# Binary classification

Maxim Raginsky

October 15, 2013

The problem of binary classification can be stated as follows. We have a random couple $Z = (X, Y)$, where $X \in \mathbb{R}^d$ is called the *feature vector* and $Y \in \{-1, 1\}$ is called the *label*[1]. In the spirit of the model-free framework, we assume that the relationship between the features and the labels is stochastic and described by an unknown probability distribution $P \in \mathscr{P}(\mathsf{Z})$, where $\mathsf{Z} = \mathbb{R}^d \times \{-1, 1\}$.

As usual, we consider the case when we are given an i.i.d. sample of length $n$ from $P$. The goal is to learn a *classifier*, i.e., a mapping $g : \mathbb{R}^d \to \{-1, 1\}$, such that the probability of classification error, $\mathbb{P}(g(X) \neq Y)$, is small. As we have seen before, the optimal choice is the *Bayes classifier*

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

where $\eta(x) \triangleq \mathbb{P}[Y = 1 | X = x]$ is the *regression function*. However, since we make no assumptions on $P$, in general we cannot hope to learn the Bayes classifier $g^*$. Instead, we focus on a more realistic goal: We fix a collection $\mathscr{G}$ of classifiers and then use the training data to come up with a hypothesis $\widehat{g}_n \in \mathscr{G}$, such that

$$\mathbb{P}(\widehat{g}_n(X) \neq Y) \approx \inf_{g \in \mathscr{G}} \mathbb{P}(g(X) \neq Y)$$

with high probability.

By way of notation, let us write $L(g)$ for the classification error of $g$, i.e., $L(g) \triangleq \mathbb{P}(g(X) \neq Y)$, and let $L^*(\mathscr{G})$ denote the smallest classification error attainable over $\mathscr{G}$:

$$L^*(\mathscr{G}) \triangleq \inf_{g \in \mathscr{G}} L(g).$$

We will assume that a minimizing $g^* \in \mathscr{G}$ exists. For future reference, we note that

$$L(g) = \mathbb{P}(g(X) \neq Y) \leq \mathbb{P}(Y g(X) < 0). \tag{2}$$

**Warning:** In what follows, we will use $C$ or $c$ to denote various absolute constants; their values may change from line to line.

## 1 Learning linear discriminant rules

One of the simplest classification rules (and one of the first to be studied) is a *linear discriminant rule*: given a nonzero vector $w = (w^{(1)}, \ldots, w^{(d)}) \in \mathbb{R}^d$ and a scalar $b \in \mathbb{R}$, let

$$g(x) \equiv g_{w,b}(x) \triangleq \begin{cases} 1, & \text{if } \langle w, x \rangle + b > 0 \\ -1, & \text{otherwise} \end{cases} \tag{3}$$

[1] The reason why we chose $\{-1, 1\}$, rather than $\{0, 1\}$, for the label space is merely convenience.

Let $\mathscr{G}$ be the class of all such linear discriminant rules as $w$ ranges over all nonzero vectors in $\mathbb{R}^d$ and $b$ ranges over all reals: $\mathscr{G} = \{g_{w,b} : w \in \mathbb{R}^d \setminus \{0\}, b \in \mathbb{R}\}$.

Given the training sample $Z^n$, let $\widehat{g}_n \in \mathscr{G}$ be the output of the ERM algorithm, i.e.,

$$\widehat{g}_n \triangleq \arg\min_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{g(X_i) \neq Y_i\}}.$$

In other words, $\widehat{g}_n$ is any classifier of the form (3) that minimizes the number of misclassifications on the training sample. Then we have the following:

**Theorem 1.** *There exists an absolute constant $C > 0$, such that for any $n \in \mathbb{N}$ and any $\delta \in (0,1)$, the bound*

$$L(\widehat{g}_n) \leq L^*(\mathscr{G}) + C\sqrt{\frac{d+1}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{4}$$

*holds with probability at least $1-\delta$.*

*Proof.* A standard argument leads to the bound

$$L(\widehat{g}_n) \leq L^*(\mathscr{G}) + 2\Delta_n(Z^n), \tag{5}$$

where

$$\Delta_n(Z^n) \triangleq \sup_{g \in \mathscr{G}} \left| L(g) - L_n(g) \right|$$

is the uniform deviation and $L_n(g)$ denotes the *empirical classification error* of $g$ on $Z^n$:

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{g(X_i) \neq Y_i\}},$$

which is the fraction of incorrectly labeled points in the training sample $Z^n$. Consider a classifier $g \in \mathscr{G}$ and define the set

$$C_g \triangleq \left\{ (x,y) \in \mathbb{R}^d \times \{-1,1\} : y \cdot (\langle w, x \rangle + b) \leq 0 \right\}.$$

Then it is easy to see that

$$L(g) = P(C_g) \qquad \text{and} \qquad L_n(g) = P_n(C_g),$$

where, as before,

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i} = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$$

is the empirical distribution of the sample $Z^n$. Let $\mathscr{C}$ denote the collection of all sets of the form $C = C_g$ for some $g \in \mathscr{G}$. Then

$$\Delta_n(Z^n) = \sup_{C \in \mathscr{C}} |P_n(C) - P(C)|.$$

Let $\mathscr{F} = \mathscr{F}_{\mathscr{C}}$ denote the class of indicator functions of the sets in $\mathscr{C}$: $\mathscr{F}_{\mathscr{C}} = \{\mathbf{1}_{\{\cdot \in C\}} : C \in \mathscr{C}\}$. Then we know that, with probability at least $1-\delta$,

$$\Delta_n(Z^n) \leq 2\mathbb{E}R_n(\mathscr{F}(Z^n)) + \sqrt{\frac{\log(1/\delta)}{2n}}, \tag{6}$$

where $R_n(\mathscr{F}(Z^n))$ is the Rademacher average of the projection of $\mathscr{F}$ onto the sample $Z^n$. Now,

$$\begin{aligned}\mathscr{F}(Z^n) &= \big\{(f(Z_1),\dots,f(Z_n)) : f \in \mathscr{F}\big\} \\ &= \big\{(\mathbf{1}_{\{Z_1\in C\}},\dots,\mathbf{1}_{\{Z_n\in C\}}) : C \in \mathscr{C}\big\}.\end{aligned}$$

