ESAIM: Probability and Statistics Will be set by the publisher
URL: http://www.emath.fr/ps/

THEORY OF CLASSIFICATION: A SURVEY OF RECENT ADVANCES *

Olivier Bousquet[1], Stéphane Boucheron[2] and Gábor Lugosi[3]

Abstract. The last few years have witnessed important new developments in the theory and practice of pattern classification. We intend to survey some of the main new ideas that have lead to these important recent developments.

Résumé. Durant ces dernières années, la théorie et la pratique de la reconnaissance des formes ont été marquées par des développements originaux. Ce survol présente certaines des principales idées novatrices qui ont conduit à ces développements importants.

1991 Mathematics Subject Classification. 62G08,60E15,68Q32.

June 17, 2004.

Contents

Keywords and phrases: Pattern Recognition, Statistical Learning Theory, Concentration Inequalities, Empirical Processes, Model Selection

* *The first and third authors acknowledge support by the PASCAL Network of Excellence under EC grant no. 506778. The work of the third author was supported by the Spanish Ministry of Science and Technology and FEDER, grant BMF2003-03324*

[1] Max Planck Institute for Biological Cybernetics, Spemannstr. 38, D-72076 Tübingen, Germany, olivier.bousquet@tuebingen.mpg.de

[2] Laboratoire de Recherche en Informatique, CNRS & Université Paris-Sud, Orsay, France, stephane.boucheron@lri.fr

[3] Pompeu Fabra University, Barcelona, Spain, lugosi@upf.es

1. Introduction

The last few years have witnessed important new developments in the theory and practice of pattern classification. The introduction of new and effective techniques of handling high-dimensional problems—such as boosting and support vector machines—have revolutionized the practice of pattern recognition. At the same time, the better understanding of the application of empirical process theory and concentration inequalities have lead to effective new ways of understanding these methods and provided a statistical explanation for their success. These new tools have also helped develop new model selection methods that are at the heart of any classification method.

The purpose of this survey is to offer an overview of some of these theoretical tools and give the main ideas of the analysis of some of the important algorithms. This survey does not attempt to be exhaustive. The selection of the topics is largely influenced by the personal taste of the authors. We also limit ourselves to describing the key ideas in a simple way, often sacrificing generality. In these cases the reader is pointed to the references for the sharpest and more general results available.

2. Basic model

The problem of pattern classification is about guessing or predicting the unknown class of an observation. An observation is often a collection of numerical and/or categorical measurements represented by a d-dimensional vector x but in some cases it may even be a curve or an image. In our model we simply assume that $x \in \mathcal{X}$ where $\mathcal{X}$ is some abstract (measurable) set. The unknown nature of the observation is called a *class*. It is denoted by y and in the simplest case takes values in the binary set $\{-1, 1\}$.

In these notes we restrict our attention to binary classification. The reason is simplicity and that the binary problem already captures many of the main features of more general problems. Even though there is much to say about multiclass classification, this survey does not cover this increasing field of research.

In classification, one creates a function $g(x) : \mathcal{X} \to \{-1, 1\}$ which represents one's guess of y given x. The mapping g is called a *classifier*. The classifier errs on x if $g(x) \neq y$.

To model the learning problem, we introduce a probabilistic setting, and let (X, Y) be an $\mathcal{X} \times \{-1, 1\}$-valued random pair. The distribution of the random pair (X, Y) may be described by the pair (μ, η), where μ is the probability measure for X (i.e., $\mu(A) = \mathbb{P}\{X \in A\}$) and $\eta(x) = \mathbb{P}\{Y = 1 | X = x\}$. The function η is called the *a posteriori probability*. We measure the performance of classifier g by its *probability of error*

$$L(g) = \mathbb{P}\{g(X) \neq Y\} \ .$$

Given η, one may easily construct a classifier with minimal probability of error. In particular, it is easy to see that if we define

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

then $L(g^*) \leq L(g)$ for any classifier g. $L^* \stackrel{\text{def}}{=} L(g^*)$ is the *Bayes risk* (or Bayes error). More precisely, it is immediate to see that

$$L(g) - L^* = \mathbb{E}\left[\mathbb{1}_{\{g(X) \neq g^*(X)\}} \, |2\eta(X) - 1|\right] \geq 0 \ . \tag{1}$$

g^* is often called the *Bayes classifier* and In the statistical model we focus on, one has access to a collection of data (X_i, Y_i), $1 \le i \le n$. We assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a sequence of independent identically distributed (*i.i.d.*) random pairs with the same distribution as that of (X, Y).

A classifier is constructed on the basis of $X_1, Y_1, \ldots, X_n, Y_n$ and is denoted by g_n. Thus, the value of Y is guessed by $g_n(X) = g_n(X; X_1, Y_1, \ldots, X_n, Y_n)$. The performance of g_n is measured by its (conditional) *probability of error*

$$L(g_n) = \mathbb{P}\{g_n(X) \neq Y | X_1, Y_1, \ldots, X_n, Y_n)\} \ .$$

The focus of the theory (and practice) of classification is to construct classifiers g_n whose probability of error is as close to L^* as possible.

Obviously, the whole arsenal of traditional parametric and nonparametric statistics may be used to attack this problem. However, the high-dimensional nature of many of the new applications (such as image recognition, text classification, micro-biological applications, etc.) leads to territories beyond the reach of traditional methods. Most new advances of statistical learning theory intent to face these new challenges.

Bibliographical remarks. Several textbooks, surveys, and research monographs have been written on pattern classification and statistical learning theory. A partial list includes Anthony and Bartlett [8], Anthony and Biggs [9], Breiman, Friedman, Olshen, and Stone [49], Devijver and Kittler [66], Devroye, Györfi, and Lugosi [67], Duda and Hart [71], Duda, Hart, and Stork [72], Fukunaga [92], Kearns and Vazirani [110], Kulkarni, Lugosi, and Venkatesh [121], Lugosi [137], McLachlan [163], Mendelson [165], Natarajan [168], Ripley [177], Vapnik [219,220], Vapnik and Chervonenkis [223], and Vidyasagar [225].

3. Empirical risk minimization and Rademacher averages

A simple and natural approach to the classification problem is to consider a class $\mathbb{C}$ of classifiers $g : \mathcal{X} \to \{-1, 1\}$ and use data-based estimates of the probabilities of error $L(g)$ to select a classifier from the class. The most natural choice to estimate the probability of error $L(g) = \mathbb{P}\{g(X) \neq Y\}$ is the error count

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

$L_n(g)$ is called the *empirical error* of the classifier g.

First we outline the basics of the theory of empirical risk minimization (i.e., the classification analog of M-estimation). Denote by g_n^* the classifier that minimizes the estimated probability of error over the class:

$$L_n(g_n^*) \le L_n(g) \qquad \text{for all } g \in \mathbb{C}.$$

Then the probability of error

$$L(g_n^*) = \mathbb{P}\left\{ g_n^*(X) \neq Y \middle| D_n \right\}$$

of the selected rule is easily seen to satisfy the elementary inequalities

$$L(g_n^*) - \inf_{g \in \mathbb{C}} L(g) \le 2 \sup_{g \in \mathbb{C}} |L_n(g) - L(g)| \ , \tag{2}$$

$$L(g_n^*) \le L_n(g_n^*) + \sup_{g \in \mathbb{C}} |L_n(g) - L(g)| \ .$$

We see that by guaranteeing that the uniform deviation $\sup_{g \in \mathbb{C}} |L_n(g) - L(g)|$ of estimated probabilities from their true values is small, we make sure that the probability of the selected classifier g_n^* is not much larger than the best probability of error in the class $\mathbb{C}$ and at the same time the empirical estimate $L_n(g_n^*)$ is also good.

Clearly, the random variable $nL_n(g)$ is binomially distributed with parameters n and $L(g)$. Thus, to obtain bounds for the success of empirical error minimization, we need to study uniform deviations of binomial random variables from their means. We formulate the problem in a somewhat more general way as follows. Let

$X_1, \ldots, X_n$ be independent, identically distributed random variables taking values in some set $\mathcal{X}$ and let $\mathcal{F}$ be a class of bounded functions $\mathcal{X} \to [-1, 1]$. Denoting expectation and empirical averages by $P(f) = \mathbb{E}f(X_1)$ and $P_n(f) = (1/n)\sum_{i=1}^n f(X_i)$, we are interested in upper bounds for the maximal deviation

$$\sup_{f\in\mathcal{F}}(P(f) - P_n(f)) \ .$$

Concentration inequalities are among the basic tools in studying such deviations. The simplest, yet quite powerful exponential concentration inequality is the *bounded differences inequality* which states that if $g : \mathcal{X}^n \to \mathbb{R}$ is a function of n variables such that for some nonnegative constants $c_1, \ldots, c_n$,

$$\sup_{\substack{x_1,\ldots,x_n,\\ x_i'\in\mathcal{X}}} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \le c_i \ , \ 1 \le i \le n$$

and $X_1, \ldots, X_n$ are independent random variables, then the random variable $Z = g(X_1, \ldots, X_n)$ satisfies

$$\mathbb{P}\left\{|Z - \mathbb{E}Z| > t\right\} \le 2e^{-2t^2/C} \ .$$

The bounded differences assumption means that if the i-th variable of g is changed while keeping all the others fixed, the value of the function cannot change by more than c_i.

Our main example for such a function is

$$Z = \sup_{f\in\mathcal{F}}(P(f) - P_n(f)) \ .$$

Obviously, Z satisfies the bounded differences assumption with $c_i = 2/n$ and therefore, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sup_{f\in\mathcal{F}}(P(f) - P_n(f)) \le \mathbb{E}\sup_{f\in\mathcal{F}}(P(f) - P_n(f)) + \sqrt{\frac{2\log\frac{1}{\delta}}{n}} \ . \tag{3}$$

This concentration result allows us to focus on the expected value that can be bounded conveniently by a simple symmetrization device. Introduce the "ghost sample" $X_1', \ldots, X_n'$, independent of the X_i and distributed identically. If $P_n'(f) = (1/n)\sum_{i=1}^n f(X_i')$ denotes the empirical averages measured on the ghost sample, then by Jensen's inequality,

$$\mathbb{E}\sup_{f\in\mathcal{F}}(P(f) - P_n(f)) = \mathbb{E}\sup_{f\in\mathcal{F}}\left(\mathbb{E}[P_n'(f) - P_n(f)|X_1, \ldots, X_n]\right) \le \mathbb{E}\sup_{f\in\mathcal{F}}\left(P_n'(f) - P_n(f)\right) \ .$$

Let now $\sigma_1, \ldots, \sigma_n$ be independent random variables with $\mathbb{P}\{\sigma_i = 1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$, independent of the X_i and X_i'. Then

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left(P_n'(f) - P_n(f)\right) = \mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(f(X_i')-f(X_i)) = \mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i(f(X_i')-f(X_i)) \le 2\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(X_i) \ .$$

Let $A \in \mathbb{R}^n$ be a bounded set of vectors $a = (a_1, \ldots, a_n)$, and introduce the quantity

$$R_n(A) = \mathbb{E}\sup_{a\in A}\frac{1}{n}\sum_{i=1}^n\sigma_i a_i \ .$$

$R_n(A)$ is called the *Rademacher average* associated to A. For a given sequence $x_1, \ldots, x_n \in \mathcal{X}$, we write $\mathcal{F}(x_1^n)$ for the class of n-vectors $(f(x_1), \ldots, f(x_n))$ with $f \in \mathcal{F}$. Thus, using this notation, we have deduced that with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} (P(f) - P_n(f)) \leq 2\mathbb{E}R_n(\mathcal{F}(X_1^n) + \sqrt{\frac{2 \log \frac{1}{\delta}}{n}} \ .$$

Noticing that the random variable $R_n(\mathcal{F}(X_1^n)$ satisfies the conditions of the bounded differences inequality, we also have

$$\sup_{f \in \mathcal{F}} (P(f) - P_n(f)) \leq 2R_n(\mathcal{F}(X_1^n)) + \sqrt{\frac{2 \log \frac{2}{\delta}}{n}} \ .$$

This is our first *data-dependent* performance bound. It involves the Rademacher average of the coordinate projection of $\mathcal{F}$ given by the data $X_1, \ldots, X_n$. Given the data, one may calculate the Rademacher average, for example, by Monte Carlo integration. Note that for a given choice of the random signs $\sigma_1, \ldots, \sigma_n$, the computation of $\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i)$ is equivalent to minimizing $-\sum_{i=1}^n \sigma_i f(X_i)$ over $f \in \mathcal{F}$ and therefore it is computationally equivalent to empirical risk minimization. $R_n(\mathcal{F}(X_1^n))$ measures the richness of the class $\mathcal{F}$ and provides a sharp estimate for the maximal deviations. In fact, one may prove that

$$\frac{1}{2}\mathbb{E}R_n(\mathcal{F}(X_1^n)) - \frac{1}{2\sqrt{n}} \leq \mathbb{E} \sup_{f \in \mathcal{F}} (P(f) - P_n(f)) \leq 2\mathbb{E}R_n(\mathcal{F}(X_1^n)))$$

(see, e.g., [217]).

Next we recall some of the simple structural properties of Rademacher averages. Let A, B be bounded subsets of $\mathbb{R}^n$ and let $c \in \mathbb{R}$ be a constant. Then the following subadditivity properties are immediate from the definition:

$$R_n(A \cup B) \leq R_n(A) + R_n(B) \ , \qquad R_n(c \cdot A) = |c| R_n(A) \ , \qquad R_n(A \oplus B) \leq R_n(A) + R_n(B)$$

where $c \cdot A = \{ca : a \in A\}$ and $A \oplus B = \{a + b : a \in A, b \in B\}$. It is also easy to see that if $A = \{a^{(1)}, \ldots, a^{(N)}\} \subset \mathbb{R}^n$ is a finite set, then

$$R_n(A) \leq \max_{j=1,\ldots,N} \|a^{(j)}\| \frac{\sqrt{2 \log N}}{n} \ . \tag{4}$$

The above inequality follows by *Hoeffding's inequality* which states that if X is a bounded zero-mean random variable taking values in an interval $[\alpha, \beta]$, then for any $s > 0$, $\mathbb{E} \exp(sX) \leq \exp\left(s^2(\beta - \alpha)^2/8\right)$. In particular, by independence,

$$\mathbb{E} \exp\left(s \frac{1}{n} \sum_{i=1}^n \sigma_i a_i \right) = \prod_{i=1}^n \mathbb{E} \exp\left(s \frac{1}{n} \sigma_i a_i \right) \leq \prod_{i=1}^n \exp\left(\frac{s^2 a_i^2}{2n^2} \right) = \exp\left(\frac{s^2 \|a\|^2}{2n^2} \right)$$

This implies that

$$\begin{aligned} e^{sR_n(A)} &= \exp\left(s\mathbb{E} \max_{j=1,\ldots,N} \frac{1}{n} \sum_{i=1}^n \sigma_i a_i^{(j)} \right) \leq \mathbb{E} \exp\left(s \max_{j=1,\ldots,N} \frac{1}{n} \sum_{i=1}^n \sigma_i a_i^{(j)} \right) \\ &\leq \sum_{j=1}^N \mathbb{E} e^{s \frac{1}{n} \sum_{i=1}^n \sigma_i a_i^{(j)}} \leq N \max_{j=1,\ldots,N} \exp\left(\frac{s^2 \|a^{(j)}\|^2}{2n^2} \right) \ . \end{aligned}$$

Taking the logarithm of both sides, dividing by s, and choosing s to minimize the obtained upper bound for $R_n(A)$, we arrive at (4). Finally, we mention two important properties of Rademacher averages. The first is

that if $\mathrm{absconv}(A) = \left\{\sum_{j=1}^N c_j a^{(j)} : N \in \mathbb{N},\ \sum_{j=1}^N |c_j| \le 1, a^{(j)} \in A\right\}$ is the absolute convex hull of A, then

$$R_n(A) = R_n(\mathrm{absconv}(A)) \tag{5}$$

as it is easily seen from the definition. The second is known as the *contraction principle*: let $\phi : \mathbb{R} \to \mathbb{R}$ be a function with $\phi(0) = 0$ and Lipschitz constant L_ϕ. Defining $\phi \circ A$ as the set of vectors of form $(\phi(a_1), \ldots, \phi(a_n)) \in \mathbb{R}^n$ with $a \in A$, we have

$$R_n(\phi \circ A) \le L_\phi R_n(A) \ .$$

Often it is useful to derive further upper bounds on Rademacher averages. As an illustration, we consider the case when $\mathcal{F}$ is a class of indicator functions. Recall that this is the case in our motivating example in the classification problem described above when each $f \in \mathcal{F}$ is the indicator function of a set of the form $\{(x, y) : \mathbb{1}_{g(x) \neq y}$. In such a case, for any collection of points $x_1^n = (x_1, \ldots, x_n)$, $\mathcal{F}(x_1^n)$ is a finite subset of $\mathbb{R}^n$ whose cardinality is denoted by $\mathbb{S}_{\mathcal{F}}(x_1^n)$ and is called the VC *shatter coefficient*. Obviously, $\mathbb{S}_{\mathcal{F}}(x_1^n) \le 2^n$. By inequality (4), we have, for all x_1^n,

$$R_n(\mathcal{F}(x_1^n)) \le \sqrt{\frac{2 \log \mathbb{S}_{\mathcal{F}}(x_1^n)}{n}} \tag{6}$$

where we used the fact that for each $f \in \mathcal{F}$, $\sum_i f(X_i)^2 \le n$. In particular,

$$\mathbb{E} \sup_{f \in \mathcal{F}} (P(f) - P_n(f)) \le 2\mathbb{E}\sqrt{\frac{2 \log \mathbb{S}_{\mathcal{F}}(X_1^n)}{n}} \ .$$

The logarithm of the VC shatter coefficient may be upper bounded in terms of a combinatorial quantity, called the VC *dimension*. If $A \subset \{-1, 1\}^n$, then the VC dimension of A is the size V of the largest set of indices $\{i_1, \ldots, i_V\} \subset \{1, \ldots, n\}$ such that for each binary V-vector $b = (b_1, \ldots, b_V) \in \{-1, 1\}^V$ there exists an $a = (a_1, \ldots, a_n) \in A$ such that $(a_{1_1}, \ldots, a_{i_V}) = b$. The key inequality establishing a relationship between shatter coefficients and VC dimension is known as *Sauer's lemma* which states that the cardinality of any set $A \subset \{-1, 1\}^n$ may be upper bounded as

$$|A| \le \sum_{i=0}^{V} \binom{n}{i} \le (n+1)^V$$

where V is the VC dimension of A. In particular,

$$\log \mathbb{S}_{\mathcal{F}}(x_1^n) \le V(x_1^n) \log(n+1)$$

where we denote by $V(x_1^n)$ the VC dimension of $\mathcal{F}(x_1^n)$. Thus, the expected maximal deviation $\mathbb{E} \sup_{f \in \mathcal{F}} (P(f) - P_n(f))$ may be upper bounded by $2\mathbb{E}\sqrt{2V(X_1^n) \log(n+1)/n}$. To obtain distribution-free upper bounds, introduce the VC dimension of a class of binary functions $\mathcal{F}$, defined by

$$V = \sup_{n, x_1^n} V(x_1^n) \ .$$

Then clearly, for all distributions one has

$$\mathbb{E} \sup_{f \in \mathcal{F}} (P(f) - P_n(f)) \le 2\sqrt{\frac{2V \log(n+1)}{n}} \ .$$

This bound is a version of what has been known as the *Vapnik-Chervonenkis inequality*. By a somewhat refined analysis (called *chaining*) the logarithmic factor can be removed resulting a bound of the form

$$\mathbb{E}\sup_{f\in\mathcal{F}}(P(f)-P_n(f))\le C\sqrt{\frac{V}{n}}$$

for a universal constant C. The VC dimension is an important combinatorial parameter of the class and many of its properties are well known. Here we just recall one useful result and refer the reader to the references for further study: let $\mathcal{G}$ be an m-dimensional vector space of real-valued functions defined on $\mathcal{X}$. The class of indicator functions

$$\mathcal{F}=\left\{f(x)=\mathbb{1}_{g(x)\ge 0}:g\in\mathcal{G}\right\}$$

has VC dimension $V\le m$.

