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

THEORY OF CLASSIFICATION: A SURVEY OF SOME RECENT ADVANCES *

Stéphane Boucheron[1], Olivier Bousquet[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 led to these recent results.

Résumé. La pratique et la théorie de la reconnaissance des formes ont connu des développements importants durant ces dernières années. Ce survol vise à exposer certaines des idées nouvelles qui ont conduit à ces développements.

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

September 27, 2005.

Contents

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

* *The 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] Laboratoire Probabilités et Modèles Aléatoires, CNRS & Université Paris VII, Paris, France, www.proba.jussieu.fr/~boucheron

[2] Pertinence SA, 32 rue des Jeûneurs, 75002 Paris, France

[3] Department of Economics, Pompeu Fabra University, Ramon Trias Fargas 25-27, 08005 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 led to effective new ways of studying 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 many classification algorithms.

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 biased 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. References and bibliographical remarks are given at the end of each section, in an attempt to avoid interruptions in the arguments.

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 space equipped with a σ-algebra. 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 $g : \mathcal{X} \to \{-1, 1\}$ 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 formalize the learning problem, we introduce a probabilistic setting, and let (X, Y) be an $\mathcal{X} \times \{-1, 1\}$-valued random pair, modeling observation and its corresponding class. The distribution of the random pair (X, Y) may be described by the probability distribution of X (given by the probabilities $\mathbb{P}\{X \in A\}$ for all measurable subsets A of $\mathcal{X}$) 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

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

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

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

(see, e.g., [72]). The optimal classifier g^* is often called the *Bayes classifier*. In the statistical model we focus on, one has access to a collection of data (X_i, Y_i), $1 \leq i \leq n$. We assume that the data D_n consists of a sequence of independent identically distributed (i.i.d.) random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ with the same distribution as that of (X, Y).

A classifier is constructed on the basis of $D_n = (X_1, Y_1, \ldots, X_n, Y_n)$ 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*

$$L(g_n) = \mathbb{P}\{g_n(X) \neq Y | D_n\} \ .$$

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 aim 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 Fukunaga [97], Duda and Hart [77], Vapnik and Chervonenkis [233], Devijver and Kittler [70], Vapnik [229,230], Breiman, Friedman, Olshen, and Stone [53], Natarajan [175], McLachlan [169], Anthony and Biggs [10], Kearns and Vazirani [117], Devroye, Györfi, and Lugosi [72], Ripley [185], Vidyasagar [235]. Kulkarni, Lugosi, and Venkatesh [128], Anthony and Bartlett [9], Duda, Hart, and Stork [78], Lugosi [144], and Mendelson [171].

3. Empirical risk minimization and Rademacher averages

A simple and natural approach to the classification problem is to consider a class $\mathcal{C}$ 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:

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

Then the probability of error

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

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

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

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

We see that by guaranteeing that the uniform deviation $\sup_{g\in\mathcal{C}} |L_n(g) - L(g)|$ 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 $\mathcal{C}$ and at the same time the empirical estimate $L_n(g_n^*)$ is also good.

It is important to note at this point that bounding the excess risk by the maximal deviation as in (2) is quite loose in many situations. In Section 5 we survey some ways of obtaining improved bounds. On the other hand, the simple inequality above offers a convenient way of understanding some of the basic principles and it is even sharp in a certain minimax sense, see Section 5.5.

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 $Pf = \mathbb{E}f(X_1)$ and $P_nf = (1/n)\sum_{i=1}^n f(X_i)$, we are interested in upper bounds for the maximal deviation

$$\sup_{f\in\mathcal{F}} (Pf - P_nf) \ .$$

Concentration inequalities are among the basic tools in studying such deviations. The simplest, yet quite powerful exponential concentration inequality is the *bounded differences inequality.*

Theorem 3.1. BOUNDED DIFFERENCES INEQUALITY. *Le $g : \mathcal{X}^n \to \mathbb{R}$ be a function of n variables such that for some nonnegative constants $c_1, \ldots, c_n$,*

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

Let $X_1, \ldots, X_n$ be independent random variables. The random variable $Z = g(X_1, \ldots, X_n)$ satisfies

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

where $C = \sum_{i=1}^n c_i^2$.

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

$$Z = \sup_{f\in\mathcal{F}} |Pf - P_nf| \ .$$

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

$$\sup_{f\in\mathcal{F}} |Pf - P_nf| \le \mathbb{E}\sup_{f\in\mathcal{F}} |Pf - P_nf| + \sqrt{\frac{2\log\frac{1}{\delta}}{n}} \ . \tag{3}$$

This concentration result allows us to focus on the expected value, which can be bounded conveniently by a simple symmetrization device. Introduce a "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,

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

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

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}|P_n'f - P_nf| &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^n (f(X_i') - f(X_i)\right|\right] \\ &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i(f(X_i') - f(X_i)\right|\right] \\ &\le 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i f(X_i)\right|\right] .\end{aligned}$$

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

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

$R_n(A)$ is called the *Rademacher average* associated with 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 the following.

Theorem 3.2. *With probability at least* $1-\delta$*,*

$$\sup_{f\in\mathcal{F}}|Pf - P_nf| \le 2\mathbb{E}R_n(\mathcal{F}(X_1^n)) + \sqrt{\frac{2\log\frac{1}{\delta}}{n}} .$$

We also have

$$\sup_{f\in\mathcal{F}}|Pf - P_nf| \le 2R_n(\mathcal{F}(X_1^n)) + \sqrt{\frac{2\log\frac{2}{\delta}}{n}} .$$

The second statement follows simply by noticing that the random variable $R_n(\mathcal{F}(X_1^n)$ satisfies the conditions of the bounded differences inequality. The second inequality 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 compute 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

$$\frac{1}{2}\mathbb{E}R_n(\mathcal{F}(X_1^n)) - \frac{1}{2\sqrt{n}} \le \mathbb{E}\sup_{f\in\mathcal{F}}|Pf - P_nf| \le 2\mathbb{E}R_n(\mathcal{F}(X_1^n)))$$

(see, e.g., van der Vaart and Wellner [227]).

Next we recall some of the simple structural properties of Rademacher averages.

Theorem 3.3. PROPERTIES OF RADEMACHER AVERAGES. *Let* A, B *be bounded subsets of* $\mathbb{R}^n$ *and let* $c \in \mathbb{R}$ *be a constant. Then*

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

where $c \cdot A = \{ca : a \in A\}$ *and* $A \oplus B = \{a + b : a \in A, b \in B\}$. *Moreover, if* $A = \{a^{(1)}, \dots, a^{(N)}\} \subset \mathbb{R}^n$ *is a finite set, then*

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

where $\|\cdot\|$ *denotes Euclidean norm. If* $absconv(A) = \left\{ \sum_{j=1}^N c_j a^{(j)} : N \in \mathbb{N},\ \sum_{j=1}^N |c_j| \le 1, a^{(j)} \in A \right\}$ *is the absolute convex hull of* A*, then*

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

Finally, the contraction principle *states that if* $\phi : \mathbb{R} \to \mathbb{R}$ *is a function with* $\phi(0) = 0$ *and Lipschitz constant* L_ϕ *and* $\phi \circ A$ *is the set of vectors of form* $(\phi(a_1), \dots, \phi(a_n)) \in \mathbb{R}^n$ *with* $a \in A$*, then*

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

PROOF. The first three properties are immediate from the definition. Inequality (4) 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 $s > 0$, $\mathbb{E} \exp(sX) \le \exp\left(s^2(\beta - \alpha)^2/8\right)$. In particular, by independence,

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

This implies that

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

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).

The identity (5) is easily seen from the definition. For a proof of the contraction principle, see Ledoux and Talagrand [133]. □

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 $\{(x, y) : g(x) \ne y\}$. In such a case, for any collection of points $x_1^n = (x_1, \dots, x_n)$, $\mathcal{F}(x_1^n)$ 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* (where VC stands for Vapnik-Chervonenkis). 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,

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

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 $(a_{i_1}, \ldots, a_{i_V}) = b$. 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

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

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

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

where we denote by $V(x_1^n)$ the VC dimension of $\mathcal{F}(x_1^n)$. Thus, the expected maximal deviation $\mathbb{E} \sup_{f \in \mathcal{F}} |Pf - P_n f|$ 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 we obtain the following version of what has been known as the *Vapnik-Chervonenkis inequality*:

Theorem 3.4. VAPNIK-CHERVONENKIS INEQUALITY. *For all distributions one has*

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

Also,

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

for a universal constant C.

The second inequality, that allows to remove the logarithmic factor, follows from a somewhat refined analysis (called *chaining*).

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

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

has VC dimension $V \leq m$.