Therefore, if we prove that $\mathscr{C}$ is a VC class, then

$$R_n(\mathscr{F}(Z^n)) \le C\sqrt{\frac{V(\mathscr{C})}{n}}.$$

But this follows from the fact that any $C \in \mathscr{C}$ has the form

$$C = \left\{(x,y) \in \mathbb{R}^d \times \{-1,1\} : \sum_{j=1}^{d} w^{(j)} y x^{(j)} + by \le 0\right\}$$

for some $w \in \mathbb{R}^d \backslash \{0\}$ and some $b \in \mathbb{R}$, and the functions $(x,y) \mapsto y$ and $(x,y) \mapsto yx^{(j)}$, $1 \le j \le d$, span a linear space of dimension no greater than $d+1$. Hence, $V(\mathscr{C}) \le d+1$, so that

$$R_n(\mathscr{F}(Z^n)) \le C\sqrt{\frac{V(\mathscr{C})}{n}} \le C\sqrt{\frac{d+1}{n}}.$$

Combining this with (5) and (6), we see that (4) holds with probability at least $1-\delta$. □

### 1.1 Generalized linear discriminant rules

In the same vein, we may consider classification rules of the form

$$g(x) = \begin{cases} 1, & \text{if } \sum_{j=1}^{k} w^{(j)} \psi_j(x) + b > 0 \\ -1, & \text{otherwise} \end{cases} \tag{7}$$

where $k$ is some positive integer (not necessarily equal to $d$), $w = (w^{(1)},\dots,w^{(k)}) \in \mathbb{R}^k$ is a nonzero vector, $b \in \mathbb{R}$ is an arbitrary scalar, and $\Psi = \{\psi_j : \mathbb{R}^d \to \mathbb{R}\}_{j=1}^k$ is some fixed "dictionary" of real-valued functions on $\mathbb{R}^d$. For a fixed $\Psi$, let $\mathscr{G}$ denote the collection of all classifiers of the form (7) as $w$ ranges over all nonzero vectors in $\mathbb{R}^k$ and $b$ ranges over all reals. Then the ERM rule is, again, given by

$$\widehat{g}_n \triangleq \inf_{g\in\mathscr{G}} L_n(g) \equiv \inf_{g\in\mathscr{G}} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{g(X_i)\neq Y_i\}}.$$

The following result can be proved pretty much along the same lines as Theorem 1:

**Theorem 2.** *There exists an absolute constant $C > 0$, such that for any $n \in \mathbb{N}$ and any $\delta \in (0,1)$, the bound*

$$L(\widehat{g}_n) \le L^*(\mathscr{G}) + C\sqrt{\frac{k+1}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{8}$$

*holds with probability at least $1-\delta$.*

### 1.2 Two fundamental issues

As Theorems 1 and 2 show, the ERM algorithm applied to the collection of all (generalized) linear discriminant rules is guaranteed to work well in the sense that the classification error of the output hypothesis will, with high probability, be close to the optimum achievable by any discriminant rule with the given structure. The same argument extends to any collection of classifiers $\mathscr{G}$, for which the "error sets" $\{(x,y) : y \cdot g(x) \le 0\}$, $g \in \mathscr{G}$, form a VC class of dimension much smaller than the sample size $n$. In other words, with high probability the difference

$$L(\widehat{g}_n) - L^*(\mathscr{G}) = L(\widehat{g}_n) - \inf_{g \in \mathscr{G}} L(g)$$

will be small. However, precisely because the VC dimension of $\mathscr{G}$ cannot be too large, the approximation properties of $\mathscr{G}$ will be limited. Another problem is computational. For instance, the problem of finding an empirically optimal linear discriminant rule is NP-hard. In other words, unless P is equal to NP, there is no hope of coming up with an efficient ERM algorithm for linear discriminant rules that would work for all feature space dimensions $d$. If $d$ is fixed, then it is possible to enumerate all projections of a given sample $Z^n$ onto the class of indicators of all halfspaces in $O(n^{d-1}\log n)$ time, which allows for an exhaustive search for an ERM solution, but the usefulness of this naive approach is limited to $d < 5$.

## 2 Risk bounds for combined classifiers via surrogate loss functions

One way to sidestep the above approximation-theoretic and computational issues is to replace the 0–1 Hamming loss function that gives rise to the probability of error criterion with some other loss function. What we gain is the ability to bound the performance of various complicated classifiers built up by combining simpler *base classifiers* in terms of the complexity (e.g, the VC dimension) of the collection of the base classifiers, as well as considerable computational advantages, especially if the problem of minimizing the empirical surrogate loss turns out to be a convex programming problem. What we lose, though, is that, in general, we will not be able to compare the generalization error of the learned classifier to the minimum classification risk. Instead, we will have to be content with the fact that the generalization error will be close to the smallest *surrogate loss*.

We will consider classifiers of the form

$$g_f(x) = \operatorname{sgn} f(x) \equiv \begin{cases} 1, & \text{if } f(x) \ge 0 \\ -1, & \text{otherwise} \end{cases} \tag{9}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is some function. From (2) we have

$$L(g_f) = \mathbb{P}(g_f(X) \ne Y) \le \mathbb{P}(Y g_f(X) < 0) = \mathbb{P}(Y f(X) < 0).$$

From now on, when dealing with classifiers of the form (9), we will write $L(f)$ instead of $L(g_f)$ to keep the notation simple. Now we introduce the notion of a surrogate loss function.