Bibliographical remarks. Uniform deviations of averages from their expectations is one of the central problems of empirical process theory. Here we merely refer to some of the comprehensive coverages, such as Dudley [78], Giné [93], Shorack and Wellner [188], Vapnik [221], van der Vaart and Wellner [217]. The use of empirical processes in classification was pioneered by Vapnik and Chervonenkis [222,223] and re-discovered 20 years later by Blumer, Ehrenfeucht, Haussler, and Warmuth [37], Ehrenfeucht, Haussler, Kearns, and Valiant [83]. For surveys see Anthony and Bartlett [8], Anthony and Biggs [9], Devroye, Györfi, and Lugosi [67], Kearns and Vazirani [110], Natarajan [168], Ripley [177], Vapnik [220,221] Vidyasagar [225].

The bounded differences inequality was formulated explicitly first by McDiarmid [160] who proved it by martingale methods (see the surveys [160], [161]), but closely related concentration results have been obtained in various ways including information-theoretic methods (see Alhswede, Gács, and Körner [1], Marton [147], [148], [149], Dembo [65], Massart [151] and Rio [175]), Talagrand's induction method [206], [202], [205] (see also Luczak and McDiarmid [136], McDiarmid [162], Panchenko [169–171]) and the so-called "entropy method", based on logarithmic Sobolev inequalities, developed by Ledoux [125], [124], see also Bobkov and Ledoux [38], Massart [152], Rio [175], Boucheron, Lugosi, and Massart [41], [42], Boucheron, Bousquet, Lugosi, and Massart [40], and Bousquet [43].

The simple symmetrization trick shown above is due to Giné and Zinn [94] but different forms of symmetrization have been at the core of obtaining related results of several flavor, see Anthony and Shawe-Taylor [10], Cannon, Ettinger, Hush, Scovel [51], Herbrich and Williamson [103], Mendelson and Philips [166], Vapnik and Chervonenkis [222,223].

The use of Rademacher averages in classification was first promoted by Koltchinskii [117] and Bartlett, Boucheron, and Lugosi [20], see also Koltchinskii and Panchenko [119,120], Bartlett and Mendelson [26], Bartlett, Bousquet, and Mendelson [21], Bousquet, Koltchinskii, and Panchenko [46], Kégl, Linder, and Lugosi [11], Mendelson [164].

Hoeffding's inequality appears in [104]. For a proof of the contraction principle we refer to Ledoux and Talagrand [126].

Sauer's lemma was proved independently by Sauer [179], Shelah [187], and Vapnik and Chervonenkis [222]. For related combinatotial results we refer to Alesker [6], Alon, Ben-David, Cesa-Bianchi, and Haussler [7], Cesa-Bianchi and Haussler [56], Frankl [85], Haussler [101], Mendelson and Vershinin [167], Szarek and Talagrand [199].

The question of how $\sup_{f\in\mathcal{F}}(P(f)-P_n(f))$ behaves has been known as the Glivenko-Cantelli problem and much has been said about it. A few key references include Alon, Ben-David, Cesa-Bianchi, and Haussler [7], Dudley [74,76,77], Dudley, Giné, and Zinn [79], Li, Long, and Srinivasan [131], Mendelson and Vershinin [167], Talagrand [200,201,203,207], Vapnik and Chervonenkis [222,224].

The VC dimension has been widely studied and many of its properties are known. We refer to Anthony and Bartlett [8], Assouad [13] Bartlett and Maass [25], Cover [59], Dudley [75,78], Goldberg and Jerrum [96], Karpinski and A. Macintyre [107], Khovanskii [111], Koiran and Sontag [114], Macintyre and Sontag [142], Steele [193], and Wenocur and Dudley [228].

4. Minimizing cost functions: some basic ideas behind boosting and svm's

The results summarized in the previous section guarantee that minimizing the empirical risk $L_n(g)$ over a class $\mathbb{C}$ of classifiers with a VC dimension much smaller than the sample size n is guaranteed to work well. This result has two fundamental problems. First, by requiring that the VC dimension be small, one imposes serious limitations on the approximation properties of the class. In particular, even though the difference between the probability of error $L(g_n)$ of the empirical risk minimizer is close to the smallest probability of error $\inf_{g\in\mathbb{C}} L(g)$ in the class, $\inf_{g\in\mathbb{C}} L(g) - L^*$ may be very large. The other problem is algorithmic: minimizing the empirical probability of misclassification $L(g)$ is very often a computationally difficult problem. Even in seemingly simple cases, for example when $\mathcal{X} = \mathbb{R}^d$ and $\mathbb{C}$ is the class of classifiers that split the space observations by a hyperplane, the minimization problem is NP hard.

4.1. **Margin-based performance bounds**

An attempt to solve both of these problems is to modify the empirical functional to be minimized by introducing a *cost function*. Next we describe the main ideas of empirical minimization of cost functionals and its analysis. We consider classifiers of the form

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

where $f : \mathcal{X} \to \mathbb{R}$ is a real-valued function. In such a case the probability of error of g may be written as

$$L(g_f) = \mathbb{P}\{\text{sgn}(f(X)) \neq Y\} \leq \mathbb{E}\mathbb{1}_{f(X)Y<0} \ .$$

To lighten notation we will simply write $L(f) = L(g_f)$. Let $\phi : \mathbb{R} \to \mathbb{R}_+$ be a nonnegative valued cost function such that $\phi(x) \geq \mathbb{1}_{x>0}$. (Typical choices of ϕ include $\phi(x) = e^x$, $\phi(x) = \log_2(1 + e^x)$, and $\phi(x) = (1 + x)_+$.) Introduce the *cost functional* and its empirical version by

$$A(f) = \mathbb{E}\phi(-f(X)Y) \qquad \text{and} \qquad A_n(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(-f(X_i)Y_i) \ .$$

Obviously, $L(f) \leq A(f)$ and $L_n(f) \leq A_n(f)$. Assume that the function f_n is chosen from a class $\mathcal{F}$ based on the data $(Z_1, \ldots, Z_n) \stackrel{\text{def}}{=} (X_1, Y_1), \ldots, (X_n, Y_n)$. Then the probability of error of the corresponding classifier may be bounded by the argument of the previous section as follows: let B denote a uniform upper bound on $\phi(-f(x)y)$. Then with probability at least $1 - \delta$,

$$\begin{aligned}
L(f_n) &\leq A(f_n) \\
&\leq A_n(f_n) + \sup_{f\in\mathcal{F}}(A(f) - A_n(f)) \\
&\leq A_n(f_n) + 2\mathbb{E}R_n(\phi \circ \mathcal{H}(Z_1^n)) + B\sqrt{\frac{2\log\frac{1}{\delta}}{n}} \\
&\qquad \text{(where } \mathcal{H} \text{ is the class of functions } \mathcal{X} \times \{-1, 1\} \to \mathbb{R} \text{ of the form } -f(x)y,\ f \in \mathcal{F}) \\
&\leq A_n(f_n) + 2L_\phi\mathbb{E}R_n(\mathcal{H}(Z_1^n)) + B\sqrt{\frac{2\log\frac{1}{\delta}}{n}} \\
&\qquad \text{(by the contraction principle cited in the previous section where } L_\phi \text{ is the Lipschitz constant of } \phi) \\
&= A_n(f_n) + 2L_\phi\mathbb{E}R_n(\mathcal{F}(X_1^n)) + B\sqrt{\frac{2\log\frac{1}{\delta}}{n}} \ .
\end{aligned}$$

Thus, the Rademacher average of the class of real-valued functions f bounds the performance of the classifier.

4.1.1. *Weighted voting schemes*

In many applications such as *boosting* and *bagging*, classifiers are combined by weighted voting schemes which means that the classification rule is obtained by means of functions f from a class

$$\mathcal{F}_\lambda = \left\{ f(x) = \sum_{j=1}^{N} c_j g_j(x) : \ N \in \mathbb{N}, \ \sum_{j=1}^{N} |c_j| \leq \lambda, g_1, \ldots, g_N \in \mathbb{C} \right\} \tag{7}$$

where $\mathbb{C}$ is a class of *base classifiers*, that is, functions defined on $\mathcal{X}$, taking values in $\{-1, 1\}$. A classifier of this form may be thought of as one that, upon observing x, takes a weighted vote of the classifiers $g_1, \ldots, g_N$ (using the weights $c_1, \ldots, c_N$) and decides according to the weighted majority. In this case, by (5) and (6) we have

$$= \lambda R_n(\mathbb{C}(X_1^n)) \leq \lambda \sqrt{\frac{2V_{\mathbb{C}} \log(n+1)}{n}}$$

where $V_{\mathbb{C}}$ is the VC dimension of the base class.

To understand the richness of classes formed by weighted averages of classifiers from a base class, just consider the simple one-dimensional example in which the base classifier $\mathbb{C}$ contains all classifiers of the form $g(x) = 2\mathbb{1}_{x \leq a} - 1$, $a \in \mathbb{R}$. Then $V_{\mathbb{C}} = 1$ and the closure of $\mathcal{F}_\lambda$ (under the L_∞ norm) is the set of all functions of total variation bounded by 2λ. Thus, $\mathcal{F}_\lambda$ is rich in the sense that any classifier may be approximated by classifiers associated to the functions in $\mathcal{F}_\lambda$. In particular, the VC dimension of the class of all classifiers induced by functions in $\mathcal{F}_\lambda$ is infinite.

Summarizing, we have obtained that if $\mathcal{F}_\lambda$ is of the form indicated above, then for any function f_n chosen from $\mathcal{F}_\lambda$ in a data-based manner, the probability of error of the associated classifier satisfies, with probability at least $1 - \delta$,

$$L(f_n) \leq A_n(f_n) + 2L_\phi \lambda \sqrt{\frac{2V_{\mathbb{C}} \log(n+1)}{n}} + B\sqrt{\frac{2\log \frac{1}{\delta}}{n}} \ . \tag{8}$$

The remarkable fact about this inequality is that the upper bound only involves the VC dimension of the class $\mathbb{C}$ of base classifiers which is typically small. The price we pay is that the first term on the right-hand side is the empirical cost functional instead of the empirical probability of error. As a first illustration, consider the example when γ is a fixed positive parameter and

$$\phi(x) = \begin{cases} 0 & \text{if } x \leq -\gamma \\ 1 & \text{if } x \geq 0 \\ 1 + x/\gamma & \text{otherwise} \end{cases}$$

In this case $B = 1$ and $L_\phi = 1/\gamma$. Notice also that $\mathbb{1}_{x>0} \leq \phi(x) \leq \mathbb{1}_{x>-\gamma}$ and therefore $A_n(f) \leq L_n^\gamma(f)$ where $L_n^\gamma(f)$ is the so-called *margin error* defined by

$$L_n^\gamma(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{f(X_i)Y_i < \gamma}.$$

Notice that for all $\gamma > 0$, $L_n^\gamma(f) \geq L_n(f)$ and the $L_n^\gamma(f)$ is increasing in γ. An interpretation of the margin error $L_n^\gamma(f)$ is that it counts, apart from the number of misclassified pairs (X_i, Y_i), also those which are well classified but only with a small "confidence" (or "margin") by f. Thus, (8) implies that, for any $\gamma > 0$, with probability at least $1 - \delta$,

$$L(f_n) \leq L_n^\gamma(f_n) + 2\frac{\lambda}{\gamma}\sqrt{\frac{2V_{\mathbb{C}} \log(n+1)}{n}} + \sqrt{\frac{2\log \frac{1}{\delta}}{n}} \ . \tag{9}$$

Notice that, as γ grows, the first term of the sum increases, while the second decreases. The bound can be very useful whenever a classifier has a small margin error for a relatively large γ (i.e., if the classifier classifies the training data well with high "confidence") since the second term only depends on the VC dimension of the small base class $\mathbb{C}$. This result has been used to explain the good behavior of some voting methods such as AdaBoost, since these methods have a tendency to find classifiers that classify the data points well with a large margin.

4.1.2. *Kernel methods*

Another popular way to obtain classification rules from a class of real-valued functions which is used in *kernel methods* such as *Support Vector Machines (SVM)* or *Kernel Fisher Discriminant (KFD)* is to consider balls of a reproducing kernel Hilbert space.

The basic idea is to use a positive definite kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, that is, a symmetric function satisfying

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

for all choices of n, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathcal{X}$. Such a function naturally generates a space of functions of the form

$$\mathcal{F} = \left\{ \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) : n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X} \right\} ,$$

which, with the inner product $\langle \sum \alpha_i k(x_i, \cdot), \sum \beta_j k(x_j, \cdot) \rangle \stackrel{\text{def}}{=} \sum \alpha_i \beta_j k(x_i, x_j)$ can be completed into a Hilbert space.

The key property is that for all $x_1, x_2 \in \mathcal{X}$ there exist elements $\phi_{x_1}, \phi_{x_2} \in \mathcal{F}$ such that $k(x_1, x_2) = \langle \phi_{x_1}, \phi_{x_2} \rangle$. This means that any linear algorithm based on computing inner products only can be extended into a non-linear version by replacing the inner products by a kernel function. The advantage is that even though the algorithm remains of low complexity, it works in a class of functions that can potentially represent any continuous function arbitrarily well (provided k is chosen appropriately).

Algorithms working with kernels usually perform minimization of a cost function on a ball of the associated reproducing kernel Hilbert space of the form

$$\mathcal{F}_\lambda = \left\{ f(x) = \sum_{j=1}^{N} c_j k(x_j, x) : \; N \in \mathbb{N}, \; \sum_{i,j=1}^{N} c_i c_j k(x_i, x_j) \leq \lambda^2, x_1, \ldots, x_N \in \mathcal{X} \right\} . \tag{10}$$

Notice that, in contrast with (7) where the constraint is of ℓ_1 type, the constraint here is of ℓ_2 type. Also, the basis functions, instead of being chosen from a fixed class, are determined by elements of $\mathcal{X}$ themselves. The consequences are mainly computational as they allow to have a number of parameters equal to the number of samples instead of the number of functions in the base class or the dimension of the input space.

An important property of functions in the reproducing kernel Hilbert space associated to k is that for all $x \in \mathcal{X}$,

$$f(x) = \langle f, k(x, \cdot) \rangle \,.$$

This is called the *reproducing property*. The reproducing property may be used to estimate precisely the Rademacher average of $\mathcal{F}_\lambda$. Indeed, denoting by $\mathbb{E}_\sigma$ expectation with respect to the Rademacher variables

$\sigma_1, \ldots, \sigma_n$, we have

$$\begin{aligned} R_n(\mathcal{F}_\lambda(X_1^n)) &= \frac{1}{n}\mathbb{E}_\sigma \sup_{\|f\|\le\lambda} \sum_{i=1}^n \sigma_i f(X_i) \\ &= \frac{1}{n}\mathbb{E}_\sigma \sup_{\|f\|\le\lambda} \sum_{i=1}^n \sigma_i \langle f, k(X_i,\cdot)\rangle \\ &= \frac{\lambda}{n}\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i k(X_i,\cdot) \right\| . \end{aligned}$$

The Kahane-Khinchine inequality states that for any vectors $a_1, \ldots, a_n$ in a Hilbert space,

$$\frac{1}{\sqrt{2}}\mathbb{E}\left\|\sum_{i=1}^n \sigma_i a_i\right\|^2 \le \left(\mathbb{E}\left\|\sum_{i=1}^n \sigma_i a_i\right\|\right)^2 \le \mathbb{E}\left\|\sum_{i=1}^n \sigma_i a_i\right\|^2 .$$

It is also easy to see that

$$\mathbb{E}\left\|\sum_{i=1}^n \sigma_i a_i\right\|^2 = \mathbb{E}\sum_{i,j=1}^n \sigma_i\sigma_j\langle a_i, a_j\rangle = \sum_{i=1}^n \|a_i\|^2 ,$$

so we obtain

$$\frac{\lambda}{n\sqrt{2}}\sqrt{\sum_{i=1}^n k(X_i,X_i)} \le R_n(\mathcal{F}_\lambda(X_1^n)) \le \frac{\lambda}{n}\sqrt{\sum_{i=1}^n k(X_i,X_i)} .$$

This is very nice as it gives a bound that can be computed very easily from the data. A reasoning similar to the one leading to (9) using the bounded differences inequality to replace the Rademacher average by its empirical version gives that, with probability at least $1-\delta$,

$$L(f_n) \le L_n^\gamma(f_n) + 2\frac{\lambda}{\gamma n}\sqrt{\sum_{i=1}^n k(X_i,X_i)} + \sqrt{\frac{2\log\frac{2}{\delta}}{n}} .$$

4.2. **Convex cost functionals**

Next we show that a proper choice of the cost function ϕ has further advantages. To this end, we consider nonnegative convex nondecreasing cost functions with $\lim_{x\to-\infty}\phi(x) = 0$ and $\phi(0) = 1$. Main examples of ϕ include the *exponential cost function* $\phi(x) = e^x$ used in AdaBoost and related boosting algorithms, the *logit cost function* $\phi(x) = \log_2(1+e^x)$, and the *hinge loss* (or *soft margin loss*) $\phi(x) = (1+x)_+$ used in support vector machines. One of the main advantages of using convex cost functions is that minimizing the empirical cost $A_n(f)$ often becomes a convex optimization problem and is therefore computationally feasible. In fact, most boosting and support vector machine classifiers may be viewed as empirical minimizers of a convex cost functional.

However, minimizing convex cost functionals have other theoretical advantages. To understand this, assume, in addition to the above, that ϕ is strictly convex and differentiable. Then it is easy to determine the function f^* minimizing the cost functional $A(f) = \mathbb{E}\phi(-Yf(X))$. Just note that for each $x \in \mathcal{X}$,

$$\mathbb{E}\left[\phi(-Yf(X))|X = x\right] = \eta(x)\phi(-f(x)) + (1-\eta(x))\phi(f(x))$$

and therefore the function f^* is given by

$$f^*(x) = \operatorname{argmin}_\alpha h_{\eta(x)}(\alpha)$$

where for each $\eta \in [0,1]$, $h_\eta(\alpha) = \eta\phi(-\alpha) + (1-\eta)\phi(\alpha)$. Note that h_η is strictly convex and therefore f^* is well defined (though it may take values $\pm\infty$ if η equals 0 or 1). The minimum is achieved for the value of α for which $h'_\eta(\alpha) = 0$, that is, when

$$\frac{\eta}{1-\eta} = \frac{\phi'(\alpha)}{\phi'(-\alpha)} \ .$$

Since ϕ' is strictly increasing, we see that the solution is positive if and only if $\eta > 1/2$. This reveals the important fact that the minimizer f^* of the functional $A(f)$ is such that the corresponding classifier $g^*(x) = 2\mathbb{1}_{f^*(x)\geq 0} - 1$ is just the *Bayes classifier*. Thus, minimizing a convex cost functional leads to an optimal classifier. For example, if $\phi(x) = e^x$ is the exponential cost function, then $f^*(x) = (1/2)\log(\eta(x)/(1-\eta(x)))$. In the case of the logit cost $\phi(x) = \log_2(1+e^x)$, we have $f^*(x) = \log(\eta(x)/(1-\eta(x)))$.

We note here that, even though the hinge loss $\phi(x) = (1+x)_+$ does not satisfy the conditions for ϕ used above (e.g., it is not strictly convex), it is easy to see that the function f^* minimizing the cost functional equals

$$f^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ -1 & \text{if } \eta(x) < 1/2 \end{cases}$$

Thus, in this case the f^* not only induces the Bayes classifier but it equals to it.

