i k(x_i, x).$$

Since  $f_{\phi(x_i)} \in \mathcal{F}$  is a function from  $\mathcal{X}$  to  $\mathbb{R}$ , this is the same as taking a linear combination of the functions, in terms of their pointwise evaluations.

So, we can think of  $\mathcal{F}$  as having a vector space structure without direct reference to  $w$ , where  $af + f'$  is defined as the function  $x \mapsto af(x) + f'(x)$ , and where  $f(x) = \langle f, f_{\phi(x)} \rangle_{\mathcal{F}}$  (also known as the **reproducing property**) – at least for any  $f$  that's a linear combination of  $f_{\phi(x_i)}$  for some  $x_i$ . This will be the basis for our construction of a *reproducing kernel Hilbert space* (RKHS) for a generic kernel.

The notation  $f_{\phi(x)}$  is a little bit cumbersome. Kernels people often use  $k(x, \cdot)$  to denote this. This notation is justified because  $k(x, \cdot)$  would normally mean the function  $t \mapsto k(x, t)$ ; but that's exactly what you get when you do  $f_{\phi(x)}(t) = \phi(x) \cdot \phi(t) = k(x, t)$ .

### 11.3 REPRODUCING KERNELS

Not every function can be a kernel: it needs to be possible to write as an inner product. So:

**DEFINITION 11.8.** A function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is a *positive definite kernel* if and only if there exists some Hilbert space  $\mathcal{G}$  and feature map  $\phi : \mathcal{X} \rightarrow \mathcal{G}$  such that  $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{G}}$ .

Notice that the space, and the map, don't need to be unique (e.g. you could always use  $-\phi$  instead of  $\phi$ ). Sometimes it's clear what such a map is: for the cubic kernel we considered above, we used  $\mathcal{G} = \mathbb{R}^4$  and  $\phi(x) = (\sqrt{c^3}, \sqrt{3c^2}x, \sqrt{3c}x^2, x^3)$ . Sometimes, though, it's not obvious for a given  $k$  whether there is such a map or not.

The definition implies that we need  $k(x, x') = k(x', x)$ , and that  $k(x, x) \geq 0$ . But those are only necessary, not sufficient.

Unfortunately people are very inconsistent about terminology around positive definiteness. For matrices, "positive semi-definite" unambiguously means the eigenvalues are nonnegative, and "strictly positive definite" unambiguously means eigenvalues are all positive, but "positive definite" might mean either. Some people get annoyed if you try to say "positive semi-definite kernel function," though.

**THEOREM 11.9** ([Aro50]). A function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is a positive definite kernel if and

only if for all  $m \geq 1$  and  $x_1, \dots, x_m \in \mathcal{X}$ , the kernel matrix 
$$\begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_m) \\ \vdots & \ddots & \vdots \\ k(x_m, x_1) & \dots & k(x_m, x_m) \end{bmatrix} \in \mathbb{R}^{m \times m}$$
 is positive semi-definite.

Recall that a positive semi-definite matrix can be equivalently characterized as:

- For all  $\alpha \in \mathbb{R}^m$ ,  $\alpha^T K \alpha \geq 0$ .
- All eigenvalues of  $K$  are nonnegative.
- $K = LL^T$  for some  $L \in \mathbb{R}^{m \times m}$ .

*Proof (sketch).* One direction is easy: if  $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{G}}$ , then

$$\alpha^T K \alpha = \sum_{i=1}^m \sum_{j=1}^m \alpha_i \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{G}} \alpha_j = \left\| \sum_{i=1}^m \alpha_i \phi(x_i) \right\|_{\mathcal{G}}^2 \geq 0.$$

To show the other direction, given a  $k$  satisfying this property, we'll construct a space  $\mathcal{F}$ : the reproducing kernel Hilbert space.