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 Shorack and Wellner [199], Giné [98], van der Vaart and Wellner [227], Vapnik [231], Dudley [83]. The use of empirical processes in classification was pioneered by Vapnik and Chervonenkis [232,233] and re-discovered 20 years later by Blumer, Ehrenfeucht, Haussler, and Warmuth [41], Ehrenfeucht, Haussler, Kearns, and Valiant [88]. For surveys see Natarajan [175], Devroye [71] Anthony and Biggs [10], Kearns and Vazirani [117], Vapnik [230,231], Devroye, Györfi, and Lugosi [72], Ripley [185], Vidyasagar [235], Anthony and Bartlett [9],

The bounded differences inequality was formulated explicitly first by McDiarmid [166] (see also the surveys [167]). The martingale methods used by McDiarmid had appeared in early work of Hoeffding [109], Azuma [18], Yurinksii [242, 243], Milman and Schechtman [174]. Closely related concentration results have been obtained in various ways including information-theoretic methods (see Ahlswede, Gács, and Körner [1], Marton [154],

[155], [156], Dembo [69], Massart [158] and Rio [183]), Talagrand's induction method [217], [213], [216] (see also McDiarmid [168], Luczak and McDiarmid [143], Panchenko [176–178]) and the so-called "entropy method", based on logarithmic Sobolev inequalities, developed by Ledoux [132], [131], see also Bobkov and Ledoux [42], Massart [159], Rio [183], Boucheron, Lugosi, and Massart [45, 46], Bousquet [47], and Boucheron, Bousquet, Lugosi, and Massart [44].

Symmetrization was at the basis of the original arguments of Vapnik and Chervonenkis [232, 233]. We learnt the simple symmetrization trick shown above from Giné and Zinn [99] but different forms of symmetrization have been at the core of obtaining related results of similar flavor, see also Anthony and Shawe-Taylor [11], Cannon, Ettinger, Hush, Scovel [55], Herbrich and Williamson [108], Mendelson and Philips [172].

The use of Rademacher averages in classification was first promoted by Koltchinskii [124] and Bartlett, Boucheron, and Lugosi [24], see also Koltchinskii and Panchenko [126,127], Bartlett and Mendelson [29], Bartlett, Bousquet, and Mendelson [25], Bousquet, Koltchinskii, and Panchenko [50], Kégl, Linder, and Lugosi [13], Mendelson [170].

Hoeffding's inequality appears in [109]. For a proof of the contraction principle we refer to Ledoux and Talagrand [133].

Sauer's lemma was proved independently by Sauer [189], Shelah [198], and Vapnik and Chervonenkis [232]. For related combinatorial results we refer to Frankl n [90], Haussler [106], Alesker [7], Alon, Ben-David, Cesa-Bianchi, and Haussler [8], Szarek and Talagrand [210], Cesa-Bianchi and Haussler [60], Mendelson and Vershynin [173], [188].

The second inequality of Theorem 3.4 is based on the method of chaining, and was first proved by Dudley [81].

The question of how $\sup_{f\in\mathcal{F}} |Pf - P_n f|$ behaves has been known as the Glivenko-Cantelli problem and much has been said about it. A few key references include Vapnik and Chervonenkis [232, 234], Dudley [79, 81, 82], Talagrand [211, 212, 214, 218], Dudley, Giné, and Zinn [84], Alon, Ben-David, Cesa-Bianchi, and Haussler [8], Li, Long, and Srinivasan [138], Mendelson and Vershynin [173].

The VC dimension has been widely studied and many of its properties are known. We refer to Cover [63], Dudley [80, 83], Steele [204], Wenocur and Dudley [238], Assouad [15], Khovanskii [118], Macintyre and Sontag [149], Goldberg and Jerrum [101], Karpinski and A. Macintyre [114], Koiran and Sontag [121], Anthony and Bartlett [9], and Bartlett and Maass [28].

4. Minimizing cost functions: some basic ideas behind boosting and support vector machines

The results summarized in the previous section reveal that minimizing the empirical risk $L_n(g)$ over a class $\mathcal{C}$ 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 $\inf_{g\in\mathcal{C}} L(g)$ in the class, $\inf_{g\in\mathcal{C}} L(g) - L^*$ 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 $\mathcal{C}$ is the class of classifiers that split the space of observations by a hyperplane, the minimization problem is NP hard.

The computational difficulty of learning problems deserves some more attention. Let us consider in more detail the problem in the case of half-spaces. Formally, we are given a sample, that is a sequence of n vectors $(x_1, \ldots, x_n)$ from $\mathbb{R}^d$ and a sequence of n labels $(y_1, \ldots, y_n)$ from $\{-1, 1\}^n$, and in order to minimize the empirical misclassification risk we are asked to find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ so as to minimize

$$\#\{k \ : \ y_k \cdot (\langle w, x_k \rangle - b) \leq 0\} \ .$$

Without loss of generality, the vectors constituting the sample are assumed to have rational coefficients, and the size of the data is the sum of the bit lengths of the vectors making the sample. Not only minimizing the number of misclassification errors has been proved to be at least as hard as solving any NP-complete problem, but even approximately minimizing the number of misclassification errors within a constant factor of the optimum has been shown to be NP-hard.

This means that, unless P =NP, we will not be able to build a computationally efficient empirical risk minimizer for half-spaces that will work for all input space dimensions. If the input space dimension d is fixed, an algorithm running in $O(n^{d-1}\log n)$ steps enumerates the trace of half-spaces on a sample of length n. This allows an exhaustive search for the empirical risk minimizer. Such a possibility should be considered with circumspection since its range of applications would extend much beyond problems where input dimension is less than 5.

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

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

To lighten notation we will simply write $L(f) = L(g_f)$. Let $\phi : \mathbb{R} \to \mathbb{R}_+$ be a nonnegative 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 $L_n(f) \leq A_n(f)$.

Theorem 4.1. *Assume that the function f_n is chosen from a class $\mathcal{F}$ based on the data $(Z_1,\ldots,Z_n) \stackrel{\text{def}}{=} (X_1,Y_1),\ldots,(X_n,Y_n)$. Let B denote a uniform upper bound on $\phi(-f(x)y)$ and let L_ϕ be the Lipschitz constant of ϕ. Then the probability of error of the corresponding classifier may be bounded, with probability at least $1-\delta$, by*

$$L(f_n) \leq A_n(f_n) + 2L_\phi \mathbb{E}R_n(\mathcal{F}(X_1^n)) + B\sqrt{\frac{2\log\frac{1}{\delta}}{n}} \,.$$

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

PROOF. The proof similar to he argument of the previous section:

$$\begin{aligned}
L(f_n) &\leq A(f_n) \\
&\leq A_n(f_n) + \sup_{f\in\mathcal{F}}(A(f) - A_n(f)) \\
&\leq A_n(f_n) + 2\mathbb{E}R_n(\phi\circ\mathcal{H}(Z_1^n)) + B\sqrt{\frac{2\log\frac{1}{\delta}}{n}} \\
&\qquad \text{(where $\mathcal{H}$ is the class of functions $\mathcal{X}\times\{-1,1\}\to\mathbb{R}$ of the form $-f(x)y$, $f\in\mathcal{F}$)} \\
&\leq A_n(f_n) + 2L_\phi\mathbb{E}R_n(\mathcal{H}(Z_1^n)) + B\sqrt{\frac{2\log\frac{1}{\delta}}{n}} \\
&\qquad \text{(by the contraction principle of Theorem 3.3)} \\
&= A_n(f_n) + 2L_\phi\mathbb{E}R_n(\mathcal{F}(X_1^n)) + B\sqrt{\frac{2\log\frac{1}{\delta}}{n}}\ .
\end{aligned}$$

□

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

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

where $\mathcal{C}$ 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

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

where $V_\mathcal{C}$ 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 class $\mathcal{C}$ contains all classifiers of the form $g(x) = 2\mathbb{1}_{x\leq a} - 1$, $a\in\mathbb{R}$. Then $V_\mathcal{C} = 1$ 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 with 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. For such large classes of classifiers it is impossible to guarantee that $L(f_n)$ exceeds the minimal risk in the class by something of the order of $n^{-1/2}$ (see Section 5.5). However, $L(f_n)$ may be made as small as the minimum of the cost functional $A(f)$ over the class plus $O(n^{-1/2})$.

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 $1-\delta$,

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

The remarkable fact about this inequality is that the upper bound only involves the VC dimension of the class $\mathcal{C}$ 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

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

In this case $B = 1$ and $L_\phi = 1/\gamma$. Notice also that $\mathbb{1}_{x>0} \le \phi(x) \le \mathbb{1}_{x>-\gamma}$ and therefore $A_n(f) \le L_n^\gamma(f)$ 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 $\gamma > 0$, $L_n^\gamma(f) \ge L_n(f)$ 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 the following margin-based bound for the risk:

Corollary 4.2. *For any $\gamma > 0$, with probability at least $1 - \delta$,*

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

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 $\mathcal{C}$. 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

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

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

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

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 $f_{x_1}, f_{x_2} \in \mathcal{F}$ such that $k(x_1, x_2) = \langle f_{x_1}, f_{x_2}\rangle$. This means that any linear algorithm based on computing inner products 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 functional on a ball of the associated reproducing kernel Hilbert space of the form

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

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.