**Definition 1.** *A* surrogate loss function *is any function* $\varphi : \mathbb{R} \to \mathbb{R}_+$, *such that*

$$\varphi(x) \ge \mathbf{1}_{\{x>0\}}. \tag{10}$$

Some examples of commonly used surrogate losses:

1. Exponential, $\varphi(x) = e^x$
2. Logit, $\varphi(x) = \log_2(1 + e^x)$
3. Hinge loss, $\varphi(x) = (x+1)_+ \equiv \max\{x+1, 0\}$

Let $\varphi$ be a surrogate loss. Then for any $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ and any $f : \mathbb{R}^d \to \mathbb{R}$ we have

$$yf(x) < 0 \qquad \Rightarrow \qquad \varphi(-yf(x)) \geq \mathbf{1}_{\{-yf(x)>0\}} = \mathbf{1}_{\{yf(x)<0\}}. \tag{11}$$

Therefore, defining the *$\varphi$-risk* of $f$ by

$$A_\varphi(f) \triangleq \mathbb{E}[\varphi(-Yf(X))]$$

and its empirical version

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

we see from (11) that

$$L(f) \leq A_\varphi(f) \qquad \text{and} \qquad L_n(f) \leq A_{\varphi,n}(f). \tag{12}$$

Now that these preliminaries are out of the way, we can state and prove the basic surrogate loss bound:

**Theorem 3.** *Consider any learning algorithm $\mathscr{A} = \{\mathscr{A}_n\}_{n=1}^\infty$, where, for each $n$, the mapping $\mathscr{A}_n$ receives the training sample $Z^n = (Z_1, \ldots, Z_n)$ as input and produces a function $\widehat{f}_n : \mathbb{R}^d \to \mathbb{R}$ from some class $\mathscr{F}$. Suppose that $\mathscr{F}$ and the surrogate loss $\varphi$ are chosen so that the following conditions are satisfied:*

1. *There exists some constant $B > 0$ such that*

$$\sup_{(x,y)\in\mathbb{R}^d\times\{-1,1\}} \sup_{f\in\mathscr{F}} \varphi(-yf(x)) \leq B$$

2. *There exists some constant $M_\varphi > 0$ such that $\varphi$ is $M_\varphi$-Lipschitz, i.e.,*

$$|\varphi(u) - \varphi(v)| \leq M_\varphi |u - v|, \qquad \forall u, v \in \mathbb{R}.$$

*Then for any $n$ and any $\delta \in (0, 1)$ the following bound holds with probability at least $1 - \delta$:*

$$L(\widehat{f}_n) \leq A_{\varphi,n}(\widehat{f}_n) + 4M_\varphi \mathbb{E}R_n(\mathscr{F}(X^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{13}$$

*Proof.* Using (12), we can write

$$\begin{aligned} L(\widehat{f}_n) &\leq A_\varphi(\widehat{f}_n) \\ &= A_{\varphi,n}(\widehat{f}_n) + A_\varphi(\widehat{f}_n) - A_{\varphi,n}(\widehat{f}_n) \\ &\leq A_{\varphi,n}(\widehat{f}_n) + \sup_{f\in\mathscr{F}} \left| A_\varphi(f) - A_{\varphi,n}(f) \right|. \end{aligned}$$

Now let $\mathscr{H}$ denote the class of functions $h : \mathbb{R}^d \times \{-1,1\} \to \mathbb{R}$ of the form $h(x,y) = -yf(x)$, $f \in \mathscr{F}$. Then

$$\begin{aligned}\sup_{f\in\mathscr{F}} \left|A_\varphi(f) - A_{\varphi,n}(f)\right| &= \sup_{f\in\mathscr{F}} \left|\mathbb{E}[\varphi(-Yf(X))] - \frac{1}{n}\sum_{i=1}^n \varphi(-Y_i f(X_i))\right| \\ &= \sup_{h\in\mathscr{H}} \left|P(\varphi\circ h) - P_n(\varphi\circ h)\right|,\end{aligned}$$

where $\varphi\circ h(z) \triangleq \varphi(h(z))$ for every $z = (x,y) \in \mathbb{R}^d \times \{-1,1\}$. Let

$$\Delta_n(Z^n) \triangleq \sup_{h\in\mathscr{H}} \left|P(\varphi\circ h) - P_n(\varphi\circ h)\right|.$$

We can now use the familiar symmetrization argument to bound

$$\mathbb{E}\Delta_n(Z^n) \le 2\mathbb{E}R_n\left(\varphi\circ\mathscr{H}(Z^n)\right), \tag{14}$$

where $\varphi\circ\mathscr{H}$ denotes the class of all functions of the form $\varphi\circ h$, $h \in \mathscr{H}$, and

$$\varphi\circ\mathscr{H}(Z^n) \triangleq \left\{\left(\varphi\circ h(Z_1),\ldots,\varphi\circ h(Z_n)\right) : h\in\mathscr{H}\right\} = \left\{\left(\varphi(h(Z_1)),\ldots,\varphi(h(Z_n))\right) : h\in\mathscr{H}\right\} \tag{15}$$

is the projection of $\varphi\circ\mathscr{H}$ onto the random sample $Z^n$. We now use a very powerful result about the Rademacher averages called the *contraction principle*, which states the following: If $\mathscr{A} \subset \mathbb{R}^n$ is a bounded set and $\varphi : \mathbb{R} \to \mathbb{R}$ is an $M_\varphi$-Lipschitz function, then

$$R_n(\varphi\circ\mathscr{A}) \le 2M_\varphi R_n(\mathscr{A}), \tag{16}$$

where $\varphi\circ\mathscr{A} \triangleq \left\{(\varphi(a_1),\ldots,\varphi(a_n)) : a = (a_1,\ldots,a_n) \in \mathscr{A}\right\}$. (The proof of the contraction principle is somewhat involved, and we do not give it here.) From (15) we immediately see that

$$\varphi\circ\mathscr{H}(Z^n) = \varphi\circ[\mathscr{H}(Z^n)].$$

Therefore, applying (16) to $\mathscr{A} = \mathscr{H}(Z^n)$ and then using the resulting bound in (14), we obtain

$$\mathbb{E}\Delta_n(Z^n) \le 4M_\varphi\mathbb{E}R_n\left(\mathscr{H}(Z^n)\right).$$