We'll start by building a "pre-Hilbert space"  $\mathcal{F}_0$ , containing functions  $\mathcal{X} \rightarrow \mathbb{R}$ . Start by defining the functions  $\varphi(x) = [x' \mapsto k(x, x')]$  for all  $x$ . Then, let  $\mathcal{F}_0$  be the set of all linear combinations of these functions,  $\sum_{i=1}^m \alpha_i \varphi(x_i)$  for any  $m \geq 0$ ,  $x_1, \dots, x_m \in \mathcal{X}$ ,  $\alpha_1, \dots, \alpha_m \in \mathbb{R}$ . Define an inner product by

$$\left\langle \sum_{i=1}^m \alpha_i \varphi(x_i), \sum_{j=1}^n \beta_j \varphi(x'_j) \right\rangle_{\mathcal{F}_0} = \sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j k(x_i, x'_j).$$

This satisfies the required linearity and nonnegativity properties to be an inner product. It also has the reproducing properties that we expect:

$$\langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}_0} = k(x, x') \quad \langle f, \varphi(x) \rangle_{\mathcal{F}_0} = f(x).$$

Notice also that this is well-defined in the sense that it's representation-independent:

$$\left\langle \sum_{i=1}^m \alpha_i \varphi(x_i), f' \right\rangle_{\mathcal{F}_0} = \sum_{i=1}^m \alpha_i \langle \varphi(x_i), f' \rangle_{\mathcal{F}_0} = \sum_{i=1}^m \alpha_i f'(x_i),$$

which doesn't depend on how we wrote  $f'$  as a linear combination, just on its values.

The only thing left is that we need  $\mathcal{F}_0$  to be complete: it's conceivable that not all Cauchy sequences have limits in this space. So, we construct the RKHS as the completion of  $\mathcal{F}_0$ : just add the limits in, defining their inner products as limits of the inner products of the sequence (which is guaranteed to exist since the sequence is Cauchy and  $\mathbb{R}$  is complete). So, not all  $f \in \mathcal{F}$  can be written as  $\sum_{i=1}^n \alpha_i \varphi(x_i)$ , but you can always get arbitrarily close (in the distance defined by  $\|\cdot\|_{\mathcal{F}}$ ) to  $f$  with things of that form.

After checking all the details work out, we've constructed a Hilbert space and a feature map for any  $k$ .  $\square$

(There are also other ways to define an RKHS; it turns out each RKHS has a unique kernel, and each kernel has a unique RKHS, though there could be more than Hilbert space aligning with the definition.)

### 11.3.1 Special case: linear kernel

If we use  $k(x, x') = x \cdot x'$  for  $x \in \mathbb{R}^d$ , then  $\varphi(x) = [x' \mapsto x' \cdot x]$  is just a linear function with weight  $x$ . Also,

$$\|\varphi(x)\|_{\mathcal{F}} = \sqrt{\langle \varphi(x), \varphi(x) \rangle_{\mathcal{F}}} = \sqrt{k(x, x)} = \|x\|.$$

So everything we've done with linear predictors can be thought of as operating in the RKHS corresponding to a linear kernel. This is often a useful thing to think about if you're looking at some complicated kernel expression: see what it'd be with a linear kernel.

## 11.4 OPTIMIZING IN THE RKHS

**THEOREM 11.10** (Representer theorem). *If  $\mathcal{F}$  is an RKHS with feature map  $\varphi$ , then for any function  $L : \mathbb{R}^m \rightarrow \mathbb{R}$  and any nondecreasing function  $R : \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}$ ,*

$$\arg \min_{f \in \mathcal{F}} L(f(x_1), \dots, f(x_m)) + R(\|f\|)$$

*contains a solution of the form  $f = \sum_{i=1}^m \alpha_i \varphi(x_i)$ , where  $S = (x_1, \dots, x_m)$ . If  $R$  is strictly increasing, all solutions are of this form.*

Notice that  $\arg \min_{f: \|f\|_{\mathcal{F}} \leq B} L_S(f)$  fits this form: use  $R(t) = \begin{cases} 0 & t \leq B \\ \infty & t > B \end{cases}$ .

*Proof.* Let  $\mathcal{F}_{\parallel}$  be the subspace of  $\mathcal{F}$  spanned by  $\{\varphi(x_i)\}_{i=1}^m$ , and  $\mathcal{F}_{\perp}$  its orthogonal complement. Then any element of  $\mathcal{F}$  can be uniquely decomposed into  $f_{\parallel} + f_{\perp}$ , where  $f_{\parallel} \in \mathcal{F}_{\parallel}$ ,  $f_{\perp} \in \mathcal{F}_{\perp}$ , and  $\langle f_{\parallel}, f_{\perp} \rangle_{\mathcal{F}} = 0$ . Now, since

$$f(x_i) = \langle f, \varphi(x_i) \rangle_{\mathcal{F}} = \langle f_{\parallel} + f_{\perp}, \varphi(x_i) \rangle_{\mathcal{F}} = \langle f_{\parallel}, \varphi(x_i) \rangle_{\mathcal{F}} + \underbrace{\langle f_{\perp}, \varphi(x_i) \rangle_{\mathcal{F}}}_0,$$