An important property of functions in the reproducing kernel Hilbert space associated with 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

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

by the Cauchy-Schwarz inequality, where $\|\cdot\|$ denotes the norm in the reproducing kernel Hilbert space. 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 the following.

Corollary 4.3. *Let f_n be any function chosen from the ball $\mathcal{F}_\lambda$. Then, 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). Assuming that h_η is differentiable, 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 [244] 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 $c = 1/\sqrt{2}$.

Theorem 4.4. EXCESS RISK OF CONVEX RISK MINIMIZERS. *Assume that f_n is chosen from a class $\mathcal{F}_\lambda$ defined in (7) by minimizing the empirical cost functional $A_n(f)$ using either the exponential of the logit cost function. Then, with probability at least $1-\delta$,*

$$L(f_n) - L^* \le 2\left(2L_\phi\lambda\sqrt{\frac{2V_\mathcal{C}\log(n+1)}{n}} + B\sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right)^{1/2} + \sqrt{2}\left(\inf_{f\in\mathcal{F}_\lambda} A(f) - A^*\right)^{1/2}$$

PROOF.

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

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 $L_\phi \le 1$ and $B = \lambda$. In both cases, if there exists a λ 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 fact that the exponent in the rate of convergence is dimension-free is remarkable. (We note here that these rates may be further improved by applying the refined techniques resumed in Section 5.3, see also [40].) 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 $\mathcal{C}$ contains all "decision stumps", that is, all classifiers of the form $s^+_{i,t}(x) = \mathbb{1}_{x^{(i)}\ge t} - \mathbb{1}_{x^{(i)}<t}$ and $s^-_{i,t}(x) = \mathbb{1}_{x^{(i)}<t} - \mathbb{1}_{x^{(i)}\ge t}$, $t \in [0,1]$, $i = 1,\ldots,d$, 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 $V_\mathcal{C} \le \lfloor 2\log_2(2d)\rfloor$. 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 $|f_1|_{TV} + \cdots + |f_d|_{TV} \le \lambda$ 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 dimension.

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 $H(\eta) = 2 \in (\eta, 1-\eta)$ and therefore inequality (11) holds with $c = 1/2$ and $s = 1$. Thus,

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

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 [112], Vu [236], Bartlett and Ben-David [26], Ben-David, Eiron, and Simon [32].

Boosting algorithms were originally introduced by Freund and Schapire (see [91], [94], and [190]), 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 [191], and Koltchinskii and Panchenko [127]. It was Breiman [51] 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 [157], Collins, Schapire, and Singer [61], Friedman, Hastie, and Tibshirani [95]. Based on this view, various versions of boosting algorithms have been shown to be consistent in different settings, see Breiman [52], Bühlmann and Yu [54], Blanchard, Lugosi, and Vayatis [40], Jiang [111], Lugosi and Vayatis [146], Mannor and Meir [152], Mannor, Meir, and Zhang [153], Zhang [244]. Inequality (8) was first obtained by Schapire, Freund, Bartlett, and Lee [191]. The analysis presented here is due to Koltchinskii and Panchenko [127].

Other classifiers based on weighted voting schemes have been considered by Catoni [57–59], Yang [241], Freund, Mansour, and Schapire [93].

Kernel methods were pioneered by Aizerman, Braverman, and Rozonoer [2–5], Vapnik and Lerner [228], Bashkirov, Braverman, and Muchnik [31], Vapnik and Chervonenkis [233], and Specht [203].

Support vector machines originate in the pioneering work of Boser, Guyon, and Vapnik [43], Cortes and Vapnik [62]. For surveys we refer to Cristianini and Shawe-Taylor [65], Smola, Bartlett, Schölkopf, and Schuurmans [201], Hastie, Tibshirani, and Friedman [104], Schölkopf and Smola [192].

The study of universal approximation properties of kernels and statistical consistency of Support Vector Machines is due to Steinwart [205–207], Lin [140, 141], Zhou [245], and Blanchard, Bousquet, and Massart [39].

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 Kimeldorf and Wahba [120] and Craven and Wahba [64] in the context of regression. Relationships between Support Vector Machines and regularization were described by Smola, Schölkopf, and Müller [202] and Evhgeniou, Pontil, and Poggio [89]. General properties of regularized algorithms in reproducing kernel Hilbert spaces are investigated by Cucker and Smale [68], Steinwart [206], Zhang [244].

Various properties of the Support Vector Machine algorithm are investigated by Vapnik [230,231], Schölkopf and Smola [192], Scovel and Steinwart [195] and Steinwart [208,209].

The fact that minimizing an exponential cost functional leads to the Bayes classifier was pointed out by Breiman [52], see also Lugosi and Vayatis [146], Zhang [244]. For a comprehensive theory of the connection between cost functions and probability of misclassification, see Bartlett, Jordan, and McAuliffe [27]. Zhang's lemma (11) appears in [244]. For various generalizations and refinements we refer to Bartlett, Jordan, and McAuliffe [27] and Blanchard, Lugosi, and Vayatis [40].

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 $p = Pf$. Recall that, as a simple consequence of (3), with probability at least $1-\delta$,

$$Pf - P_n f \le \sqrt{\frac{2\log\frac{1}{\delta}}{n}}\,. \tag{12}$$

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

$$Pf - P_n f \le \sqrt{\frac{2\mathsf{Var}(f)\log\frac{1}{\delta}}{n}} + \frac{2\log\frac{1}{\delta}}{3n}\,. \tag{13}$$

Since f takes its values in $\{0,1\}$, $\mathsf{Var}(f) = Pf(1-Pf) \le Pf$ which shows that when Pf is small, (13) is much better than (12).

5.1.1. *General inequalities*

Next we exploit the phenomenon described above to obtain sharper performance bounds for empirical risk minimization. Note that if we consider the difference $Pf - P_n f$ uniformly over the class $\mathcal{F}$, the largest deviations are obtained by functions that have a large variance (i.e., Pf is close to $1/2$). An idea is to scale each function by dividing it by $\sqrt{Pf}$ 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. If $nt^2 \le 2$,

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

Next we introduce Rademacher random variables, obtaining, by simple symmetrization,

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

(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 $X'_1,\ldots,X'_n$.

Summarizing, we obtain the following inequalities:

Theorem 5.1. *Let $\mathcal{F}$ be a class of functions taking binary values in $\{0,1\}$. For any $\delta\in(0,1)$, with probability at least $1-\delta$, all $f\in\mathcal{F}$ satisfy*

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

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

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

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

$$\sup_{f\in\mathcal{F}}\frac{Pf-P_nf}{Pf+P_nf+s/2}\leq 2\sqrt{\frac{\log\mathbb{S}_\mathcal{F}(X_1^{2n})+\log\frac{4}{\delta}}{sn}}\tag{14}$$

and the same is true if P and P_n are permuted. Another consequence of Theorem 5.1 with interesting applications is the following. For all $t\in(0,1]$, with probability at least $1-\delta$,

$$\forall f\in\mathcal{F}, P_nf\leq(1-t)Pf\quad\text{implies}\quad Pf\leq 4\frac{\log\mathbb{S}_\mathcal{F}(X_1^{2n})+\log\frac{4}{\delta}}{t^2n}\,.\tag{15}$$

In particular, setting $t=1$,

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

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 from the second inequality of Theorem 5.1 that, with probability at least $1-\delta$, for all $f\in\mathcal{F}$,

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

Corollary 5.2. *Let g_n^* be the empirical risk minimizer in a class $\mathcal{C}$ of* VC *dimension V. Then, with probability at least $1-\delta$,*

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

Consider first the extreme situation when there exists a classifier in $\mathcal{C}$ which classifies without error. This also means that for some $g' \in \mathcal{C}$, $Y = g'(X)$ with probability one. This is clearly a quite restrictive assumption, only satisfied in very special cases. Nevertheless, the assumption that $\inf_{g\in\mathcal{C}} L(g) = 0$ has been commonly used in computational learning theory, perhaps because of its mathematical simplicity. In such a case, clearly $L_n(g_n^*) = 0$, so that we get, with probability at least $1-\delta$,

$$L(g_n^*) - \inf_{g\in\mathcal{C}} L(g) \le 4\frac{2V\log(n+1) + \log\frac{4}{\delta}}{n}\,. \tag{16}$$

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 these two cases as follows: for simplicity, assume that there is a classifier g' in $\mathcal{C}$ such that $L(g') = \inf_{g\in\mathcal{C}} L(g)$. Then we have

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

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

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

which, together with Corollary 5.2, yields:

Corollary 5.3. *There exists a constant C such that, with probability at least $1-\delta$,*

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

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 Pf and P_nf) and its expectation, namely, $\mathsf{Var}(f) \le Pf$. This is the key property that allows one to obtain faster rates of convergence for $L(g_n^*) - \inf_{g\in\mathcal{C}} L(g)$.

