

# Binary classification, Part 2

Maxim Raginsky

October 2, 2014

## 1 Kernel machines

Another powerful way of building complicated classifiers from simple functions is by means of *kernels*. Kernel methods are popular in machine learning for a variety of reasons, not the least of which is that any algorithm that operates in a Euclidean space and relies only on the computation of inner products between feature vectors can be modified to work with any suitably well-behaved kernel.

To start with, let us define what we mean by a kernel. We will stick to Euclidean feature spaces, although everything works out for arbitrary separable metric spaces.

**Definition 1.** Let  $X$  be a closed subset of  $\mathbb{R}^d$ . A real-valued function  $K : X \times X \rightarrow \mathbb{R}$  is called a Mercer kernel provided the following conditions are met:

1. It is symmetric, i.e.,  $K(x, x') = K(x', x)$  for any  $x, x' \in X$ .
2. It is continuous, i.e., if  $\{x_n\}$  is a sequence of points in  $X$  converging to a point  $x$ , then

$$\lim_{n \rightarrow \infty} K(x_n, x') = K(x, x'), \quad \forall x' \in X.$$

3. It is positive semidefinite, i.e., for all  $\alpha_1, \dots, \alpha_n \in \mathbb{R}$  and all  $x_1, \dots, x_n \in X$ ,

$$\sum_{i,j=1}^n \alpha_i \alpha_j K(x_i, x_j) \geq 0. \tag{1}$$

**Remark 1.** Another way to interpret the positive semidefiniteness condition is as follows. For any  $n$ -tuple  $x^n = (x_1, \dots, x_n) \in X^n$ , define the  $n \times n$  kernel Gram matrix

$$G_K(x^n) \triangleq [K(x_i, x_j)]_{i,j=1}^n.$$

Then (1) is equivalent to saying that  $G_K(x^n)$  is positive semidefinite in the usual sense, i.e., for any vector  $v \in \mathbb{R}^n$  we have

$$\langle v, G_K(x^n) v \rangle \geq 0.$$

**Remark 2.** From now on, we will just say “kernel,” but always mean “Mercer kernel.”

Here are some examples of kernels:

1. With  $X = \mathbb{R}^d$ ,  $K(x, x') = \langle x, x' \rangle$ , the usual Euclidean inner product.

2. A more general class of kernels based on the Euclidean inner product can be constructed as follows. Let  $X = \{x \in \mathbb{R}^d : \|x\| \leq R\}$ ; choose any sequence  $\{a_j\}_{j=0}^{\infty}$  of nonnegative reals such that

$$\sum_{j=0}^{\infty} a_j R^{2j} < \infty.$$

Then

$$K(x, x') = \sum_{j=0}^{\infty} a_j \langle x, x' \rangle^j$$

is a kernel.

3. Let  $X = \mathbb{R}^d$ , and let  $k : \mathbb{R}^d \rightarrow \mathbb{R}$  be a continuous function, which is *reflection-symmetric*, i.e.,  $k(-x) = k(x)$  for all  $x$ . Then  $K(x, x') \triangleq k(x - x')$  is a kernel provided the Fourier transform of  $k$ ,

$$\widehat{k}(\xi) \triangleq \int_{\mathbb{R}^d} e^{-i\langle \xi, x \rangle} k(x) dx,$$

is nonnegative. A prime example is the *Gaussian kernel*, induced by the function  $k(x) = e^{-\gamma \|x\|^2}$ .

In all of the above cases, the first two properties of a Mercer kernel are easy to check. The third, i.e., positive semidefiniteness, requires a bit more work. For details, consult Section 2.5 of the book by Cucker and Zhou [CZ07].

The importance of kernels in machine learning stems from the fact that we can use them to represent (or approximate) arbitrarily complicated continuous functions on the feature space  $X$ . In order to take full advantage of this representational power, we must take a detour into the theory of *Hilbert spaces*.

## 1.1 A crash course on Hilbert spaces

Hilbert spaces are a powerful generalization of the usual Euclidean space with an inner product; once we have an inner product, we can introduce the notion of an *angle* and, consequently, orthogonality. Moreover, a Hilbert space has certain favorable convergence properties, so we can speak about (unique) linear projections of their elements onto closed linear subspaces. Let us make these ideas precise.