the  $L$  component only depends on  $f_{\parallel}$ . Also,

$$\|f\|_{\mathcal{F}}^2 = \|f_{\parallel}\|_{\mathcal{F}}^2 + \|f_{\perp}\|_{\mathcal{F}}^2 + 2 \underbrace{\langle f_{\parallel}, f_{\perp} \rangle_{\mathcal{F}}}_0 = \|f_{\parallel}\|_{\mathcal{F}}^2 + \|f_{\perp}\|_{\mathcal{F}}^2.$$

Thus, having a nonzero value of  $f_{\perp}$  does not change  $L$ , and cannot help  $R$ . If  $R$  is strictly increasing, it can only hurt the overall objective.  $\square$

That is, *any* problem will have a solution of the form  $w = \sum_i \alpha_i \varphi(x_i)$ . This allows us to reduce optimization in  $\mathcal{F}$  – potentially infinite-dimensional – to optimization over  $\alpha \in \mathbb{R}^m$ .

### 11.4.1 Example: kernel ridge regression

Consider the problem

$$\min_{h \in \mathcal{F}} L_S^{sq}(h) + \lambda \|h\|_{\mathcal{F}}^2 \tag{11.1}$$

for  $\lambda > 0$ . First off, with a linear kernel, this becomes just plain ridge regression  $\min_w L_S^{sq}(x \mapsto w \cdot x) + \lambda \|w\|^2$ .

We know that all solutions will be of the form  $\sum_{i=1}^m \alpha_i \varphi(x_i)$ , so (11.1) is equivalent to

$$\min_{\alpha \in \mathbb{R}^m} L_S^{sq} \left( \sum_i \alpha_i \varphi(x_i) \right) + \lambda \left\| \sum_i \alpha_i \varphi(x_i) \right\|_{\mathcal{F}}^2. \tag{11.2}$$

The second term here is just

$$\left\| \sum_i \alpha_i \varphi(x_i) \right\|_{\mathcal{F}}^2 = \sum_{i,j} \alpha_i k(x_i, x_j) \alpha_j = \alpha^{\top} \mathbf{K}_{|S_x} \alpha,$$

where  $K|_{S_x} \in \mathbb{R}^{m \times m}$  is the kernel matrix on  $S_x$ , as in Theorem 11.9. For the first term, notice that

$$\sum_i \alpha_i k(x_i, x_j) = \alpha^\top K|_{S_x} e_j$$

where  $e_j \in \mathbb{R}^m$  is the  $j$ th standard basis vector. Then

$$L_S^{sq} \left( \sum_i \alpha_i \varphi(x_i) \right) = \frac{1}{m} \sum_i \left( \alpha^\top K|_{S_x} e_i - y_i \right)^2 = \frac{1}{m} \|K\alpha - y\|_{\mathbb{R}^m}^2.$$

Thus the overall problem is

$$\begin{aligned} \hat{\alpha} &\in \arg \min_{\alpha} \frac{1}{m} \alpha^\top K|_{S_x} K|_{S_x} \alpha - \frac{2}{m} y^\top K|_{S_x} \alpha + \frac{1}{m} y^\top y + \lambda \alpha^\top K|_{S_x} \alpha \\ &= \arg \min_{\alpha} \alpha^\top K|_{S_x} (K|_{S_x} + m\lambda I) \alpha - 2y^\top K|_{S_x} \alpha. \end{aligned}$$

Setting the gradient to zero gives that we want

$$K|_{S_x} (K|_{S_x} + m\lambda I) \alpha = K|_{S_x} y,$$

which is achieved by

$$\hat{\alpha} = (K|_{S_x} + m\lambda I)^{-1} y.$$

When  $\lambda > 0$  this inverse is guaranteed to exist, since  $K|_{S_x}$  is positive semidefinite, so  $K|_{S_x} + m\lambda$  has all eigenvalues at least  $m\lambda$ .

We can also make predictions on an arbitrary test point with

$$\left\langle \sum_i \hat{\alpha}_i \varphi(x_i), \varphi(x) \right\rangle_{\mathcal{F}} = \sum_i \hat{\alpha}_i k(x_i, x) = \hat{\alpha} \cdot \begin{bmatrix} k(x_1, x) \\ \vdots \\ k(x_m, x) \end{bmatrix}.$$

It's worth checking for yourself that this agrees with standard ridge regression. (You might have to use the [Woodbury matrix identity](#) to line them up, since usual expressions for ridge regression invert a  $d \times d$  matrix instead of an  $m \times m$  one. In 340, we called this version the "other normal equations.")