In particular, in the ideal situation mentioned above, when $\inf_{g\in\mathcal{C}} L(g) = 0$, the difference $L(g_n^*) - \inf_{g\in\mathcal{C}} L(g)$ may be much smaller than the worst-case difference $\sup_{g\in\mathcal{C}}(L(g) - L_n(g))$. 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 $L(g_n^*) - \inf_{g\in\mathcal{C}} L(g)$, 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 $\mathcal{C}$, that is, such that $L(g') = \inf_{g\in\mathcal{C}} L(g)$. 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 the loss class $\mathcal{F}$ 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 in $\mathcal{F}$ given by the inequality

$$\mathsf{Var}(f) \le \left(\frac{Pf}{h}\right)^{\alpha} \tag{17}$$

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

$$Pf \leq P_n f + \sqrt{\frac{2(Pf/h)^\alpha \log \frac{N}{\delta}}{n}} + \frac{4 \log \frac{N}{\delta}}{3n} .$$

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

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

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

$$L(g_n^*) - \inf_{g \in \mathcal{G}} L(g) \leq \left(2 \frac{\log \frac{N}{\delta}}{nh^\alpha} \right)^{\frac{1}{2-\alpha}} . \tag{18}$$

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 whether (17) is a reasonable assumption. As the simplest possible example, assume that the Bayes classifier g^* belongs to the class $\mathcal{C}$ (i.e., $g' = g^*$) and the a posteriori probability function η is bounded away from $1/2$, that is, there exists a positive constant h such that for all $x \in \mathcal{X}$, $|2\eta(x) - 1| > h$. Note that the assumption $g' = g^*$ is very restrictive and is unlikely to be satisfied in "practice," especially if the class $\mathcal{C}$ is finite, as it is assumed in this discussion. The assumption that η is bounded away from zero may also appear to be quite specific. However, the situation described here may serve as a first illustration of a nontrivial example when fast rates may be achieved. Since $|\mathbb{1}_{g(X) \neq Y} - \mathbb{1}_{g^*(X) \neq Y}| \leq \mathbb{1}_{g(X) \neq g^*(X)}$, the conditions stated above and (1) imply that

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

Thus (17) holds with $\beta = 1/h$ and $\alpha = 1$ which shows that, with probability at least $1 - \delta$,

$$L(g_n) - L^* \leq C \frac{\log \frac{N}{\delta}}{hn} . \tag{19}$$

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 $\mathcal{C}$ 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 [72, 239, 240]. 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.

The condition of η being bounded away from $1/2$ may be significantly relaxed and generalized. Indeed, in the context of discriminant analysis, Mammen and Tsybakov [151] and Tsybakov [221] formulated a useful condition that has been adopted by many authors. Let $\alpha \in [0, 1)$. Then the Mammen-Tsybakov condition may be stated by any of the following three equivalent statements:

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

We refer to this as the *Mammen-Tsybakov 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. The case $\alpha = 1$ in (1) is realized when there exists an $s > 0$ such that $|2\eta(X) - 1| > s$ almost surely (which is just the extreme 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 $\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y}$. 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 (18) for a finite class of functions.

The sharper bounds, established in this section and the next, come at the price of the assumption that the Bayes classifier is in the class $\mathcal{C}$. Because of this, it is difficult to compare the fast rates achieved with the slower rates proved in Section 3. On the other hand, noise conditions like the Mammen-Tsybakov condition may be used to get improvements even when g^* is not contained in $\mathcal{C}$. In these cases the "approximation error" $L(g') - L^*$ also needs to be taken into account, and the situation becomes somewhat more complex. We return to these issues in Sections 5.3.5 and 8.

5.3. **Localization**

The purpose of this section is to generalize the simple argument of the previous section to more general classes $\mathcal{C}$ 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 [212] 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 5.4. *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 $1 - \delta$, for any $\theta > 0$,*

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

which, for $\theta = 1$ translates to

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

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 5.4. We have, with probability at least $1-\delta$, for all $f \in \mathcal{F}$ with $\mathsf{Var}(f) \le r$,

$$Pf - P_n f \le 2\mathbb{E}\left[\sup_{f\in\mathcal{F}:\mathsf{Var}(f)\le r}(Pf - P_n f)\right] + C\sqrt{\frac{r\log\frac{1}{\delta}}{n}} + C\frac{\log\frac{1}{\delta}}{n}\,. \tag{20}$$

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 $\mathcal{F} = \{(x,y) \mapsto \mathbb{1}_{g(x)\neq y} - \mathbb{1}_{g^*(x)\neq y} : g \in \mathcal{C}\}$ and assume that $g^* \in \mathcal{C}$ and the Mammen-Tsybakov noise condition is satisfied in the extreme case, that is, $|2\eta(x) - 1| > s > 0$ for all $x \in \mathcal{X}$, so for all $f \in \mathcal{F}$, $\mathsf{Var}(f) \le \frac{1}{s}Pf$.

Inequality (20) thus implies that, with probability at least $1-\delta$, all $g \in \mathcal{C}$ satisfy

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

In particular, for the empirical risk minimizer g_n we have, with probability at least $1-\delta$,

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

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 that only contains functions with error less than that of g_n, and thus we would obtain something like

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

This indicates that the quantity that should appear as an upper bound of $L(g_n) - L^*$ is something like $\max\{r : r \le \tilde{\psi}(r/s)\}$. 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 offers a detailed argument.

5.3.3. *Localization: rigorous argument*

Let us introduce the loss class $\mathcal{F} = \{(x,y) \mapsto \mathbb{1}_{g(x)\neq y} - \mathbb{1}_{g^*(x)\neq y} : g \in \mathcal{C}\}$ and 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 $f \in \mathcal{F}^*$, $Pf \ge 0$. Also, denoting by f_n the function in $\mathcal{F}$ corresponding to the empirical risk minimizer g_n, we have $P_n f_n \le 0$.

Let $T : \mathcal{F} \to \mathbb{R}^+$ be a function such that for all $f \in \mathcal{F}$, $\mathsf{Var}(f) \le T^2(f)$ and also for $\alpha \in [0,1]$, $T(\alpha f) \le \alpha T(f)$. 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 average indexed by the star-hull of $\mathcal{F}$:

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

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

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

Of course, ψ and w are non-negative and non-decreasing. Moreover, the maps $x \mapsto \psi(x)/x$ and $w(x)/x$ are non-increasing. Indeed, for $\alpha \ge 1$,

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

This entails that ψ and w are continuous on $]0,1]$. In the sequel, we will also use $w^{-1}(x) \stackrel{\text{def}}{=} \max\{u : w(u)\le x\}$, so for $r>0$, we have $w(w^{-1}(r)) = r$. Notice also that $\psi(1)\le 1$ and $w(1)\le 1$. The analysis below uses the additional assumption that $x\mapsto w(x)/\sqrt{x}$ is also non-increasing. This can be enforced by substituting $w'(r)$ for $w(r)$ where $w'(r) = \sqrt{r}\sup_{r'\ge r} w(r')/\sqrt{r'}$.

The purpose of this section is to prove the following theorem which provides sharp distribution-dependent learning rates when the Bayes classifier g^* belongs to $\mathcal{C}$. In Section 5.3.5 an extension is proposed.

Theorem 5.5. *Let $r^*(\delta)$ denote the minimum of 1 and of the solution of the fixed-point equation*

$$r = 4\psi(w(r)) + w(r)\sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{n}\,.$$

Let ε^* denote the solution of the fixed-point equation

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

Then, if $g^*\in\mathcal{C}$, with probability at least $1-\delta$, the empirical risk minimizer g_n satisfies

$$\max\left(L(g_n)-L^*, L_n(g^*)-L_n(g_n)\right) \le r^*(\delta)\,, \tag{21}$$

and

$$\max\left(L(g_n)-L^*, L_n(g^*)-L_n(g_n)\right) \le 2\left(16\varepsilon^* + \left(2\frac{(w(\varepsilon^*))^2}{\varepsilon^*}+8\right)\frac{\log\frac{1}{\delta}}{n}\right)\,. \tag{22}$$

Remark 5.6. Both ψ and w may be replaced by convenient upper bounds. This will prove useful when deriving data-dependent estimates of these distribution-dependent risk bounds.

Remark 5.7. Inequality (22) follows from Inequality (21) by observing that $\varepsilon^*\le r^*(\delta)$, and using the fact that $x\mapsto w(x)/\sqrt{x}$ and $x\mapsto\psi(x)/x$ are non-increasing. This shows that $r^*(\delta)$ satisfies the following inequality:

$$r \le \sqrt{r}\left(4\sqrt{\varepsilon^*} + \frac{w(\varepsilon^*)}{\sqrt{\varepsilon^*}}\sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right) + \frac{8\log\frac{1}{\delta}}{n}\,.$$

Inequality (22) follows by routine algebra.

