Chapter 6
VC Dimension, Fat-Shattering Dimension, Rademacher Averages, and Their Applications

Vladimir V. V'yugin

Abstract We consider several complexity measures which capture the difficulty of learning under the i.i.d. assumption. Among these measures are growth function and VC dimension, covering number and fat-shattering dimension, and Rademacher complexity from statistical learning theory. Relationships among these complexity measures, their connection to learning, and tools for bounding them are provided. For each complexity measure, a uniform upper bound on the generalization error of classification problems is presented.

6.1 Introduction

The goal of statistical learning theory is to study, in a statistical framework, the properties of learning algorithms. The set of methods for assessing the quality of classification and regression schemes is called *generalization theory*. In particular, most results in this field take the form of error bounds. This survey chapter introduces the techniques that are used to obtain such results.

In the statistical theory of machine learning, we refer to some underlying probability distribution generating data. We assume that each training or test example is generated at random from a fixed but unknown to us probability distribution and that the data is independently and identically distributed (i.i.d.).

A step aside from the classical theory is that the distribution generating the data may be unknown, and we cannot even estimate its parameters. In this case, the bounds of classification (or regression) errors should be distribution independent. We refer to such a bound as a *generalization error*.

In this theory, the estimates of classification error can be computed, provided that the training was carried out on a large enough random training sample and its resulting classification function agreed with the training set.

V.V. V'yugin (✉)
Institute for Information Transmission Problems, Bol'shoi Karetnyi Pereulok 19,
Moscow, GSP-4 127994, Russia
e-mail: vyugin@iitp.ru

V. Vovk et al. (eds.), *Measures of Complexity*,
DOI 10.1007/978-3-319-21852-6_6

The most important parameter of such an assessment is the *capacity* or complexity of a class of classification functions. Usually in assessing classification errors the length of a training set and the capacity of a class of classification functions are in competition—the longer the training set the greater the capacity of the class of hypotheses which can be used.

We discuss three measures of capacity and the corresponding parameters—growth function and VC dimension (Sect. 6.2), covering number and fat-shattering dimension (Sect. 6.3), and Rademacher averages (Sect. 6.4). Each section concludes with a uniform upper bound on the generalization error in terms of the corresponding complexity. The first of them—VC dimension (and growth function) was introduced by Vapnik and Chervonenkis [11], Vapnik [12] and serves as a starting point for further research in this area. A disadvantage of this characteristic is that for some important classes of classifiers (for example, for separating hyperplanes) it depends on the dimension of the objects' space. Methods based on fat-shattering dimension and Rademacher averages lead to dimension-free bounds. The first of them is tighter but based on the assumption that objects are located in a restricted area. The second one is free from assumptions about the data location area.

In this chapter we consider only the batch setting. For online versions of these notions see Rakhlin et al. [6] and Chap. 15 of this volume.

6.2 Vapnik–Chervonenkis Generalization Theory

A generalization theory presents upper bounds for classification error of a classifier defined using a random training sample. Statistical learning theory uses a hypothesis on the existence of a probabilistic mechanism generating the observed data. In classification or regression problems, these data are pairs (x_i, y_i) of objects and their labels generated sequentially according to some probability distribution unknown to us. We do not try to find parameters of this distribution. We suppose only that pairs (x_i, y_i) are i.i.d. (independently and identically distributed) with respect to this distribution. Methods used in statistical learning theory are uniform with respect to all probability distributions from this very broad class.

A classifier (or regression function) is constructed from a training sample using methods of optimization. A class of classification functions can be very broad—from the class of all separating hyperplanes in n-dimensional Euclidian space to a class of arbitrary n-dimensional manifolds that are mapped using kernel methods to hyperplanes in more general spaces. No probability distributions are used in algorithms computing values of these classifiers.

In this section, let $\mathcal{X}$ be a set of objects equipped with a σ-algebra of Borel sets and a probability distribution P. Also, let $D = \{-1, +1\}$ be a set of labels of elements of $\mathcal{X}$.

Let $S = ((x_1, y_1), \dots, (x_l, y_l))$ be a training sample, where $x_i \in \mathcal{X}$ and $y_i \in \{-1, 1\}$ for $1 \leq i \leq l$. In probabilistic analysis, we suppose that the training sample S is a vector random variable consisting of random variables (x_i, y_i), $i = 1, \dots, l$.

Let a classifier $h : \mathcal{X} \to \{-1, 1\}$ be given. Its *classification error* (risk functional) is defined as

$$\mathrm{err}_P(h) = P\{(x, y) : h(x) \neq y\},$$

that is as the probability of a wrong classification. The classifier h *agrees* with a sample $S = ((x_1, y_1), \dots, (x_l, y_l))$ if $h(x_i) = y_i$ for all $1 \leq i \leq l$.

A simple and natural approach to the classification problem is to consider a class of classifiers h and use data-based estimates of the probabilities of error $\mathrm{err}_P(h)$ to select a classifier from the class. The most natural choice to estimate the probability of error $\mathrm{err}_P(h)$ is the error count

$$\mathrm{err}_S(h) = \frac{1}{l}\, |\{i : h(x_i) \neq y_i, 1 \leq i \leq l\}|,$$

which is called the *empirical error* of the classifier h on a sample S. Here $|A|$ is the cardinality of a finite set A.

We start with the simplest special case. Assume that a classifier h agrees with a sample S, i.e., $\mathrm{err}_S(h) = 0$. For any $\epsilon > 0$ we have

$$\begin{aligned} P\{S : \mathrm{err}_S(h) = 0 \,\&\, \mathrm{err}_P(h) > \epsilon\} &= \prod_{i=1}^{l} P\{h(x_i) = y_i\} \\ &= \prod_{i=1}^{l} (1 - P\{h(x_i) \neq y_i\}) = (1 - \mathrm{err}_P(h))^l \leq e^{-l\epsilon}. \end{aligned} \tag{6.1}$$

Let H be a class of classification hypotheses. For a finite class H, by (6.1), we have the bound:

$$P^l\{S : (\exists h \in H)(\mathrm{err}_S(h) = 0 \,\&\, \mathrm{err}_P(h) > \epsilon)\} \leq |H|\, e^{-l\epsilon}. \tag{6.2}$$

For an infinite class H of classifiers a similar bound can be obtained using Vapnik–Chervonenkis generalization theory. In this case the cardinality of a finite class is replaced by the *growth function* of the infinite class H:

$$B_H(l) = \max_{(x_1, x_2, \dots, x_l)} |\{(h(x_1), h(x_2), \dots, h(x_l)) : h \in H\}|\,.$$

The set $\{x_1, \dots, x_l\}$ is *shattered* by the class H if $\{(h(x_1), \dots, h(x_l)) : h \in H\} = \{-1, 1\}^l$. As follows from the definition, $B_H(l) \leq 2^l$ for all l, and if there exists a sample of length l that is shattered by H, then $B_H(l) = 2^l$.