*People sometimes call this transformed version a dual form, especially e.g. for kernel ridge regression. While "dual" isn't necessarily a strictly defined term, note that it's not a Lagrange dual.*

We often won't be able to solve things in closed form like we can for kernel ridge regression. But the representer theorem will still be helpful for any problem of the right form; we just still might have to run an optimization algorithm like gradient descent on the  $\alpha$  variables.

## 11.5 OTHER KERNELS

The most common kernel people use is the Gaussian kernel, also called the "square exponential" or "exponentiated quadratic" by some communities:

$$k(x, x') = \exp\left(-\frac{1}{2\sigma^2} \|x - x'\|^2\right).$$

My preferred way to prove this is a kernel goes through the following construction:

**PROPOSITION 11.11.** *Let  $k, k_1, k_2, \dots$  be positive definite kernels on  $\mathcal{X}$ . Then the following are all also positive definite kernels:*

1.  $\gamma k = (x, x') \mapsto \gamma k(x, x')$  for any  $\gamma > 0$ .

2.  $k_1 + k_2 = (x, x') \mapsto k_1(x, x') + k_2(x, x')$ .
3.  $k_1 k_2 = (x, x') \mapsto k_1(x, x') k_2(x, x')$ .
4.  $k^n = (x, x') \mapsto k(x, x')^n$  for any nonnegative integer  $n$ .
5.  $k_\infty = (x, x') \mapsto \lim_{n \rightarrow \infty} k_n(x, x')$ , when the limit always exists.
6.  $e^k = (x, x') \mapsto \exp(k(x, x'))$ .
7.  $(x, x') \mapsto f(x)k(x, x')f(x')$  for any function  $f : \mathcal{X} \rightarrow \mathbb{R}$ .
8.  $(x, x') \mapsto k'(f(x), f(x'))$  for any function  $f : \mathcal{X} \rightarrow \mathcal{X}'$  and  $k'$  a kernel on  $\mathcal{X}'$ .

*Proof.* Let  $\varphi, \varphi_1, \varphi_2, \dots$  be the feature maps for these kernels, and  $K, K_1, K_2, \dots$  the kernel matrices for arbitrary  $(x_1, \dots, x_m) \in \mathcal{X}^m$ .

1. Use the feature map  $x \mapsto \sqrt{\gamma} \phi$ .
2.  $\alpha^\top (K_1 + K_2) \alpha = \alpha^\top K_1 \alpha + \alpha^\top K_2 \alpha \geq 0$ .
3. This is called the **Schur product theorem**. Define independent multivariate normal random vectors  $V \sim \mathcal{N}(0, K_1)$  and  $W \sim \mathcal{N}(0, K_2)$ . Let  $V \odot W$  be the elementwise product of  $V$  and  $W$ ; this has covariance matrix  $K_1 \odot K_2$ , and covariances must be psd.
4. Iteratively apply the previous property; also,  $k^0$  has feature map  $x \mapsto 1$ .
5.  $\alpha^\top K_\infty \alpha = \alpha^\top [\lim_{n \rightarrow \infty} K_n] \alpha = \lim_{n \rightarrow \infty} \alpha^\top K_n \alpha \geq 0$ .
6. Use  $\exp(k(x, x')) = \lim_{N \rightarrow \infty} \sum_{n=0}^N \frac{1}{n!} k(x, x')^n$  and the previous properties.
7. Use the feature map  $x \mapsto f(x) \varphi(x)$ .
8. Use the feature map  $x \mapsto \varphi'(f(x))$ . □

To get the Gaussian kernel, notice that

$$\exp\left(-\frac{1}{2\sigma^2} \|x - x'\|^2\right) = \exp\left(-\frac{1}{2\sigma^2} \|x\|^2\right) \exp\left(\frac{1}{\sigma^2} x \cdot x'\right) \exp\left(-\frac{1}{2\sigma^2} \|x'\|^2\right)$$

and apply the properties above.

The Gaussian kernel is universal; you can prove this fairly immediately via Stone-Weierstrass (Theorem 10.10).