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

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

At the end of the proof, we will consider $r = w(r^*(\delta))$ or $r = w(\varepsilon^*)$. But for a while we will work with a generic value of r. This will serve to motivate the choice of $r^*(\delta)$.

We thus apply Talagrand's inequality (Theorem 5.4) 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 that, on an event E that has 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 on 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

$$Pf - P_n f \leq \frac{w(Pf) \vee r}{r} \left(4\psi(r) + r\sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n} \right). \tag{23}$$

Then either $w(Pf) \leq r$ which implies $Pf \leq w^{-1}(r)$ or $w(Pf) \geq r$. In this latter case,

$$Pf \leq P_n f + \frac{w(Pf)}{r} \left(4\psi(r) + r\sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n} \right). \tag{24}$$

Moreover, as we have assumed that $x \mapsto w(x)/\sqrt{x}$ is non-increasing, we also 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 $x \leq A^2 + 2B$),

$$Pf \leq 2P_n f + \frac{1}{w^{-1}(r)} \left(4\psi(r) + r\sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n} \right)^2. \tag{25}$$

Since the function f_n corresponding to the empirical risk minimizer satisfies $P_n f_n \leq 0$, we obtain that, on the event E,

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

To minimize the right-hand side, we look for the value of r which makes the two quantities in the maximum equal, that is, $w(r^*(\delta))$ if $r^*(\delta)$ is smaller than 1 (otherwise the first statement in the theorem is trivial).

Now, taking $r = w(r^*(\delta))$ in (24), as $0 \le Pf_n \ge r^*(\delta)$, we also have

$$\begin{aligned} -P_n f_n &\le w(r^*(\delta)) \left(4\frac{\psi(w(r^*(\delta)))}{w(r^*(\delta))} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3w(r^*(\delta))n} \right) \\ &= r^*(\delta)\,. \end{aligned}$$

This proves the first part of Theorem 5.5.

□

5.3.4. *Consequences*

To understand the meaning of Theorem 5.5, consider the case $w(x) = (x/h)^{\alpha/2}$ with $\alpha \le 1$. Observe that such a choice of w is possible under the Mammen-Tsybakov noise condition. Moreover, if we assume that $\mathcal{C}$ is a VC class with VC-dimension V, then it can be shown (see, e.g., Massart [160], Bartlett, Bousquet, and Mendelson [25], [125]) that

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

so that ε^* is upper bounded by

$$C^{2/(2-\alpha)}\left(\frac{V\log n}{n\,h^\alpha}\right)^{1/(2-\alpha)}.$$

We can plug this upper bound into inequality (22). Thus, with probability larger than $1-\delta$,

$$L(g_n) - L^* \le 4\left(\frac{1}{nh^\alpha}\right)^{1/(2-\alpha)}\left(8(C^2V\log n)^{1/(2-\alpha)} + \left((C^2V\log n)^{(\alpha-1)/(2-\alpha)} + 4\right)\log\frac{1}{\delta}\right).$$

5.3.5. *An extended local analysis*

In the preceding sections, we assumed that the Bayes classifier g^* belongs to the class $\mathcal{C}$ and in the description of the consequences that $\mathcal{C}$ is a VC class (and is, therefore, relatively "small").

As already pointed out, in realistic settings, it is more reasonable to assume that the Bayes classifier is only approximated by $\mathcal{C}$. Fortunately, the above-described analysis, the so-called peeling device, is robust and extends to the general case. In the sequel we assume that g' minimizes $L(g)$ over $g \in \mathcal{C}$, be we do not assume that $g' = g^*$.

The loss class $\mathcal{F}$, its star-hull $\mathcal{F}^*$ and the function ψ are defined as in Section 5.3.3, that is,

$$\mathcal{F} = \left\{(x,y) \mapsto \mathbb{1}_{g(x)\neq y} - \mathbb{1}_{g^*(x)\neq y} : g \in \mathcal{C}\right\}.$$

Notice that for $f \in \mathcal{F}$ or $f \in \mathcal{F}^*$, we still have $Pf \ge 0$. Also, denoting by f_n the function in $\mathcal{F}$ corresponding to the empirical risk minimizer g_n, and by f' the function in $\mathcal{F}$ corresponding to g', we have $P_n f_n - P_n f' \le 0$.

Let $w(\cdot)$ be defined as in Section 5.3.3, that is, the smallest function satisfying $w(r) \ge \sup_{f\in\mathcal{F}, Pf\le r}\sqrt{\mathsf{Var}[f]}$ such that $w(r)/\sqrt{r}$ is non-increasing. Let again ϵ^* be defined as the positive solution of $r = \psi(w(r))$.

Theorem 5.8. *For any $\delta > 0$, let $r^*(\delta)$ denote the solution of*

$$r = 4\psi(w(r)) + 2w(r)\sqrt{\frac{2\log\frac{2}{\delta}}{n}} + \frac{16\log\frac{2}{\delta}}{3n}$$

and ε^* the positive solution of equation $r = \psi(w(r))$. Then for any $\theta > 0$, with probability at least $1-\delta$, the empirical risk minimizer g_n satisfies

$$L(g_n) - L(g') \le \theta\left(L(g') - L(g^*)\right) + \frac{(1+\theta)^2}{4\theta}r^*(\delta)\,,$$

and

$$L(g_n) - L(g') \le \theta\,(L(g') - L(g^*)) + \frac{(1+\theta)^2}{4\theta}\left(32\varepsilon^* + \left(4\frac{w^2(\varepsilon^*)}{\varepsilon^*} + \frac{32}{3}\right)\frac{\log\frac{2}{\delta}}{n}\right).$$

Remark 5.9. When $g' = g^*$, the bound in this theorem has the same form as the upper bound in (22).

Remark 5.10. The second bound in the Theorem follows from the first one in the same way as Inequality (22) follows from Inequality (21). In the proof, we focus on the first bound.

The proof consists mostly of replacing the observation that $L_n(g_n) \le L_n(g^*)$ in the proof of Theorem 5.5 by $L_n(g_n) \le L_n(g')$.

PROOF. Let r denote a positive real. Using the same approach as in the proof of Theorem 5.5, that is, by applying Talagrand's inequality to the reweighted star-hull $\mathcal{F}^*$, we get that with probability larger than $1-\delta$, for all $f \in \mathcal{F}$ such that $Pf \ge r$,

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

while we may also apply Bernstein's inequality to $-f'$ and use the fact that $\sqrt{\mathsf{Var}(f')} \le w(Pf)$ for all $f \in \mathcal{F}$:

$$P_n f' - Pf' \le \sqrt{\mathsf{Var}(f')}\sqrt{\frac{2\log\frac{2}{\delta}}{n}} + \frac{8\log\frac{2}{\delta}}{3n} \le (w(Pf) \vee r)\sqrt{\frac{2\log\frac{2}{\delta}}{n}} + \frac{8\log\frac{2}{\delta}}{3n}.$$

Adding the two inequalities, we get that, with probability larger than $1-\delta$, for all $f \in \mathcal{F}$

$$(Pf - Pf') + (P_n f' - P_n f) \le \frac{w(Pf) \vee r}{r}\left(4\psi(r) + 2r\sqrt{\frac{2\log\frac{2}{\delta}}{n}} + \frac{16\log\frac{2}{\delta}}{3n}\right).$$

If we focus on $f = f_n$, then the two terms in the left-hand-side are positive. Now we substitute $w(r^*(\delta))$ for r in the inequalities. Hence, using arguments that parallel the derivation of (25) we get that, on an event that has probability larger than $1-\delta$, we have either $Pf_n \le r^*(\delta)$ or at least

$$Pf_n - Pf' \le \frac{\sqrt{Pf_n}}{\sqrt{r^*(\delta)}}\left(4\psi(w(r^*(\delta))) + 2w(r^*(\delta))\sqrt{\frac{2\log\frac{2}{\delta}}{n}} + \frac{16\log\frac{2}{\delta}}{3n}\right) = \sqrt{Pf_n}\sqrt{r^*(\delta)}\,.$$

Standard computations lead to the first bound in the Theorem. □

Remark 5.11. The bound of Theorem 5.8 helps identify situations where taking into account noise conditions improves on naive risk bounds. This is the case when the approximation bias is of the same order of magnitude as the estimation bias. Such a situation occurs when dealing with a plurality of models, see Section 8.

Remark 5.12. The bias term $L(g') - L(g^*)$ shows up in Theorem 5.8 because we do not want to assume any special relationship between $\mathsf{Var}[\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g'(X)\neq Y}]$ and $L(g) - L(g')$. Such a relationship may exists when dealing with convex risks and convex models. In such a case, it is usually wise to take advantage of it.