To obtain inequalities for the probability of error of classifiers based on minimization of empirical cost functionals, we need to establish a relationship between the excess probability of error $L(f) - L^*$ and the corresponding excess cost functional $A(f) - A^*$ where $A^* = A(f^*) = \inf_f A(f)$. Here we recall a simple inequality of Zhang [232] which states that if the function $H : [0,1] \to \mathbb{R}$ is defined by $H(\eta) = \inf_\alpha h_\eta(\alpha)$ and the cost function ϕ is such that for some positive constants $s \geq 1$ and $c \geq 0$

$$\left|\frac{1}{2} - \eta\right|^s \leq c^s(1 - H(\eta)) \ , \qquad \eta \in [0,1] \ ,$$

then for any function $f : \mathcal{X} \to \mathbb{R}$,

$$L(f) - L^* \leq 2c\left(A(f) - A^*\right)^{1/s} \ . \tag{11}$$

(The simple proof of this inequality is based on the expression (1) and elementary convexity properties of h_η.) In the special case of the exponential and logit cost functions $H(\eta) = 2\sqrt{\eta(1-\eta)}$ and $H(\eta) = -\eta\log_2\eta - (1-\eta)\log_2(1-\eta)$, respectively. In both cases it is easy to see that the condition above is satisfied with $s = 2$ and $c = \sqrt{2}$. Thus, in both of these cases, we have that if f_n is chosen from a class $\mathcal{F}_\lambda$ defined in (7) then

$$\begin{aligned}
L(f_n) - L^* &\leq 2\sqrt{2}\left(A(f_n) - A^*\right)^{1/2} \\
&\leq 2\sqrt{2}\left(A(f_n) - \inf_{f\in\mathcal{F}_\lambda} A(f)\right)^{1/2} + 2\sqrt{2}\left(\inf_{f\in\mathcal{F}_\lambda} A(f) - A^*\right)^{1/2} \\
&\leq 4\sqrt{2}\left(\sup_{f\in\mathcal{F}_\lambda} |A(f) - A_n(f)|\right)^{1/2} + 2\sqrt{2}\left(\inf_{f\in\mathcal{F}_\lambda} A(f) - A^*\right)^{1/2} \\
&\quad \text{(just like in (2))} \\
&\leq 4\sqrt{2}\left(2L_\phi\lambda\sqrt{\frac{2V_{\mathbb{C}}\log(n+1)}{n}} + B\sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right)^{1/2} + 2\sqrt{2}\left(\inf_{f\in\mathcal{F}_\lambda} A(f) - A^*\right)^{1/2}
\end{aligned} \tag{12}$$

with probability at least $1-\delta$, where at the last step we used the same bound for $\sup_{f\in\mathcal{F}_\lambda} |A(f) - A_n(f)|$ as in (8). Note that for the exponential cost function $L_\phi = e^\lambda$ and $B = \lambda$ while for the logit cost $L_\phi \leq 1$ and $B = \lambda$. In both cases, if λ is sufficiently large so that $\inf_{f\in\mathcal{F}_\lambda} A(f) = A^*$ then the approximation error disappears and we obtain $L(f_n) - L^* = O\left(n^{-1/4}\right)$. The dimension-free nature of this rate of convergence is remarkable. (We

note here that these rates may be further improved by applying the refined techniques resumed in Section 5.3, see also [36].) It is an interesting approximation-theoretic challenge to understand what kind of functions f^* may be obtained as a convex combination of base classifiers and, more generally, to describe approximation properties of classes of functions of the form (7).

Next we describe a simple example when the above-mentioned approximation properties are well understood. Consider the case when $\mathcal{X} = [0,1]^d$ and the base class $\mathbb{C}$ contains all "decision stumps", that is, all classifiers of the form $s^+_{i,t}(x) = \mathbb{1}_{x^{(i)} \geq t} - \mathbb{1}_{x^{(i)} < t}$ and $s^-_{i,t}(x) = \mathbb{1}_{x^{(i)} < t} - \mathbb{1}_{x^{(i)} \geq t}$, $t \in [0,1]$, $i = 1, \ldots, d$, where $x^{(i)}$ denotes the i-th coordinate of x. In this case the VC dimension of the base class is easily seen to be bounded by $V_{\mathbb{C}} \leq \lfloor 2 \log_2(2d) \rfloor$. Also it is easy to see that the closure of $\mathcal{F}_\lambda$ with respect to the supremum norm contains all functions f of the form

$$f(x) = f_1(x^{(1)}) + \cdots + f_d(x^{(d)})$$

where the functions $f_i : [0,1] \to \mathbb{R}$ are such that $|f_1|_{TV} + \cdots + |f_d|_{TV} \leq \lambda$ where $|f_i|_{TV}$ denotes the total variation of the function f_i. Therefore, if f^* has the above form, we have $\inf_{f \in \mathcal{F}_\lambda} A(f) = A(f^*)$. Recalling that the function f^* optimizing the cost $A(f)$ has the form

$$f^*(x) = \frac{1}{2} \log \frac{\eta(x)}{1 - \eta(x)}$$

in the case of the exponential cost function and

$$f^*(x) = \log \frac{\eta(x)}{1 - \eta(x)}$$

in the case of the logit cost function, we see that boosting using decision stumps is especially well fitted to the so-called additive logistic model in which η is assumed to be such that $\log(\eta/(1-\eta))$ is an additive function (i.e., it can be written as a sum of univariate functions of the components of x). Thus, when η permits an additive logistic representation then the rate of convergence of the classifier is fast and has a very mild dependence on the distribution.

Consider next the case of the hinge loss $\phi(x) = (1+x)_+$ often used in Support Vector Machines and related kernel methods. In this case $H(\eta) = 2\max(\eta, 1-\eta)$ and therefore inequality (11) holds with $c = 2$ and $s = 1$. Thus,

$$L(f_n) - L^* \leq 4\,(A(f_n) - A^*)$$

and the analysis above leads to even better rates of convergence. However, in this case $f^*(x) = 2\mathbb{1}_{\eta(x) \geq 1/2} - 1$ and approximating this function by weighted sums of base functions may be more difficult than in the case of exponential and logit costs. Once again, the approximation-theoretic part of the problem is far from being well understood, and it is difficult to give recommendations about which cost function is more advantageous and what base classes should be used.

Bibliographical remarks. For results on the algorithmic difficulty of empirical risk minimization, see Johnson and Preparata [106], Bartlett and Ben-David [23], Ben-David, Eiron, and Simon [28], Vu [226].

Boosting algorithms were originally introduced by Freund and Schapire (see [86], [89], and [180]), as adaptive aggregation of simple classifiers contained in a small "base class". The analysis based on the observation that AdaBoost and related methods tend to produce large-margin classifiers appears in Schapire, Freund, Bartlett, and Lee [181], and Koltchinskii and Panchenko [120]). It was Breiman [47] who observed that boosting performs gradient descent optimization of an empirical cost function different from the number of misclassified samples, see also Mason, Baxter, Bartlett, and Frean [150], Collins, Schapire, and Singer [57], Friedman, Hastie, and Tibshirani [90]. Based on this view, various versions of boosting algorithms have been shown to be consistent in different settings, see Blanchard, Lugosi, and Vayatis [36], Breiman [48], Bühlmann and Yu [50], Jiang [105], Lugosi and Vayatis [139], Mannor and Meir [145], Mannor, Meir, and Zhang [146], Zhang [232]. Inequality (8) was first obtained by Schapire, Freund, Bartlett, and Lee [181]. The analysis presented here is due to Koltchinskii and Panchenko [120].

Other classifiers based on weighted voting schemes have been considered by Freund, Mansour, and Schapire [88], Catoni [53–55], Yang [231].

Support vector machines originate in the pioneering work of Boser, Guyon, and Vapnik [39], Cortes and Vapnik [58], but some of the ideas can be traced back to Vapnik and Lerner [218], Aizerman, Braverman, and Rozonoer [2–4], Bashkirov, Braverman, and Muchnik [27], Specht [192], and Vapnik and Chervonenkis [223]. For surveys we refer to Cristianini and Shawe-Taylor [61], Hastie, Tibshirani, and Friedman [99], Schölkopf and Smola [182], Smola, Bartlett, Schölkopf, and Schuurmans [190].

The study of universal approximation properties of kernels and statistical consistency of Support Vector Machines is due to Steinwart [194–196] and Lin [133, 134].

We have considered the case of minimization of a loss function on a ball of the reproducing kernel Hilbert space. However, it is computationally more convenient to formulate the problem as the minimization of a regularized functional of the form

$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \phi(-Y_i f(X_i)) + \lambda\|f\|^2 .$$

The standard Support Vector Machine algorithm then corresponds to the choice of $\phi(x) = (1+x)_+$.

Kernel based regularization algorithms were studied by Craven and Wahba [60] and Kimeldorf and Wahba [113], in the context of regression. Relationships between Support Vector Machines and regularization were described by Evgeniou, Pontil and Poggio [84] and Smola, Schölkopf and Müller [191]. General properties of regularized algorithms in reproducing kernel Hilbert spaces are investigated by Cucker and Smale [64], Steinwart [195], Zhang [232].

Various properties of the Support Vector Machine algorithm are investigated by Vapnik [220, 221], Schölkopf and Smola [182], Scovel and Steinwart [185] and Steinwart [197, 198].

The fact that minimizing an exponential cost functional leads to the Bayes classifier was pointed out by Breiman [48], see also Lugosi and Vayatis [139], Zhang [232]. For a comprehensive theory of the connection between cost functions and probability of misclassification, see Bartlett, Jordan, and McAuliffe [24]. Zhang's lemma (11) appears in [232]. For various generalizations and refinements we refer to Bartlett, Jordan, and McAuliffe [24] and Blanchard, Lugosi, and Vayatis [36].

5. Tighter bounds for empirical risk minimization

This section is dedicated to the description of some refinements of the ideas described in the earlier sections. What we have seen so far only used "first-order" properties of the functions that we considered, namely their boundedness. It turns out that using "second-order" properties, like the variance of the functions, many of the above results can be made sharper.

5.1. **Relative Deviations**

In order to understand the basic phenomenon, let us go back to the simplest case in which one has a fixed function f with values in $\{0,1\}$. In this case, $P_n(f)$ is an average of independent Bernoulli random variables with parameter $p = P(f)$. Recall that, as a simple consequence of (3), with probability at least $1-\delta$,

$$P(f) - P_n(f) \le \sqrt{\frac{2\log\frac{1}{\delta}}{n}} . \tag{13}$$

This is basically tight when $P(f) = 1/2$, but can be significantly improved when $P(f)$ is small. Indeed, Bernstein's inequality gives, with probability at least $1-\delta$,

$$P(f) - P_n(f) \le \sqrt{\frac{2\mathsf{Var}(f)\log\frac{1}{\delta}}{n}} + \frac{2\log\frac{1}{\delta}}{3n} . \tag{14}$$

Since f takes its values in $\{0,1\}$, $\mathsf{Var}(f) = P(f)(1-P(f)) \leq P(f)$ which shows that when $P(f)$ is small, (14) is much better than (13).

5.1.1. *General Inequalities*

Next we to exploit the phenomenon described above to obtain sharper performance bounds for empirical risk minimization. Note that if we consider the difference between $P(f) - P_n(f)$ uniformly over the class $\mathcal{F}$, the largest deviations are obtained by functions that have a large variance (i.e., $P(f)$ is close to $1/2$). The idea is to scale each function by dividing it by $\sqrt{P(f)}$ so that they all behave in a similar way. Thus, we bound the quantity

$$\sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}}\,.$$

The first step consists in symmetrization of the tail probabilities:

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}} \geq t\right\} \leq 2\mathbb{P}\left\{\sup_{f\in\mathcal{F}} \frac{P'_n f - P_n f}{\sqrt{(P_n f + P'_n f)/2}} \geq t\right\}\,.$$

Next we introduce Rademacher random variables

$$\cdots = 2\mathbb{E}\left[\mathbb{P}_\sigma\left\{\sup_{f\in\mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X'_i) - f(X_i))}{\sqrt{(P_n f + P'_n f)/2}} \geq t\right\}\right]$$

(where $\mathbb{P}_\sigma$ is the conditional probability, given the X_i and X'_i). The last step uses tail bounds for individual functions and a union bound over $\mathcal{F}(X_1^{2n})$, where X_1^{2n} denotes the union of the initial sample X_1^n and of the extra symmetrization sample $X'_1, \ldots, X'_n$.

Finally, we obtain that for $\delta \in (0,1)$, with probability at least $1-\delta$, all $f \in \mathcal{F}$ satisfy

$$\frac{Pf - P_n f}{\sqrt{Pf}} \leq 2\sqrt{\frac{\log \mathbb{S}_{\mathcal{F}}(X_1^{2n}) + \log\frac{4}{\delta}}{n}}\,. \tag{15}$$

Also, with probability at least $1-\delta$, for all $f \in \mathcal{F}$,

$$\frac{P_n f - Pf}{\sqrt{P_n f}} \leq 2\sqrt{\frac{\log \mathbb{S}_{\mathcal{F}}(X_1^{2n}) + \log\frac{4}{\delta}}{n}}\,. \tag{16}$$

As a consequence, we have that for all $s > 0$, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} \frac{P(f) - P_n(f)}{P(f) + P_n(f) + s/2} \leq 2\sqrt{\frac{\log \mathbb{S}_{\mathcal{F}}(X_1^{2n}) + \log\frac{4}{\delta}}{sn}} \tag{17}$$

and the same is true if P and P_n are permuted. Another consequence of (15) and (16) with interesting applications is the following. For all $t \in (0,1]$, with probability at least $1-\delta$,

$$\forall f \in \mathcal{F}, P_n(f) \leq (1-t)P(f) \text{ implies } P(f) \leq 4\frac{\log \mathbb{S}_{\mathcal{F}}(X_1^{2n}) + \log\frac{4}{\delta}}{t^2 n}\,. \tag{18}$$

In particular, setting $t = 1$,

$$\forall f \in \mathcal{F}, P_n(f) = 0 \text{ implies } P(f) \leq 4\frac{\log \mathbb{S}_{\mathcal{F}}(X_1^{2n}) + \log\frac{4}{\delta}}{n}\,.$$

5.1.2. *Applications to Empirical Risk Minimization*

It is easy to see that, for non-negative numbers $A, B, C \geq 0$, the fact that $A \leq B\sqrt{A} + C$ entails $A \leq B^2 + B\sqrt{C} + C$ so that we obtain, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,

$$P(f) \leq P_n(f) + 2\sqrt{P_n(f)\frac{\log \mathbb{S}_{\mathcal{F}}(X_1^{2n}) + \log \frac{4}{\delta}}{n}} + 4\frac{\log \mathbb{S}_{\mathcal{F}}(X_1^{2n}) + \log \frac{4}{\delta}}{n}\,.$$

If we apply this to the empirical risk minimizer g_n^* in a class $\mathbb{C}$ of VC dimension V, we obtain, with probability at least $1 - \delta$,

$$L(g_n^*) \leq L_n(g_n^*) + 2\sqrt{L_n(g_n^*)\frac{2V\log(n+1) + \log \frac{4}{\delta}}{n}} + 4\frac{2V\log(n+1) + \log \frac{4}{\delta}}{n}\,. \tag{19}$$

Consider first the extreme situation when there exists a classifier in $\mathbb{C}$ which classifies without error. This also means that for some $g' \in \mathbb{C}$, $Y = g'(X)$ with probability one, a quite restrictive assumption. Nevertheless, the assumption that $\inf_{g\in\mathbb{C}} L(g) = 0$ is common in computational learning theory. In such a case, clearly $L_n(g_n^*) = 0$, so that we get, with probability at least $1 - \delta$,

$$L(g_n^*) - \inf_{g\in\mathbb{C}} L(g) \leq 4\frac{2V\log(n+1) + \log \frac{4}{\delta}}{n}\,. \tag{20}$$

The main point here is that the upper bound obtained in this special case is of smaller order of magnitude than in the general case ($O(V \ln n/n)$ as opposed to $O\left(\sqrt{V \ln n/n}\right)$.) One can actually obtain a version which interpolates between those two cases as follows: For simplicity, assume that there is a classifier g' in $\mathbb{C}$ such that $L(g') = \inf_{g\in\mathbb{C}} L(g)$. Then we have

$$L_n(g_n^*) \leq L_n(g') = L_n(g') - L(g') + L(g')\,.$$

Using Bernstein's inequality, we get, with probability $1 - \delta$,

$$L_n(g_n^*) - L(g') \leq \sqrt{\frac{2L(g')\log \frac{1}{\delta}}{n}} + \frac{2\log \frac{1}{\delta}}{3n}\,,$$

which, together with (19), yields that for some constant C, with probability at least $1 - \delta$,

$$L(g_n^*) - \inf_{g\in\mathbb{C}} L(g) \leq C\left(\sqrt{\inf_{g\in\mathbb{C}} L(g)\frac{V\log n + \log \frac{1}{\delta}}{n}} + \frac{V\log n + \log \frac{1}{\delta}}{n}\right)\,. \tag{21}$$

5.2. **Noise and Fast Rates**

We have seen that in the case where f takes values in $\{0, 1\}$ there is a nice relationship between the variance of f (which controls the size of the deviations between $P(f)$ and $P_n(f)$) and its expectation, namely, $\mathsf{Var}(f) \leq P(f)$. This is the key property that allows one to obtain faster rates of convergence for $L(g_n^*) - \inf_{g\in\mathbb{C}} L(g)$.

In particular, in the ideal situation mentioned above, where $\inf_{g\in\mathbb{C}} L(g) = 0$, the difference $L(g_n^*) - \inf_{g\in\mathbb{C}} L(g)$ may be much smaller than the difference between $L(g_n^*)$ and $L_n(g_n^*)$ or even than $L(g') - L_n(g')$. This actually happens in many cases, whenever the distribution satisfies certain conditions. Next we describe such conditions and show how the finer bounds can be derived.

The main idea is that, in order to get precise rates for $L(g_n^*) - \inf_{g\in\mathbb{C}} L(g)$, we consider functions of the form $\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g'(X)\neq Y}$ where g' is a classifier minimizing the loss in the class $\mathbb{C}$, that is, such that $L(g') = \inf_{g\in\mathbb{C}} L(g)$. Note that functions of this form are no longer non-negative.

To illustrate the basic ideas in the simplest possible setting, consider the case when $\mathcal{F}$ is a finite set of N functions of the form $\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g'(X)\neq Y}$. In addition, we assume that there is a relationship between the variance and the expectation of the functions is $\mathcal{F}$ given by the inequality

$$\mathsf{Var}(f) \leq c(P(f))^\alpha \tag{22}$$

for some $c > 0$ and $\alpha \in (0, 1]$. By Bernstein's inequality and a union bound over the elements of $\mathbb{C}$, we have that, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,

$$P(f) \leq P_n(f) + \sqrt{\frac{2c(P(f))^\alpha \log\frac{N}{\delta}}{n}} + \frac{4\log\frac{N}{\delta}}{3n}\,.$$

As a consequence, using the fact that $L_n(g_n^*) - L_n(g') \leq 0$, we have with probability at least $1 - \delta$,

$$L(g_n^*) - L(g') \leq \sqrt{\frac{2c(L(g_n^*) - L(g'))^\alpha \log\frac{N}{\delta}}{n}} + \frac{4\log\frac{N}{\delta}}{3n}\,.$$

Solving this inequality for $L(g_n^*) - L(g')$ finally gives that with probability at least $1 - \delta$,

$$L(g_n^*) - \inf_{g\in\mathcal{G}} L(g) \leq C\left(\frac{\log\frac{N}{\delta}}{n}\right)^{\frac{1}{2-\alpha}}\,. \tag{23}$$

Note that the obtained rate is then faster than $n^{-1/2}$ whenever $\alpha > 0$. In particular, for $\alpha = 1$ we get n^{-1} as in the ideal case.

It now remains to show that (22) is a reasonable assumption. As an example, assume that the Bayes classifier g^* belongs to the class $\mathbb{C}$ (i.e., $g' = g^*$) and the a posteriori probability function η is bounded away from $1/2$, that is, there exists a positive constant s such that for all $x \in \mathcal{X}$, $|2\eta(x) - 1| > s$. In the sequel we will refer to this condition as *Massart's noise condition*. Since $|\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y}| \leq \mathbb{1}_{g(X)\neq g^*(X)}$, Massart's noise condition and (1) imply that

$$\mathsf{Var}(f) \leq \mathbb{E}\mathbb{1}_{g(X)\neq g^*(X)} \leq \frac{1}{s}\mathbb{E}|2\eta(X) - 1|\mathbb{1}_{g(X)\neq g^*(X)} = \frac{1}{s}(L(g) - L^*)\,.$$

Thus (22) holds with $c = 1/s$ and $\alpha = 1$ which shows that with probability $1 - \delta$,

$$L(g_n) - L^* \leq C\frac{\log\frac{N}{\delta}}{sn}\,. \tag{24}$$

Thus, the empirical risk minimizer has a significantly better performance than predicted by the results of the previous section whenever the Bayes classifier is in the class $\mathbb{C}$ and the a posteriori probability η stays away from $1/2$. The behavior of η in the vicinity of $1/2$ has been known to play an important role in the difficulty of the classification problem, see [67, 229, 230]. Roughly speaking, if η has a complex behavior around the critical threshold $1/2$, then one cannot avoid estimating η, which is a typically difficult nonparametric regression problem. However, the classification problem is significantly easier than regression if η is far from $1/2$ with a large probability.