The following theorem (Vapnik–Chervonenkis, Sauer, Shelah) is the main result of the theory of VC dimension.

Theorem 6.1 *For any class H of indicator functions, one of the following two conditions holds:*

- *$B_H(l) = 2^l$ for all l, i.e., for each l an ordered sample of length l shattered by H exists.*
- *There exists a sample of maximal length d that is shattered by H. In this case $B_H(l) = 2^l$ for $l \le d$ and $B_H(l) \le \sum\limits_{i=0}^{d} \binom{l}{i} \le \left(\frac{el}{d}\right)^d$ for $l > d$.*

In other words, the function $G_H(l) = \ln B_H(l)$ is linear for all l or becomes logarithmic: $O(d \ln l)$ for all $l > d$. For example, it cannot be $O(l^r)$ for $0 < r < 1$.

The number d is called the VC dimension (Vapnik–Chervonenkis dimension) of H; VC dimension is infinite in the first case.

The main result of Vapnik–Chervonenkis generalization theory is an analogue of the inequality (6.2) for infinite class H:

Theorem 6.2 *For $l > 2/\epsilon$, the following upper bound is valid:*

$$P^l\{S : (\exists\, h \in H)(\mathrm{err}_S(h) = 0 \,\&\, \mathrm{err}_P(h) > \epsilon)\} \le 2B_H(2l)e^{-\epsilon l/4}. \tag{6.3}$$

The PAC-learning form of this result is as follows.

Corollary 6.1 *Assume that a class H of classifiers has a finite VC dimension d and a critical probability $0 < \delta < 1$ of accepting a wrong classification hypothesis $h \in H$ agreeing with a training sample S is given.*

Then with probability $\ge 1-\delta$ any classifier $h_S \in H$ defined by a training sample S and agreeing with it has the classification error

$$\mathrm{err}_P(h_S) \le \frac{4}{l}\left(d \ln \frac{2el}{d} + \ln \frac{2}{\delta}\right)$$

for $l \ge d$.

These results can be generalized for the case of learning with mistakes.

Theorem 6.3 *For $l > 2/\epsilon$, the following upper bound is valid:*

$$P^l\{S : (\exists\, h \in H)(\mathrm{err}_P(h) - \mathrm{err}_S(h) > \epsilon)\} \le 4B_H(2l)e^{-\epsilon^2 l/2}.$$

The PAC-learning form is as follows.

Corollary 6.2 *Assume that a class H of classifiers has a finite VC dimension d. Then for any $0 < \delta < 1$, with probability $\ge 1 - \delta$, for any $h \in H$ the following inequality holds:*

$$\mathrm{err}_P(h) \le \mathrm{err}_S(h) + \sqrt{\frac{2}{l}\left(d \ln \frac{2el}{d} + \ln \frac{4}{\delta}\right)},$$

where $l \ge d$.

For the proof, we refer the reader to Vapnik and Chervonenkis [11], Vapnik [12], Bousquet et al. [4], and so on.

6.3 Margin-Based Performance Bounds for Classification

Let $\mathcal{F}$ be a class of real valued functions with domain $\mathcal{X}$, and let $S = ((x_1, y_1), \dots, (x_l, y_l))$ be a sample of length l. A function $f \in \mathcal{F}$ defines the classifier:

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

It can be shown that the VC dimension of the class of all classifiers defined by the separating linear hyperplanes in the Euclidian space $\mathcal{R}^n$ is equal to $n+1$. In practice, the length of a sample can be less than n, and bounds on the classification error like (6.3) are useless in this case. By this reasoning, Theorem 6.2 and Corollary 6.1 can have only a theoretical meaning. This drawback is connected with a too poor method used for separation of the data. Separating training sample with arbitrarily small thresholds, we lose the predictive performance of our classification algorithms. Also, we do not restrict the space where our training sample is located.

In what follows we will consider methods of separation with a given positive threshold γ and will suppose that the points generated by the probability distributions are located in some ball in the Euclidian space $\mathcal{R}^n$ of a given radius R. Using γ and R as the new parameters, we will define a new dimension-free notion of the capacity of the functional class $\mathcal{F}$. We present new upper bounds for the classification error which can have some practical meaning.

For a function $f \in \mathcal{F}$ we define its margin on an example (x_i, y_i) to be $\gamma_i = y_i f(x_i)$. The functional margin of a training set $S = ((x_1, y_1), \dots, (x_l, y_l))$ is defined to be: $m_S(f) = \min\limits_{i=1,\dots,l} \gamma_i$. If $\gamma_i > 0$ then the classification by means of f is right. It holds $m_S(f) > 0$ if and only if the function f classifies all examples from the sample S right and with a positive threshold.

Let $\epsilon > 0$. A finite set $\mathcal{B}$ of functions is called an *ϵ-cover* of a functional class $\mathcal{F}$ on a set $X = \{x_1, \dots, x_l\}$ if for any $f \in \mathcal{F}$ a function $g \in \mathcal{B}$ exists such that $|f(x_i) - g(x_i)| < \epsilon$ for all $i = 1, \dots, l$. Define the covering number of a class $\mathcal{F}$ on a set X:

$$\mathcal{N}(\epsilon, \mathcal{F}, X) = \min\{|\mathcal{B}| : \mathcal{B} \text{ is an } \epsilon\text{-cover of } \mathcal{F}\}.$$

Define the *covering number* $\mathcal{N}(\epsilon, \mathcal{F}, l)$ of a class $\mathcal{F}$ as the maximum of all covering numbers of the class $\mathcal{F}$ on sets X such that $|X| = l$:

$$\mathcal{N}(\epsilon, \mathcal{F}, l) = \max_{|X|=l} \mathcal{N}(\epsilon, \mathcal{F}, X).$$

Let $\mathrm{err}_S(f)$ be the empirical error of a classifier h_f on the training set $S = ((x_1, y_1), \dots, (x_l, y_l))$. This number is equal to the fraction in S of all examples (x_i, y_i) such that $h_f(x_i) \neq y_i$.