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 arguments 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 $A(f) - A^*$. According to [27] Zhang's lemma (11) may be improved under the Mammen-Tsybakov noise conditions to yield

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

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 Pm_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 $1 - \delta$,

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

where the constant C depends on the cost functional and the VC dimension V of the base class $\mathcal{C}$. Combining this with the above improvement of Zhang's lemma, one obtains significant improvements of the performance bound of Theorem 4.4.

5.5. **Minimax lower bounds**

The purpose of this section is to investigate the accuracy of the bounds obtained in the previous sections. 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 $\mathcal{C}$ 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 $\mathcal{C}$, 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.

To make the exposition simpler, we only consider classes of functions 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, $L_{\mathcal{C}} = \inf_{g\in\mathcal{C}} L(g)$. In particular, we seek lower bounds for

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

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 $\mathcal{C}$ 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 $L_{\mathcal{C}}$ kept at a fixed value between zero and 1/2. We will see that the bounds depend on n, V the VC dimension of

$\mathcal{C}$, and $L_{\mathcal{C}}$ jointly. As it turns out, the situations for $L_{\mathcal{C}} > 0$ and $L_{\mathcal{C}} = 0$ are quite different. Also, the fact that the noise is controlled (with the Mammen-Tsybakov noise conditions) has an important influence.

Integrating the deviation inequalities such as Corollary 5.3, we have that for any class $\mathcal{C}$ of classifiers with VC dimension V, a classifier g_n minimizing the empirical risk satisfies

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

and also

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

Let $\mathcal{C}$ 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 $L_{\mathcal{C}} = 0$. Then, for every classification rule g_n based upon $X_1, Y_1, \ldots, X_n, Y_n$, and $n \ge V - 1$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}L(g_n) \ge \frac{V-1}{2en}\left(1 - \frac{1}{n}\right) . \tag{26}$$

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

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

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

$$\sup_{P \in \mathcal{P}} \mathbb{E}(L(g_n) - L) \ge \sqrt{\frac{L(V-1)}{32n}} \quad \text{if} \quad n \ge \frac{V-1}{8} \max\left(\frac{2}{(1-2L)^2}, \frac{1}{L}\right) . \tag{27}$$

In the extreme case of the Mammen-Tsybakov noise condition, that is, when $\sup_x |2\eta(x) - 1| \ge h$ for some positive h, we have seen that the rate can be improved and that we essentially have, when g_n is the empirical error minimizer,

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

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

Let $\mathcal{C}$ 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

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

and assume that $|\eta(X) - 1/2| \ge h$ almost surely where $s > 0$ is a constant. Then, for every classification rule g_n based upon $X_1, Y_1, \ldots, X_n, Y_n$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}(L(g_n) - L^*) \ge C\left(\sqrt{\frac{V}{n}} \wedge \frac{V}{nh}\right) . \tag{28}$$

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 $\mathcal{C}$ is such a class, under the above conditions, one can improve (28) to get

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

Bibliographical remarks. Inequality (12) is known as Hoeffding's inequality [109], while (13) is referred to as Bernstein's inequality [34]. The constants shown here in Bernstein's inequality actually follow from an inequality due to Bennett [33]. Theorem 5.1 and their corollaries (16), and Corollary 5.3 are due to Vapnik and Chervonenkis [232, 233]. The proof sketched here is due to Anthony and Shawe-Taylor [11]. Regarding the corollaries of this result, (14) is due to Pollard [181] and (15) is due to Haussler [105]. Breiman, Friedman, Olshen, and Stone [53] also derive inequalities similar, in spirit, to (14).

The fact that the variance can be related to the expectation and that this can be used to get improved rates has been known for a while the context of regression function estimation or other statistical problems (see [110], [226], [227] and references therein). For example, asymptotic results based on this were obtained by van de Geer [224]. For regression, Birgé and Massart [36] and Lee, Bartlett and Williamson [134] proved exponential inequalities The fact that this phenomenon also occurs in the context of discriminant analysis and classification, under conditions on the noise (sometimes called margin) has been pointed out by Mammen and Tsybakov [151], (see also Polonik [182] and Tsybakov [220] for similar elaborations on related problems like excess-mass maximization or density level-sets estimation). Massart [160] showed how to use optimal noise conditions to improve model selection by penalization.

Talagrand's inequality for empirical processes first appeared in [212]. For various improvements, see Ledoux [132], Massart [159], Rio [184]. The version presented in Theorem 5.4 is an application of the refinement given by Bousquet [47]. Variations on the theme and detailed proofs appeared in [48].

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, which goes back to Huber [110], is used, for example, by van de Geer [224–226]. Another approach consists in weighting the class and was used by Vapnik and Chervonenkis [232] in the special case of binary valued functions and extended by Pollard [181], for example. Combining this approach with concentration inequalities was proposed by Massart [160] 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 [227]. More recently non-asymptotic deviation inequalities involving this quantity were obtained, essentially in the work of Massart [160] and Koltchinskii and Panchenko [126]. Both approaches use a version of the peeling technique, but the one of Massart uses in addition a weighting approach. More recently, Mendelson [171] 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 [25] where the peeling and star-hull approach are compared.

It is pointed out in recent results of Bartlett, Mendelson, and Philips [30] and Koltchinskii [125] that sharper and simpler bounds may be obtained by taking Rademacher averages over level sets of the excess risk rather than on $L_2(P)$ balls.

Empirical estimates of the fixed point of type ε^* were studied by Koltchinskii and Panchenko [126] in the zero error case. In a related work, Lugosi and Wegkamp [147] obtain bounds in terms of empirically estimated localized Rademacher complexities without noise conditions. In their approach, the complexity of a subclass of $\mathcal{C}$ containing only classifiers with a small empirical risk is used to obtain sharper bounds. A general result, applicable under general noise conditions, was proven by Bartlett, Bousquet and Mendelson [25].

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 [30]. Applications of results like Theorem 5.5 to classification with VC classes of functions were investigated by Massart and Nédélec [162].

Properties of convex loss functions were investigated by Lin [139], Steinwart [206], and Zhang [244]. The improvement of Zhang's lemma under the Mammen-Tsybakov noise condition is due to Bartlett, Jordan and McAuliffe [27] who establish more general results. For a further improvement we refer to Blanchard, Lugosi, and Vayatis [40]. The cited improved rates of convergence for $A(f_n) - A^*$ is also taken from [27] and [40] which is based on bounds derived by Blanchard, Bousquet, and Massart [39]. The latter reference also investigates the special cost function $(1+x)_+$ under the extreme case $\alpha = 1$ of the Mammen-Tsybakov noise condition, see also Bartlett, Jordan and McAuliffe [27], Steinwart [195].

Massart [160] gives a version of Theorems 5.5 and 5.8 for the case $w(r) = c\sqrt{r}$ and arbitrary bounded loss functions which is extended for general w in Bartlett, Jordan and McAuliffe [27] and Massart and Nédélec [162]. Bartlett, Bousquet and Mendelson [25] give an empirical version of Theorem 5.5 in the case $w(r) = c\sqrt{r}$.

The lower bound (26) was proved by Vapnik and Chervonenkis [233], see also Haussler, Littlestone, and Warmuth [107], Blumer, Ehrenfeucht, Haussler, and Warmuth [41]. Inequality (27) is due to Audibert [17] who improves on a result of Devroye and Lugosi [73], see also Simon [200] for related results. The lower bounds under conditions on the noise are due to Massart and Nédélec [162]. Related results under the Mammen-Tsybakov noise condition for large classes of functions (i.e., with polynomial growth of entropy) are given in the work of Mammen and Tsybakov [151] and Tsybakov [221]. Other minimax results based on growth rate of entropy numbers of the class of function are obtained in the context of classification by Yang [239, 240]. 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 does not depend on the sample size. Such results are called strong minimax lower bounds and were investigated by Antos and Lugosi [14] and Schuurmans [193].

6. PAC-bayesian bounds

We now describe the so-called PAC-bayesian approach to derive error bounds. (PAC is an acronym for "probably approximately correct.") The distinctive feature of this approach is that one assumes that the class $\mathcal{C}$ 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 $\mathcal{C}$ (called the posterior). Throughout this section we assume that the class $\mathcal{C}$ is at most countably infinite.