Furthermore, letting $\sigma^n$ be an i.i.d. Rademacher tuple independent of $Z^n$, we have

$$\begin{aligned}R_n(\mathscr{H}(Z^n)) &= \frac{1}{n}\mathbb{E}_{\sigma^n}\left[\sup_{h\in\mathscr{H}}\left|\sum_{i=1}^n \sigma_i h(Z_i)\right|\right] \\ &= \frac{1}{n}\mathbb{E}_{\sigma^n}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i Y_i f(X_i)\right|\right] \\ &= \frac{1}{n}\mathbb{E}_{\sigma^n}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i f(X_i)\right|\right] \\ &\equiv R_n(\mathscr{F}(X^n)),\end{aligned}$$

which leads to

$$\mathbb{E}\Delta_n(Z^n) \le 4M_\varphi\mathbb{E}R_n\left(\mathscr{F}(X^n)\right). \tag{17}$$

Now, since every function $\varphi \circ h$ is bounded between 0 and $B$, the function $\Delta_n(Z^n)$ has bounded differences with $c_1 = \ldots = c_n = B/n$. Therefore, from (17) and from McDiarmid's inequality, we have for every $t > 0$ that

$$\mathbb{P}\Big(\Delta_n(Z^n) \geq 4M_\varphi \mathbb{E}R_n(\mathscr{F}(X^n)) + t\Big) \leq \mathbb{P}\Big(\Delta_n(Z^n) \geq \mathbb{E}\Delta_n(Z^n) + t\Big) \leq e^{-2nt^2/B^2}.$$

Choosing $t = B\sqrt{(2n)^{-1}\log(1/\delta)}$, we see that

$$\Delta_n(Z^n) \leq 4M_\varphi \mathbb{E}R_n(\mathscr{F}(X^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}}$$

with probability at least $1-\delta$. Therefore, since

$$L(\widehat{f}_n) \leq A_{\varphi,h}(\widehat{f}_n) + \Delta_n(Z^n),$$

we see that (13) holds with probability at least $1-\delta$. □

What the above theorem tells us is that the performance of the learned classifier $\widehat{f}_n$ is controlled by the Rademacher average of the class $\mathscr{F}$, and we can always arrange it to be relatively small. We will now look at several specific examples.

## 3 Weighted linear combination of classifiers

Let $\mathscr{G} = \{g : \mathbb{R}^d \to \{-1,1\}\}$ be a class of *base classifiers* (not to be confused with *Bayes* classifiers!), and consider the class

$$\mathscr{F}_\lambda \triangleq \left\{ f = \sum_{j=1}^N c_j g_j : N \in \mathbb{N}, \sum_{j=1}^N |c_j| \leq \lambda; g_1, \ldots, g_N \in \mathscr{G} \right\},$$

where $\lambda > 0$ is a tunable parameter. Then for each $f = \sum_{j=1}^N c_j g_j \in \mathscr{F}_\lambda$ the corresponding classifier $g_f$ of the form (9) is given by

$$g_f(x) = \operatorname{sgn}\left( \sum_{j=1}^N c_j g_j(x) \right).$$

A useful way of thinking about $g_f$ is that, upon receiving a feature $x \in \mathbb{R}^d$, it computes the outputs $g_1(x), \ldots, g_N(x)$ of the $N$ base classifiers from $\mathscr{G}$ and then takes a weighted "majority vote" – indeed, if we had $c_1 = \ldots = c_N = \lambda/N$, then $\operatorname{sgn}(g_f(x))$ would precisely correspond to taking the majority vote among the $N$ base classifiers. Note, by the way, that the number of base classifiers is not fixed, and can be learned from the data.

Now, Theorem 3 tells us that the performance of any learning algorithm that accepts a training sample $Z^n$ and produces a function $\widehat{f}_n \in \mathscr{F}_\lambda$ is controlled by the Rademacher average $R_n(\mathscr{F}_\lambda(X^n))$. It turns out, moreover, that we can relate it to the Rademacher average of the base class $\mathscr{G}$. To start, note that

$$\mathscr{F}_\lambda = \lambda \cdot \operatorname{absconv}\mathscr{G},$$

where

$$\operatorname{absconv}\mathscr{G} = \left\{ \sum_{j=1}^N c_j g_j : N \in \mathbb{N}; \sum_{j=1}^N c = |c_j| \leq 1; g_1, \ldots, g_N \in \mathscr{G} \right\}$$

is the absolute convex hull of $\mathscr{G}$. Therefore

$$R_n(\mathscr{F}_\lambda(X^n)) = \lambda \cdot R_n(\mathscr{G}(X^n)).$$

Now note that the functions in $\mathscr{G}$ are binary-valued. Therefore, assuming that the base class $\mathscr{G}$ is a VC class, we will have

$$R_n(\mathscr{G}(X^n)) \le C\sqrt{\frac{V(\mathscr{G})}{n}}.$$

Combining these bounds with the bound of Theorem 3, we conclude that for any $\widehat{f}_n$ selected from $\mathscr{F}_\lambda$ based on the training sample $Z^n$, the bound

$$L(\widehat{f}_n) \le A_{\varphi,n}(\widehat{f}_n) + CM_\varphi\sqrt{\frac{V(\mathscr{G})}{n}} + B\sqrt{\frac{\log(1/\delta)}{2n}}$$

will hold with probability at least $1-\delta$, where $B$ is the uniform upper bound on $\varphi(-yf(x))$, $f \in \mathscr{F}_\Lambda$, $(x,y) \in \mathbb{R}^d \times \{-1,1\}$ and $M_\varphi$ is the Lipschitz constant of the surrogate loss $\varphi$.