Let P be a probability distribution on $\mathcal{X} \times \{-1, 1\}$ generating elements of the sample S. Then the classification error of the classifier h_f can be written as

$$\mathrm{err}_P(f) = P\{h_f(x) \neq y\}.$$

The following theorem is an analogue of Theorem 6.2.

Theorem 6.4 *For any $\epsilon > 0$, $\gamma > 0$, and $l > 2/\epsilon$,*

$$P^l\{S : (\exists\, f \in \mathcal{F})(m_S(f) \geq \gamma \,\&\, \mathrm{err}_P(f) > \epsilon)\} \leq 2\mathcal{N}(\gamma/2, \mathcal{F}, 2l)e^{-\epsilon l/4}.$$

The PAC-learning form of this result is as follows.

Corollary 6.3 *Let a class $\mathcal{F}$ of real functions and numbers $\gamma > 0$, $\delta > 0$ be given. Then for any probability distribution P on $\mathcal{X} \times \{-1, 1\}$, with probability $1 - \delta$, any function $f \in \mathcal{F}$ with margin bound $m_S(f) > \gamma$ on a random sample S has classification error*

$$\mathrm{err}_P(f) \leq \frac{4}{l}\left(\log \mathcal{N}(\gamma/2, \mathcal{F}, 2l) + \log \frac{2}{\delta}\right)$$

for all l.[1]

We define the fat-shattering dimension of a class $\mathcal{F}$ of functions. Let $\gamma > 0$. A set $X = \{x_1, \ldots, x_l\}$ of objects is called *γ-shattered* if numbers $r_1, \ldots, r_l$ exist such that for any subset $E \subseteq X$ a function $f_E \in \mathcal{F}$ exists such that $f_E(x_i) \geq r_i + \gamma$ if $x_i \in E$ and $f_E(x_i) < r_i - \gamma$ if $x_i \notin E$ for all i.

The *fat-shattering* dimension $\mathrm{fat}_\gamma(\mathcal{F})$ of a class $\mathcal{F}$ is equal to the cardinality of the maximal γ-shattered set X. The fat-shattering dimension of the class $\mathcal{F}$ depends on the parameter $\gamma > 0$. A class $\mathcal{F}$ has infinite fat-shattering dimension if there are γ-shattered sets of arbitrarily large size.

Covering and Packing numbers. Consider these notions from a more general position. Let $(\mathcal{D}, d)$ be a metric space with a metric $d(x, y)$ which defines the distance between any two elements $x, y \in \mathcal{X}$.

Let $A \subseteq \mathcal{D}$, $B \subseteq A$, and α be a positive number. The set B is called an *α-cover* of the set A if for any $a \in A$ a $b \in B$ exists such that $d(a, b) < \alpha$. A *covering number* of the set A is a function:

$$\mathcal{N}_d(\alpha, A) = \min\{|B| : B \text{ is an } \alpha\text{-covering of } A\}. \tag{6.4}$$

We say that the set $B \subseteq \mathcal{D}$ is *α-separated* if $d(a, b) > \alpha$ for any $a, b \in B$ such that $a \neq b$. A *packing number* of the set A is a function

$$\mathcal{M}_d(\alpha, A) = \max\{|B| : B \subseteq A \text{ is } \alpha\text{-separated}\}. \tag{6.5}$$

The covering number and the packing number are closely related.

[1] By $\log r$ we mean logarithm to base 2.

Lemma 6.1 *For any $A \subseteq \mathcal{D}$ and $\alpha > 0$,*

$$\mathcal{M}_d(2\alpha, A) \leq \mathcal{N}_d(\alpha, A) \leq \mathcal{M}_d(\alpha, A).$$

The main purpose of this section is to present an outline of the proof of Theorem 6.5. To carry this out, we need to further develop our dimension theory for functions with a finite number of values.

Let $\mathcal{X}$ be a set and $B = \{0, 1, \ldots, b\}$ be a finite set. Also, let $\mathcal{F} \subseteq B^{\mathcal{X}}$ be a class of functions with domain $\mathcal{X}$ and range in the finite set B. Consider a metric on $\mathcal{F}$:

$$l(f, g) = \sup_{x \in \mathcal{X}} |f(x) - g(x)|.$$

Any two functions $f, g \in \mathcal{F}$ are said to be *separated* (2-separated) if $l(f, g) > 2$. In other words, an $x \in \mathcal{X}$ exists such that $|f(x) - g(x)| > 2$. A class $\mathcal{F}$ is said to be *pairwise separated* if any two different functions $f, g \in \mathcal{F}$ are separated.

Let $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ be a linearly ordered set—a sample—and $\mathcal{F} \subseteq B^{\mathcal{X}}$. We say that the class $\mathcal{F}$ *strongly shatters* the set X if there exists a collection $s = \{s_1, \ldots, s_n\}$ of elements of B such that for all $E \subseteq X$ a function $f_E \in \mathcal{F}$ exists such that $f_E(x_i) \geq s_i + 1$ if $x_i \in E$ and $f_E(x_i) \leq s_i - 1$ if $x_i \notin E$, for all i. In this case we also say that $\mathcal{F}$ strongly shatters the set X according to s. The *strong dimension* of $\mathcal{F}$, denoted $\mathrm{Sdim}(\mathcal{F})$, is the size of the largest set strongly shattered by $\mathcal{F}$.

We will shift our attention from real-valued functions $f : \mathcal{X} \to [0, 1]$ to ones taking values in a finite set by a simple discretization. For any real $\alpha > 0$ define $f^{\alpha}(x) = \left[\frac{f(x)}{\alpha}\right]$ for all x, where $[r]$ is the closest integer to r such that $|r - [r]| \leq \frac{1}{2}$. If the number r is located in the middle of the interval between two integer numbers we define $[r]$ using some tie-breaking rule. Define $\mathcal{F}^{\alpha} = \{f^{\alpha} : f \in \mathcal{F}\}$.

Clearly, the range of any function f^{α} is a subset of the set $\{0, 1, \ldots, \lfloor 1/\alpha \rfloor\}$.

The covering number $\mathcal{N}_d(\alpha, A)$ and the packing number $\mathcal{M}_d(\alpha, A)$ were defined by (6.4) and (6.5).

Let us define a specific metric on the class $\mathcal{F}$ connected with the set $X = \{x_1, \ldots, x_n\}$: $l_X(f, g) = \max_{1 \leq i \leq n} |f(x_i) - g(x_i)|$. Consider the corresponding covering and packing numbers:

$$\mathcal{N}(\alpha, \mathcal{F}, X) = \mathcal{N}_{l_X}(\alpha, \mathcal{F}),$$
$$\mathcal{M}(\alpha, \mathcal{F}, X) = \mathcal{M}_{l_X}(\alpha, \mathcal{F}).$$

The following lemma relates the combinatorial dimensions and packing numbers of the classes $\mathcal{F}$ and $\mathcal{F}^{\alpha}$.