Tsybakov [209] formulated a useful generalization of Massart's noise condition that has been adopted by many authors. Let $\alpha \in [0,1]$. Then Tsybakov's condition may be stated by any of the following three equivalent statements:

$$(1) \quad \exists \beta > 0,\ \forall g \in \{0,1\}^{\mathcal{X}},\ \mathbb{E}\mathbb{1}_{g(X)\neq g^*(X)} \leq \beta (L(g) - L^*)^\alpha$$
$$(2) \quad \exists c > 0,\ \forall A \subset \mathcal{X},\ \int_A dP(x) \leq c(\int_A |2\eta(x) - 1| dP(x))^\alpha$$
$$(3) \quad \exists B > 0,\ \forall t \geq 0,\ \mathbb{P}\left\{|2\eta(X) - 1| \leq t\right\} \leq B t^{\frac{\alpha}{1-\alpha}}\ .$$

We refer to this as *Tsybakov's noise condition*. The proof that these statements are equivalent is straightforward, and we omit it, but we comment on the meaning of these statements. Notice first that α has to be in $[0,1]$ because

$$L(g) - L^* = \mathbb{E}\left[|2\eta(X) - 1| \mathbb{1}_{g(X)\neq g^*(X)}\right] \leq \mathbb{E}\mathbb{1}_{g(X)\neq g^*(X)}\,.$$

Also, when $\alpha = 0$ these conditions are void, while when $\alpha = 1$ they imply that there exists an $s > 0$ such that $|2\eta(X) - 1| > s$ almost surely (which is just Massart's noise condition we considered above).

The most important consequence of these conditions is that they imply a relationship between the variance and the expectation of functions of the form $\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{t(X)\neq Y}$. Indeed, we obtain

$$\mathbb{E}\left[(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y})^2\right] \leq c(L(g) - L^*)^\alpha\,.$$

This is thus enough to get (23) for a finite class of functions.

5.3. **Localization**

The purpose of this section is to generalize the simple argument of the previous section to more general classes $\mathbb{C}$ of classifiers. This generalization reveals the importance of the modulus of continuity of the empirical process as a measure of complexity of the learning problem.

5.3.1. *Talagrand's Inequality*

One of the most important recent developments in empirical process theory is a concentration inequality for the supremum of an empirical process first proved by Talagrand [201] and refined later by various authors. This inequality is at the heart of many key developments in statistical learning theory. Here we recall the following version:

Theorem 5.1. *Let $b > 0$ and set $\mathcal{F}$ to be a set of functions from $\mathcal{X}$ to $\mathbb{R}$. Assume that all functions in $\mathcal{F}$ satisfy $Pf - f \leq b$. Then, with probability at least $1 - \delta$,*

$$\sup_{f\in\mathcal{F}} (P(f) - P_n(f)) \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} (P(f) - P_n(f))\right] + \sqrt{\frac{2(\sup_{f\in\mathcal{F}} \mathsf{Var}(f)) \log\frac{1}{\delta}}{n}} + \frac{4b\log\frac{1}{\delta}}{3n}\,.$$

5.3.2. *Localization: informal argument*

We first explain informally how Talagrand's inequality can be used in conjunction with noise conditions to yield improved results. Start by rewriting the inequality of Theorem 5.1. Letting $r = \sup_{f\in\mathcal{F}} \mathsf{Var}(f)$ we have, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,

$$P(f) - P_n(f) \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}:\mathsf{Var}(f)\leq r} (P(f) - P_n(f))\right] + C\sqrt{\frac{r\log\frac{1}{\delta}}{n}} + C\frac{\log\frac{1}{\delta}}{n}\,. \tag{25}$$

Denote the right-hand side of the above inequality by $\tilde{\psi}(r)$. Note that $\tilde{\psi}$ is an increasing nonnegative function.

Consider the class of functions $\mathcal{F} = \{(x,y) \mapsto \mathbb{1}_{g(x)\neq y} - \mathbb{1}_{g^*(x)\neq y} : g \in \mathbb{C}\}$ and assume, for simplicity, that $g^* \in \mathbb{C}$ and Massart's noise condition is satisfied with constant s, so that for all $f \in \mathcal{F}$, $\mathsf{Var}(f) \leq \frac{1}{s} Pf$.

Inequality (25) thus implies that, with probability at least $1-\delta$, all $g \in \mathbb{C}$ satisfy

$$L(g) - L^* \leq L_n(g) - L_n(g^*) + \tilde{\psi}\left(\frac{1}{s}\sup_{g\in\mathbb{C}} L(g) - L^*\right).$$

In particular, we have, with probability at least $1-\delta$,

$$L(g_n) - L^* \leq \tilde{\psi}\left(\frac{1}{s}\sup_{g\in\mathbb{C}} L(g) - L^*\right).$$

For the sake of an informal argument, assume that we somehow know beforehand what $L(g_n)$ is. Then we can 'apply' the above inequality to a subclass which only contains functions with error less than that of g_n, and thus we would obtain something like

$$L(g_n) - L^* \leq \tilde{\psi}\left(\frac{1}{s}\left(L(g_n) - L^*\right)\right).$$

This indicates that the quantity that should appear as an upper bound of $L(g_n) - L^*$ is something like $\max\{r : r \leq \tilde{\psi}(r/s)\}$. We will see that the smallest allowable value is actually the solution of $r = \tilde{\psi}(r/s)$. The reason why this bound can improve the rates is that in many situations, $\tilde{\psi}(r)$ is of order $\sqrt{r/n}$. In this case the solution r^* of $r = \tilde{\psi}(r/s)$ satisfies $r^* \approx 1/(sn)$ thus giving a bound of order $1/n$ for the quantity $L(g_n) - L^*$.

The argument sketched here, once made rigorous, applies to possibly infinite classes with a complexity measure that captures the size of the empirical process in a small ball (i.e., restricted to functions with small variance). The next section is offers a detailed argument.

5.3.3. *Localization: rigorous argument*

Let $\mathcal{F} = \{(x,y) \mapsto \mathbb{1}_{g(x)\neq y} - \mathbb{1}_{g^*(x)\neq y} : g \in \mathbb{C}\}$ and introduce the *star-hull* of $\mathcal{F}$ defined by $\mathcal{F}^* = \{\alpha f : \alpha \in [0,1],\, f \in \mathcal{F}\}$.

Notice that for $f \in \mathcal{F}$ or $f \in \mathcal{F}^*$, $P(f) \geq 0$. Also, denoting by f_n the function in $\mathcal{F}$ corresponding to the empirical risk minimizer g_n, we have $P_n(f_n) \leq 0$.

Let $T : \mathcal{F} \to \mathbb{R}^+$ be a function such that for all $f \in \mathcal{F}$, $\mathsf{Var}(f) \leq T^2(f)$ and also for $\alpha \in [0,1]$, $T(\alpha f) \leq \alpha T(f)$. An important example is $T(f) = \sqrt{Pf^2}$.

Introduce the following two functions which characterize the properties of the problem of interest (i.e., the loss function, the distribution, and the class of functions). The first one is a sort of modulus of continuity of the Rademacher indexed by the star-hull of $\mathcal{F}$:

$$\psi(r) = \mathbb{E}R_n\{f \in \mathcal{F}^* : T(f) \leq r\}.$$

The second one is the modulus of continuity of the variance (or rather its upper bound T) with respect to the expectation:

$$w(r) = \sup_{f\in\mathcal{F}^* : Pf \leq r} T(f).$$

In fact, both of these functions may be replaced by convenient upper bounds. Of course, ψ and w are non-negative and non-decreasing. Moreover, the map $x \mapsto \psi(x)/x$ is non-increasing. Indeed, for $\alpha \geq 1$,

$$\begin{aligned}
\psi(\alpha x) &= \mathbb{E}R_n\{f \in \mathcal{F}^* : T(f) \leq \alpha x\} \\
&\leq \mathbb{E}R_n\{f \in \mathcal{F}^* : T(f/\alpha) \leq x\} \\
&\leq \mathbb{E}R_n\{\alpha f : f \in \mathcal{F}^*,\, T(f) \leq x\} = \alpha\psi(x).
\end{aligned}$$

The analysis below uses the additional assumption that $x \mapsto w(x)/\sqrt{x}$ is also non-increasing. Below we indicate in which cases this is satisfied.

The main idea is to weight the functions in $\mathcal{F}$ in order to have a handle on their variance (which is the key to making good use of Talagrand's inequality). To do this, consider

$$\mathcal{G}_r = \left\{ \frac{rf}{T(f) \vee r} : f \in \mathcal{F}, Pf \geq r \right\}.$$

We thus apply Talagrand's inequality to this class of functions. Noticing that $Pg - g \leq 2$ and $\mathsf{Var}(g) \leq r^2$ for $g \in \mathcal{G}_r$, we obtain, with probability at least $1 - \delta$,

$$Pf - P_n f \leq \frac{T(f) \vee r}{r} \left(2\mathbb{E} \sup_{g \in \mathcal{G}_r} (Pg - P_n g) + r\sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n} \right).$$

As shown in Section 3, we can upper bound the expectation in the right-hand side by $2\mathbb{E}R_n(\mathcal{G}_r)$. Notice that for $f \in \mathcal{G}_r$, $T(f) \leq r$ and also $\mathcal{G}_r \subset \mathcal{F}^*$ which implies that

$$R_n(\mathcal{G}_r) \leq R_n\{f \in \mathcal{F}^* : T(f) \leq r\}.$$

We thus obtain

$$Pf - P_n f \leq \frac{T(f) \vee r}{r} \left(4\psi(r) + r\sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n} \right).$$

Using the definition of w, this yields

$$Pf - P_n f \leq \frac{w(Pf) \vee r}{r} \left(4\psi(r) + r\sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n} \right).$$

Then either $w(Pf) \leq r$ which implies $Pf \leq w^{-1}(r)$ (where $w^{-1}(x) \stackrel{\text{def}}{=} \max\{u : w(u) \leq x\}$, so that we have $w(w^{-1}(r)) \leq r$ and $w^{-1}(w(u)) \geq u$), or $w(Pf) \geq r$. In this latter case,

$$Pf \leq P_n f + w(Pf) \left(4\psi(r)/r + \sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3rn} \right).$$

By assumption we have

$$w(Pf) \leq \frac{r\sqrt{Pf}}{\sqrt{w^{-1}(r)}},$$

so that finally (using the fact that $x \leq A\sqrt{x} + B$ implies $x \leq A^2 + 2B$),

$$Pf \leq 2P_n f + \frac{1}{w^{-1}(r)} \left(4\psi(r) + r\sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n} \right)^2.$$

Since we are interested in a function f_n such that $P_n f_n \leq 0$, we obtain that, with probability at least $1 - \delta$,

$$Pf_n \leq \max\left(w^{-1}(r), \frac{1}{w^{-1}(r)} \left(4\psi(r) + r\sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n} \right)^2 \right).$$

To minimize the obtained upper bound, we find a value of r which makes the two quantities in the maximum equal. It is convenient to introduce the value ε^* as the solution of the fixed-point equation

$$\varepsilon = \psi(w(\varepsilon))\,.$$

As $\psi(x)/x$ is non-increasing, we get, for $r \geq w(\varepsilon^*)$, $\psi(r) \leq r\varepsilon^*/w(\varepsilon^*)$. Hence, we obtain that for all $r \geq w(\varepsilon^*)$, with probability at least $1-\delta$,

$$L(g_n) - L^* \leq \max\left(w^{-1}(r), \frac{1}{w^{-1}(r)} \left(4r\frac{\varepsilon^*}{w(\varepsilon^*)} + r\sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n} \right)^2 \right)\,. \tag{26}$$

5.3.4. *Consequences*

To understand the meaning of (26), consider the case $w(x) = cx^{\alpha/2}$ with $\alpha \leq 1$. Observe that such a choice of w is possible under Tsybakov's noise condition. In this case $w^{-1}(r) = (r/c)^{2/\alpha}$, which leads to an equation of the form

$$r^{2/\alpha} = Cr\left((\varepsilon^*)^{1-\alpha/2} + \sqrt{\frac{\log\frac{1}{\delta}}{n}} \right) + C'\frac{\log\frac{1}{\delta}}{n}\,.$$

Clearly, there exists $r \geq w(\varepsilon^*)$ such that this is satisfied and

$$r \leq C\left((\varepsilon^*)^{1-\alpha/2} + \sqrt{\frac{\log\frac{1}{\delta}}{n}} \right)^{\frac{\alpha}{2-\alpha}}\,,$$

so that finally, we have

$$L(g_n) - L^* \leq C\left((\varepsilon^*)^{1-\alpha/2} + \sqrt{\frac{\log\frac{1}{\delta}}{n}} \right)^{\frac{2}{2-\alpha}}$$

This is of order $O(n^{-1/(2-\alpha)})$ similarly to (23) provided ε^* is of order at most $O(n^{-1/(2-\alpha)})$. For VC classes of functions, it can be shown (see, e.g., [153] or [21]) that

$$\psi(x) \leq Cx\sqrt{\frac{V}{n}\log n}$$

so that

$$\frac{\varepsilon^*}{w(\varepsilon^*)} \leq C\sqrt{\frac{V}{n}\log n}\,.$$

Thus, we indeed obtain an improved rate and the improvement is possible even if $L^* > 0$.

In the special case when $\alpha = 1$ (i.e., under Massart's noise condition), one has $w(x) = \sqrt{x/s}$ where s is such that $|\eta - 1/2| \geq s$. Applying (26) thus gives that with probability at least $1-\delta$,

$$L(g_n) - L^* \leq C\frac{V\log n + \log\frac{1}{\delta}}{ns}\,,$$

which, combined with the bound obtained without noise conditions finally gives

$$L(g_n) - L^* \leq C\left(\frac{V\log n + \log\frac{1}{\delta}}{ns} \wedge \sqrt{\frac{V + \log\frac{1}{\delta}}{n}} \right)\,.$$

5.4. **Cost Functions**

The refined bounds described in the previous section may be carried over to the analysis of classification rules based on the empirical minimization of a convex cost functional $A_n(f) = (1/n)\sum_{i=1}^n \phi(-f(X_i)Y_i)$, over a class $\mathcal{F}$ of real-valued functions as is the case in many popular algorithms including certain versions of boosting and SVM's. The refined bounds improve the ones described in Section 4.

Most of the argument described in the previous section work in this framework as well, provided the loss function is Lipschitz and there is a uniform bound on the functions $(x, y) \mapsto \phi(-f(x)y)$. However, some extra steps are needed to obtain the results. On the one hand, one relates the excess misclassification error $L(f) - L^*$ to the excess loss $A(f) - A^*$. Zhang's lemma (11) may be improved under Tsybakov's noise condition to yield

$$L(f) - L(f^*) \leq \left(\frac{2^s c}{\beta^{1-s}}(A(f) - A^*)\right)^{1/(s-s\alpha+\alpha)} .$$

On the other hand, considering the class of functions

$$\mathcal{M} = \{m_f(x, y) = \phi(-yf(x)) - \phi(-yf^*(x)) : f \in \mathcal{F}\},$$

one has to relate $\mathsf{Var}(m_f)$ to $P(m_f)$, and finally compute the modulus of continuity of the Rademacher process indexed by $\mathcal{M}$. We omit the often somewhat technical details and direct the reader to the references for the detailed arguments.

As an illustrative example recall the case when $\mathcal{F} = \mathcal{F}_\lambda$ is defined as in (7). Then, the empirical minimizer f_n of the cost functional $A_n(f)$ satisfies, with probability at least $1 - \delta$,

$$A(f_n) - A^* \leq C\left(n^{-\frac{1}{2}\cdot\frac{V+2}{V+1}} + \frac{\log(1/\delta)}{n}\right)$$

where the constant C depends on the cost functional and the VC dimension V of the base class $\mathbb{C}$. Combining this with the above improvement of Zhang's lemma, one obtains significant improvements of the performance bound (12).

5.5. **Minimax Lower Bounds**

The purpose of this section is to investigate how good the bounds obtained in the previous sections for empirical risk minimization are. We seek answers for the following questions: Are these upper bounds (at least up to the order of magnitude) tight? Is there a much better way of selecting a classifier than minimizing the empirical error?

Let us formulate exactly what we are interested in. Let $\mathbb{C}$ be a class of decision functions $g : \mathbb{R}^d \to \{0, 1\}$. The training sequence $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ is used to select the classifier $g_n(X) = g_n(X, D_n)$ from $\mathbb{C}$, where the selection is based on the data D_n. We emphasize here that g_n can be an arbitrary function of the data, we do not restrict our attention to empirical error minimization, where g_n is a classifier in $\mathbb{C}$ that minimizes the number errors committed on the data D_n.

To make the exposition simpler, we only consider classes of functions (where the target is supposed to lie) with finite VC dimension. As before, we measure the performance of the selected classifier by the difference between the error probability $L(g_n)$ of the selected classifier and that of the best in the class, $L_{\mathbb{C}} = \inf_{g\in\mathbb{C}} L(g)$. In particular, we seek lower bounds for

$$\sup \mathbb{E}L(g_n) - L_{\mathbb{C}},$$

where the supremum is taken over all possible distributions of the pair (X, Y). A lower bound for this quantity means that no matter what our method of picking a rule from $\mathbb{C}$ is, we may face a distribution such that our method performs worse than the bound.

Actually, we investigate a stronger problem, in that the supremum is taken over all distributions with $L_{\mathbb{C}}$ kept at a fixed value between zero and 1/2. We will see that the bounds depend on n, V the VC dimension of $\mathbb{C}$, and $L_{\mathbb{C}}$ jointly. As it turns out, the situations for $L_{\mathbb{C}} > 0$ and $L_{\mathbb{C}} = 0$ are quite different. Also, the fact that the noise is controlled (with Massart's or Tsybakov's noise conditions) has an important influence.

Integrating the deviation inequalities such as (21), we have that for any class $\mathbb{C}$ of classifiers with VC dimension V, a classifier g_n minimizing the empirical risk satisfies

$$\mathbb{E}L(g_n) - L_{\mathbb{C}} \leq O\left(\sqrt{\frac{L_{\mathbb{C}} V_{\mathbb{C}} \log n}{n}} + \frac{V_{\mathbb{C}} \log n}{n}\right),$$

and also

$$\mathbb{E}L(g_n) - L_{\mathbb{C}} \leq O\left(\sqrt{\frac{V_{\mathbb{C}}}{n}}\right).$$

Let $\mathbb{C}$ be a class of classifiers with VC dimension V. Let $\mathcal{P}$ be the set of all distributions of the pair (X, Y) for which $L_{\mathbb{C}} = 0$. Then, for every discrimination rule g_n based upon $X_1, Y_1, \ldots, X_n, Y_n$, and $n \geq V - 1$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}L(g_n) \geq \frac{V-1}{2en}\left(1 - \frac{1}{n}\right). \tag{27}$$

This can be generalized as follows. Let $\mathbb{C}$ be a class of discrimination functions with VC dimension $V \geq 2$. Let $\mathcal{P}$ be the set of all probability distributions of the pair (X, Y) for which for fixed $L \in (0, 1/2)$,

$$L = \inf_{g \in \mathbb{C}} L(g).$$

Then, for every discrimination rule g_n based upon $X_1, Y_1, \ldots, X_n, Y_n$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}(L(g_n) - L) \geq \sqrt{\frac{L(V-1)}{24n}} e^{-8} \quad \text{if } n \geq \frac{V-1}{2L} \max(9, 1/(1-2L)^2). \tag{28}$$

Under Massart's noise condition with parameter s, we have seen that the rate can be improved and that we essentially have, when g_n is the empirical error minimizer,

$$\mathbb{E}(L(g_n) - L^*) \leq C\left(\sqrt{\frac{V}{n}} \wedge \frac{V \log n}{ns}\right),$$

no matter what L^* is, provided $L^* = L_{\mathbb{C}}$. There also exist lower bounds under these circumstances.