**Definition 2.** A real vector space  $V$  is an inner product space if there exists a function  $\langle \cdot, \cdot \rangle_V : V \times V \rightarrow \mathbb{R}$ , which is:

1. *Symmetric:*  $\langle v, v' \rangle_V = \langle v', v \rangle_V$  for all  $v, v' \in V$
2. *Linear:*  $\langle \alpha v_1 + \beta v_2, v' \rangle_V = \alpha \langle v_1, v' \rangle_V + \beta \langle v_2, v' \rangle_V$  for all  $\alpha, \beta \in \mathbb{R}$  and all  $v_1, v_2, v' \in V$
3. *Positive definite:*  $\langle v, v \rangle_V \geq 0$  for all  $v \in V$ , and  $\langle v, v \rangle_V = 0$  if and only if  $v = 0$

Let  $(V, \langle \cdot, \cdot \rangle_V)$  be an inner product space. Then we can define a *norm* on  $V$  via

$$\|v\|_V \triangleq \sqrt{\langle v, v \rangle_V}.$$

It is easy to check that this is, indeed, a norm —

1. It is homogeneous: for any  $v \in V$  and any  $\alpha \in \mathbb{R}$ ,

$$\|\alpha v\|_V = \sqrt{\langle \alpha v, \alpha v \rangle_V} = \sqrt{\alpha^2 \langle v, v \rangle_V} = |\alpha| \sqrt{\langle v, v \rangle_V} = |\alpha| \cdot \|v\|_V$$

2. It satisfies the triangle inequality: for any  $v, v' \in V$ ,

$$\|v + v'\|_V \leq \|v\|_V + \|v'\|_V. \quad (2)$$

To prove this, we first need to establish another key property of  $\|\cdot\|_V$ : the *Cauchy–Schwarz inequality*, which generalizes its classical Euclidean counterpart and says that

$$|\langle v, v' \rangle_V| \leq \|v\|_V \|v'\|_V. \quad (3)$$

To prove (3), we start with the observation that  $\|v - \lambda v'\|_V^2 = \langle v - \lambda v', v - \lambda v' \rangle_V \geq 0$  for any  $\lambda \in \mathbb{R}$ . Expanding this, we get

$$\langle v - \lambda v', v - \lambda v' \rangle_V = \lambda^2 \|v'\|_V^2 - 2\lambda \langle v, v' \rangle_V + \|v\|_V^2 \geq 0.$$

This is a quadratic function of  $\lambda$ , and from the above we see that its graph does not cross the horizontal axis. Therefore, we must have

$$4|\langle v, v' \rangle_V|^2 \leq 4\|v\|_V^2 \|v'\|_V^2 \iff |\langle v, v' \rangle_V| \leq \|v\|_V \|v'\|_V.$$

Now we can write

$$\begin{aligned} (\|v\|_V + \|v'\|_V)^2 &= \|v\|_V^2 + 2\|v\|_V \|v'\|_V + \|v'\|_V^2 \\ &\geq \|v\|_V^2 + 2\langle v, v' \rangle_V + \|v'\|_V^2 \\ &= \langle v, v \rangle_V + \langle v, v' \rangle_V + \langle v', v \rangle_V + \langle v', v' \rangle_V \\ &= \langle v + v', v + v' \rangle_V \\ &\equiv \|v + v'\|_V^2, \end{aligned}$$

where the first step uses the Cauchy–Schwarz inequality, the second step uses the definition of  $\|\cdot\|_V$  and the symmetry of  $\langle \cdot, \cdot \rangle_V$ , the third step uses the linearity of  $\langle \cdot, \cdot \rangle_V$ , and the final step is, again, by definition. Since all norms are nonnegative, we can take square roots of both sides to get the triangle inequality.

3. Finally,  $\|v\|_V \geq 0$ , and  $\|v\|_V = 0$  if and only if  $v = 0$  – this is obvious from definitions.

Thus, an inner product space can be equipped with a norm that has certain special properties (mainly, the Cauchy–Schwarz inequality, since a lot of useful things follow from it alone). Now that we have a norm, we can talk about *convergence* of sequences in  $V$ :

**Definition 3.** Let  $\{v_n\}_{n=1}^\infty$  be a sequence of elements of  $V$ . We say that it converges to  $v \in V$  if

$$\lim_{n \rightarrow \infty} \|v_n - v\|_V = 0. \quad (4)$$

**Remark 3.** This definition is valid for any norm on  $V$ , not necessarily a norm induced by an inner product.

Any norm-convergent sequence has the property that, as  $n$  gets larger, its elements get closer and closer to one another. Specifically, suppose that  $\{v_n\}$  converges to  $v$ . Then (4) implies that for any  $\varepsilon > 0$  we can choose  $n$  large enough, so that  $\|v_n - v\|_V < \varepsilon/2$  for all  $m \geq n$ . But the triangle inequality gives

$$\|v_n - v_m\|_V \leq \|v_n - v\|_V + \|v_m - v\|_V < \varepsilon, \quad \forall m \geq n.$$

In other words, we have

$$\lim_{m \rightarrow \infty} \|v_n - v_m\| = 0.$$

Since this holds for every  $n$ , we can write

$$\lim_{\min(m,n) \rightarrow \infty} \|v_n - v_m\| = 0. \tag{5}$$

Any sequence  $\{v_n\}$  that has the property (5) is called a *Cauchy sequence*. We have just proved that any convergent sequence is Cauchy. However, the converse is not necessarily true: a Cauchy sequence does not have to be convergent. This motivates the following definition:

**Definition 4.** A normed space  $(V, \|\cdot\|_V)$  is complete if any Cauchy sequence  $\{v_n\}$  of its elements is convergent. If the norm  $\|\cdot\|_V$  is induced by an inner product, then we say that  $V$  is a Hilbert space.