Note that the above bound involves only the VC dimension of the *base class*, which is typically small. On the other hand, the class $\mathscr{F}_\lambda$ obtained by forming weighted combinations of classifiers from $\mathscr{G}$ is extremely rich, and will generally have infinite VC dimension! But there is a price we pay: The first term is the empirical surrogate loss $A_{\varphi,n}(\widehat{f}_n)$, rather than the empirical classification error $L_n(\widehat{f}_n)$. However, it is possible to choose the surrogate $\varphi$ in such a way that $A_{\varphi,n}(\cdot)$ can be bounded in terms of a quantity *related* to the number of misclassified training examples. Here is an example.

Fix a positive parameter $\gamma > 0$ and consider

$$\varphi(x) = \begin{cases} 0, & \text{if } x \le -\gamma \\ 1, & \text{if } x \ge 0 \\ 1 + x/\gamma, & \text{otherwise} \end{cases}$$

This is a valid surrogate loss with $B = 1$ and $M_\varphi = 1/\gamma$, but in addition we have $\varphi(x) \le \mathbf{1}_{\{x>-\gamma\}}$, which implies that $\varphi(-yf(x)) \le \mathbf{1}_{\{yf(x)<\gamma\}}$. Therefore, for any $f$ we have

$$A_{\varphi,n}(f) = \frac{1}{n}\sum_{i=1}^n \varphi(-Y_i f(X_i)) \le \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i f(X_i)<\gamma\}}. \tag{18}$$

The quantity

$$L_n^\gamma(f) \triangleq \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i f(X_i)<\gamma\}} \tag{19}$$

is called the *margin error* of $f$. Notice that:

- For any $\gamma > 0$, $L_n^\gamma(f) \ge L_n(f)$
- The function $\gamma \mapsto L_n^\gamma(f)$ is increasing.

Notice also that we can write

$$L_n^\gamma(f) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i f(X_i)<0\}} + \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{0\le Y_i f(X_i)<\gamma\}},$$

where the first term is just $L_n(f)$, while the second term is the number of training examples that were classified correctly, but only with small "margin" (the quantity $Yf(X)$ is often called the *margin* of the classifier $f$).

**Theorem 4** (Margin-based risk bound for weighted linear combinations)**.** *For any $\gamma > 0$, the bound*

$$L(\widehat{f}_n) \le L_n^{\gamma}(\widehat{f}_n) + \frac{C\lambda}{\gamma}\sqrt{\frac{V(\mathscr{G})}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \tag{20}$$

*holds with probability at least* $1-\delta$.

**Remark 1.** Note that the first term on the right-hand side of (20) increases with $\gamma$, while the second term decreases with $\gamma$. Hence, if the learned classifier $\widehat{f}_n$ has a small margin error for a large $\gamma$, i.e., it classifies the training samples well and with high "confidence," then its generalization error will be small.

# 4 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 2.** *Let* $\mathsf{X}$ *be a closed subset of* $\mathbb{R}^d$*. A real-valued function* $K : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ *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* $x, x' \in \mathsf{X}$.
2. *It is* continuous*, i.e., if* $\{x_n\}$ *is a sequence of points in* $\mathsf{X}$ *converging to a point* $x$*, then*

$$\lim_{n\to\infty} K(x_n, x') = K(x, x'), \qquad \forall\, x' \in \mathsf{X}.$$

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

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \ge 0. \tag{21}$$

**Remark 2.** Another way to interpret the positive semidefiniteness condition is as follows. For any $n$-tuple $x^n = (x_1, \ldots, x_n) \in \mathsf{X}^n$, define the $n \times n$ *kernel Gram matrix*

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

Then (21) 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

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

**Remark 3.** From now on, we will just say "kernel," but always mean "Mercer kernel."

Here are some examples of kernels:

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

2. A more general class of kernels based on the Euclidean inner product can be constructed as follows. Let $\mathsf{X} = \{x \in \mathbb{R}^d : \|x\| \le R\}$; 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 $\mathsf{X} = \mathbb{R}^d$, and let $k : \mathbb{R}^d \to \mathbb{R}$ 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 $\mathsf{X}$. In order to take full advantage of this representational power, we must take a detour into the theory of *Hilbert spaces*.

### 4.1 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 3.** *A real vector space* $\mathsf{V}$ *is an* inner product space *if there exists a function* $\langle \cdot, \cdot \rangle_{\mathsf{V}} : \mathsf{V} \times \mathsf{V} \to \mathbb{R}$, *which is:*

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

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

$$\|v\|_{\mathsf{V}} \triangleq \sqrt{\langle v, v \rangle_{\mathsf{V}}}.$$

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

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

$$\|\alpha v\|_{\mathsf{V}} = \sqrt{\langle \alpha v, \alpha v \rangle_{\mathsf{V}}} = \sqrt{\alpha^2 \langle v, v \rangle_{\mathsf{V}}} = |\alpha| \sqrt{\langle v, v \rangle_{\mathsf{V}}} = |\alpha| \cdot \|v\|_{\mathsf{V}}$$

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

$$\|v + v'\|_{\mathsf{V}} \le \|v\|_{\mathsf{V}} + \|v'\|_{\mathsf{V}}. \tag{22}$$