Let $\mathbb{C}$ be a class of classifiers with VC dimension V. Let $\mathcal{P}$ be the set of all probability distributions of the pair (X, Y) for which

$$\inf_{g \in \mathbb{C}} L(g) = L^*,$$

and Massart's noise condition is satisfied, that is, $|\eta(X) - 1/2| \geq s$ almost surely where $s > 0$ is a constant. Then, for every discrimination rule g_n based upon $X_1, Y_1, \ldots, X_n, Y_n$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}(L(g_n) - L^*) \geq C\left(\sqrt{\frac{V}{n}} \wedge \frac{V}{ns}\right). \tag{29}$$

Thus, there is a small gap between upper and lower bounds (essentially of a logarithmic factor). This gap can be reduced when the class of functions is rich enough, where richness means that there exists some d such that

all dichotomies of size d can be realized by functions in the class. When $\mathbb{C}$ is such a class, under the above conditions, one can improve (29) to get

$$\sup_P \mathbb{E}(L(g_n) - L^*) \geq K(1-h)\frac{d}{ns}\left(1 + \log\frac{ns^2}{d}\right) \quad \text{if} \quad s \geq \sqrt{d/n}\,.$$

Bibliographical remarks. Inequality (13) is known as Hoeffding's inequality [104], while (14) is referred to as Bernstein's inequality [30]. The constants shown here in Bernstein's inequality actually follow from an inequality due to Bennett [29]. The results (15),(16) and their corollaries (20),(21) are due to Vapnik and Chervonenkis [222, 223]. The proof sketched here is due to Anthony and Shawe-Taylor [10]. Regarding the corollaries of this result, (17) is due to Pollard [174] and (18) is due to Haussler [100]. The fact that the variance can be related to the expectation and that this can be used to get improved rate has been known for a long time in the context of regression function estimation or other statistical problems. Asymptotic results based on this were obtained for example by van de Geer [212]. Birgé and Massart [32] and Lee, Bartlett and Williamson [127] proved exponential inequalities for regression. The fact that that this phenomenon also occurs in the context of classification, under conditions on the noise has been pointed out by Massart [154] and Mammen and Tsybakov [144].

Talagrand's inequality for empirical processes first appeared in [201], for various improvements see Ledoux [125], Massart [152], Rio [176]. The version presented in Theorem 5.1 is an application of the refinement given by Bousquet [43]. Variations on the theme and detailed proofs appeared in [44].

Several methods have been developed in order to obtain sharp rates for empirical error minimization (or M-estimation). A classical trick is the so-called *peeling* technique where the idea is to cut the class of interest into several pieces (according to the variance of the functions) and to apply deviation inequalities separately to each sub-class. This technique is used, for example, by van de Geer [212, 213, 215]. Another approach consists in weighting the class and was used by Vapnik and Chervonenkis [222] in the special case of binary valued functions and extended by Pollard [174], for example. Combining this approach with concentration inequalities was proposed by Massart [154] and this is the approach we have taken here.

The fixed point of the modulus of continuity of the empirical process has been known to play a role in the asymptotic behavior of M-estimators. More recently non-asymptotic deviation inequalities involving this quantity were obtained, essentially in the work of Massart [154] and Koltchinskii and Panchenko [119]. Both approaches use a version of the peeling technique, but the one of Massart uses in addition a weighting approach. More recently, Mendelson [165] obtained similar results using a weighting technique but a peeling into two subclasses only. The main ingredient was the introduction of the star-hull of the class (as we do it here). This approach was further extended in [21] were the peeling and star-hull approach are compared.

Empirical estimates of the fixed point of type ε^* were studied by Koltchinskii and Panchenko [119] in the zero error case. In a related work, Lugosi and Wegkamp [140] obtain bounds in terms of empirically estimated localized Rademacher complexities without noise conditions. In their approach, the complexity of a subclass of $\mathbb{C}$ containing only classifiers with a small empirical risk is used to obtain sharper bounds. A general result, applicable under noise conditions, was proven by Bartlett, Bousquet and Mendelson [21].

Replacing the inequality by an equality in the definition of ψ (thus making the quantity smaller) can yield better rates for certain classes as shown by Bartlett and Mendelson [22]. Applications of results like (26) to classification with VC classes of functions were investigated by Massart and Nédélec [156].

Properties of convex loss functions were investigated by Lin [132], Steinwart [195], and Zhang [232]. The improvement of Zhang's lemma under Tsybakov's noise condition is due to Bartlett, Jordan and McAuliffe [24] who establish more general results. For a further improvement we refer to Blanchard, Lugosi, and Vayatis [36]. The cited improved rates of convergence for $A(f_n) - A^*$ is also taken from [36] which is based on bounds derived by Blanchard, Bousquet, and Massart [35]. [35] also investigates the special cost function $(1+x)_+$ under Massart's noise condition, see also Bartlett, Jordan and McAuliffe [24], Steinwart [185].

Massart [154] gives a version of (26) for the case $w(r) = c\sqrt{r}$ and arbitrary loss functions which is extended for general w in Massart and Nédélec [156] and Bartlett, Jordan and McAuliffe [24]. Bartlett, Bousquet and Mendelson [21] give an empirical version of (26) in the case $w(r) = c\sqrt{r}$.

The lower bound (27) was proved by Vapnik and Chervonenkis [223], see also Haussler, Littlestone, and Warmuth [102], Blumer, Ehrenfeucht, Haussler, and Warmuth [37]. (28) is due to Devroye and Lugosi [68], see also Simon [189] for related results. The lower bounds under conditions on the noise are due to Massart and Nédélec [156]. Similar results under Tsybakov's noise condition for large classes of functions (i.e., with polynomial growth of entropy) are given in the work of Mammen and Tsybakov [144] and Tsybakov [209]. Other minimax results based on growth rate of entropy numbers of the class of function are obtained in the context of classification by Yang [229, 230]. We notice that the distribution which achieves the supremum in the lower bounds typically depends on the sample size. It is thus reasonable to require the lower bounds to be derived in such a way that P cannot depend on the sample size. Such results are called strong minimax lower bounds and were investigated by Antos and Lugosi [12] and Schuurmans [183].

6. PAC-Bayesian Bounds

We now describe the so-called PAC-Bayesian approach to get error bounds. The distinctive feature of this approach is that one assumes that the class $\mathbb{C}$ is endowed with a fixed probability measure π (called the prior) and that the output of the classification algorithm is not a single function but rather a probability distribution ρ over the class $\mathbb{C}$ (called the posterior).

Given this probability distribution ρ, the error is measured under expectation with respect to ρ. In other words, the quantities of interest are $\rho L(g) \stackrel{\text{def}}{=} \int L(g) d\rho(g)$ and $\rho L_n(g) \stackrel{\text{def}}{=} \int L_n(g) d\rho(g)$. This models classifiers whose output is *randomized*, which means that for $x \in \mathcal{X}$, the prediction at x is a random variable taking values in $\{0,1\}$ and being equal to one with probability $\rho g(x) \stackrel{\text{def}}{=} \int g(x) d\rho(g)$. It is important to notice that ρ is allowed to depend on the training data.

We first show how to get results relating $\rho L(g)$ and $\rho L_n(g)$ using basic techniques and deviation inequalities. A preliminary remark is that if ρ does not depend on the training sample, then $\rho L_n(g)$ is simply a sum of independent random variables whose expectation is $\rho L(g)$ so that Hoeffding's inequality applies trivially.

So the difficulty comes when ρ depends on the data. By Hoeffding's inequality, for the class $\mathcal{F} = \{\mathbb{1}_{g(x)\neq y} : g \in \mathbb{C}\}$, one easily gets that for each fixed $f \in \mathcal{F}$,

$$\mathbb{P}\left\{\exists f \in \mathcal{F} : P(f) - P_n(f) \geq \sqrt{\frac{\log(1/\delta)}{2n}}\right\} \leq \delta\,. \tag{30}$$

One can then obtain a *weighted union bound* as follows

$$\begin{aligned}\mathbb{P}\left\{\exists f \in \mathcal{F} : P(f) - P_n(f) \geq \sqrt{\frac{\log(1/(\pi(f)\delta))}{2n}}\right\} &\leq \sum_{f\in\mathcal{F}} \mathbb{P}\left\{\exists f \in \mathcal{F} : P(f) - P_n(f) \geq \sqrt{\frac{\log(1/(\pi(f)\delta))}{2n}}\right\} \\ &\leq \sum_{f\in\mathcal{F}} \pi(f)\delta = \delta\,,\end{aligned}$$

so that we obtain that with probability at least $1-\delta$,

$$\forall f \in \mathcal{F},\ P(f) - P_n(f) \leq \sqrt{\frac{\log(1/\pi(f)) + \log(1/\delta)}{2n}}\,. \tag{31}$$

It is interesting to notice that now the bound depends on the actual function f being considered and not just on the set $\mathcal{F}$. Now, observe that $(\exists f \in \mathcal{F},\ I(f) \geq 0) \Leftrightarrow (\exists \rho,\ \rho I(f) \geq 0)$ where ρ denotes an arbitrary probability

measure on $\mathcal{F}$ so that we can take the expectation of (31) with respect to ρ and use Jensen's inequality. This gives with probability at least $1-\delta$,

$$\forall \rho,\ \rho(P(f) - P_n(f)) \leq \sqrt{\frac{K(\rho,\pi) + H(\rho) + \log(1/\delta)}{2n}}\,.$$

Rewriting this in terms of the class $\mathbb{C}$, we get that, with probability at least $1-\delta$,

$$\forall \rho,\ \rho L(g) - \rho L_n(g) \leq \sqrt{\frac{K(\rho,\pi) + H(\rho) + \log(1/\delta)}{2n}}\,. \tag{32}$$

The left-hand side is the difference between true and empirical errors of a randomized classifier which uses ρ as weights for choosing the decision function (independently of the data). On the right-hand side the entropy H of the distribution ρ (which is small when ρ is concentrated on a few functions) and the Kullback-Leibler divergence K between ρ and the prior distribution π appear.

It turns out that the entropy term is not necessary. The PAC-Bayes bound is a refined version of the above which is proved using convex duality of the relative entropy. The starting point is the following inequality which follows from convexity properties of the Kullback-Leibler divergence (or relative entropy): for any random variable X_f,

$$\rho X_f \leq \inf_{\lambda>0} \frac{1}{\lambda}\left(\log \pi e^{\lambda X_f} + K(\rho,\pi)\right)\,.$$

This inequality is applied to the random variable $X_f = (P(f) - P_n(f))_+^2$ and this means that we have to upper bound $\pi e^{\lambda (P(f)-P_n(f))_+^2}$. We use Markov's inequality and Fubini's theorem to get

$$\mathbb{P}\left\{\pi e^{\lambda X_f} \geq \epsilon\right\} \leq \epsilon^{-1} \pi \mathbb{E} e^{\lambda X_f}\,.$$

Now for a given $f \in \mathcal{F}$,

$$\begin{aligned}
\mathbb{E} e^{\lambda (P(f)-P_n(f))_+^2} &= 1 + \int_1^\infty \mathbb{P}\left\{e^{\lambda (P(f)-P_n(f))_+^2} \geq t\right\} dt \\
&= 1 + \int_0^\infty \mathbb{P}\left\{\lambda (P(f)-P_n(f))_+^2 \geq t\right\} e^t dt \\
&= 1 + \int_0^\infty \mathbb{P}\left\{P(f)-P_n(f) \geq \sqrt{t/\lambda}\right\} e^t dt \\
&\leq 1 + \int_0^\infty e^{-2nt/\lambda + t} dt = 2n
\end{aligned}$$

where we have chosen $\lambda = 2n-1$ in the last step. With this choice of λ we obtain

$$\mathbb{P}\left\{\pi e^{\lambda X_f} \geq \epsilon\right\} \leq \frac{2n}{\epsilon}\,.$$

Choosing $\epsilon = 2n\delta^{-1}$, we finally obtain that with probability at least $1-\delta$,

$$\frac{1}{2n-1} \log \pi e^{\lambda (P(f)-P_n(f))_+^2} \leq \frac{1}{2n-1} \log(2n/\delta)\,.$$

The resulting bound has the following form. With probability at least $1-\delta$,

$$\forall \rho,\ \rho L(g) - \rho L_n(g) \leq \sqrt{\frac{K(\rho,\pi) + \log(2n) + \log(1/\delta)}{2n-1}}\,. \tag{33}$$

This should be compared to (32). The main difference is that the entropy of ρ has disappeared and we now have a logarithmic factor instead (which is usually dominated by the other terms). To some extent, one can consider that the PAC-Bayes bound is a refined union bound where the gain happens when ρ is not concentrated on a single function (or more precisely ρ has entropy larger than $\log n$).

A natural question is whether one can take advantage of PAC-Bayesian bounds to obtain bounds for deterministic classifiers (returning a single function and not a distribution) but this is not possible with (33) when the space $\mathcal{F}$ is uncountable. Indeed, the main drawback of PAC-Bayesian bounds is that the complexity term blows up when ρ is concentrated on a single function, which corresponds to the deterministic case. Hence, they cannot be used directly to recover bounds of the type discussed in previous sections. One way to avoid this problem is to allow the prior to depend on the data. In that case, one can work conditionally on the data (using a double sample trick) and in certain circumstances, the coordinate projection of the class of functions is finite so that the complexity term remains bounded.

Another approach to bridge the gap between the deterministic and randomized cases is to consider successive approximating sets (similar to ϵ-nets) of the class of functions and to apply PAC-Bayesian bounds to each of them. This goes in the direction of chaining or generic chaining.

Bibliographical remarks. The PAC-Bayesian bound (33) was derived by McAllester [157] and later extended in [158, 159]. Langford and Seeger [123] gave an easier proof and some refinements. The symmetrization and conditioning approach was first suggested by Catoni and studied in [53–55]. The chaining idea appears in the work of Kolmogorov [115, 116] and was further developed by Dudley [73] and Pollard [173]. It was generalized by Talagrand [204] and a detailed account of recent developments is given in [208]. The chaining approach to PAC-Bayesian bounds appears in Audibert and Bousquet [14].

7. Stability

The idea of stability is to directly consider the quantity of interest when one studies the error of a given algorithm. More precisely, given a classifier g_n, our aim is to bound $L(g_n) - L_n(g_n)$.

Under certain circumstances, this random quantity is concentrated around its expectation. In that case, one directly obtains a bound from a concentration inequality.

A simple example of such an approach is the following. We consider the case of real-valued classifiers, when the classifier g_n is obtained by thresholding at zero a real-valued function $f_n : \mathcal{X} \to \mathbb{R}$. Given a set of data $(X_1, Y_1), \dots, (X_n, Y_n)$, let us denote by f_n^i the function that is learned from the data after replacing (X_i, Y_i) by an arbitrary pair (x_i', y_i'). Let ϕ be a cost function as defined in Section 4 and assume that for any set of data, any replacement pair and any x, y,

$$|\phi(-yf_n(x)) - \phi(-yf_n^i(x))| \leq \beta \,,$$

for some $\beta > 0$ and that $\phi(-yf(x))$ is bounded by some constant $M > 0$. This is called the *uniform stability* condition. Under this condition, it is easy to see that

$$\mathbb{E}\left[A(f_n) - A_n(f_n)\right] \leq \beta$$

(where the functionals A and A_n are defined in Section 4). Moreover, by the bounded difference inequality, one easily obtains that with probability at least $1 - \delta$,

$$A(f_n) - A_n(f_n) \leq \beta + (2n\beta + M)\sqrt{\frac{\log(1/\delta)}{2n}} \,.$$

Of course, to be of interest, this bound has to be such that β is a non-increasing function of n such that $\sqrt{n}\beta \to 0$ as $n \to \infty$.

This turns out to be the case for regularization-based algorithms such as the support vector machine. Hence one can obtain error bounds for such algorithms using the stability approach. We omit the details and refer the interested reader to the bibliographical remarks for further reading.

Bibliographical remarks. The idea of using stability of a learning algorithm to obtain error bounds was first exploited by Devroye and Wagner [69, 70]. Kearns and Ron [109] investigated it further and introduced formally several measures of stability. Bousquet and Elisseeff [45] obtained exponential bounds under restrictive conditions on the algorithm, using the notion of *uniform stability*. These conditions were relaxed by Kutin and Niyogi [122]. The link between stability and consistency of the empirical error minimizer was studied by Poggio, Rifkin, Mukherjee and Niyogi [172].

8. Model selection

8.1. **Basic concepts**

When facing a concrete classification problem, choosing the right set $\mathbb{C}$ of possible classifiers is a key to success. If $\mathbb{C}$ is so large that it can approximate arbitrarily well any measurable classifier, then $\mathbb{C}$ is susceptible to overfitting and is not suitable for empirical risk minimization, or empirical ϕ-risk minimization. On the other hand, if $\mathbb{C}$ is a small class, for example a class with finite VC dimension, $\mathbb{C}$ will be unable to approximate in any reasonable sense ($L_1(P)$ or Hausdorff distance) a large set of measurable classification rules. Such a dilemma is by no way specific to classification problems, similar phenomena occur in regression problems.

In regression estimation problems one sometimes considers a large set $\mathcal{S}$ of possible targets like Sobolev spaces, and decomposes it into classes $(\mathcal{S}_\theta)_{\theta\in\Theta}$ defined by some smoothness criteria (e.g., by bounds on Sobolev or Besov norms). In non-parametric regression problems with additive noise under known integrability properties, the nuisance parameter is the smoothness of the target function. Note that in classification, a target is defined by a joint distribution on $\mathcal{X}\times\{0,1\}$. The target defines the Bayes classifier g^*, but the same Bayes classifier may be associated with many different targets. In classification problems, the smoothness criteria should reflect the complexity of the Bayes classifier and perhaps margin conditions (see Section 5.3). Thus, model a theory of model selection for classification analog of that developed for regression requires a deep understanding of approximation properties of characteristic functions of sets. Unfortunately, as of writing this survey, no approximation theory tailored to the needs of classification has reached the level of maturity of approximation theory in functional analysis.

Assume that $\mathcal{S}=\cup_{\theta\in\Theta}\mathcal{S}_\theta$, and that for each $\theta\in\Theta$, for each sample size n, there exists an integer $k(\theta)$ such that $g^*_{n,k}$ minimizes the worst-case excess risk, that is,

$$\sup_{P\in\mathcal{S}_\theta}\mathbb{E}\left[L(g^*_{n,k(\theta)})-L^*\right]=\min_{k\in\mathbb{N}}\sup_{P\in\mathcal{S}_\theta}\mathbb{E}\left[L(g^*_{n,k})-L^*\right].$$

An inference procedure outputting a classifier $\tilde{g}_n$ is said to be *adaptive* with respect to $\mathcal{S}$ if there exists a bounded sequence $C_n(\theta)$ such that for all $P\in\mathcal{S}$,

$$\mathbb{E}\left[L(\tilde{g}_n)-L^*\right]\le C_n(\theta)\mathbb{E}\left[L(g^*_{n,k(\theta)})-L^*\right].$$

If, moreover, the sequence $C_n(\theta)$ converges to 1, then the model selection procedure is said to enjoy *exact asymptotic adaptation in the minimax sense*. Inference procedures that satisfy such a property in regression problems have been known for a while for large sets of target functions. However, the classification problem appears to be significantly more complex as it is detailed below.