There is a standard procedure of starting with an inner product and the corresponding normed space and then *completing* it by adding the limits of all Cauchy sequences. We will not worry too much about this procedure. Here are a few standard examples of Hilbert spaces:

1. The Euclidean space  $V = \mathbb{R}^d$  with the usual inner product

$$\langle v, v' \rangle = \sum_{j=1}^d v_j v'_j.$$

The corresponding norm is the familiar  $\ell_2$  norm,  $\|v\| = \sqrt{\langle v, v \rangle}$ .

2. More generally, if  $A$  is a positive definite  $d \times d$  matrix, then the inner product

$$\langle v, v' \rangle_A \triangleq \langle v, Av' \rangle$$

induces the  $A$ -weighted norm  $\|v\|_A \triangleq \sqrt{\langle v, v \rangle_A} = \sqrt{\langle v, Av \rangle}$ , which makes  $\mathbb{R}^d$  into a Hilbert space. The preceding example is a special case with  $A = I_d$ , the  $d \times d$  identity matrix.

3. The space  $L^2(\mathbb{R}^d)$  of all *square-integrable* functions  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , i.e.,

$$\int_{\mathbb{R}^d} f^2(x) dx < \infty,$$

is a Hilbert space with the inner product

$$\langle f, g \rangle_{L^2(\mathbb{R}^d)} \triangleq \int_{\mathbb{R}^d} f(x)g(x) dx$$

and the corresponding norm

$$\|f\|_{L^2(\mathbb{R}^d)} \triangleq \sqrt{\int_{\mathbb{R}^d} f^2(x) dx}.$$

4. Let  $(\Omega, \mathcal{B}, P)$  be a probability space. Then the space  $L^2(P)$  space of all real-valued random variables  $X : \Omega \rightarrow \mathbb{R}$  with finite second moment, i.e.,

$$\mathbb{E}X^2 = \int_{\Omega} X^2(\omega)P(d\omega) < +\infty,$$

is a Hilbert space with the inner product

$$\langle X, X' \rangle_{L^2(P)} \triangleq \mathbb{E}[XX'] = \int_{\Omega} X(\omega)X'(\omega)P(d\omega)$$

and the corresponding norm

$$\|X\|_{L^2(P)} \triangleq \sqrt{\int_{\Omega} |X(\omega)|^2 P(d\omega)} \equiv \sqrt{\mathbb{E}X^2}.$$

From now on, we will denote a typical Hilbert space by  $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ ; the induced norm will be denoted by  $\|\cdot\|_{\mathcal{H}}$ .

An enormous advantage of working with Hilbert spaces is the availability of the notion of *orthogonality* and *orthogonal projection*. Two elements  $h, g$  of a Hilbert space  $\mathcal{H}$  are said to be *orthogonal* if  $\langle h, g \rangle_{\mathcal{H}} = 0$ .

Now consider a closed linear subspace  $\mathcal{H}_1$  of  $\mathcal{H}$ , where “closed” means that the limit of any convergent sequence  $\{h_n\}$  of elements of  $\mathcal{H}_1$  is also contained in  $\mathcal{H}_1$ . Then we have the following basic facts:

**Theorem 1.** *Let  $\mathcal{H}_1^{\perp}$  be the set of all  $h^{\perp} \in \mathcal{H}$ , such that  $\langle g, h^{\perp} \rangle_{\mathcal{H}} = 0$  for all  $g \in \mathcal{H}_1$ . Then:*

1.  $\mathcal{H}_1^{\perp}$  is also a closed linear subspace of  $\mathcal{H}$ .
2. Any element  $g$  of  $\mathcal{H}$  can be uniquely decomposed as  $g = h + h^{\perp}$ , where  $h \in \mathcal{H}_1$  and  $h^{\perp} \in \mathcal{H}_1^{\perp}$ .
3. Define the orthogonal projection  $\Pi: \mathcal{H} \rightarrow \mathcal{H}_1$  onto  $\mathcal{H}_1$  through

$$\Pi g \triangleq h \quad \text{if } g = h + h^{\perp} \text{ with } h \in \mathcal{H}_1, h^{\perp} \in \mathcal{H}_1^{\perp}.$$

Then  $\Pi$  has the following properties:

- (a) It is a linear operator.
- (b)  $\Pi^2 = \Pi$ , i.e.,  $\Pi(\Pi g) = \Pi g$  for any  $g \in \mathcal{H}$ .
- (c) If  $g \in \mathcal{H}_1$ , then  $\Pi g = g$ .
- (d) For any  $g \in \mathcal{H}$  and any  $h \in \mathcal{H}_1$ ,

$$\langle \Pi g, h \rangle_{\mathcal{H}} = \langle g, h \rangle_{\mathcal{H}}.$$

- (e) For any  $g \in \mathcal{H}$ ,  $h = \Pi g \in \mathcal{H}_1$  is the unique solution of the optimization problem

$$\text{minimize } \|h - g\| \text{ subject to } h \in \mathcal{H}_1.$$

**Remark 4.** It is important for  $\mathcal{H}_1$  to be a *closed* linear subspace of  $\mathcal{H}$  for the above results to hold.