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

$$|\langle v, v' \rangle_{\mathsf{V}}| \le \|v\|_{\mathsf{V}} \|v'\|_{\mathsf{V}}. \tag{23}$$

To prove (23), we start with the observation that $\|v - \lambda v'\|_{\mathsf{V}}^2 = \langle v - \lambda v', v - \lambda v' \rangle_{\mathsf{V}} \ge 0$ for any $\lambda \in \mathbb{R}$. Expanding this, we get

$$\langle v - \lambda v', v - \lambda v' \rangle_{\mathsf{V}} = \lambda^2 \|v'\|_{\mathsf{V}}^2 - 2\lambda \langle v, v' \rangle_{\mathsf{V}} + \|v\|_{\mathsf{V}}^2 \ge 0.$$

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

$$4|\langle v, v' \rangle_{\mathsf{V}}|^2 \le 4\|v\|_{\mathsf{V}}^2 \|v'\|_{\mathsf{V}}^2 \qquad \Longleftrightarrow \qquad |\langle v, v' \rangle_{\mathsf{V}}| \le \|v\|_{\mathsf{V}} \|v'\|_{\mathsf{V}}.$$

Now we can write

$$\begin{aligned} \big(\|v\|_{\mathsf{V}} + \|v'\|_{\mathsf{V}}\big)^2 &= \|v\|_{\mathsf{V}}^2 + 2\|v\|_{\mathsf{V}}\|v'\|_{\mathsf{V}} + \|v'\|_{\mathsf{V}}^2 \\ &\ge \|v\|_{\mathsf{V}}^2 + 2\langle v, v' \rangle_{\mathsf{V}} + \|v'\|_{\mathsf{V}}^2 \\ &= \langle v, v \rangle_{\mathsf{V}} + \langle v, v' \rangle_{\mathsf{V}} + \langle v', v \rangle_{\mathsf{V}} + \langle v', v' \rangle_{\mathsf{V}} \\ &= \langle v + v', v + v' \rangle_{\mathsf{V}} \\ &\equiv \|v + v'\|_{\mathsf{V}}^2, \end{aligned}$$

where the first step uses the Cauchy–Schwarz inequality, the second step uses the definition of $\|\cdot\|_{\mathsf{V}}$ and the symmetry of $\langle \cdot, \cdot \rangle_{\mathsf{V}}$, the third step uses the linearity of $\langle \cdot, \cdot \rangle_{\mathsf{V}}$, 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, $\|v\|_{\mathsf{V}} \ge 0$, and $\|v\|_{\mathsf{V}} = 0$ 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 $\mathsf{V}$:

**Definition 4.** *Let* $\{v_n\}_{n=1}^{\infty}$ *be a sequence of elements of* $\mathsf{V}$. *We say that it* converges *to* $v \in \mathsf{V}$ *if*

$$\lim_{n \to \infty} \|v_n - v\|_{\mathsf{V}} = 0. \tag{24}$$

**Remark 4.** This definition is valid for any norm on $\mathsf{V}$, 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 (24) implies that for any $\varepsilon > 0$ we can choose $n$ large enough, so that $\|v_n - v\|_{\mathsf{V}} < \varepsilon/2$ for all $m \ge n$. But the triangle inequality gives

$$\|v_n - v_m\|_{\mathsf{V}} \le \|v_n - v\|_{\mathsf{V}} + \|v_m - v\|_{\mathsf{V}} < \varepsilon, \qquad \forall m \ge n.$$

In other words, we have

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

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

$$\lim_{\min(m,n)\to\infty} \|v_n - v_m\| = 0. \tag{25}$$

Any sequence $\{v_n\}$ that has the property (25) 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 5.** *A normed space* $(\mathsf{V}, \|\cdot\|_{\mathsf{V}})$ *is* complete *if any Cauchy sequence* $\{v_n\}$ *of its elements is convergent. If the norm* $\|\cdot\|_{\mathsf{V}}$ *is induced by an inner product, then we say that* $\mathsf{V}$ *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 $\mathsf{V} = \mathbb{R}^d$ 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 $f : \mathbb{R}^d \to \mathbb{R}$, 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 $(\Omega, \mathscr{B}, P)$ be a probability space. Then the space $L^2(P)$ space of all real-valued random variables $X : \Omega \to \mathbb{R}$ 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 $(\mathscr{H}, \langle \cdot, \cdot \rangle_{\mathscr{H}})$; the induced norm will be denoted by $\| \cdot \|_{\mathscr{H}}$.

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 $\mathscr{H}$ are said to be *orthogonal* if $\langle h, g \rangle_{\mathscr{H}} = 0$.

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

**Theorem 5.** *Let $\mathscr{H}_1^\perp$ be the set of all $h^\perp \in \mathscr{H}$, such that $\langle g, h^\perp \rangle_{\mathscr{H}} = 0$ for all $g \in \mathscr{H}_1$. Then:*

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

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

*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 $g \in \mathscr{H}$.*
   *(c) If $g \in \mathscr{H}_1$, then $\Pi g = g$.*
   *(d) For any $g \in \mathscr{H}$ and any $h \in \mathscr{H}_1$,*

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

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

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

**Remark 5.** It is important for $\mathscr{H}_1$ to be a *closed* linear subspace of $\mathscr{H}$ for the above results to hold.