Even in the regression estimation framework, other, weaker, definitions of adaptivity have been considered. If $\ell(n)$ denotes a slowly varying function (like, e.g., $\log^k(n)$ for some positive k), an inference procedure outputting $\tilde{g}_n$ is said to be adaptive up to $\ell(n)$ with respect to $\mathcal{S}$ if there exists a bounded sequence $C_n(\theta)$ such that, for

all $P \in \mathcal{S}$,

$$\mathbb{E}\left[L(\tilde{g}_n) - L^*\right] \leq C_n(\theta)\ell(n)\mathbb{E}\left[L(g^*_{n,k(\theta)}) - L^*\right] .$$

Different notions of adaptivity in well-understood frameworks, provide us with guidelines about the possible ways to tackle adaptivity issues in classification.

A closely related model selection method is based on Vapnik's *structural risk minimization* principle. Here one considers a possibly infinite collection of classes of classifiers $\mathbb{C}_1, \mathbb{C}_2, \ldots,$.

For each model $\mathbb{C}_k$, let $g^*_{n,k}$ minimize the empirical classification risk over $\mathbb{C}_k$. The model selection problem may be stated as follows: given the data D_n, select a good hypothesis $g^*_{n,k}$ among the minimizers of empirical risk. Now it remains to build a sound decision selection procedure, to assess this procedure on real-life applications, and to prove its efficiency in well-defined theoretical settings. A lot of work remains to be done to reconcile theory and practice.

Recent approaches to model selection in classification problems are shaped by the acknowledged fact that the behavior of excess risk does not depend entirely on the Bayes classifier g^* and on the class of possible classifiers $\mathbb{C}$, but also on the relation between excess risk and the variance of the empirical process indexed by $\mathbb{C}$ which is often governed by noise conditions, see Section 5.3.

Bibliographical remarks. Early work on model selection in the context of regression or prediction with squared loss can be found in Mallows [143], Akaike [5]. Mallows introduced the C_p criterion in [143]. Grenander [97] discusses the use of regularization in statistical inference. Vapnik and Chervonenkis [223] proposed the structural risk minimization approach to model selection in classification, see also [141], [219–221]. This has been refined to allow random penalties estimated from the training data by Lugosi and Nobel [138], see also Bartlett, Boucheron, and Lugosi [20], Lugosi and Wegkamp [140]. A general and influential approach to non-parametric inference through penalty-based model selection is described in Barron, Birgé and Massart [16], see also Birgé and Massart [33], [34]. These papers provide a profound account of the use of sharp bounds on the excess risk for model selection via penalization. In particular, these papers pioneered the use of sharp concentration inequalities in solving model selection problems, see also [15, 52] for illustrations in regression and density estimation.

A recent account of inference methods in non-parametric settings can be found in Tsybakov [211].

The search for simplicity as a model selection method has often been successful in density estimation and data compression problems. It should be noted that when logarithmic loss functions are used, Ockham's razor provides the basis of a sound bias-variance decomposition. Model selection for data compression or density estimation relatively to the Kullback-Leibler loss was investigated by Schwarz [184] and Rissanen [178], the former introduced the BIC criterion (Bayesian Information Criterion), the latter introduced the MDL criterion (Minimum Description Length Criterion), see also Kieffer [112] for a general perspective on the related problem of code-based model selection. The recent papers by Csiszár and Shields [63] and Csiszár [62] emphasize the relevance of tight non-asymptotic exponential inequalities on the excess risk when investigating the asymptotic consistency of BIC and MDL. Barron [17], Barron and Cover [19], [18] investigate model selection in the framework of discrete models for density estimation and regression.

Kernel methods and nearest-neighbor rules have been used to design universal learning rules and in some sense bypass the model selection problem. We refer to Devroye, Györfi and Lugosi [67] for exposition and references.

Hall [98] and many other authors use resampling techniques to perform model selection.

8.2. **Naive model selection through penalization**

We start with describing a naive approach that uses ideas exposed at the first part of this survey. Penalty-based model selection chooses the model $\hat{k}$ that minimizes

$$L_n(g^*_{n,k}) + \mathrm{pen}(n,k) ,$$

among all models $(\mathbb{C}_k)_{k\in\mathbb{N}}$. Denote the selected model by $\hat{k}$. In other words, the selected classifier is $g^*_{n,\hat{k}}$. Throughout this section, $\mathrm{pen}(n,k)$ is a positive, possibly data-dependent, quantity. The intuition behind using penalties is that as large models tend to overfit, and are thus prone to producing excessively small empirical risks, they should be penalized. Determining the right amount of penalization s crucial to the success of the method of minimization of penalized empirical risk.

As usual, our aim is to get an upper-bound on $L(g^*_{n,\hat{k}}) - L^*$. From the definition of the selection criterion, we have for all k,

$$L(g^*_{n,\hat{k}}) - L^* \le L(g^*_{n,k}) - L^* - \left(L(g^*_{n,k}) - L_n(g^*_{n,k}) - \mathrm{pen}(n,k)\right) + \left(L(g^*_{n,\hat{k}}) - L_n(g^*_{n,\hat{k}}) - \mathrm{pen}(n,\hat{k})\right) . \quad (34)$$

Taking expectations, we get

$$\begin{aligned}
\mathbb{E}\left[L(g^*_{n,\hat{k}}) - L^*\right] &\le \mathbb{E}\left[L(g^*_{n,k}) - L^*\right] - \mathbb{E}\left[\left(L(g^*_{n,k}) - L_n(g^*_{n,k})\right) + \mathrm{pen}(n,k)\right] \\
&\quad + \mathbb{E}\left[\left(L(g^*_{n,\hat{k}}) - L_n(g^*_{n,\hat{k}}) - \mathrm{pen}(n,\hat{k})\right)\right] \\
&\le \mathbb{E}\left[L(g^*_{n,k}) - L^* + \mathrm{pen}(n,k)\right] + \mathbb{E}\left[\sup_k \left(L(g^*_{n,k}) - L_n(g^*_{n,k}) - \mathrm{pen}(n,k)\right)\right] \\
&\le \mathbb{E}\left[L(g^*_{n,k}) - L^* + \mathrm{pen}(n,k)\right] + \mathbb{E}\left[\sup_k \left(\sup_{g\in\mathbb{C}_k} \left(L(g) - L_n(g)\right) - \mathrm{pen}(n,k)\right)\right] .
\end{aligned}$$

In view of this inequality, it is natural to look for penalties $\mathrm{pen}(n,k)$ such that both

$$\sup_k \left(\sup_{g\in\mathbb{C}_k} \left(L(g) - L_n(g)\right) - \mathrm{pen}(n,k)\right)$$

and $(\mathrm{pen}(n,k))$ remain as small as possible. Recalling that $\left(\sup_{g\in\mathbb{C}_k} \left(L(g) - L_n(g)\right)\right)$ is sharply concentrated around its mean, it is sensible to choose

$$\mathrm{pen}(n,k) = \mathbb{E}\left[\sup_{g\in\mathbb{C}_k} \left(L(g) - L_n(g)\right)\right] + \sqrt{\frac{\log k}{n}} .$$

Indeed, by the union bound,

$$\mathbb{E}\left[\sup_k \left(\sup_{g\in\mathbb{C}_k} \left(L(g) - L_n(g)\right) - \mathrm{pen}(n,k)\right)\right] \le \sum_k \mathbb{E}\left[\left(\sup_{g\in\mathbb{C}_k} \left(L(g) - L_n(g)\right) - \mathrm{pen}(n,k)\right)_+\right] .$$

Using the bounded differences inequality for each k,

$$\mathbb{P}\left\{\sup_{g\in\mathbb{C}_k} \left(L(g) - L_n(g)\right) \ge \mathrm{pen}(n,k) + \delta\right\} \le \exp\left(-2n\left(\sqrt{\frac{\log k}{n}} + \delta\right)^2\right) \le \frac{1}{k^2}\exp(-2n\delta^2)$$

which implies

$$\mathbb{E}\left[\left(\sup_{g\in\mathbb{C}_k} \left(L(g) - L_n(g)\right) - \mathrm{pen}(n,k)\right)_+\right] \le \frac{1}{k^2}\sqrt{\frac{1}{2n}} .$$

Summing over all k, we get

$$\mathbb{E}\left[\sup_k \left(\sup_{g\in\mathbb{C}_k} \left(L(g) - L_n(g)\right) - \mathrm{pen}(n,k)\right)\right] \le \sqrt{\frac{2}{n}} .$$

Hence, this elementary reasoning leads to the *oracle inequality*

$$\mathbb{E}\left[L(g^*_{n,\hat{k}}) - L^*\right] \leq \inf_k \left(L(g^*_k) - L^* + 3\mathbb{E}\left[\sup_{g\in\mathbb{C}_k} (L(g) - L_n(g))\right] + \sqrt{\frac{\log k}{n}} \right) + \sqrt{\frac{2}{n}}\,.$$

Note that the penalty suggested above is unrealistic as it assumes the knowledge of the true underlying distribution. Therefore, it should be replaced by either a distribution-free penalty or a data-dependent quantity. Distribution-free penalties necessarily lead to highly conservative bounds. In recent years, several data-driven penalization procedures have been proposed. Such procedures are motivated according to computational or to statistical arguments. Here we only focus on statistical arguments. Rademacher averages, as presented in Section 3 are by now regarded as a standard basis for designing data-driven penalties.

In the sequel $\mathcal{F}_k = \{\mathbb{1}_{g(x)\neq y} : g \in \mathbb{C}_k\}$ denotes the loss class associated with $\mathbb{C}_k$. Let $\mathrm{pen}(n,k)$ be defined as

$$\mathrm{pen}(n,k) = 3R_n(\mathcal{F}_k) + \sqrt{\frac{\log k}{n}} + \frac{18\log k}{n}\,. \tag{35}$$

Rademacher averages are sharply concentrated as they not only satisfy the bounded-difference inequality, but also the "Bernstein-like" inequalities

$$\begin{aligned} \mathrm{Var}\left(R_n(\mathcal{F}_k)\right) &\leq \frac{1}{n}\mathbb{E}\left[R_n(\mathcal{F}_k)\right] \\ \mathbb{P}\left\{R_n(\mathcal{F}_k) \leq \mathbb{E}\left[R_n(\mathcal{F}_k)\right] - \epsilon\right\} &\leq \exp\left(-\frac{n\epsilon^2}{2\mathbb{E}[R_n(\mathcal{F}_k)]}\right)\,. \end{aligned}$$

This sharp tail-behavior can be exploited as follows:

$$\begin{aligned} &\mathbb{P}\left\{\sup_{g\in\mathbb{C}_k} (L(g) - L_n(g)) \geq \mathrm{pen}(n,k) + 2\delta\right\} \\ &\leq \mathbb{P}\left\{\sup_{g\in\mathbb{C}_k} (L(g) - L_n(g)) \geq \mathbb{E}\left[\sup_{g\in\mathbb{C}_k} (L(g) - L_n(g))\right] + \sqrt{\frac{\log k}{n}} + \delta\right\} \\ &\quad + \mathbb{P}\left\{R_n(\mathcal{F}_k) \leq \frac{2}{3}\mathbb{E}\left[R_n(\mathcal{F}_k)\right] - \frac{18\log k}{3n} - \frac{\delta}{3}\right\} \\ &\leq \frac{1}{k^2}\exp(-2n\delta^2) + \exp\left(-n\frac{\left(\mathbb{E}\left[R_n(\mathcal{F}_k)/3\right] + \frac{18\log k}{3n} + \frac{\delta}{3}\right)^2}{2\mathbb{E}\left[R_n(\mathcal{F}_k)\right]}\right) \\ &\quad \frac{1}{k^2}\exp(-2n\delta^2) + \exp\left(-n\frac{\left(\frac{18\log k}{3n} + \frac{\delta}{3}\right)}{3}\right) \\ &\leq \frac{1}{k^2}\exp(-2n\delta^2) + \frac{1}{k^2}\exp\left(-\frac{n\delta}{9}\right)\,. \end{aligned}$$

Integrating by parts and summing with respect to k leads to the oracle inequality

$$\mathbb{E}\left[L(g^*_{n,\hat{k}}) - L^*\right] \leq \inf_k \left(L(g^*_k) - L^* + 3\mathbb{E}\left[R_n(\mathcal{F}_k)\right] + \sqrt{\frac{\log k}{n}} + \frac{18k}{\log n} \right) + \sqrt{\frac{2\pi}{n}} + \frac{18}{n}\,. \tag{36}$$

Hence, the price to pay for using data-dependent penalties is negligible with compared to the size of the penalty used in the ideal distribution-dependent scenario. This is due to the fact that the typical fluctuations of

Rademacher averages are much smaller than the supremum of the empirical risk. As a matter of fact, using the Bernstein-like inequality to handle the fluctuations of Rademacher averages is an overkill. We could have simply used the bounded differences inequality in order to deal with the fluctuations of Rademacher averages. Nevertheless, this overkill allows us to point out that the naive way of analyzing penalization and calibrating penalties does not lead to very satisfactory oracle inequalities. Indeed, if noise conditions were favorable, and if we were told in advance which is the right model, the excess risk should decrease like $\frac{1}{n}$. The penalty defined by (35) may not be of the same order of magnitude as $\mathbb{E}\left[L\left(g_{n,k}^* - L_k^*\right)\right]$.

Bibliographical remarks. Data-dependent penalties were suggested by Lugosi and Nobel [138], and in the closely related "luckiness" framework introduced by Shawe-Taylor, Bartlett, Williamson, and Anthony [186]. Penalization based on Rademacher averages was suggested by Bartlett, Boucheron, and Lugosi [20] and Koltchinskii [117]. For refinements and further development, see also Lugosi and Wegkamp [140], Freund [87], Herbrich and Williamson [103], Mendelson and Philips [166], Bartlett, Bousquet and Mendelson [21], Bousquet, Koltchinskii and Panchenko [46]. Koltchinskii and Panchenko [119], Lozano [135].

Kearns, Mansour, Ng, and Ron [108], provides an early attempt to compare model selection criteria originating in structural risk minimization theory, MDL, and the performance of hold-out estimates of overfitting. This paper introduced the *interval problem* where empirical risk minimization and model selection can be performed in a computationally efficient way. The latter problem has also been investigated in [20, 91, 135].

The proof that Rademacher averages, empirical VC-entropy and empirical VC-dimension are sharply concentrated around their mean can be found in [41] and [42].

Lugosi and Wegkamp [140] propose a refined penalization scheme based on localized Rademacher complexities that reconciles bounds presented in this section and the results described by Koltchinskii and Panchenko [119] when the optimal risk is null.

Fromont [91] shows that Rademacher averages are actually a special case of weighted bootstrap estimates of the supremum of empirical processes, and shows how a large collection of variants of bootstrap estimates can be used in model selection for classification. We refer to Giné [95] and Efron et al. [80–82] for basic result on the bootstrap.

8.3. **Adaptive model selection under Massart's noise conditions**

In Section 5 we saw that upper bounding $L(g_{n,k}^*) - L(g_k^*)$ by $2\sup_{g\in\mathbb{C}_k}(L(g) - L_n(g))$ does not lead to sharp bounds. Sharper bounds could be obtained by bounding the excess risk by the increment of a centered empirical process. This line of reasoning becomes relevant when dealing with model selection problems in classification. It can be summarized in the following way:

(1) The difference between the risk of the selected classifier and the risk of any contender can be upper bounded by the increment of a suitable empirical process and the difference between the penalties associated to the selected models and its contender. This upper bound holds for any sample, and follows directly from the definition of the model selection procedure.
(2) With overwhelming probability, the increment of the centered empirical process can be upper bounded by an expression involving the excess-risk of the selected classifier and the excess risk of the contender. This leads to an inequality involving the excess risk of the selected classifier and the excess risk of the contender that holds with high probability. The tools that had been successfully used to establish sharp rates of convergence in Section 5.3, that is, Talagrand's concentration inequality and the peeling device, prove efficient again.
(3) The oracle inequality just follows by taking advantage of the choice of the penalties and taking expectations.

Let us now carry out the first step of this program. Let c denote a positive real number. For any class $\mathbb{C}_k$,

$$\begin{aligned} L\left(g^*_{n,\hat{k}}\right) - L^* &= (1+c)\left(L_n\left(g^*_{n,k}\right) - L_n(g^*) + \mathrm{pen}(n,\hat{k})\right) \\ &\quad + \left(L\left(g^*_{n,\hat{k}}\right) - (1+c)L_n\left(g^*_{n,\hat{k}}\right) - L^* + (1+c)L_n(g^*) - (1+c)\mathrm{pen}(n,\hat{k})\right) \\ &\leq (1+c)\left(L_n\left(g^*_k\right) - L_n(g^*) + \mathrm{pen}(n,k)\right) \\ &\quad + \sup_{k'}\left(L\left(g^*_{n,k'}\right) - (1+c)L_n\left(g^*_{n,k'}\right) - L(g^*) + (1+c)L_n(g^*) - (1+c)\mathrm{pen}(n,k')\right) \quad (37) \\ &\quad \text{(by the definition of } \hat{k}. \end{aligned}$$

Note that (37) differs from (34) in two respects: the last term on the right-hand side is the increment of an empirical process instead of a supremum, and an offset factor $(1+c)$ has been introduced. The latter may be taken arbitrarily close to 1, it will show up in oracle inequalities and seems to be unavoidable if small penalties (that is penalties that are of smaller order than $1/\sqrt{n}$) are to be chosen. Now, taking expectations on both sides,

$$\begin{aligned} &\mathbb{E}\left[L\left(g^*_{n,\hat{k}}\right) - L^*\right] \\ &\leq (1+c)\left(L(g^*_k) - L^* + \mathbb{E}\left[\mathrm{pen}(n,k)\right]\right) \\ &\quad + \mathbb{E}\left[\sup_{k'}\left(L\left(g^*_{n,k'}\right) - (1+c)L_n\left(g^*_{n,k'}\right) - L(g^*) + (1+c)L_n(g^*) - (1+c)\mathrm{pen}(n,k')\right)\right] \\ &\leq (1+c)\left(L(g^*_k) - L^* + \mathbb{E}\left[\mathrm{pen}(n,k)\right]\right) \\ &\quad + \mathbb{E}\left[\sum_{k'}\left(L\left(g^*_{n,k'}\right) - (1+c)L_n\left(g^*_{n,k'}\right) - L(g^*) + (1+c)L_n(g^*) - (1+c)\mathrm{pen}(n,k')\right)\right]. \quad (38) \end{aligned}$$

Relation (38) looks like (34), but the two relations differ to the same extent as Section 3 and Section 5 differ. (38) now suggests what kind of penalties should and could be used in order to achieve some kind of non-asymptotic adaptivity. Indeed, $\mathbb{E}\left[\mathrm{pen}(n,k)\right]$ should be at least of the order of $\mathbb{E}\left[L(g^*_{n,k}) - L(g^*)\right]$, that is, of the order of the excess risk in the k-th model.

We will have to check whether this is enough to ensure that the last summand remains small with respect to $\mathrm{pen}(n,k)$. In order to use the union bound, a term of order $(\log k)/n$, will be incorporated into $\mathrm{pen}(n,k)$. In the next paragraph we check that, under fixed noise conditions, it is possible to define distribution-dependent penalties that lead to some kind of adaptivity. Finally, the construction of the corresponding data-dependent penalties will be sketched.

8.3.1. *Model selection under Massart's noise conditions*

For the sake of simplicity, we work out the main ideas first under Massart's noise conditions. Recall that under such, conditions

$$w(r) = \sup_{g: L(g) \leq L^* + r} \sqrt{\mathsf{Var}\left(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y}\right)} \leq \sqrt{\frac{1}{h}\left(L(g) - L^*\right)}.$$

Such conditions are enforced when $1/\left|1 - 2\eta(\cdot)\right|$ is upper bounded by $1/h$.

Just as we did for the naive model selection procedure, we proceed in two steps:

(1) definition of sensible distribution-dependent penalties, assuming knowledge of $w(\cdot)$ and of the complexities of the classes $\mathbb{C}_k$.
(2) derivation of data-dependent penalties. This step will be exemplified in different ways. In the simplest setting, $w(\cdot)$ is assumed to be known. Then attempts to circumvent the lack of knowledge of $w(\cdot)$ will be presented.