## 1.2 Reproducing kernel Hilbert spaces

Now let us return to our original goal. Suppose we have a fixed kernel  $K$  on our feature space  $X$  (which we assume to be a closed subset of  $\mathbb{R}^d$ ). Let  $\mathcal{L}_K(X)$  be the *linear span* of the set  $\{K(x', \cdot) : x' \in X\}$ , i.e., the set of all functions  $f : X \rightarrow \mathbb{R}$  of the form

$$f(x) = \sum_{j=1}^N c_j K(x_j, x) \quad (6)$$

for all possible choices of  $N \in \mathbb{N}$ ,  $c_1, \dots, c_N \in \mathbb{R}$ , and  $x_1, \dots, x_N \in X$ . It is easy to see that  $\mathcal{L}_K(X)$  is a *vector space*: for any two functions  $f, f'$  of the form (6), their sum is also of that form; if we multiply any  $f \in \mathcal{L}_K(X)$  by a scalar  $c \in \mathbb{R}$ , we will get another element of  $\mathcal{L}_K(X)$ ; and the zero function is clearly in  $\mathcal{L}_K(X)$ . It turns out that, for any (Mercer) kernel  $K$ , we can *complete*  $\mathcal{L}_K(X)$  into a *Hilbert space* of functions that can potentially represent *any* continuous function from  $X$  into  $\mathbb{R}$ , provided  $K$  is chosen appropriately.

The following result is essential (for the proof, see Section 2.4 of Cucker and Zhou [CZ07]):

**Theorem 2.** *Let  $X$  be a closed subset of  $\mathbb{R}^d$ , and let  $K : X \times X \rightarrow \mathbb{R}$  be a Mercer kernel. Then there exists a unique Hilbert space  $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K)$  of real-valued functions on  $X$  with the following properties:*

1. *For all  $x \in X$ , the function  $K_x(\cdot) \triangleq K(x, \cdot)$  is an element of  $\mathcal{H}_K$ , and  $\langle K_x, K_{x'} \rangle_K = K(x, x')$  for all  $x, x' \in X$ .*
2. *The linear space  $\mathcal{L}_K(X)$  is dense in  $\mathcal{H}_K$ , i.e., for any  $f \in \mathcal{H}_K$  and any  $\varepsilon > 0$  there exist some  $N \in \mathbb{N}$ ,  $c_1, \dots, c_N \in \mathbb{R}$ , and  $x_1, \dots, x_N \in X$ , such that*

$$\left\| f - \sum_{j=1}^N c_j K_{x_j} \right\|_K < \varepsilon.$$

3. *For all  $f \in \mathcal{H}_K$  and all  $x \in X$ ,*

$$f(x) = \langle K_x, f \rangle_K. \quad (7)$$

Moreover, the functions in  $\mathcal{H}_K$  are continuous. The Hilbert space  $\mathcal{H}_K$  is called the *Reproducing Kernel Hilbert Space (RKHS)* associated with  $K$ ; the property (7) is referred to as the *reproducing kernel property*.

**Remark 5.** The reproducing kernel property essentially states that the value of any function  $f \in \mathcal{H}_K$  at any point  $x \in X$  can be “extracted” by projecting  $f$  onto the function  $K_x(\cdot) = K(x, \cdot)$ , i.e., a copy of the kernel  $K$  “centered” at the point  $x$ . It is easy to prove when  $f \in \mathcal{L}_K(X)$ . Indeed, if  $f$  has the form (6), then

$$\begin{aligned} \langle f, K_x \rangle_K &= \left\langle \sum_{j=1}^N c_j K_{x_j}, K_x \right\rangle_K \\ &= \sum_{j=1}^N c_j \langle K_{x_j}, K_x \rangle_K \\ &= \sum_{j=1}^N c_j K(x_j, x) \\ &= f(x). \end{aligned}$$

Since any  $f \in \mathcal{H}_K$  can be expressed as a limit of functions from  $\mathcal{L}_K(X)$ , the proof of (7) for a general  $f$  follows by continuity.