Lemma 6.2 *Let $\mathcal{F} \subseteq B^{\mathcal{X}}$ and $\alpha > 0$. Then*

$$\mathrm{Sdim}(\mathcal{F}^{\alpha}) \leq \mathrm{fat}_{\alpha/2}(\mathcal{F}), \tag{6.6}$$

$$\mathcal{M}(\alpha, \mathcal{F}, X) \leq \mathcal{M}(2, \mathcal{F}^{\alpha/2}, X). \tag{6.7}$$

We can now state the main result of the theory of combinatorial dimension—the Alon, Ben-David, Cesa-Bianchi, and Haussler theorem [1].

Theorem 6.5 *Let $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ and $\alpha \in [0,1]$. Denote $d = \mathrm{fat}_{\alpha/4}(\mathcal{F})$. Then*

$$\mathcal{N}(\alpha, \mathcal{F}, n) \le 2\left(n\left(\frac{2}{\alpha}+1\right)^2\right)^{\lceil d\log(\frac{2en}{d\alpha})\rceil}.$$

The following lemma is the main technical part of the proof of Theorem 6.5.

Lemma 6.3 *Let $|\mathcal{X}| = n$ and $B = \{0,1,\dots,b\}$. Also, let $\mathcal{F} \subseteq B^{\mathcal{X}}$ and $d =$ $\mathrm{Sdim}(\mathcal{F})$. Then $\mathcal{M}_l(2,\mathcal{F}) \le 2(n(b+1)^2)^{\lceil \log y\rceil}$, where $y = \sum\limits_{i=1}^{d}\binom{n}{i}b^i$.*

Using the fact that the covering number does not exceed the packing number, inequality (6.7) of Lemmas 6.2 and 6.3, we obtain the following chain of inequalities: $\mathcal{N}(\alpha,\mathcal{F},n) = \sup\limits_{|X|=n}\mathcal{N}(\alpha,\mathcal{F},X) \le \sup\limits_{|X|=n}\mathcal{M}(\alpha,\mathcal{F},X) \le \sup\limits_{|X|=n}\mathcal{M}(2,\mathcal{F}^{\alpha/2},X) =$ $\mathcal{M}_l(2,\mathcal{F}^{\alpha/2}) \le 2(n(b+1)^2)^{\lceil\log y\rceil}$, where $b = \lceil\frac{2}{\alpha}\rceil$.

Note that the class $\mathcal{F}^{\alpha/2}$ satisfies the assumption of Lemma 6.3 for $b = \lceil\frac{2}{\alpha}\rceil$.

From inequality (6.6) of Lemma 6.2, the inequality $d' \le \mathrm{fat}_{\alpha/4}(\mathcal{F}) = d$ follows. Hence, $y \le \sum\limits_{i=1}^{d}\binom{n}{i}b^i \le b^d\sum\limits_{i=1}^{d}\binom{n}{i} \le b^d\left(\frac{en}{d}\right)^d$. In particular, $\log y \le d\log\left(\frac{ben}{d}\right)$.

The following corollary is a reformulation of this theorem with a little attenuation of estimates.

Corollary 6.4 *Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to [a,b]$, where $a < b$. For $0 <$ $\gamma < 1$ denote $d = \mathrm{fat}_{\gamma/4}(\mathcal{F})$. Then*

$$\log\mathcal{N}(\gamma,\mathcal{F},l) \le 1 + d\log\frac{2el(b-a)}{d\gamma}\log\frac{4l(b-a)^2}{\gamma^2}.$$

Corollaries 6.3 and 6.4 imply the following

Corollary 6.5 *Let $\mathcal{F}$ be a class of real functions with the range $[-1,1]$, $\gamma > 0$, $\delta > 0$, and P be a probability distribution generating a sample S. Then, with probability $1-\delta$, any hypothesis $f \in \mathcal{F}$ with the margin bound $m_S(f) \ge \gamma$ has classification error*

$$\mathrm{err}_P(f) \le \frac{4}{l}\left(d\log\frac{16el}{d\gamma}\log\frac{128l}{\gamma^2} + \log\frac{2}{\delta}\right),$$

where $d = \mathrm{fat}_{\gamma/8}(\mathcal{F})$.

A dimension-free upper bound on the fat-shattering dimension can be obtained for the class of all (homogeneous) linear functions on $\mathcal{R}^n$ with restricted domain.

Theorem 6.6 *Let $X = \{\bar{x} : |\bar{x}| \le R\}$ be a ball of radius R in n-dimensional Euclidian space and $\mathcal{F}$ be the class of all homogeneous linear functions $f(\bar{x}) = (\bar{w} \cdot \bar{x})$, where $\|\bar{w}\| \le 1$ and $\bar{x} \in X$. Then*

$$\mathrm{fat}_\gamma(\mathcal{F}) \le \left(\frac{R}{\gamma}\right)^2.$$

Substituting the bound of Theorem 6.6 into the bound of Corollary 6.5, we obtain the final theorem:

Theorem 6.7 *Consider the classification problem by use of linear homogeneous functions $f(\bar{x}) = (\bar{w} \cdot \bar{x})$, where $\bar{x} \in \mathcal{R}^n$ and $\|\bar{w}\| \le 1$.*

Let a number $\gamma > 0$ and a probability distribution P concentrated in the ball of radius R and centered at the origin be given. Also, let a sample $S = ((\bar{x}_1, y_1), \dots, (\bar{x}_l, y_l))$ be generated by the probability distribution P. Then, with probability $1 - \delta$, any classification hypothesis f with margin bound $m_S(f) \ge \gamma$ has classification error

$$\mathrm{err}_P(f) \le \frac{4}{l}\left(\frac{64R^2}{\gamma^2} \log \frac{el\gamma}{4R} \log \frac{128Rl}{\gamma^2} + \log \frac{2}{\delta}\right). \tag{6.8}$$

The bounds of Theorems 6.6 and 6.7 form the basis of the theory of dimension-free bounds of classification errors.

Inseparable training sample. Now we extend the upper bound (6.8) to the case where a training sample is not completely separated by a classification function. This estimate serves as a basis for setting the corresponding optimization problem of constructing an optimal classifier.