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 equals 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 $\mathcal{F} = \{\mathbb{1}_{g(x) \neq y} : g \in \mathcal{C}\}$, one easily gets that for each fixed $f \in \mathcal{F}$,

$$\mathbb{P}\left\{ Pf - P_n f \geq \sqrt{\frac{\log(1/\delta)}{2n}} \right\} \leq \delta\,.$$

For any positive weights $\pi(f)$ with $\sum_{f \in \mathcal{F}} \pi(f) = 1$, one may write a *weighted union bound* as follows

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

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

$$\forall f \in \mathcal{F},\ Pf - P_n f \leq \sqrt{\frac{\log(1/\pi(f)) + \log(1/\delta)}{2n}}\,. \tag{29}$$

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 for any functional I, $(\exists f \in \mathcal{F},\, I(f) \geq 0) \Leftrightarrow (\exists \rho,\, \rho I(f) \geq 0)$ where ρ denotes an arbitrary probability measure on $\mathcal{F}$ so that we can take the expectation of (29) with respect to ρ and use Jensen's inequality. This gives, with probability at least $1-\delta$,

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

where $K(\rho,\pi)$ denotes the Kullback-Leibler divergence between ρ and π and $H(\rho)$ is the entropy of ρ. Rewriting this in terms of the class $\mathcal{C}$, we get that, with probability at least $1-\delta$,

$$\forall \rho,\; \rho L(g) - \rho L_n(g) \leq \sqrt{\frac{K(\rho,\pi) + H(\rho) + \log(1/\delta)}{2n}}\,. \tag{30}$$

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 $X_f = (Pf - P_n f)_+^2$ and this means that we have to upper bound $\pi e^{\lambda (Pf - P_n f)_+^2}$. 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 $f \in \mathcal{F}$,

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

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 $1-\delta$,

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

The resulting bound has the following form.

Theorem 6.1. PAC-BAYESIAN BOUND. *With probability at least* $1-\delta$,

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

This should be compared to (30). 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 Theorem 6.1 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 of Theorem 6.1 was derived by McAllester [163] and later extended in [164, 165]. Langford and Seeger [130] and Seeger [196] gave an easier proof and some refinements. The symmetrization and conditioning approach was first suggested by Catoni and studied in [57–59]. The chaining idea appears in the work of Kolmogorov [122, 123] and was further developed by Dudley [81] and Pollard [180]. It was generalized by Talagrand [215] and a detailed account of recent developments is given in [219]. The chaining approach to PAC-bayesian bounds appears in Audibert and Bousquet [16]. Audibert [17] offers a thorough study of PAC-bayesian results.

7. Stability

Given a classifier g_n, one of the fundamental problems is to obtain estimates for the magnitude of the difference $L(g_n) - L_n(g_n)$ between the true risk of the classifier and its estimate $L_n(g_n)$, measured on the same data on which the classifier was trained. $L_n(g_n)$ is often called the *resubstitution estimate* of $L(g_n)$.

It has been pointed out by various authors that the size of the difference $L(g_n) - L_n(g_n)$ is closely related to the "stability" of the classifier g_n. Several notions of stability have been introduced, aiming at capturing this idea. Roughly speaking, a classifier g_n is "stable" if small perturbations in the data do not have a big effect on the classifier. Under a proper notion of stability, concentration inequalities may be used to obtain estimates for the quantity of interest.

A simple example of such an approach is the following. 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 data $(X_1, Y_1), \ldots, (X_n, Y_n)$, 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 differences inequality, one easily obtains that with probability at least $1-\delta$,

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

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 Rogers and Wagner [187], Devroye and Wagner [74, 75]. Kearns and Ron [116] investigated it further and introduced formally several measures of stability. Bousquet and Elisseeff [49] obtained exponential bounds under restrictive conditions on the algorithm, using the notion of *uniform stability*. These conditions were relaxed by Kutin and Niyogi [129]. The link between stability and consistency of the empirical error minimizer was studied by Poggio, Rifkin, Mukherjee and Niyogi [179].

8. Model selection

8.1. Oracle inequalities

When facing a concrete classification problem, choosing the right set $\mathcal{C}$ of possible classifiers is a key to success. If $\mathcal{C}$ is so large that it can approximate arbitrarily well any measurable classifier, then $\mathcal{C}$ is susceptible to overfitting and is not suitable for empirical risk minimization, or empirical ϕ-risk minimization. On the other hand, if $\mathcal{C}$ is a small class, for example a class with finite VC dimension, $\mathcal{C}$ will be unable to approximate in any reasonable sense a large set of measurable classification rules.

In order to achieve a good balance between estimation error and approximation error, a variety of techniques have been considered. In the remainder of the paper, we will focus on the analysis of model selection methods which could be regarded as heirs of the structural risk minimization principle of Vapnik and Chervonenkis.

Model selection aims at getting the best of different worlds simultaneously. Consider a possibly infinite collection of classes of classifiers $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k, \ldots$ Each class is called a model. Our guess is that some of these models contain reasonably good classifiers for the pattern recognition problem we are facing. Assume that for each of these models, we have a learning algorithm that picks a classification rule $g^*_{n,k}$ from $\mathcal{C}_k$ when given the sample D_n. The model selection problem may then be stated as follows: select among $(g^*_{n,k})_k$ a "good" classifier.

Notice here that the word "selection" may be too restrictive. Rather that selecting some special $g^*_{n,k}$, we may consider combining them using a voting scheme and use a boosting algorithm where the base class would just be the (data-dependent) collection $(g^*_{n,k})_k$. For the sake of brevity, we will just focus on model selection in the narrow sense.

In an ideal world, before we see the data D_n, a benevolent oracle with the full knowledge of the noise conditions and of the Bayes classifier would tell us which model (say $\tilde{k}$) minimizes the expected excess risk $\mathbb{E}[L(g^*_{n,k}) - L^*]$, if such a model exists in our collection. Then we could use our learning rule for this most promising model with the guarantee that

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

But as the most promising model $\tilde{k}$ depends on the learning problem and may even not exist, there is no hope to perfectly mimic the behavior of the benevolent and powerful oracle. What statistical learning theory has tried hard to do is to approximate the benevolent oracle in various ways.

It is important to think about what could be reasonable upper bounds on the right-hand side of the preceding oracle inequality. It seems reasonable to incorporate a factor C at least as large as 1 and additive terms of the form $C'\gamma(k,n)/n$ where $\gamma(\cdot)$ is a slowly growing function of its arguments and to ask for

$$\mathbb{E}\left[L(g^*_{n,\hat{k}}) - L^*\right] \leq C \inf_k \left(\mathbb{E}\left[L(g^*_{n,k}) - L^*\right] + C' \frac{\gamma(k,n)}{n}\right), \tag{31}$$

where $\hat{k}$ is the index of the model selected according to empirical evidence.

Let $L^*_k = \inf_{g \in \mathcal{C}_k} L(g)$ for each model index k. In order to understand the role of $\gamma(\cdot)$, it is useful to split $\mathbb{E}[L(g^*_{n,k}) - L^*]$ in a bias term $L^*_k - L^*$ and a "variance" term $\mathbb{E}[L(g^*_{n,k}) - L^*_k]$. The last inequality translates into

$$\mathbb{E}\left[L(g^*_{n,\hat{k}}) - L^*\right] \leq C \inf_k \left(L^*_k - L^* + \mathbb{E}\left[L(g^*_{n,k}) - L^*_k\right] + C' \frac{\gamma(k,n)}{n}\right).$$

The term $C'\frac{\gamma(k,n)}{n}$ should ideally be at most of the same order of magnitude as $L(g^*_{n,k}) - L^*_k$.

To make the roadmap more detailed, we may invoke the robust analysis of the performance of empirical risk minimization sketched in Section 5.3.5. Recall that $w(\cdot)$ was defined in such a way that $\sqrt{\mathsf{Var}(\mathbb{1}_{g \neq g^*})} \leq w\left(L(g) - L^*\right)$ for all classifiers $g \in \cup_k \mathcal{C}_k$ and such that $w(r)/\sqrt{r}$ is non-increasing. Explicit constructions of $w(\cdot)$ were possible under the Mammen-Tsybakov noise conditions.

To take into account the plurality and the richness of models, for each model $\mathcal{C}_k$, let ψ_k be defined as

$$\psi_k(r) = \mathbb{E}R_n \left\{ f \in \mathcal{F}^*_k : \sqrt{\mathrm{Var}(f)} \leq r \right\}$$

where $\mathcal{F}^*_k$ is the star-hull of the loss class defined by $\mathcal{C}_k$ (see Section 5.3.5). For each k, let ϵ^*_k be defined as the positive solution of $r = \psi_k(w(r))$. Then, viewing Theorem 5.8, we can get sensible upper bounds on the excess risk for each model and we may look for oracle inequalities of the form

$$\mathbb{E}\left[L(g^*_{n,\hat{k}}) - L^*\right] \leq C \inf_k \left(L^*_k - L^* + C'\left(\epsilon^*_k + \frac{w^2(\epsilon^*_k)}{n\,\epsilon^*_k}\frac{\log k}{n}\right)\right). \tag{32}$$

The right-hand side is then of the same order of magnitude as the infimum of the upper bounds on the excess risk described in Section 5.3.

8.2. A glimpse at model selection methods

As we now have a clear picture of what we are after, we may look for methods suitable to achieve this goal. The model selection problem looks like a multiple hypotheses testing problem: we have to test many pairs of hypotheses where the null hypothesis is $L(g^*_{n,k}) \leq L(g^*_{n,k'})$ against the alternative $L(g^*_{n,k}) > L(g^*_{n,k'})$. Depending on the scenario, we may or may not have fresh data to test these pairs of hypotheses. Whatever the situation, the tests are not independent. Furthermore there does not seem to be any obvious way to combine possibly conflicting answers.