Now we pick a kernel  $K$  on our feature space and consider classifiers of the form

$$g_f(x) = \text{sgn } f(x) \equiv \begin{cases} 1, & \text{if } f(x) \geq 0 \\ -1, & \text{otherwise} \end{cases}$$

with the underlying  $f$  taken from a suitable subset of the RKHS  $\mathcal{H}_K$ . One choice, which underlies such things as the Support Vector Machine, is to take a ball in  $\mathcal{H}_K$ : given some  $\lambda > 0$ , let

$$\mathcal{F}_\lambda \triangleq \{f \in \mathcal{H}_K : \|f\|_K \leq \lambda\}.$$

This set is the closure (in the  $\|\cdot\|_K$  norm) of the convex set

$$\left\{ \sum_{j=1}^N c_j K_{x_j} : N \in \mathbb{N}; c_1, \dots, c_N \in \mathbb{R}; x_1, \dots, x_N \in \mathcal{X}; \sum_{i,j=1}^N c_i c_j K(x_i, x_j) \leq \lambda^2 \right\} \subset \mathcal{L}_K(\mathcal{X}),$$

and is itself convex. Now, as we already know, the performance of any learning algorithm that chooses an element  $\hat{f}_n \in \mathcal{F}_\lambda$  in a data-dependent way is controlled by the Rademacher average  $R_n(\mathcal{F}_\lambda(X^n))$ . It turns out that this Rademacher average is fairly easy to estimate. Indeed, using the reproducing kernel property (7) and then the linearity of the inner product  $\langle \cdot, \cdot \rangle_K$ , we can write

$$\begin{aligned} R_n(\mathcal{F}_\lambda(X^n)) &= \frac{1}{n} \mathbb{E}_{\sigma^n} \sup_{f: \|f\|_K \leq \lambda} \left| \sum_{i=1}^n \sigma_i f(X_i) \right| \\ &= \frac{1}{n} \mathbb{E}_{\sigma^n} \sup_{f: \|f\|_K \leq \lambda} \left| \sum_{i=1}^n \sigma_i \langle f, K_{X_i} \rangle_K \right| \\ &= \frac{1}{n} \mathbb{E}_{\sigma^n} \sup_{f: \|f\|_K \leq \lambda} \left| \left\langle f, \sum_{i=1}^n \sigma_i K_{X_i} \right\rangle_K \right| \end{aligned}$$

Now, using the Cauchy–Schwarz inequality (3), it is not hard to show that

$$\sup_{f: \|f\|_K \leq \lambda} |\langle f, g \rangle_K| = \lambda \|g\|_K$$

for any  $g \in \mathcal{H}_K$ . Therefore,

$$R_n(\mathcal{F}_\lambda(X^n)) = \frac{\lambda}{n} \mathbb{E}_{\sigma^n} \left\| \sum_{i=1}^n \sigma_i K_{X_i} \right\|_K.$$

Now we exploit the following easily proved fact: for any  $n$  functions  $g_1, \dots, g_n \in \mathcal{H}_K$ ,

$$\mathbb{E}_{\sigma^n} \left\| \sum_{i=1}^n \sigma_i g_i \right\|_K \leq \sqrt{\sum_{i=1}^n \|g_i\|_K^2}. \quad (8)$$

The proof of this is in two steps: First, we use the concavity of the square root to write

$$\mathbb{E}_{\sigma^n} \sqrt{\left\| \sum_{i=1}^n \sigma_i g_i \right\|_K^2} \leq \sqrt{\mathbb{E} \left\| \sum_{i=1}^n \sigma_i g_i \right\|_K^2}.$$

Then we expand the squared norm:

$$\left\| \sum_{i=1}^n \sigma_i g_i \right\|_K^2 = \left\langle \sum_{i=1}^n \sigma_i g_i, \sum_{i=1}^n \sigma_i g_i \right\rangle_K = \sum_{i,j=1}^n \sigma_i \sigma_j \langle g_i, g_j \rangle_K.$$

And finally we take the expectation over  $\sigma^n$  and use the fact that  $\mathbb{E}[\sigma_i \sigma_j] = 1$  if  $i = j$  and 0 otherwise to get

$$\mathbb{E} \left\| \sum_{i=1}^n \sigma_i g_i \right\|_K^2 = \sum_{i=1}^n \langle g_i, g_i \rangle_K = \sum_{i=1}^n \|g_i\|_K^2.$$

Hence, we obtain

$$R_n(\mathcal{F}_\lambda(X^n)) \leq \frac{\lambda}{n} \sqrt{\sum_{i=1}^n \langle K_{X_i}, K_{X_i} \rangle_K} = \frac{\lambda}{n} \sqrt{\sum_{i=1}^n K(X_i, X_i)}.$$

Finally, taking the expectation w.r.t.  $X^n$  and once more using concavity of the square root, we have

$$\mathbb{E} R_n(\mathcal{F}_\lambda(X^n)) \leq \frac{\lambda \sqrt{\mathbb{E} K(X, X)}}{\sqrt{n}}.$$