Assume again that for each $k \in \mathbb{N}$, there exists a non-decreasing positive function ϕ_k such that $\phi_k(x)/x$ is non-increasing, and

$$\mathbb{E}\left[\sup_{g\in\mathbb{C}_k,\mathsf{Var}(g-g^*)\leq r^2} |L(g) - L_n(g) - L(g^*) + L_n(g^*)|\right] \leq \phi_k(r)\,.$$

This assumption is not innocuous, we refer to the bibliographical remarks for comments. Let ε_k^* be defined as the solution of the equation $r = \phi_k(w(r))$.

Let δ denote a small positive quantity.

In order to take advantage of the master equation (38), it will prove fruitful to control reweighted empirical processes. The reweighting technique used in this section differ slightly from the technique used in Section 5.3, though the differences are not essential.

Henceforth, let us assume that $c < h$, where h lower bounds $|1 - 2\eta(\cdot)|$. Let K denote a positive number larger than 1 that will be chosen according the supposedly known value of h. In the sequel, for example, we will use $K = (c+1)/(ch)$. Fix δ for a moment and let $y_{k'}$ be defined as

$$y_{k'} = 2K\left[\left(4\frac{\varepsilon_{k'}^*}{w(\varepsilon_{k'}^*)}\right) + \sqrt{\frac{4\log(1/\delta)}{n}} + \frac{4\log(1/\delta)}{3n}\right].$$

For any $g \in \mathbb{C}_{k'}$, the weight of g is defined as

$$\omega_{k'}(g) = P(g^* - g)^2 + y_{k'}^2\,.$$

and the reweighted empirical process indexed by $\mathbb{C}_{k'}$ is defined as

$$g \mapsto \frac{L_n\left(g_k^*\right) - L\left(g_k^*\right) - L_n\left(g\right) + L\left(g\right)}{\omega_{k'}(g)}\,.$$

We will be interested in the supremum of this process

$$V_{k'} = \sup_{g\in\mathbb{C}_{k'}} \frac{L_n\left(g^*\right) - L\left(g^*\right) - L_n\left(g\right) + L\left(g\right)}{\omega_{k'}(g)}$$

Note that $V_{k'}$ is just a supremum of a reweighted centered empirical process. The increments of V_k' are upper bounded by $2/(ny_{k'}^2)$ and for each g,

$$\mathsf{Var}\left[\frac{L_n\left(g_k^*\right) - L\left(g_k^*\right) - L_n\left(g\right) + L\left(g\right)}{\omega_{k'}(g)}\right] \leq \frac{1}{2ny_{k'}^2}\,.$$

Invoking Talagrand's inequality (5.1), we have, with probability $1-\delta$,

$$V_{k'} \leq 2\mathbb{E}\left[V_{k'}\right] + \sqrt{\frac{4\log(1/\delta)}{ny_{k'}^2}} + \frac{4\log(1/\delta)}{3ny_{k'}^2}\,.$$

An appropriate version of the peeling device shows that, for all k',

$$\mathbb{E}\left[V_{k'}\right] \leq \frac{4\varepsilon_{k'}^*}{y_{k'}w(\varepsilon_{k'}^*)}\,.$$

Hence, combining the last two inequalities, we have, with probability at least $1-\delta$,

$$\begin{aligned} V_{k'} &\leq \frac{2}{y_{k'}}\left(4\frac{\varepsilon^*_{k'}}{w(\varepsilon^*_{k'})}\right)+\sqrt{\frac{4\log(1/\delta)}{ny^2_{k'}}}+\frac{4\log(1/\delta)}{3ny^2_{k'}} \\ &\leq \frac{1}{K}=\frac{ch}{h+1}\,, \end{aligned}$$

where the second inequality follows from the definitions of $y_{k'}$ and K. Using this bound, the definition of $V_{k'}$, and the fact that $c<h$:

$$\begin{aligned} L\left(g^*_{n,k'}\right)-L^*+(1+c)\left(L_n\left(g^*\right)-L_n\left(g^*_{n,k'}\right)\right) &\leq chy^2_{k'} \\ &\leq 192\frac{(1+c)^2}{ch}\left[\left(\frac{\varepsilon^*_{k'}}{w(\varepsilon^*_{k'})}\right)^2+\frac{\log(1/\delta)}{3n}\right] \\ &\leq 192\frac{(1+c)^2}{c}\left[\varepsilon^*_{k'}+\frac{\log(1/\delta)}{3nh}\right]. \end{aligned}$$

Now, if $\mathrm{pen}(n,k')$ is defined by

$$\mathrm{pen}(n,k')=192\frac{(1+c)}{c}\left(\varepsilon^*_{k'}+\frac{\log k'^2}{3hn}\right), \tag{39}$$

then the event

$$L\left(g^*_{n,k'}\right)-L^*+(1+c)\left(L_n\left(g^*\right)-L_n\left(g^*_{n,k'}\right)\right)-(1+c)\mathrm{pen}(n,k')\geq x$$

has probability less than

$$\frac{1}{k'^2}\exp\left(-n\frac{3chx}{192(1+c)^2}\right).$$

Thus, we may conclude that

$$\sum_{k'}\mathbb{E}\left[\left(L\left(g^*_{n,k'}\right)-(1+c)L_n\left(g^*_{n,k'}\right)-L(g^*)+(1+c)L_n(g^*)-(1+c)\mathrm{pen}(n,k')\right)\right]\leq 2\frac{192(1+c)^2}{3chn}\,.$$

Hence, under Massart's noise conditions, if the above-defined ideal penalties were used, the following oracle inequality would hold:

$$\mathbb{E}\left[L(g^*_{\hat{k}})-L^*\right]\leq(1+c)\inf_k\left(L(g^*_k)-L^*+192\frac{(1+c)}{c}\left(\varepsilon^*_k+\frac{\log k^2}{3hn}\right)\right)+\frac{128(1+c)^2}{chn}\,. \tag{40}$$

Up to the $1+c$ factor, this penalization scheme represents an ideal bias-variance decomposition. In this sense, the model selection procedure described in [153] is an asymptotically adaptive scheme in the minimax sense (but not an exact one).

8.3.2. *Data-dependent penalties*

In order to get an effective penalty calibration procedure, we need to develop an estimation procedure for $(\varepsilon_{k'})_{k'}$. Such a procedure has been proposed by Bartlett, Bousquet and Mendelson in [21] following ideas initially introduced by Koltchinskii and Panchenko [119]. Here, we sketch the main ideas that lead to sharp data-dependent penalties.

Note that in order to turn the above-described penalization scheme into a data-dependent procedure, it is enough to have, with probability larger than $1-\frac{1}{k^2}\exp(-c'n)$, an upper bound $\hat{\varepsilon}_k$ on ε^*_k. Suppose that $\hat{\phi}_k$ is

defined in such a way that $\hat{\phi}_k$ is positive, non-decreasing, $\hat{\phi}_k(x)/x$ is non-increasing and $\hat{\phi}_k(w(\varepsilon_k^*)) > \phi_k(w(\varepsilon_k^*))$, then the solution $\hat{\varepsilon}_k$ of the equation $r = \hat{\phi}_k(w(r))$ satisfies $\hat{\varepsilon}_k > \varepsilon_k^*$. Note that if there are no a priori guarantees that $\hat{\phi}(x)/x$ is non-increasing, it is enough to replace $\hat{\phi}(x)$ by $x \sup_{x' \geq x} \hat{\phi}(x')/(x')$.

Now, in order to define an adequate approximation $\hat{\phi}_k$ of ϕ_k, we may take advantage of the concentration properties of conditional Rademacher averages and of empirical processes. Instead of

$$\mathbb{E}\left[\sup_{g \in \mathbb{C}_k, \mathsf{Var}(g-g^*) \leq r^2} |L(g) - L_n(g) - L(g^*) + L_n(g^*)|\right] \leq \phi_k(r)\,,$$

we consider

$$\hat{\phi}_k(r) = \mathbb{E}_\sigma\left[\sup_{g \in \mathbb{C}_k, L_n(g) - L_n(g^*_{n,k}) \leq 2r^2} \left|\sum_{i=1}^n \sigma_i \mathbb{1}_{g(X_i) \neq Y_i}\right|\right]\,,$$

where expectation is taken with respect to the Rademacher random variables σ_i. If r is not too small, then as a consequence of Talagrand's inequality, the empirically defined set

$$\left\{g \in \mathbb{C}_k, L_n(g) - L_n(g^*_{n,k}) \leq 2r^2\right\}$$

contains the set

$$\left\{g \in \mathbb{C}_k \ : \ \mathsf{Var}(g - g^*) \leq r^2\right\}.$$

The concentration properties of conditional Rademacher averages now imply that, with high probability,

$$\mathbb{E}_\sigma\left[\sup_{g \in \mathbb{C}_k, \mathsf{Var}(g-g^*) \leq r^2} \left|\sum_{i=1}^n \sigma_i \mathbb{1}_{g(X_i) \neq Y_i}\right|\right] \geq \phi_k(r)\,.$$

It is useful to realize that, provided $\phi_k(x)/x$ and $\hat{\phi}_k(x)/x$ are non-increasing, in order to have $\hat{\epsilon}_k^* \geq \epsilon_k^*$, it is enough that with high probability $\phi_k(w(\epsilon^*)) \leq \hat{\phi}_k(w(\epsilon^*))$.

Bibliographical remarks. The results described in this section are based on Massart [153]. Massart describes how carefully using sharp concentration inequalities for empirical VC dimension and entropy allows to estimate the excess risk in classification problems and how such estimates can be used to build adaptive model selection procedures under suitable noise conditions. In [153] data-dependent estimates of the ε_k^* rely on data-dependent estimates of the average VC dimension and on classical results connecting empirical L_2 entropy numbers and the VC dimension.

Bartlett, Bousquet and Mendelson [21], Lugosi and Wegkamp [140], and more recently Koltchinskii [118] explore the applications of localized Rademacher complexity. They show how localized Rademacher complexities can be used in order to estimate the excess risk and provide an alternative to Massart's approach [153]. Koltchinskii [118] discusses thoroughly different views at sharp data-dependent excess risk estimation procedures.

8.4. **Adaptive model selection under unknown noise conditions**

Under Massart's noise conditions, the relationship between ε_k^* and the excess risk in $\mathbb{C}_k$ is completely specified provided h is known. If h is unknown, let alone if $w(r)$ is unknown, defining data-dependent penalties remains a challenge. Indeed, the computation of localized Rademacher averages allows to approximate the functions $\phi_k(\cdot)$ by data-dependent functions $\hat{\phi}_k(\cdot)$. Unfortunately, in general, this is not enough to approximate the solution of equation $r = \phi_k(w(r))$. However, under certain conditions on the behavior of $w(\cdot)$, an estimate of the desired form may be constructed. For example, one may assume that $w(r)$ behaves like $(r/h)^{\alpha/2}$ for some unknown constants $\alpha \in (0, 1]$, and h.

Pre-testing (comparison) methods attempt to turn around this difficulty in the following framework. The sequence of models $(\mathbb{C}_k)_k$ is assumed to be nested (and finite): $\mathbb{C}_k \subseteq \mathbb{C}_{k+1}$. Moreover, it is assumed that there exists a k^*, such that $g^* \in \mathbb{C}_{k^*}$.

Now recall that if $k \geq k^*$, we have a fairly precise idea of the behavior of the random variables $L(g_k^*) - L(g^*)$, $L_n(g^*) - L_n(g_k^*)$, $d_2^2(g^*, g_k^*)$, and $d_{2,n}^2(g^*, g_k^*)$.

On the other hand, if $g^* \in \mathbb{C}$, we have the relation

$$w^2(r) \leq \sup \left\{ d_2^2(g, g') \ : \ g, g' \in \mathbb{C}, \ L(g) \vee L(g') \leq r \right\} \leq 2w^2(r) \,. \tag{41}$$

This relation tells us that if we can estimate the diameters of the level sets of the excess risk in the class, then we can also estimate $w()$ from above.

It is a quite routine exercise to check that, provided r is large enough, with high probability, the squared empirical diameter of sets is a close approximation of their squared diameter (this is a straightforward consequence of concentration inequalities for suprema of empirical processes indexed by bounded positive functions). It has already been observed that empirical level sets of the excess risk approximately coincide with level sets of the excess risk.

Bringing all these ideas together, it is possible to estimate w from empirical data as soon as the model under consideration contains the Bayes classifier. If we have a finite collection of embedded models whose union is guaranteed to contain the Bayes classifier, it is always possible to estimate $w()$ or rather $\phi_k(w())$ from the data by monitoring the diameters of the level sets of the empirical risk in the largest model. Such a procedure would allow to define a penalty-based model selection method that would be as adaptive to the noise conditions that is to $w()$ as the pre-testing method (or comparison method) described in [210].

Bibliographical remarks. Pre-testing procedures were first proposed by Lepskii [129], [130], [128] for performing model selection in a regression context. They are also discussed by Birgé [31]. Their use in model selection for classification was pioneered by Tsybakov [209]. An early account of ratio-type concentration inequalities can be found in Chapter V of [214].

The testing procedure presented in [210] can be considered as a comparison of $L_n(g_{n,k}^*) - L_n(g_{n,k'}^*)$ and

$$R_n \left(\{ \mathbb{1}_{g(X) \neq Y} - \mathbb{1}_{g_{n,k'}^*(X) \neq Y} \ : \ g \in \mathbb{C}_{k'}, d_{2,n}(g, g_{n,k'}^*) \leq d_{2,n}(g_{n,k}^*, g_{n,k'}^*) \} \right) ,$$

where $k' > k$. If the excess empirical risk is upper bounded by the conditional Rademacher average, this strongly suggests that $L_n(g_{n,k}^*) - L_n(g_{n,k'}^*)$ behaves like the increment of a centered empirical process and thus that $L(g_{n,k}^*)$ and $L(g_{n,k'}^*)$ are not essentially different. Tsybakov's comparison procedure implicitly computes a Rademacher complexity in an empirical L_2 ball centered around a minimizer of the empirical risk. Without having to compute the fixed point of $\phi_k(w(\cdot))$ it provides adaptivity under some restrictive but relevant conditions.

Recently, Bartlett and Mendelson [22], and Koltchinskii [118] went one step further and pointed out that there is no need to estimate separately complexity and noise conditions: what matters is $\phi(w(\cdot))$. In order to estimate the latter quantity, it makes sense to compute localized Rademacher complexities in the level sets of the empirical risk. In his recent work Koltchinskii [118] also revisits comparison-based methods using concentration inequalities and provides a unified account of penalty-based and comparison-based model selection techniques in classification.

Van de Geer and Tsybakov [216] recently pointed out that in some special cases penalty-based model selection can achieve adaptivity to the noise conditions.

8.5. **Revisiting hold-out estimates**

Designing and assessing the above-described model selection policies requires a good command of empirical processes theory. This partly explains why re-sampling techniques like ten-fold cross-validation tend to be favored by practitioners. Moreover, there is no simple way to reduce the computation of local Rademacher averages to empirical risk minimization, while re-sampling methods do not suffer from such a drawback: according

to the computational complexity perspective, carrying out ten-fold cross-validation is not harder than empirical risk minimization. Obtaining non-asymptotic oracle inequalities of such cross-validation methods remains to be a challenge.

The simplest cross validation method is hold-out. It consists in splitting the sample in two parts: a training set of length $n-m$ and a test set of length m. Let us denote by $L'_m(g)$ the average loss of g on the test set. The line of reasoning that allowed to analyze penalty-based model selection under known noise conditions works again:

$$\begin{aligned} L(g^*_{n,\hat{k}}) - L^* &\leq (1+c)\left(L_m(g^*_{n,\hat{k}}) - L_m(g^*) + \mathrm{pen}(n,\hat{k})\right) \\ &\quad + \left(L(g^*_{n,\hat{k}}) - L^* - (1+c)L_m(g^*_{n,\hat{k}}) + (1+c)L_m(g^*) - (1+c)\mathrm{pen}(n,k)\right) \\ &\leq (1+c)\left(L_m(g^*_{n,k}) - L_m(g^*) + \mathrm{pen}(n,k)\right) \\ &\quad + \sup_{k'}\left(L(g^*_{n,k'}) - L^* - (1+c)L_m(g^*_{n,k'}) + (1+c)L_m(g^*) - (1+c)\mathrm{pen}(n,k')\right) . \end{aligned}$$

Let ε denote the solution of the equation $w(x) = \sqrt{x}$. Fix k' for now, and assume that $L(g^*_{n,k'}) - L^* > \varepsilon$. By Bernstein's inequality,

$$\begin{aligned} &\mathbb{P}\left\{L(g^*_{n,k'}) - L^* - (1+c)L_m(g^*_{n,k'}) + (1+c)L_m(g^*) - (1+c)\mathrm{pen}(n,k') \geq x\right\} \\ &\leq \exp\left(-\frac{cm}{2(2c+1)}\left(c(L(g^*_{n,k'}) - L^*)/(c+1) + \mathrm{pen}(n,k') + x\right)\right) \\ &\leq \exp\left(-\frac{cm}{2(2c+1)}\left(\mathrm{pen}(n,k') + x\right)\right) . \end{aligned}$$

If $\mathrm{pen}(n,k') > 2(2c+1)\frac{\log k'^2}{m}$, combining the above-described inequalities and resorting again to integration by parts:

$$\mathbb{E}\left[L(g^*_{n,\hat{k}}) - L^*\right] \leq (1+c)\inf_k\left(L(g^*_{n,k}) - L^* + \varepsilon + \mathrm{pen}(n,k)\right) + \frac{2(2c+1)}{cm} .$$

Bibliographical remarks. Hastie, Tibshirani and Friedman [99] provide an application-oriented discussion of model selection strategies. They provide an argument in defense of the hold-out methodology. An early account of the hold-out strategy can be found in [138], and in [20]. A sharp use of hold-out estimates in an adaptive regression framework is described by Wegkamp in [227]. Better constants and exponential inequalities were recently pointed out by P. Massart [155].

References

[1] R. Ahlswede, P. Gács, and J. Körner. Bounds on conditional probabilities with applications in multi-user communication. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 34:157–177, 1976. (correction in 39:353–354,1977).

[2] M.A. Aizerman, E.M. Braverman, and L.I. Rozonoer. The method of potential functions for the problem of restoring the characteristic of a function converter from randomly observed points. *Automation and Remote Control*, 25:1546–1556, 1964.

[3] M.A. Aizerman, E.M. Braverman, and L.I. Rozonoer. The probability problem of pattern recognition learning and the method of potential functions. *Automation and Remote Control*, 25:1307–1323, 1964.

[4] M.A. Aizerman, E.M. Braverman, and L.I. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:917–936, 1964.

[5] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[6] S. Alesker. A remark on the Szarek-Talagrand theorem. *Combinatorics, Probability, and Computing*, 6:139–144, 1997.

[7] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44:615–631, 1997.