Let a class $\mathcal{F}$ of functions of type $\mathcal{X} \to \mathcal{R}$ be given. Their domain $\mathcal{X}$ is a subset of $\mathcal{R}^n$. Any such function $f \in \mathcal{F}$ defines a classifier:

$$h(x) = \begin{cases} 1 & \text{if } f(x) \ge 0, \\ -1 & \text{otherwise.} \end{cases}$$

Let a sample $S = ((x_1, y_1), \dots, (x_l, y_l))$ be given and $\gamma_i = y_i f(x_i)$ be the margin of an example $(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$ with respect to a function $f \in \mathcal{F}$.

We define the *margin slack variable* of an example $(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$ with respect to a function $f \in \mathcal{F}$ and target margin $\gamma > 0$ to be the quantity

$$\xi_i = \max\{0, \gamma - y_i f(\bar{x}_i)\}.$$

This is the amount by which the function f fails to achieve margin γ for the example (x_i, y_i).

A vector $\bar{\xi} = (\xi_1, \dots, \xi_l)$ is called the *margin slack vector* of a training set $S = ((x_1, y_1), \dots, (x_l, y_l))$. By definition $y_i f(x_i) + \xi_i \ge \gamma$ for all i.

If the norm of vector $\bar{\xi}$ is positive the training sample is inseparable by the classifier $f(\bar{x})$ with a threshold of $\gamma > 0$. Theorem 6.7 is not directly applicable in this case.

However, in the case of a linear classifier in Euclidian space $\mathcal{R}^n$ we can replace this problem by an equivalent one in a space of higher dimension, where a modified training set is separable. The corresponding result of Shawe-Taylor and Cristianini [5, 8] is presented in the following theorem.

Theorem 6.8 *Let $\gamma > 0$ and $\mathcal{L}$ be a class of all linear homogeneous functions $f(\bar{x}) = (\bar{w} \cdot \bar{x})$, where $\|\bar{w}\| \le 1$. Also, let P be a probability distribution on $\mathcal{X} \times \{-1, 1\}$ with support a ball of radius R centered at the origin and there is no discrete probability on misclassified training points.*

Then for any $\delta > 0$, with probability $1 - \delta$, any classifier $f \in \mathcal{L}$ has a generalization error

$$\mathrm{err}_P(f) \le \frac{c}{l}\left(\frac{R^2 + \|\bar{\xi}\|^2}{\gamma^2}\log^2 l + \log\frac{1}{\delta}\right),$$

where $\bar{\xi}$ is the margin slack vector with respect to f and a target margin $\gamma > 0$ and c is a constant.

6.4 Rademacher Averages

In this section, we consider another definition of the capacity of a class of functions, Rademacher averages. Let $z^l = (z_1, \dots, z_l)$ be a sample of unlabeled examples whose elements belong to some set $\mathcal{X}$ structured as a probability space, and P be a probability distribution on $\mathcal{X}$. Assume that the elements of z^l are generated in the i.i.d. manner according to the probability distribution P. Also let $\mathcal{F}$ be a class of real-valued functions defined on $\mathcal{X}$.

Let $\sigma_1, \dots, \sigma_l$ be i.i.d. Bernoulli variables taking values $+1$ and -1 with equal probability: $B_{1/2}(\sigma_i = 1) = B_{1/2}(\sigma_i = -1) = 1/2$ for all $1 \le i \le l$. Such variables are called *Rademacher variables*.

Define the *empirical Rademacher average* of the class $\mathcal{F}$ as the random variable (that is a function of random variables $z_1, \dots, z_l$)

$$\tilde{\mathcal{R}}_l(\mathcal{F}) = E_\sigma\left(\sup_{f \in \mathcal{F}} \frac{1}{l}\sum_{i=1}^{l} \sigma_i f(z_i)\right).$$

The *Rademacher average* of the class $\mathcal{F}$ is defined as

$$\mathcal{R}_l(\mathcal{F}) = E_{P^l}(\tilde{\mathcal{R}}_l(\mathcal{F})) = E_{P^l} E_\sigma\left(\sup_{f \in \mathcal{F}} \frac{1}{l}\sum_{i=1}^{l} \sigma_i f(z_i)\right).$$

By definition the Rademacher average is the mathematical expectation of the empirical Rademacher average with respect to probability distribution P.

Rademacher averages give us a powerful tool to obtain uniform convergence bounds. We present some properties of Rademacher averages, which will be used for obtaining the uniform upper bounds of the generalization error.

Assume that the elements of a sample $z^l = (z_1, \dots, z_l)$ are generated i.i.d. by some probability distribution P. By definition the empirical mean of a function f on the sample z^l equals

$$\hat{E}_{z^l}(f) = \frac{1}{l}\sum_{i=1}^{l} f(z_i).$$

The true mathematical expectation of the function f is equal to $E_P(f) = \int f(z)P(dz)$.

Theorem 6.9 *The following uniform bounds over class $\mathcal{F}$ are valid:*

- *Bound on the difference between the empirical and true expectations*

$$E_{z^l \sim P^l}\left(\sup_{f \in \mathcal{F}}\left(E_P(f) - \hat{E}_{z^l}(f)\right)\right) \le 2\mathcal{R}_l(\mathcal{F}). \tag{6.9}$$

- *Bounds on the difference between the expectation of the function and the sample mean of this function: for any $\delta > 0$, with probability $1 - \delta$, for all $f \in \mathcal{F}$,*

$$E_P(f) \le \hat{E}_{z^l}(f) + 2\mathcal{R}_l(\mathcal{F}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}},$$

$$E_P(f) \le \hat{E}_{z^l}(f) + 2\tilde{\mathcal{R}}_l(\mathcal{F}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2l}}.$$

- *Rademacher complexity of composition: assume that ϕ be an L-Lipschitz continuous function, i.e., $|\phi(x) - \phi(y)| \le L|x - y|$ for all x and y. Then*

$$\tilde{\mathcal{R}}_l(\phi(\mathcal{F})) \le L\tilde{\mathcal{R}}_l(\mathcal{F}), \tag{6.10}$$

$$\mathcal{R}_l(\phi(\mathcal{F})) \le L\mathcal{R}_l(\mathcal{F}). \tag{6.11}$$

Proof (of inequality (6.9)). Given a random sample $z^l = (z_1, \dots, z_l)$, let $\tilde{z}^l = (\tilde{z}_1, \dots, \tilde{z}_l)$ be a "ghost sample." This means that random variables $\tilde{z}_i$ are independent of each other and of z_i, $i = 1, \dots, l$, and have the same distribution as the latter.