### 1.3 Empirical risk minimization in an RKHS

Another advantage of working with kernels is that, in many cases, a minimizer of empirical risk over a sufficiently regular subset of an RKHS will have the form of a linear combination of kernels centered at the training feature points. The results ensuring this are often referred to in the literature as *representer theorems*. Here is one such result (due, in a slightly different form, to Schölkopf, Herbrich, and Smola [SHS01]), sufficiently general for our purposes:

**Theorem 3** (The generalized representer theorem). *Let  $X$  be a closed subset of  $\mathbb{R}^d$  and let  $Y$  be a subset of the reals. Consider a nonnegative loss function  $\ell : Y \times Y \rightarrow \mathbb{R}^+$ . Let  $K$  be a Mercer kernel on  $X$ , and let  $\mathcal{H}_K$  be the corresponding RKHS.*

*Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be an i.i.d. sample from some distribution  $P = P_{XY}$  on  $X \times Y$ , let  $\mathcal{H}_n$  be the closed linear subspace of  $\mathcal{H}_K$  spanned by  $\{K_{X_i} : 1 \leq i \leq n\}$ , and let  $\Pi_n$  denote the orthogonal projection onto  $\mathcal{H}_n$ . Let  $\mathcal{F}$  be a subset of  $\mathcal{H}_K$ , such that  $\Pi_n(\mathcal{F}) \subseteq \mathcal{F}$ . Then*

$$\inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) = \inf_{f \in \Pi_n(\mathcal{F})} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)), \quad (9)$$

*and if a minimizer of the left-hand side of (9) exists, then it can be taken to have the form*

$$\hat{f}_n = \sum_{i=1}^n c_i K_{X_i} \quad (10)$$

*for some  $c_1, \dots, c_n \in \mathbb{R}$ .*

**Remark 6.** Note that both the subspace  $\mathcal{H}_n$  and the corresponding orthogonal projection  $\Pi_n$  are *random objects*, since they depend on the random features  $X^n$ .



*Proof.* Since  $K_{X_i} \in \mathcal{H}_n$  for every  $i$ , by Theorem 1 we have

$$\langle f, K_{X_i} \rangle_K = \langle \Pi_n f, K_{X_i} \rangle_K, \quad \forall f \in \mathcal{H}_K.$$

Moreover, from the reproducing kernel property (7) we deduce that

$$f(X_i) = \langle f, K_{X_i} \rangle_K = \langle \Pi_n f, K_{X_i} \rangle_K = \Pi_n f(X_i).$$

Therefore, for every  $f \in \mathcal{F}$  we can write

$$\frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, \Pi_n f(X_i)).$$

This implies that

$$\inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) = \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, \Pi_n f(X_i)) = \inf_{g \in \Pi_n(\mathcal{F})} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, g(X_i)). \quad (11)$$

Now suppose that  $\hat{f}_n \in \mathcal{F}$  achieves the infimum on the left-hand side of (11). Then its projection  $\Pi_n \hat{f}_n$  onto  $\mathcal{H}_n$  achieves the infimum on the right-hand side. Moreover, since  $\Pi_n(\mathcal{F}) \subseteq \mathcal{F}$  by hypothesis, we may conclude that  $f = \Pi_n(f)$ , i.e.,  $f \in \mathcal{H}_n$ . Since every element of  $\mathcal{H}_n$  has the form (10), the theorem is proved.  $\square$

In the classification setting, we may take  $Y = \{-1, +1\}$  and consider the problem of minimizing the empirical surrogate loss

$$A_{\varphi, n}(f) = \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i f(X_i))$$

over the ball  $\mathcal{F}_\lambda$  in a suitable RKHS  $\mathcal{H}_K$ . By the above theorem, we may write this problem in the following form:

$$\min_{c_1, \dots, c_n \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \varphi \left( -Y_i \sum_{j=1}^n c_j K(X_i, X_j) \right) \quad (12a)$$

$$\text{subject to } \sum_{i, j=1}^n c_i c_j K(X_i, X_j) \leq \lambda^2 \quad (12b)$$

Suppose the surrogate loss function  $\varphi$  is convex. Then the objective function in (12) is convex as well, and the decision variables  $c_1, \dots, c_n \in \mathbb{R}$  are subject to a quadratic constraint. Thus, (12) is an instance of a *quadratically constrained convex program* (QCCP). Moreover, when  $\varphi$  is such that the objective is *quadratic* in  $c_1, \dots, c_n$ , then we have a *quadratically constrained quadratic problem* (QCQP), which can be solved very efficiently using interior point methods. For detailed background see the text of Boyd and Vandenberghe [BV04]. Many popular machine learning algorithms can be cast in the form (12). For instance, if we let  $\varphi$  be the hinge loss  $\varphi(u) = (u+1)_+$ , then (12) corresponds to the *Support Vector Machine* (SVM) algorithm — more precisely, the SVM is the *scalarized* version of (12), i.e., it has the form

$$\min_{c_1, \dots, c_n \in \mathbb{R}} \left\{ \frac{1}{n} \sum_{i=1}^n \left( 1 - Y_i \sum_{j=1}^n c_j K(X_i, X_j) \right)_+ + \tau \sum_{i, j=1}^n c_i c_j K(X_i, X_j) \right\}$$

for some regularization parameter  $\tau > 0$ .