[8] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.
[9] M. Anthony and N. Biggs. *Computational Learning Theory*. Cambridge Tracts in Theoretical Computer Science (30). Cambridge University Press, 1992.
[10] M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1993.
[11] A. Antos, B. Kégl, T. Linder, and G. Lugosi. Data-dependent margin-based generalization bounds for classification. *Journal of Machine Learning Research*, 3:73–98, 2002.
[12] A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Machine learning*, 30:31–56, 1998.
[13] P. Assouad. Densité et dimension. *Annales de l'Institut Fourier*, 33:233–282, 1983.
[14] J.-Y. Audibert and O. Bousquet. Pac-bayesian generic chaining. In L. Saul S. Thrun and B. Schlkopf, editors, *Advances in Neural Information Processing Systems 16*, Cambridge, Mass., 2004. MIT Press.
[15] Yannick Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.
[16] A. Barron, L. Birg, and P. Massart. Risks bounds for model selection via penalization. *Probab. Theory Relat. Fields*, 113:301–415, 1999.
[17] A.R. Barron. Logically smooth density estimation. Technical Report TR 56, Department of Statistics, Stanford University, 1985.
[18] A.R. Barron. Complexity regularization with application to artificial neural networks. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 561–576. NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991.
[19] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.
[20] P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2001.
[21] P. Bartlett, O. Bousquet, and S. Mendelson. Localized Rademacher complexities. In *Proceedings of the 15th annual conference on Computational Learning Theory*, pages 44–48, 2002.
[22] P. Bartlett and S. Mendelson. Empirical minimization. Preprint, 2004.
[23] P. L. Bartlett and S. Ben-David. Hardness results for neural network approximation problems. *Theoretical Computer Science*, 284:53–66, 2002.
[24] P.L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *manuscript*, 2003.
[25] P.L. Bartlett and W. Maass. Vapnik-Chervonenkis dimension of neural nets. In Michael A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 1188–1192. MIT Press, 2003. Second Edition.
[26] P.L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
[27] O. Bashkirov, E.M. Braverman, and I.E. Muchnik. Potential function algorithms for pattern recognition learning machines. *Automation and Remote Control*, 25:692–695, 1964.
[28] S. Ben-David, N. Eiron, and H.-U. Simon. Limitations f learning via embeddings in euclidean half saces. *Journal of Machine Learning Research*, 3:441–461, 2002.
[29] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.
[30] S.N. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.
[31] L. Birgé. An alternative point of view on Lepski's method. In *State of the art in probability and statistics (Leiden, 1999)*, volume 36 of *IMS Lecture Notes Monogr. Ser.*, pages 113–133. Inst. Math. Statist., Beachwood, OH, 2001.
[32] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.
[33] L. Birgé and P. Massart. From model selection to adaptive estimation. In E. Torgersen D. Pollard and G. Yang, editors, *Festschrift for Lucien Le Cam: Research papers in Probability and Statistics*, pages 55–87. Springer, New York, 1997.
[34] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4:329–375, 1998.
[35] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. Manuscript, 2004.
[36] G. Blanchard, G. Lugosi, and N. Vayatis. On the rates of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4:861–894, 2003.
[37] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.
[38] S. Bobkov and M. Ledoux. Poincaré's inequalities and Talagrands's concentration phenomenon for the exponential distribution. *Probability Theory and Related Fields*, 107:383–400, 1997.
[39] B. Boser, I. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152. Association for Computing Machinery, New York, NY, 1992.
[40] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *The Annals Probability*, 2004. to appear.

[41] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000.
[42] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *The Annals Probability*, 31:1583–1614, 2003.
[43] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris*, 334:495–500, 2002.
[44] O. Bousquet. Concentration inequalities for sub-additive functions using the entropy method. In C. Houdré E. Giné and D. Nualart, editors, *Stochastic Inequalities and Applications*. Birkhauser, 2003.
[45] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.
[46] O. Bousquet, V. Koltchinskii, and D. Panchenko. Some local measures of complexity of convex hulls and generalization bounds. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, pages 59–73. Springer, 2002.
[47] L. Breiman. Arcing classifiers. *Annals of Statistics*, 26:801–849, 1998.
[48] L. Breiman. Some infinite theory for predictor ensembles. Technical Report 577, Statistics Department, UC Berkeley, 2000.
[49] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth International, Belmont, CA, 1984.
[50] P. Bühlmann and B. Yu. Boosting with the l_2-loss: Regression and classification. *manuscript*, 2001.
[51] A. Cannon, J.M. Ettinger, D. Hush, and C. Scovel. Machine learning with data dependent hypothesis classes. *J. Mach. Learn. Res.*, 2:335–358, 2002.
[52] G. Castellan. Density estimation via exponential model selection. *IEEE Trans. Inform. Theory*, 49(8):2052–2060, 2003.
[53] O. Catoni. Randomized estimators and empirical complexity for pattern recognition and least square regression. preprint PMA-677.
[54] O. Catoni. *Statistical learning theory and stochastic optimization*. Springer-Verlag, 2001. Probability summer school, Saint Flour 2001, to be published.
[55] O. Catoni. Localized empirical complexity bounds and randomized estimators, 2003. Preprint.
[56] N. Cesa-Bianchi and D. Haussler. A graph-theoretic generalization of the sauer-shelah lemma. *Discrete Applied Mathematics*, 86:27–35, 1998.
[57] M. Collins, R.E. Schapire, and Y. Singer. Logistic regression, AdaBoost and Bregman distances. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000.
[58] C. Cortes and V.N. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.
[59] T.M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–334, 1965.
[60] P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31:377–403, 1979.
[61] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, UK, 2000.
[62] I. Csiszár. Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory*, 48:1616–1628, 2002.
[63] I. Csiszár and P. Shields. The consistency of the BIC Markov order estimator. *Annals of Statistics*, 28:1601–1619, 2000.
[64] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, pages 1–50, January 2002.
[65] A. Dembo. Information inequalities and concentration of measure. *Annals of Probability*, 25:927–939, 1997.
[66] P.A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice-Hall, Englewood Cliffs, NJ, 1982.
[67] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.
[68] L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28:1011–1018, 1995.
[69] L. Devroye and T. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.
[70] L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.
[71] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley, New York, 1973.
[72] R.O. Duda, P.E. Hart, and D.G.Stork. *Pattern Classification*. John Wiley and Sons, 2000.
[73] R. M. Dudley. A course on empirical processes. *Lecture Notes in Mathematics*, 1097:2–142, 1984.
[74] R.M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929, 1978.
[75] R.M. Dudley. Balls in R^k do not cut all subsets of $k+2$ points. *Advances in Mathematics*, 31 (3):306–308, 1979.
[76] R.M. Dudley. Empirical processes. In *Ecole de Probabilité de St. Flour 1982*. Lecture Notes in Mathematics #1097, Springer-Verlag, New York, 1984.
[77] R.M. Dudley. Universal Donsker classes and metric entropy. *Annals of Probability*, 15:1306–1326, 1987.
[78] R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge, 1999.
[79] R.M. Dudley, E. Giné, and J. Zinn. Uniform and universal Glivenko-Cantelli classes. *Journal of Theoretical Probability*, 4:485–510, 1991.

[80] B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.
[81] B. Efron. The jackknife, the bootstrap, and other resampling plans. *SIAM, Philadelphia*, 1982.
[82] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1994.
[83] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989.
[84] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 171–203, Cambridge, MA, 2000. MIT Press.
[85] P. Frankl. On the trace of finite sets. *Journal of Combinatorial Theory, Series A*, 34:41–45, 1983.
[86] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285, 1995.
[87] Y. Freund. Self bounding learning algorithms. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 127–135, 1998.
[88] Y. Freund, Y. Mansour, and R. E. Schapire. Generalization bounds for averaged classifiers (how to be a bayesian without believing). *The Annals of Statistics*, 2004.
[89] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.
[90] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:337–374, 2000.
[91] M. Fromont. *Some problems related to model selection: adaptive tests and bootstrap calibration of penalties*. Thèse de doctorat, Université Paris-Sud, december 2003.
[92] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 1972.
[93] E. Giné. Empirical processes and applications: an overview. *Bernoulli*, 2:1–28, 1996.
[94] E. Giné and J. Zinn. Some limit theorems for empirical processes. *Annals of Probability*, 12:929–989, 1984.
[95] Evarist Gine. Lectures on some aspects of the bootstrap. In *Lectures on probability theory and statistics (Saint-Flour, 1996)*, volume 1665 of *Lecture Notes in Math.*, pages 37–151. Springer, Berlin, 1997.
[96] P. Goldberg and M. Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parametrized by real numbers. *Machine Learning*, 18:131–148, 1995.
[97] Ulf Grenander. *Abstract inference*. John Wiley & Sons Inc., New York, 1981.
[98] Peter Hall. Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.*, 11(4):1156–1174, 1983.
[99] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer Verlag, New York, 2001.
[100] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.
[101] D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69:217–232, 1995.
[102] D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0,1\}$ functions from randomly drawn points. In *Proceedings of the 29th IEEE Symposium on the Foundations of Computer Science*, pages 100–109. IEEE Computer Society Press, Los Alamitos, CA, 1988.
[103] R. Herbrich and R.C. Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3:175–212, 2003.
[104] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
[105] W. Jiang. Process consistency for adaboost. *Annals of Statistics*, to appear, 2003.
[106] D.S. Johnson and F.P. Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6:93–107, 1978.
[107] M. Karpinski and A. Macintyre. Polynomial bounds for VC dimension of sigmoidal and general pfaffian neural networks. *Journal of Computer and System Science*, 54, 1997.
[108] M. Kearns, Y. Mansour, A.Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. In *Proceedings of the Eighth Annual ACM Workshop on Computational Learning Theory*, pages 21–30. Association for Computing Machinery, New York, 1995.
[109] M. J. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.
[110] M.J. Kearns and U.V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, Massachusetts, 1994.
[111] A. G. Khovanskii. *Fewnomials*. Translations of Mathematical Monographs, vol. 88, American Mathematical Society, 1991.
[112] J.C. Kieffer. Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Trans. Inform. Theory*, 39:893–902, 1993.
[113] G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41:495–502, 1970.

[114] P. Koiran and E.D. Sontag. Neural networks with quadratic VC dimension. *Journal of Computer and System Science*, 54, 1997.
[115] A. N. Kolmogorov. On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk SSSR*, 114:953–956, 1957.
[116] A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. *American Mathematical Society Translations, Series 2*, 17:277–364, 1961.
[117] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47:1902–1914, 2001.
[118] V. Koltchinskii. Localized rademacher complexities. Manuscript, september 2003.
[119] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In E. Giné, D.M. Mason, and J.A. Wellner, editors, *High Dimensional Probability II*, pages 443–459, 2000.
[120] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.
[121] S. Kulkarni, G. Lugosi, and S. Venkatesh. Learning pattern classification—a survey. *IEEE Transactions on Information Theory*, 44:2178–2206, 1998. Information Theory: 1948–1998. Commemorative special issue.
[122] S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. In *UAI-2002: Uncertainty in Artificial Intelligence*, 2002.
[123] J. Langford and M. Seeger. Bounds for averaging classifiers. CMU-CS 01-102, Carnegie Mellon University, 2001.
[124] M. Ledoux. Isoperimetry and gaussian analysis. In P. Bernard, editor, *Lectures on Probability Theory and Statistics*, pages 165–294. Ecole d'Eté de Probabilités de St-Flour XXIV-1994, 1996.
[125] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1997. `http://www.emath.fr/ps/`.
[126] M. Ledoux and M. Talagrand. *Probability in Banach Space*. Springer-Verlag, New York, 1991.
[127] W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.
[128] O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3):929–947, 1997.
[129] O.V. Lepskiĭ. A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470, 1990.
[130] O.V. Lepskiĭ. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659, 1991.
[131] Y. Li, P.M. Long, and A. Srinivasan. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62:516–527, 2001.
[132] Y. Lin. A note on margin-based loss functions in classification. Technical Report 1029r, Department of Statistics, University of Wisconsin, Madison, 1999.
[133] Y. Lin. Some asymptotic properties of the support vector machine. Technical Report 1044r, Department of Statistics, University of Wisconsin, Madison, 1999.
[134] Y. Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.
[135] F. Lozano. Model selection using rademacher penalization. In *Proceedings of the Second ICSC Symposia on Neural Computation (NC2000)*. ICSC Adademic Press, 2000.
[136] Malwina J. Luczak and Colin McDiarmid. Concentration for locally acting permutations. *Discrete Mathematics*, page to appear, 2003.
[137] G. Lugosi. Pattern classification and learning theory. In L. Györfi, editor, *Principles of Nonparametric Learning*, pages 5–62. Springer, Wien, 2002.
[138] G. Lugosi and A. Nobel. Adaptive model selection using empirical complexities. *Annals of Statistics*, 27:1830–1864, 1999.
[139] G. Lugosi and N. Vayatis. On the bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, 2003, to appear.
[140] G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *Annals of Statistics*, page to appear, 2004.
[141] G. Lugosi and K. Zeger. Concept learning using complexity regularization. *IEEE Transactions on Information Theory*, 42:48–54, 1996.
[142] A. Macintyre and E.D. Sontag. Finiteness results for sigmoidal "neural" networks. In *Proceedings of the 25th Annual ACM Symposium on the Theory of Computing*, pages 325–334. Association of Computing Machinery, New York, 1993.
[143] C.L. Mallows. Some comments on c_p. *IEEE Technometrics*, 15:661–675, 1997.
[144] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27(6):1808–1829, 1999.
[145] S. Mannor and R. Meir. Weak learners and improved convergence rate in boosting. In *Advances in Neural Information Processing Systems 13: Proc. NIPS'2000*, 2001.
[146] S. Mannor, R. Meir, and T. Zhang. The consistency of greedy algorithms for classification. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, 2002.

[147] K. Marton. A simple proof of the blowing-up lemma. *IEEE Transactions on Information Theory*, 32:445–446, 1986.
[148] K. Marton. Bounding $\bar{d}$-distance by informational divergence: a way to prove measure concentration. *Annals of Probability*, 24:857–866, 1996.
[149] K. Marton. A measure concentration inequality for contracting Markov chains. *Geometric and Functional Analysis*, 6:556–571, 1996. Erratum: 7:609–613, 1997.
[150] L. Mason, J. Baxter, P.L. Bartlett, and M. Frean. Functional gradient techniques for combining hypotheses. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 221–247. MIT Press, Cambridge, MA, 1999.
[151] P. Massart. Optimal constants for Hoeffding type inequalities. Technical report, Mathematiques, Université de Paris-Sud, Report 98.86, 1998.
[152] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 28:863–884, 2000.
[153] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciencies de Toulouse*, IX:245–303, 2000.
[154] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX:245–303, 2000.
[155] P. Massart. *Ecole d'Eté de Probabilité de Saint-Flour XXXIII*, chapter Concentration inequalities and model selection. LNM. Springer-Verlag, 2003.
[156] P. Massart and E. Nédélec. Risk bounds for statistical learning. Preprint, 2004.
[157] D. A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 230–234. ACM Press, 1998.
[158] D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*. ACM Press, 1999.
[159] D. A. McAllester. PAC-bayesian stochastic model selection. *To appear in the Machine Learning Journal*, 2001.
[160] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge, 1989.
[161] C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, New York, 1998.
[162] C. McDiarmid. Concentration for independent permutations. *Combinatorics, Probability, and Computing*, 2:163–178, 2002.
[163] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley, New York, 1992.
[164] S. Mendelson. Improving the sample complexity using global data. *IEEE Trans. Inform. Theory*, 48:1977–1991, 2002.
[165] S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. Smola, editors, *Advanced Lectures in Machine Learning*, LNCS 2600, pages 1–40. Springer, 2003.
[166] S. Mendelson and P. Philips. On the importance of "small" coordinate projections. *Journal of Machine Learning Research*, to appear, 2004.
[167] S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Inventiones Mathematicae*, 152:37–55, 2003.
[168] B.K. Natarajan. *Machine Learning: A Theoretical Approach*. Morgan Kaufmann, San Mateo, CA, 1991.
[169] D. Panchenko. A note on Talagrand's concentration inequality. *Electronic Communications in Probability*, 6, 2001.
[170] D. Panchenko. Some extensions of an inequality of Vapnik and Chervonenkis. *Electronic Communications in Probability*, 7, 2002.
[171] D. Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *Annals of Probability*, to appear, 2003.
[172] T. Poggio, S. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428:419–422, 2004.
[173] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.
[174] D. Pollard. Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics*, 22:271–278, 1995.
[175] E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probability Theory and Related Fields*, 119:163–175, 2001.
[176] E. Rio. Une inegalité de Bennett pour les maxima de processus empiriques. In *Colloque en l'honneur de J. Bretagnolle, D. Dacunha-Castelle et I. Ibragimov*, 2001.
[177] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
[178] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.
[179] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory Series A*, 13:145–147, 1972.
[180] R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.
[181] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:1651–1686, 1998.
[182] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[183] D. Schuurmans. Characterizing rational versus exponential learning curves. In *Computational Learning Theory: Second European Conference. EuroCOLT'95*, pages 272–286. Springer Verlag, 1995.
[184] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.
[185] C. Scovel and I. Steinwart. Fast rates for support vector machines. Los Alamos National Laboratory Technical Report LA-UR 03-9117, 2003.
[186] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.
[187] S. Shelah. A combinatorial problem: Stability and order for models and theories in infinity languages. *Pacific Journal of Mathematics*, 41:247–261, 1972.
[188] G.R. Shorack and J. Wellner. *Empirical Processes with Applications in Statistics*. Wiley, New York, 1986.
[189] H.U. Simon. General lower bounds on the number of examples needed for learning probabilistic concepts. In *Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory*, pages 402–412. Association for Computing Machinery, New York, 1993.
[190] A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors. *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 2000.
[191] A. J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.
[192] D.F. Specht. Probabilistic neural networks and the polynomial Adaline as complementary techniques for classification. *IEEE Transactions on Neural Networks*, 1:111–121, 1990.
[193] J.M. Steele. Existence of submatrices with all possible columns. *Journal of Combinatorial Theory, Series A*, 28:84–88, 1978.
[194] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, pages 67–93, 2001.
[195] I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Transactions on Information Theory*, 2002.
[196] I. Steinwart. Support vector machines are universally consistent. *Journal of Complexity*, 18:768–791, 2002.
[197] I. Steinwart. On the optimal parameter choice in ν-support vector machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1274–1284, 2003.
[198] I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.
[199] S.J. Szarek and M. Talagrand. On the convexified sauer-shelah theorem. *Journal of Combinatorial Theory, Series B*, 69:183–192, 1997.
[200] M. Talagrand. The Glivenko-Cantelli problem. *Annals of Probability*, 15:837–870, 1987.
[201] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22:28–76, 1994.
[202] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.*, 81:73–205, 1995.
[203] M. Talagrand. The Glivenko-Cantelli problem, ten years later. *Journal of Theoretical Probability*, 9:371–384, 1996.
[204] M. Talagrand. Majorizing measures: the generic chaining. *Annals of Probability*, 24:1049–1103, 1996. (Special Invited Paper).
[205] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996.
[206] M. Talagrand. A new look at independence. *Annals of Probability*, 24:1–34, 1996. (Special Invited Paper).
[207] M. Talagrand. Vapnik-Chervonenkis type conditions and uniform Donsker classes of functions. *The Annals of Probability*, 31:1565–1582, 2003.
[208] M. Talagrand. The generic chaining: upper and lower bounds for stochastic processes. Manuscript, 2004.
[209] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *C. R. Acad. Sci. Paris*, to appear, 2001.
[210] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, to appear, 2003.
[211] A. B. Tsybakov. *Introduction l'estimation non-paramtrique*. Springer, 2004.
[212] S. Van de Geer. A new approach to least-squares estimation, with applications. *Annals of Statistics*, 15:587–602, 1987.
[213] S. Van de Geer. Estimating a regression function. *Annals of Statistics*, 18:907–924, 1990.
[214] S. van de Geer. *Applications of empirical process theory*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2000.
[215] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, UK, 2000.
[216] S. van de Geer and A. Tsybakov. Square root penalty: adaptation to the margin in classification and in edge estimation. *submitted*, 2003.
[217] A.W. van der Waart and J.A. Wellner. *Weak convergence and empirical processes*. Springer-Verlag, New York, 1996.
[218] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780, 1963.
[219] V.N. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.
[220] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.
[221] V.N. Vapnik. *Statistical Learning Theory*. John Wiley, New York, 1998.
[222] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[223] V.N. Vapnik and A.Ya. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.
[224] V.N. Vapnik and A.Ya. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and its Applications*, 26:821–832, 1981.
[225] M. Vidyasagar. *A Theory of Learning and Generalization*. Springer, New York, 1997.
[226] V. Vu. On the infeasibility of traininng neural networks with small mean squared error. *IEEE Transactions on Information Theory*, 44:2892–2900, 1998.
[227] Marten Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1):252–273, 2003.
[228] R.S. Wenocur and R.M. Dudley. Some special Vapnik-Chervonenkis classes. *Discrete Mathematics*, 33:313–318, 1981.
[229] Y. Yang. Minimax nonparametric classification. I. Rates of convergence. *IEEE Trans. Inform. Theory*, 45(7):2271–2284, 1999.
[230] Y. Yang. Minimax nonparametric classification. II. Model selection for adaptation. *IEEE Trans. Inform. Theory*, 45(7):2285–2292, 1999.
[231] Y. Yang. Adaptive estimation in pattern recognition by combining different procedures. *Statistica Sinica*, 10:1069–1089, 2000.
[232] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 2003.