The following chain of equalities and inequalities is valid:

$$E_{z^l \sim P^l}\left(\sup_{f \in \mathcal{F}}\left(E_P(f) - \frac{1}{l}\sum_{i=1}^{l} f(z_i)\right)\right)$$

$$= E_{z^l \sim P^l}\left(\sup_{f \in \mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l} E_{\tilde{z}_i \sim P}(f(\tilde{z}_i) - f(z_i))\right)\right)$$

$$\leq E_{z^l\sim P^l}\left(E_{\tilde{z}^l\sim P^l}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}(f(\tilde{z}_i)-f(z_i))\right)\right)\right)$$

$$= E_{z^l\tilde{z}^l\sim P^{2l}}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}(f(\tilde{z}_i)-f(z_i))\right)\right) \tag{6.12}$$

$$= E_{z^l\tilde{z}^l\sim P^{2l}}E_{\sigma\sim B_{1/2}}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i(f(\tilde{z}_i)-f(z_i))\right)\right)$$

$$\leq E_{\tilde{z}^l\sim P^l}E_{\sigma\sim B_{1/2}}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(\tilde{z}_i)\right)\right)$$

$$+ E_{z^l\sim P^l}E_{\sigma\sim B_{1/2}}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(z_i)\right)\right) = 2\mathcal{R}_l(\mathcal{F}). \tag{6.13}$$

We are allowed to insert σ_i in line (6.13) since the mathematical expectation of the supremum in (6.12) is invariant under the transposition of any variables z_i and $\tilde{z}_i$; this is why we can insert the symbol of mathematical expectation $E_{\sigma\sim B_{1/2}}$ in (6.13). □

Proof (of inequalities (6.10) and (6.11)). Let $z^l = (z_1, \ldots, z_l)$ be a random sample distributed according to a probability distribution P, $\sigma_1, \ldots, \sigma_l$ be i.i.d. Bernoulli random variables taking values in the set $\{-1, +1\}$, and let P^l be the probability distribution on the set of all such sequences of length l induced by P.

The transformations given below are valid for mathematical expectations $E = E_\sigma$ and $E = E_{P^l}E_\sigma$. Thus we will prove both inequalities (6.10) and (6.11) simultaneously.

By definition the (empirical) Rademacher average of the class $\phi(\mathcal{F})$ is equal to

$$\mathcal{R}_l(\phi(\mathcal{F})) = E\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi(f(z_i))\right). \tag{6.14}$$

For simplicity, we assume that $L = 1$.[2] We need to prove that

$$\mathcal{R}_l(\phi(\mathcal{F})) \leq \mathcal{R}_l(\mathcal{F}) = E\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(z_i)\right). \tag{6.15}$$

We make the transition from (6.14) to (6.15) step by step. At each step, we consider a sequence of auxiliary functions $(\phi_1, \ldots, \phi_l)$, where each function ϕ_i is ϕ or the identity function I.

At the first step all the functions are ϕ: $\phi_i = \phi$ for all i, at the last step all these functions are identity functions $\phi_i = I$ for all i.

[2] One can replace the function ϕ by ϕ/L.

We also assume that at each step, except the last one, $\phi_1 = \phi$. In the transition to the next step the next function $\phi_i = \phi$ will be replaced by the identity function: $\phi'_i = I$. This will be achieved by the following chain of equalities and inequalities:

$$\begin{aligned}
& E\left(\sup_{f\in\mathcal{F}} \frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi_i(f(z_i))\right) \\
&= \frac{1}{2l}E\left(\sup_{f\in\mathcal{F}}\left(\phi(f(z_1)) + \sum_{i=2}^{l}\sigma_i\phi_i(f(z_i))\right) + \sup_{f\in\mathcal{F}}\left(-\phi(f(z_1)) + \sum_{i=2}^{l}\sigma_i\phi_i(f(z_i))\right)\right) \\
&= \frac{1}{2l}E\left(\sup_{f,f'\in\mathcal{F}}\left(\phi(f(z_1)) + \sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)) - \phi(f'(z_1)) + \sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i))\right)\right) \\
&\le \frac{1}{2l}E\left(\sup_{f,f'\in\mathcal{F}}\left(\left|f(z_1) - f'(z_1)\right| + \sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)) + \sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i))\right)\right) \\
&= \frac{1}{2l}E\left(\sup_{f,f'\in\mathcal{F}}\left(f(z_1) - f'(z_1) + \sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)) + \sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i))\right)\right) \\
&\le \frac{1}{2l}E\left(\sup_{f\in\mathcal{F}}\left(f(z_1) + \sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)) + \sup_{f'\in\mathcal{F}}\left(-f'(z_1) + \sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i))\right)\right)\right) \\
&= E\left(\sup_{f\in\mathcal{F}} \frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi'_i(f(z_i))\right), \qquad (6.16)
\end{aligned}$$

where the collection of functions $\phi'_1, \ldots, \phi'_l$ contains one more identity function than the previous collection $\phi_1, \ldots, \phi_l$.

In transition from the first line to the second one, we take the mathematical expectation over σ_1; after that one can still consider E as the expectation over the whole set σ, because now the variable σ_1 is absent. In transition from the third line to the fourth one, we have used the observation that the supremum is achieved by non-negative values of the difference $\phi(f(z_1)) - \phi(f'(z_1))$, so we can replace it by its absolute value. After that, Lipschitz's condition has used for $L = 1$. A similar reasoning was used in the transition from the fourth line to the fifth one. The transition from the fifth line to the sixth one was done by the same reasoning as the transition from the first line to the second one.

Applying several times the chain of transformations (6.16) we obtain the expression

$$E\left(\sup_{f\in\mathcal{F}} \frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi'_i(f(z_i))\right), \qquad (6.17)$$

where all ϕ'_i are identity functions, and so the sum (6.17) is equal to $\mathcal{R}_l(\mathcal{F})$.

The first line of the chain (6.16) is equal to $\mathcal{R}_l(\phi(\mathcal{F}))$ for $E = E_{P^l}E_\sigma$ or to $\tilde{\mathcal{R}}_l(\phi(\mathcal{F}))$ for $E = E_\sigma$. Thus, the inequalities (6.10) and (6.11) are satisfied. □

The connection of the Rademacher average with other known measures of capacity of classes of functions—the growth function $B_{\mathcal{F}}(l)$ and the covering number $\mathcal{N}(\alpha, \mathcal{F}, l)$—is presented in the following theorem.