The Gaussian is *not* always the best kernel, particularly in high dimensions. Functions in  $\mathcal{F}$  for a Gaussian kernel are very smooth; the Matérn kernel is preferred in some settings where rougher functions are expected. Another good general-purpose kernel is the *distance kernel* [SSGF13]

$$k(x, x') = \rho(x, O) + \rho(x', O) - \rho(x, x')$$

where  $\rho$  is a (semi)metric, and  $O \in \mathcal{X}$  is some arbitrary center point, perhaps 0. This kernel isn't actually universal [SSGF13, Proposition 35], but it is "almost universal" and works well in various settings.

If you have a good (e.g. deep) feature extractor  $\psi$ , using a kernel of the form  $k(\psi(x), \psi(x'))$  can often be a good idea. This usually won't be universal, but that usually doesn't matter for the particular problem you're looking at.

### 11.5.1 Some properties

**PROPOSITION 11.12.** Consider a kernel  $k$  with RKHS  $\mathcal{F}$ . Then

$$\text{Rad}\left(\left\{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \leq B\right\}\Big|_{S_x}\right) \leq \frac{B}{\sqrt{m}} \sqrt{\frac{1}{m} \sum_{i=1}^m k(x_i, x_i)}.$$

*Proof.* The analysis in Section 5.2.2 carries through exactly when replacing  $x_i$  with  $k(x_i, \cdot) \in \mathcal{F}$ , in which case  $\|\phi(x_i)\|^2 = \langle k(x_i, \cdot), k(x_i, \cdot) \rangle_{\mathcal{F}} = k(x_i, x_i)$ .  $\square$

For many kernels, such as the Gaussian,  $k(x, x) = 1$  no matter the choice of  $x$ . This makes it even simpler to handle than for the linear case, since we don't care about the data distribution.

This is a case where Rademacher analyses are *much* better than straightforward uses of covering numbers, since for infinite-dimensional kernels like the Gaussian the covering number of the sphere is infinite [Isr15].

**PROPOSITION 11.13.** Let  $f \in \mathcal{F}$ , the RKHS with kernel  $k$ . Then

$$|f(x)| \leq \|f\|_{\mathcal{F}} \sqrt{k(x, x)} \quad |f(x) - f(x')| \leq \|f\|_{\mathcal{F}} \sqrt{k(x, x) + k(x', x') - 2k(x, x')}.$$

*Proof.* We have by the representer property and Cauchy-Schwartz that

$$|f(x)| = |\langle f, \varphi(x) \rangle_{\mathcal{F}}| \leq \|f\|_{\mathcal{F}} \|\varphi(x)\|_{\mathcal{F}}.$$

Similarly,

$$|f(x) - f(x')| = |\langle f, \varphi(x) - \varphi(x') \rangle_{\mathcal{F}}| \leq \|f\|_{\mathcal{F}} \sqrt{k(x, x) + k(x', x') - 2k(x, x')}. \quad \square$$

Many more properties of this kind are available. For *shift-invariant* kernels,  $k(x, x') = \kappa(x - x')$ , a lot is available via Fourier properties of  $\kappa$ .

We've only scratched the surface here. We'll touch on kernels again through the rest of the course, but if you want more, Chapter 7 of [Bach24] goes in some more depth, and [SC08] is a classic very deep/mathematically thorough reference. Bayesian-oriented people might also want to see connections to Gaussian Processes [RW06; KHSS18], which are very much "almost the same thing" from a slightly different point of view.

### REFERENCES

- [Aro50] Nachman Aronszajn. [Theory of Reproducing Kernels](#). *Transactions of the American Mathematical Society* 68.3 (May 1950), pages 337–404.
- [Bach24] Francis Bach. [Learning Theory from First Principles](#). Draft version. August 2024.
- [Isr15] Robert Israel. [Can the ball  \$B\(0, r\_0\)\$  be covered with a finite number of balls of radius  \$< r\_0\$](#) . Mathematics Stack Exchange. April 1, 2015.
- [KHSS18] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. [Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences](#). 2018. arXiv: 1807.02582.
- [RW06] Carl Edward Rasmussen and Christopher K. I. Williams. [Gaussian Processes for Machine Learning](#). MIT Press, 2006.

- 
- [SC08] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.
- [SSGF13] Dino Sejdinovic, Bharath K. Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. [Equivalence of distance-based and RKHS-based statistics in hypothesis testing](#). *Annals of Statistics* 41.5 (October 2013), pages 2263–2291.