## 2 Convex risk minimization

Choosing a convex surrogate loss function  $\varphi$  has many advantages in general. First of all, we may arrange things in such a way that minimizing the surrogate loss  $A_\varphi(f)$  over all measurable  $f : \mathcal{X} \rightarrow \mathbb{R}$  is equivalent to determining the Bayes classifier

$$g^*(x) \triangleq \begin{cases} 1, & \text{if } \eta(x) > 1/2 \\ -1, & \text{otherwise} \end{cases} \quad (13)$$

**Theorem 4.** *Let  $P = P_{XY}$  be the joint distribution of the feature  $X \in \mathbb{R}^d$  and the binary label  $Y \in \{-1, +1\}$ , and let  $\eta(x) = \mathbb{P}[Y = 1|X = x]$  be the corresponding regression function. Consider a surrogate loss function  $\varphi$ , which is strictly convex and differentiable. Then the unique minimizer of the surrogate loss  $A_\varphi(f) = \mathbb{E}[\varphi(-Y f(X))]$  over all (measurable) functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  has the form*

$$f^*(x) = \operatorname{argmin}_{u \in \mathbb{R}} h_{\eta(x)}(u),$$

where for each  $\eta \in [0, 1]$  we have  $h_\eta(u) \triangleq \eta\varphi(-u) + (1 - \eta)\varphi(u)$ . Moreover,  $f^*(x)$  is positive if and only if  $\eta(x) > 1/2$ , i.e, the induced sign classifier  $g_{f^*}(x) = \operatorname{sgn}(f^*(x))$  is the Bayes classifier (13).

*Proof.* By the law of iterated expectation,

$$A_\varphi(f) = \mathbb{E}[\varphi(-Y f(X))] = \mathbb{E}[\mathbb{E}[\varphi(-Y f(X))|X]].$$

Hence,

$$\begin{aligned} \inf_f A_\varphi(f) &= \inf_f \mathbb{E}[\mathbb{E}[\varphi(-Y f(X))|X]] \\ &= \mathbb{E}\left[\inf_{u \in \mathbb{R}} \mathbb{E}[\varphi(-Y u)|X = x]\right]. \end{aligned}$$

For every  $x \in \mathcal{X}$ , we have

$$\begin{aligned} \mathbb{E}[\varphi(-Y u)|X = x] &= \mathbb{P}[Y = 1|X = x]\varphi(-u) + \mathbb{P}[Y = -1|X = x]\varphi(u) \\ &= \eta(x)\varphi(-u) + (1 - \eta)\varphi(u) \\ &\equiv h_{\eta(x)}(u). \end{aligned}$$

Since  $\varphi$  is strictly convex and differentiable, so is  $h_\eta$  for every  $\eta \in [0, 1]$ . Therefore,  $\inf_{u \in \mathbb{R}} h_\eta(u)$  exists, and is achieved by a unique  $u^*$ ; in particular,

$$f^*(x) = \operatorname{argmin}_{u \in \mathbb{R}} h_{\eta(x)}(u).$$

To find the  $u^*$  minimizing  $h_\eta$ , we differentiate  $h_\eta$  w.r.t.  $u$  and set the derivative to zero. Since

$$h'_\eta(u) = -(1 - \eta)\varphi'(-u) + \eta\varphi'(u),$$

the point of minimum  $u^*$  is the solution to the equation

$$\frac{\varphi'(u)}{\varphi'(-u)} = \frac{\eta}{1 - \eta}.$$

Suppose  $\eta > 1/2$ ; then

$$\frac{\varphi'(u)}{\varphi'(-u)} > 1.$$

Since  $\varphi$  is strictly convex, its derivative  $\varphi'$  is strictly increasing. Hence,  $u^* > -u^*$  which implies that  $u^* > 0$ . Conversely, if  $u^* \leq 0$ , then  $u^* \leq -u^*$ , so  $\varphi'(u^*) \leq \varphi'(-u^*)$ , which means that  $\eta/(1-\eta) \leq 1$ , i.e.,  $\eta \leq 1/2$ . Thus, we conclude that  $f^*(x)$ , which is the minimizer of  $h_{\eta(x)}$ , is positive if and only if  $\eta(x) > 1/2$ , i.e.,  $\text{sgn}(f^*(x))$  is the Bayes classifier.  $\square$