Theorem 6.10 *The following inequalities are valid:*

- *The Rademacher average and the growth function:*
  Let $\mathcal{F}$ be a class of indicator functions taking values in the set $\{-1, +1\}$. Then

$$\mathcal{R}_l(\mathcal{F}) \le \sqrt{\frac{2\ln B_{\mathcal{F}}(l)}{m}}$$

  for all l.
- *The empirical Rademacher average and the covering number:*

$$\tilde{\mathcal{R}}_l(\mathcal{F}) \le \inf_\alpha \left(\sqrt{\frac{2\ln \mathcal{N}(\alpha, \mathcal{F}, z^l)}{l}} + \alpha\right).$$

- *The Rademacher average and the covering number:*

$$\mathcal{R}_l(\mathcal{F}) \le \inf_\alpha \left(\sqrt{\frac{2\ln \mathcal{N}(\alpha, \mathcal{F}, l)}{l}} + \alpha\right).$$

For more information see Bartlett and Mendelson [2], Bartlett et al. [3], and Shawe-Taylor and Cristianini [9].

Rademacher averages and generalization error. Now, we present upper bounds on the generalization error for classification functions defined by threshold functions from RKHS (reproducing kernel Hilbert space). On kernels in statistical learning theory see Scholkopf and Smola [7], Steinwart [10], Shawe-Taylor and Cristianini [9].

Let $\mathcal{F}$ be a Hilbert space of functions defined on some set $\mathcal{X}$. We also assume that this space is RKHS, i.e., it is generated by a reproducing kernel $K(x, y)$. Any function $f \in \mathcal{F}$ is represented as a scalar product $f(x) = (f \cdot \phi(x))$, where $\phi(x) = K(x, \cdot)$.

An example of such an RKHS can be defined by a mapping $\phi : \mathcal{R}^n \to \mathcal{R}^N$. Let $\mathcal{F}$ be a space of functions $f(\bar{x}) = (\bar{w} \cdot \phi(\bar{x}))$, where $\bar{x} \in \mathcal{R}^n$, $\bar{w} \in \mathcal{R}^N$ and $(\bar{w} \cdot \bar{w}')$ is the dot product in $\mathcal{R}^N$. The norm of f is defined as $\|f\| = \|\bar{w}\|$, and the scalar product of functions f and $g(\bar{x}) = (\bar{w}' \cdot \phi(\bar{x}))$ is defined as $(f \cdot g) = (\bar{w} \cdot \bar{w}')$. The function $K(\bar{x}, \bar{y}) = (\phi(\bar{x}) \cdot \phi(\bar{y}))$ is the corresponding kernel.

Any function $f \in \mathcal{F}$ defines the classifier

$$h(x) = \begin{cases} 1 & \text{if } f(x) \ge 0, \\ -1 & \text{otherwise.} \end{cases}$$

Let $\mathcal{F}_1 = \{f \in \mathcal{F} : \|f\| \le 1\}$. In the example above, $\mathcal{F}_1$ is the class of functions $f(\bar{x}) = (\bar{w} \cdot \phi(\bar{x}))$ such that $\|\bar{w}\| \le 1$.

Assume that a training set $S = ((x_1, y_1), \dots, (x_l, y_l))$ is given, where $x_i \in \mathcal{X}$ and $y_i \in \{-1, 1\}$.

Let $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n$ be the Gram matrix defined by the values of the kernel on objects of the sample S; $\mathrm{tr}(\mathbf{K}) = \sum_{i=1}^l K(x_i, x_i)$ is the trace of the matrix $\mathbf{K}$.

Now we estimate the empirical Rademacher average of the class $\mathcal{F}_1$ relative to the training set S.

Theorem 6.11 *The empirical Rademacher average of the class $\mathcal{F}_1$ relative to the training set $S = ((x_1, y_1), \dots, (x_l, y_l))$ satisfies the inequality:*

$$\tilde{\mathcal{R}}_l(\mathcal{F}_1) \le \frac{1}{l}\sqrt{\mathrm{tr}(\mathbf{K})}. \tag{6.18}$$

Proof The following chain of equalities and inequalities is valid:

$$\begin{aligned}
\tilde{\mathcal{R}}_l(\mathcal{F}_1) &= E_\sigma\left(\sup_{f\in\mathcal{F}_1} \frac{1}{l}\sum_{i=1}^l \sigma_i f(x_i)\right) \\
&= E_\sigma\left(\sup_{\|f\|\le 1}\left(f \cdot \frac{1}{l}\sum_{i=1}^l \sigma_i \phi(x_i)\right)\right) \\
&\le \frac{1}{l}E_\sigma\left(\left\|\sum_{i=1}^l \sigma_i \phi(\bar{x}_i)\right\|\right) \\
&= \frac{1}{l}E_\sigma\left(\left(\sum_{i=1}^l \sigma_i \phi(x_i) \cdot \sum_{i=1}^l \sigma_i \phi(x_i)\right)^{1/2}\right) \\
&\le \frac{1}{l}\left(E_\sigma\left(\sum_{i,j=1}^l \sigma_i\sigma_j K(x_i, x_j)\right)\right)^{1/2} \\
&= \frac{1}{l}\left(\sum_{i=1}^l K(x_i, x_i)\right)^{1/2}.
\end{aligned}$$

Here in the transition from the second line to the third the Cauchy–Schwarz inequality was used, and in the transition from the third line to the fourth the definition of the norm vector was used. In the transition from the fourth line to the fifth Jensen's inequality was used, in the transition from the fifth line to the sixth, we have used the independence of the random variables σ_i and equalities $E(\sigma_i^2) = 1$ and $E(\sigma_i\sigma_j) = E(\sigma_i)E(\sigma_j) = 0$ for $i \ne j$. The theorem is proved. □

Let $S = ((x_1, y_1), \dots, (x_l, y_l))$ be a sample and $\gamma_i = y_i f(x_i)$ be the margin of an example $(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$ with respect to a function $f \in \mathcal{F}$.

Given a number $\gamma > 0$, let $\xi_i = \max\{0, \gamma - y_i f(x_i)\}$ be the margin slack variable for a function f and $\bar{\xi} = (\xi_1, \dots, \xi_l)$ be the corresponding margin slack vector.

Define an auxiliary function $f(x, y) = -yf(x)$ and the corresponding class of functions with domain $\mathcal{X} \times \{-1, 1\}$:

$$\mathcal{F}_2 = \{f(x, y) : f(x, y) = -yf(x), f \in \mathcal{F}_1\}.$$

Let

$$\chi(x) = \begin{cases} 1 & \text{if } x \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

Also, let $\text{sign}(r) = 1$ for $r \ge 0$ and $\text{sign}(r) = -1$ otherwise.