## 4.2 Reproducing kernel Hilbert spaces

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

$$f(x) = \sum_{j=1}^{N} c_j K(x_j, x) \tag{26}$$

for all possible choices of $N \in \mathbb{N}$, $c_1, \dots, c_N \in \mathbb{R}$, and $x_1, \dots, x_N \in \mathsf{X}$. It is easy to see that $\mathscr{L}_K(\mathsf{X})$ is a *vector space*: for any two functions $f, f'$ of the form (26), their sum is also of that form; if we multiply any $f \in \mathscr{L}_K(\mathsf{X})$ by a scalar $c \in \mathbb{R}$, we will get another element of $\mathscr{L}_K(\mathsf{X})$; and the zero function is clearly in $\mathscr{L}_K(\mathsf{X})$. It turns out that, for any (Mercer) kernel $K$, we can *complete* $\mathscr{L}_K(\mathsf{X})$ into a *Hilbert space* of functions that can potentially represent *any* continuous function from $\mathsf{X}$ 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 6.** *Let $\mathsf{X}$ be a closed subset of $\mathbb{R}^d$, and let $K : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ be a Mercer kernel. Then there exists a unique Hilbert space $(\mathscr{H}_K, \langle\cdot,\cdot\rangle_K)$ of real-valued functions on $\mathsf{X}$ with the following properties:*

1. *For all $x \in \mathsf{X}$, the function $K_x(\cdot) \triangleq K(x,\cdot)$ is an element of $\mathscr{H}_K$, and $\langle K_x, K_{x'}\rangle_K = K(x,x')$ for all $x, x' \in \mathsf{X}$.*
2. *The linear space $\mathscr{L}_K(\mathsf{X})$ is* dense *in $\mathscr{H}_K$, i.e., for any $f \in \mathscr{H}_K$ and any $\varepsilon > 0$ there exist some $N \in \mathbb{N}$, $c_1, \dots, c_N \in \mathbb{R}$, and $x_1, \dots, x_N \in \mathsf{X}$, such that*
$$\left\| f - \sum_{j=1}^{N} c_j K_{x_j} \right\|_K < \varepsilon.$$
3. *For all $f \in \mathscr{H}_K$ and all $x \in \mathsf{X}$,*
$$f(x) = \langle K_x, f\rangle_K. \tag{27}$$

*Moreover, the functions in $\mathscr{H}_K$ are continuous. The Hilbert space $\mathscr{H}_K$ is called the* Reproducing Kernel Hilbert Space (RKHS) *associated with $K$; the property (27) is referred to as the* reproducing kernel property.

**Remark 6.** The reproducing kernel property essentially states that the value of any function $f \in \mathscr{H}_K$ at any point $x \in \mathsf{X}$ 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 $f \in \mathscr{L}_K(\mathsf{X})$. Indeed, if $f$ has the form (26), 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 $f \in \mathscr{H}_K$ can be expressed as a limit of functions from $\mathscr{L}_K(\mathsf{X})$, the proof of (27) for a general $f$ follows by continuity.

Now we pick a kernel $K$ on our feature space and consider classifiers of the form (9) with the underlying $f$ taken from a suitable subset of the RKHS $\mathscr{H}_K$. One choice, which underlies such things as the Support Vector Machine, is to take a ball in $\mathscr{H}_K$: given some $\lambda > 0$, let

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

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

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

and is itself convex. Now, as we already know, the performance of any learning algorithm that chooses an element $\widehat{f}_n \in \mathscr{F}_\lambda$ in a data-dependent way is controlled by the Rademacher average $R_n(\mathscr{F}_\lambda(X^n))$. It turns out that this Rademacher average is fairly easy to estimate. Indeed, using the reproducing kernel property (27) and then the linearity of the inner product $\langle \cdot, \cdot \rangle_K$, we can write

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

Now, using the Cauchy–Schwarz inequality (23), it is not hard to show that

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

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

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

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

$$\mathbb{E}_{\sigma^n} \left\| \sum_{i=1}^n \sigma_i g_i \right\|_K \leq \sqrt{\sum_{i=1}^n \|g_i\|_K^2}. \tag{28}$$

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

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

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

$$\mathbb{E} R_n(\mathscr{F}(X^n)) \le \frac{\lambda \sqrt{\mathbb{E} K(X,X)}}{\sqrt{n}}.$$

### 4.3 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 7** (The generalized representer theorem)**.** *Let $\mathsf{X}$ be a closed subset of $\mathbb{R}^d$ and let $\mathsf{Y}$ be a subset of the reals. Consider a nonnegative loss function $\ell : \mathsf{Y} \times \mathsf{Y} \to \mathbb{R}^+$. Let $K$ be a Mercer kernel on $\mathsf{X}$, and let $\mathscr{H}_K$ be the corresponding RKHS.*

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

$$\inf_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) = \inf_{f \in \Pi_n(\mathscr{F})} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)), \tag{29}$$

*and if a minimizer of the left-hand side of (29) exists, then it can be taken to have the form*

$$\widehat{f}_n = \sum_{i=1}^n c_i K_{X_i} \tag{30}$$

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

**Remark 7.** Note that both the subspace $\mathscr{H}_K^{X^n}$ and the corresponding orthogonal projection $\Pi_n$ are *random objects*, since they depend on the random features $X^n$.

*Proof.* Since $K_{X_i} \in \mathscr{H}_K^{X^n}$ for every $i$, by Theorem 5 we have

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

Moreover, from the reproducing kernel property (27) 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 $f \in \mathscr{F}$ we can write

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

This implies that

$$\inf_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) = \inf_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i, \Pi_n f(X_i)\right) = \inf_{g\in\Pi_n(\mathscr{F})} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, g(X_i)). \tag{31}$$

Now suppose that $\widehat{f}_n \in \mathscr{F}$ achieves the infimum on the left-hand side of (31). Then its projection $\Pi_n \widehat{f}_n$ onto $\mathscr{H}_K^{X^n}$ achieves the infimum on the right-hand side. Moreover, since $\Pi_n(\mathscr{F}) \subseteq \mathscr{F}$ by hypothesis, we may conclude that $f = \Pi_n(f)$, i.e., $f \in \mathscr{H}_K^{X^n}$. Since every element of $\mathscr{H}_K^{X^n}$ has the form (30), the theorem is proved. □