Secondly, under some additional regularity conditions it is possible to relate the minimum surrogate loss

$$A_\varphi^* \triangleq \inf_f A_\varphi(f)$$

to the Bayes rate

$$L^* = \inf_f \mathbb{P}(Y \neq f(X)) :$$

**Theorem 5.** Assume that the surrogate loss function  $\varphi$  satisfies the conditions of our basic surrogate bound, and that there exist positive constants  $s \geq 1$  and  $c$ , such that the inequality

$$L(f) - L^* \leq c \left( A_\varphi(f) - A_\varphi^* \right)^{1/s} \quad (14)$$

holds for any measurable function  $f : \mathcal{X} \rightarrow \mathbb{R}$ . Consider the learning algorithm that minimizes empirical surrogate loss over some class  $\mathcal{F}$ :

$$\hat{f}_n = \arg \min_{f \in \mathcal{F}} A_{\varphi,n}(f) = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i f(X_i)). \quad (15)$$

Then

$$L(\hat{f}_n) - L^* \leq 2^{1/s} c \left( 4M_\varphi \mathbb{E} R_n(\mathcal{F}(X^n)) + B \sqrt{\frac{\log(1/\delta)}{2n}} \right)^{1/s} + c \left( \inf_{f \in \mathcal{F}} A_\varphi(f) - A_\varphi^* \right)^{1/s} \quad (16)$$

with probability at least  $1 - \delta$ .

*Proof.* We have the following:

$$L(\hat{f}_n) - L^* \leq c \left( A_\varphi(\hat{f}_n) - A_\varphi^* \right)^{1/s} \quad (17)$$

$$= c \left( A_\varphi(\hat{f}_n) - \inf_{f \in \mathcal{F}} A_\varphi(f) + \inf_{f \in \mathcal{F}} A_\varphi(f) - A_\varphi^* \right)^{1/s} \quad (18)$$

$$\leq c \left( A_\varphi(\hat{f}_n) - \inf_{f \in \mathcal{F}} A_\varphi(f) \right)^{1/s} + c \left( \inf_{f \in \mathcal{F}} A_\varphi(f) - A_\varphi^* \right)^{1/s} \quad (19)$$

$$\leq 2^{1/s} c \left( \sup_{f \in \mathcal{F}} |A_{\varphi,n}(f) - A_\varphi(f)| \right)^{1/s} + c \left( \inf_{f \in \mathcal{F}} A_\varphi(f) - A_\varphi^* \right)^{1/s} \quad (20)$$

$$\leq 2^{1/s} c \left( 4M_\varphi \mathbb{E} R_n(\mathcal{F}(X^n)) + B \sqrt{\frac{\log(1/\delta)}{2n}} \right)^{1/s} + c \left( \inf_{f \in \mathcal{F}} A_\varphi(f) - A_\varphi^* \right)^{1/s} \quad \text{w.p. } \geq 1 - \delta, \quad (21)$$

where:

- (17) follows from (14);
- (19) follows from the inequality  $(a + b)^{1/s} \leq a^{1/s} + b^{1/s}$  that holds for all  $a, b \geq 0$  and all  $s \geq 1$
- (20) and (21) follow from the same argument as the one used in the proof of the basic surrogate bound.

This completes the proof. □

**Remark 7.** Condition (14) is often easy to check. For instance, Zhang [Zha04] proved that it is satisfied, provided the inequality

$$\left| \frac{1}{2} - \eta \right|^s \leq (2c)^s \left( 1 - \inf_u h_\eta(u) \right) \quad (22)$$

holds for all  $\eta \in [0, 1]$ . For instance, (22) holds for the exponential loss  $\varphi(u) = e^u$  and the logit loss  $\varphi(u) = \log_2(1 + e^u)$  with  $s = 2$  and  $c = 2\sqrt{2}$ ; for the hinge loss  $\varphi(u) = (u + 1)_+$ , (22) holds with  $s = 1$  and  $c = 4$ .

What Theorem 5 says is that, assuming the expected Rademacher average  $\mathbb{E}R_n(\mathcal{F}(X^n)) = O(1/\sqrt{n})$ , the difference between the generalization error of the Convex Risk Minimization algorithm (15) and the Bayes rate  $L^*$  is, with high probability, bounded by the combination of two terms: the  $O(n^{-1/2s})$  “estimation error” term and the  $(\inf_{f \in \mathcal{F}} A_\varphi(f) - A_\varphi^*)^{1/s}$  “approximation error” term. If the hypothesis space  $\mathcal{F}$  is rich enough, so that  $\inf_{f \in \mathcal{F}} A_\varphi(f) = A_\varphi^*$ , then the difference between  $L(\hat{f}_n)$  and  $L^*$  is, with high probability, bounded as  $O(1/n^{-2s})$ , *independently* of the dimension  $d$  of the feature space.

## References

- [BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [CZ07] F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.
- [SHS01] B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In D. Helmbold and B. Williamson, editors, *Computational Learning Theory*, volume 2111 of *Lecture Notes in Computer Science*, pages 416–426. Springer, 2001.
- [Zha04] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004.