Assume that examples (x_i, y_i) of the training set S are generated i.i.d. by some probability distribution P. It is easy to verify that $P\{(x, y) : y \ne \text{sign}(f(x))\} \le E_P(\chi(-yf(x)))$. Let $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n$ be the Gram matrix defined by the kernel and the training set S.

The following theorem gives an upper bound for the generalization error of the classifier defined by the kernel K.

Theorem 6.12 *For any $\delta > 0$ and l, with probability $1 - \delta$, for any function $f \in \mathcal{F}_1$,*

$$P\{y \ne \text{sign}(f(x))\} \le \frac{1}{l\gamma} \sum_{i=1}^{l} \xi_i + \frac{2}{l\gamma} \sqrt{\text{tr}(\mathbf{K})} + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2l}}. \tag{6.19}$$

Note that the right side of (6.19) is a random variable.

Proof Define the auxiliary function

$$g(r) = \begin{cases} 1 & \text{if } r > 0, \\ 1 + r/\gamma & \text{if } -\gamma \le r \le 0, \\ 0 & \text{otherwise.} \end{cases}$$

Since $g(r) \ge \chi(r)$ for all r, and by Theorem 6.9, with probability $1 - \delta$,

$$\begin{aligned} E_P(\chi(f(x, y))) &\le E_P(g(f(x, y))) \\ &\le \tilde{E}_S(g(f(x, y))) + 2\tilde{\mathcal{R}}_l(g \circ \mathcal{F}_2) + 3\sqrt{\frac{\ln(2/\delta)}{2l}}. \end{aligned} \tag{6.20}$$

By definition of the margin slack variable $g(-y_i f(x_i)) \le \xi_i/\gamma$ for $1 \le i \le l$.

Let us bound the empirical Rademacher average of the class $\mathcal{F}_2$:

$$\begin{aligned}\tilde{\mathcal{R}}_l(\mathcal{F}_2) &= E_\sigma\left(\sup_{f\in\mathcal{F}_2}\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(x_i, y_i)\right)\\ &= E_\sigma\left(\sup_{f\in\mathcal{F}_1}\frac{1}{l}\sum_{i=1}^{l}\sigma_i y_i f(x_i)\right)\\ &= E_\sigma\left(\sup_{f\in\mathcal{F}_1}\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(x_i)\right)\\ &= \tilde{\mathcal{R}}_l(\mathcal{F}_1) \le \frac{1}{l}\sqrt{\mathrm{tr}(K)}.\end{aligned}$$

Since the function g is Lipschitz continuous with the constant $L = 1/\gamma$, we have, by Theorem 6.9, $\tilde{\mathcal{R}}_l(g\circ\mathcal{F}_2) \le \tilde{\mathcal{R}}_l(\mathcal{F}_2)/\gamma = \tilde{\mathcal{R}}_l(\mathcal{F}_1)/\gamma$. By definition for any $f \in \mathcal{F}_2$

$$\tilde{E}_S(g\circ f) = \frac{1}{l}\sum_{i=1}^{l} g(-y_i f(\bar{x}_i)) \le \frac{1}{l\gamma}\sum_{i=1}^{l}\xi_i.$$

By the inequalities (6.20) and (6.18) of Theorem 6.11, with probability $1-\delta$,

$$E_P(\chi(f(x,y))) \le \frac{1}{l\gamma}\sum_{i=1}^{l}\xi_i + \frac{2}{l\gamma}\sqrt{\mathrm{tr}(\mathbf{K})} + 3\sqrt{\frac{\ln(2/\delta)}{2l}}.$$

The theorem is proved. □

Let us turn to the example of $\mathcal{F}$ given above. Unlike the bound (6.8) obtained in the theory of fat-shattering dimension, the bound (6.19) has best constants and does not require prior knowledge of the radius of a ball containing vectors of the training sample.

The bound (6.19) is worse than a similar estimate obtained using the fat-shattering dimension. Let $\|\bar{x}_i\| \le R$ for all $1 \le i \le l$. For small values, the order of the variable $\frac{2}{l\gamma}\sqrt{\mathrm{tr}(\mathbf{K})} \le \frac{2}{l\gamma}\sqrt{lR^2} = 2\sqrt{\frac{R^2}{l\gamma^2}}$ is much greater than the order of the leading term of the bound (6.8) of Theorem 6.7, which is approximately $O\left(\frac{R^2}{l\gamma^2}\right)$.

Acknowledgments This work was partially supported by the Russian Foundation for Fundamental Research (13-01-12447 and 13-01-00521).

References

1. Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D.: Scale-sensitive dimensions, uniform convergence, and learnability. J. ACM **44**(4), 615–631 (1997)
2. Bartlett, P., Mendelson, S.: Rademacher and Gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3**, 463–482 (2002)
3. Bartlett, P., Bousquet, O., Mendelson, S.: Local Rademacher complexities. Ann. Stat. **33**, 1497–1537 (2005)
4. Bousquet, O., Boucheron, S., Lugosi, G.: Introduction to statistical learning theory. In: Bousquet, O., von Luxburg, U., Ratsch, R. (eds.) Advanced Lectures on Machine Learning. Lecture Notes in Computer Science, vol. 3176, pp. 169–207. Springer, Berlin (2004)
5. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge (2000)
6. Rakhlin, A., Sridharan, K., Tewari, A.: Online learning: beyond regret. In: Proceedings of the 24th Annual Conference on Learning Theory, JMLR Workshop and Conference Proceedings, vol. 19, pp. 559–594 (2011). Longer version available as arXiv:1011.3168
7. Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press, Cambridge (2002)
8. Shawe-Taylor, J., Cristianini, N.: Margin distribution bounds on generalization. In: Proceedings of the European Conference on Computational Learning Theory, EuroCOLT'99. Lecture Notes in Computer Science, vol. 1572, pp. 263–273 (1999)
9. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
10. Steinwart, I.: On the influence of the kernel on the consistency of support vector machines. J. Mach. Learn. Res. **2**, 67–93 (2001)
11. Vapnik, V.N., Chervonenkis, A.Y.: Теория распознавания образов: Статистические проблемы обучения (Theory of Pattern Recognition: Statistical Problems of Learning: in Russian). Nauka, Moscow (1974). German translation: Theorie der Zeichenerkennung, transl. K.G. Stöckel and B. Schneider, ed. S. Unger and B. Fritzsch, Akademie Verlag, Berlin (1979)
12. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)