In the classification setting, we may take $\mathsf{Y} = \{-1, +1\}$ 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 $\mathscr{F}_\lambda$ in a suitable RKHS $\mathscr{H}_K$. By the above theorem, we may write this problem in the following form:

$$\min_{c_1,\ldots,c_n\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \varphi\left(-Y_i \sum_{j=1}^{n} c_j K(X_i, X_j)\right) \tag{32a}$$

$$\text{subject to } \sum_{i,j=1}^{n} c_i c_j K(X_i, X_j) \le \lambda^2 \tag{32b}$$

Suppose the surrogate loss function $\varphi$ is convex. Then the objective function in (32) is convex as well, and the decision variables $c_1, \ldots, c_n \in \mathbb{R}$ are subject to a quadratic constraint. Thus, (32) is an instance of a *quadratically constrained convex program* (QCCP). Moreover, when $\varphi$ is such that the objective is *quadratic* in $c_1, \ldots, c_n$, 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 (32). For instance, if we let $\varphi$ be the hinge loss $\varphi(u) = (u+1)_+$, then (32) corresponds to the *Support Vector Machine* (SVM) algorithm — more precisely, the SVM is the *scalarized* version of (32), i.e., it has the form

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

for some regularization parameter $\tau > 0$.

## 5 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 $f : \mathsf{X} \to \mathbb{R}$ is equivalent to determining the Bayes classifier (1):

**Theorem 8.** *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 $A_\varphi(f) = \mathbb{E}[\varphi(-Yf(X))]$ over all (measurable) functions $f : \mathsf{X} \to \mathbb{R}$ has the form*

$$f^*(x) = \arg\min_{u \in \mathbb{R}} h_{\eta(x)}(u),$$

*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 (1).*

*Proof.* By the law of iterated expectation,

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

Hence,

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

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

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

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,

$$f^*(x) = \arg\min_{u \in \mathbb{R}} h_{\eta(x)}(u).$$

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 $u^* \le 0$, then $u^* \le -u^*$, so $\varphi'(u^*) \le \varphi'(-u^*)$, which means that $\eta/(1-\eta) \le 1$, i.e., $\eta \le 1/2$. 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., $\operatorname{sgn}(f^*(x))$ is the Bayes classifier. □

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 9.** *Assume that the surrogate loss function $\varphi$ satisfies the conditions of Theorem 3, and that there exist positive constants $s \geq 1$ and $c$, such that the inequality*

$$L(f) - L^* \leq c\left(A_\varphi(f) - A_\varphi^*\right)^{1/s} \tag{33}$$

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

$$\widehat{f}_n = \arg\min_{f \in \mathscr{F}} A_{\varphi,n}(f) = \arg\min_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i f(X_i)). \tag{34}$$

*Then*

$$L(\widehat{f}_n) - L^* \leq 2^{1/s} c \left( 4M_\varphi \mathbb{E} R_n(\mathscr{F}(X^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}} \right)^{1/s} + c\left( \inf_{f \in \mathscr{F}} A_\varphi(f) - A_\varphi^* \right)^{1/s} \tag{35}$$

*with probability at least $1 - \delta$.*

*Proof.* We have the following:

$$\begin{aligned}
L(\widehat{f}_n) - L^* &\leq c\left(A_\varphi(\widehat{f}_n) - A_\varphi^*\right)^{1/s} && (36)\\
&= c\left(A_\varphi(\widehat{f}_n) - \inf_{f \in \mathscr{F}} A_\varphi(f) + \inf_{f \in \mathscr{F}} A_\varphi(f) - A_\varphi^*\right)^{1/s} && (37)\\
&\leq c\left(A_\varphi(\widehat{f}_n) - \inf_{f \in \mathscr{F}} A_\varphi(f)\right)^{1/s} + c\left(\inf_{f \in \mathscr{F}} A_\varphi(f) - A_\varphi^*\right)^{1/s} && (38)\\
&\leq 2^{1/s} c \left( \sup_{f \in \mathscr{F}} \left| A_{\varphi,n}(f) - A_\varphi(f) \right| \right)^{1/s} + c\left(\inf_{f \in \mathscr{F}} A_\varphi(f) - A_\varphi^*\right)^{1/s} && (39)\\
&\leq 2^{1/s} c \left( 4M_\varphi \mathbb{E} R_n(\mathscr{F}(X^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}} \right)^{1/s} \\
&\qquad + c\left(\inf_{f \in \mathscr{F}} A_\varphi(f) - A_\varphi^*\right)^{1/s} \qquad \text{w.p. } \geq 1 - \delta, && (40)
\end{aligned}$$

where:

- (36) follows from (33);
- (38) 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 $s \geq 1$

- (39) and (40) follow from the same argument as the one used in the proof of Theorem 3.

This completes the proof. □

**Remark 8.** Condition (33) is often easy to check. For instance, Zhang [Zha04] proved that it is satisfied, provided the inequality

$$\left|\frac{1}{2}-\eta\right|^s \le (2c)^s\left(1-\inf_u h_\eta(u)\right) \tag{41}$$

holds for all $\eta \in [0,1]$. For instance, (41) 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 $\varphi(u) = (u+1)_+$, (41) holds with $s=1$ and $c=4$.

What Theorem 9 says is that, assuming the expected Rademacher average $\mathbb{E}R_n(\mathscr{F}(X^n)) = O(1/\sqrt{n})$, the difference between the generalization error of the Convex Risk Minimization algorithm (34) 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 $(\inf_{f\in\mathscr{F}} A_\varphi(f) - A_\varphi^*)^{1/s}$ "approximation error" term. If the hypothesis space $\mathscr{F}$ is rich enough, so that $\inf_{f\in\mathscr{F}} A_\varphi(f) = A_\varphi^*$, then the difference between $L(\widehat{f}_n)$ 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.