Most data-intensive model selection methods we are aware of can be described in the following way: for each pair of models $\mathcal{C}_k$ and $\mathcal{C}_{k'}$, a threshold $\tau(k,k',D_n)$ is built and model $\mathcal{C}_k$ is favored with respect to model $\mathcal{C}'_k$ if

$$L_n(g^*_{n,k}) - L_n(g^*_{n,k'}) \leq \tau(k,k',D_n).$$

The threshold $\tau(\cdot,\cdot,\cdot)$ may or may not depend on the data. Then the results of the many pairwise tests are combined in order to select a model.

Model selection by penalization may be regarded as a simple instance of this scheme. In the penalization setting, the threshold $\tau(k,k',D_n)$ is the difference between two terms that depend on the models:

$$\tau(k,k',D_n) = \mathrm{pen}(n,k') - \mathrm{pen}(n,k).$$

The selected index $\hat{k}$ minimizes the penalized empirical risk

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

Such a scheme is attractive since the combination of the results of the pairwise tests is extremely simple. As a matter of fact, it is not necessary to perform all pairwise tests, it is enough to find the index that minimizes the penalized empirical risk. Nevertheless, performing model selection using penalization suffers from some drawbacks: it will become apparent below that the ideal penalty that should be used in order to mimic the benevolent oracle, should, with high probability, be of the order of

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

As seen in Section 5.3.5, the sharpest bounds we can get on the last quantity depend on noise conditions, model complexity and on the model approximation capability $L^*_k - L^*$. Although noise conditions and model complexity can be estimated from the data (notwithstanding computational problems), estimating the model bias $L^*_k - L^*$ seems to be beyond the reach of our understanding. In fact, estimating L^* is known to be a difficult statistical problem, see Devroye, Györfi, and Lugosi [72], Antos, Devroye, and Györfi [12].

As far as classification is concerned, model selection by penalization may not put the burden where it should be. If we allow the combination of the results of pairwise tests to be somewhat more complicated than a simple search for the minimum in a list, we may avoid the penalty calibration bottleneck. In this respect, the so-called pre-testing method has proved to be quite successful when models are nested. The cornerstone of the pre-testing methods consists of the definition of the threshold $\tau(k, k', D_n)$ for $k \le k'$ that takes into account the complexity of $\mathcal{C}_k$, as well as the noise conditions. Instead of attempting an unbiased estimation of the excess risk in each model as the penalization approach, the pre-testing approach attempts to estimate differences between excess risks.

But whatever promising the pre-testing method may look like, it will be hard to convince practitioners to abandon cross-validation and other resampling methods. Indeed, a straightforward analysis of the hold-out approach to model selection suggests that hold-out enjoys almost all the desirable features of any foreseeable model selection method.

The rest of this section is organized as follows. In Subsection 8.3 we illustrate how the results collected in Sections 3 and 4 can be used to design simple penalties and derive some easy oracle inequalities that capture classical results concerning structural risk minimization. It will be obvious that these oracle inequalities are far from being satisfactory. In Subsection 8.4, we point out the problems that have to be faced in order to calibrate penalties using the refined and robust analysis of empirical risk minimization given in Section 5.3.5. In Subsection 8.6, we rely on these developments to illustrate the possibilities of pre-testing methods and we conclude in Subsection 8.7 by showing how hold-out can be analyzed and justified by resorting to a robust version of the elementary argument given at the beginning of Subsection 5.2.

Bibliographical remarks. Early work on model selection in the context of regression or prediction with squared loss can be found in Mallows [150], Akaike [6]. Mallows introduced the C_p criterion in [150]. Grenander [102] discusses the use of regularization in statistical inference. Vapnik and Chervonenkis [233] proposed the structural risk minimization approach to model selection in classification, see also Vapnik [229–231], Lugosi and Zeger [148].

The concept of oracle inequality was advocated by Donoho and Johnstone [76]. A thorough account of the concept of oracle inequality can be found in Johnstone [113].

Barron [21], Barron and Cover [23], [22] investigate model selection using complexity regularization which is a kind of penalization in the framework of discrete models for density estimation and regression. A general and influential approach to non-parametric inference through penalty-based model selection is described in Barron, Birgé and Massart [20], see also Birgé and Massart [37], [38]. 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 Baraud [19], Castellan [56] for illustrations in regression and density estimation.

A recent account of inference methods in non-parametric settings can be found in Tsybakov [222].

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 [72] for exposition and references.

Hall [103] and many other authors use resampling techniques to perform model selection.

8.3. Naive 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 $(\mathcal{C}_k)_{k\in\mathbb{N}}$. In other words, the selected classifier is $g^*_{n,\hat{k}}$. As in the preceding 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.

The naive penalties considered in this section are estimates of the expected amount of overfitting $\mathbb{E}[\sup_{g\in\mathcal{C}_k} L(g) - L_n(g)]$. Taking the expectation as a penalty 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 may lead to highly conservative bounds. The reason is that since a distribution-free upper bound holds for all distributions, it is necessarily loose in special cases when the distribution is such that the expected maximal deviation is small. This may occur, for example, if the distribution of the data is concentrated on a small-dimensional manifold and in many other cases. In recent years, several data-driven penalization procedures have been proposed. Such procedures are motivated according to computational or to statistical considerations. 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.

Theorem 8.1. *For each k, let $\mathcal{F}_k = \{\mathbb{1}_{g(x)\neq y} : g \in \mathcal{C}_k\}$ denote the loss class associated with $\mathcal{C}_k$. Let $\mathrm{pen}(n,k)$ be defined by*

$$\mathrm{pen}(n,k) = 3R_n(\mathcal{F}_k) + \sqrt{\frac{\log k}{n}} + \frac{18\log k}{n}\,. \tag{33}$$

Let $\hat{k}$ be defined as $\arg\min\Big(L_n(g^*_{n,k}) + \mathrm{pen}(n,k)\Big)$. Then

$$\mathbb{E}\left[L(g^*_{n,\hat{k}}) - L^*\right] \le \inf_k \left(L(g^*_k) - L^* + 3\mathbb{E}\left[R_n(\mathcal{F}_k)\right] + \sqrt{\frac{\log k}{n}} + \frac{18\log k}{n}\right) + \sqrt{\frac{2\pi}{n}} + \frac{18}{n}\,. \tag{34}$$

Inequality (34) has the same form as the generic oracle inequality (31). The multiplicative constant in front of the infimum is optimal since it is equal to 1. At first glance, the additive term might seem quite satisfactory: if noise conditions are not favorable, $\mathbb{E}\left[R_n(\mathcal{F}_k)\right]$ is of the order of the excess risk in the k-th model. On the other hand, in view of the oracle inequality (32) we are looking for, this inequality is loose when noise conditions are favorable, for example, when the Mammen-Tsybakov conditions are enforced with some exponent $\alpha > 0$.

In the sequel, we will sometimes uses the following property. Rademacher averages are sharply concentrated: they not only satisfy the bounded differences inequality, but also "Bernstein-like" inequalities, given in the next lemma.

Lemma 8.2. *Let $\mathcal{F}_k$ denote a class of functions with values in $[-1,1]$, and $R_n(\mathcal{F}_k)$ the corresponding conditional Rademacher averages. Then*

$$\begin{aligned}
\mathrm{Var}\left(R_n(\mathcal{F}_k)\right) &\leq \frac{1}{n}\mathbb{E}\left[R_n(\mathcal{F}_k)\right] \\
\mathbb{P}\left\{R_n(\mathcal{F}_k) \geq \mathbb{E}\left[R_n(\mathcal{F}_k)\right] + \epsilon\right\} &\leq \exp\left(-\frac{n\epsilon^2}{2(\mathbb{E}[R_n(\mathcal{F}_k)] + \varepsilon/3)}\right) \\
\mathbb{P}\left\{R_n(\mathcal{F}_k) \leq \mathbb{E}\left[R_n(\mathcal{F}_k)\right] - \epsilon\right\} &\leq \exp\left(-\frac{n\epsilon^2}{2\mathbb{E}[R_n(\mathcal{F}_k)]}\right).
\end{aligned}$$

PROOF OF THEOREM 8.1. By the definition of the selection criterion, we have for all k,

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

Taking expectations, we get

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

The tail bounds for Rademacher averages given in Lemma 8.2 can then be exploited as follows:

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{g\in\mathcal{C}_k}\left(L(g) - L_n(g)\right) \geq \mathrm{pen}(n,k) + 2\delta\right\} \\
&\leq \mathbb{P}\left\{\sup_{g\in\mathcal{C}_k}\left(L(g) - L_n(g)\right) \geq \mathbb{E}\left[\sup_{g\in\mathcal{C}_k}\left(L(g) - L_n(g)\right)\right] + \sqrt{\frac{\log k}{n}} + \delta\right\} \\
&\quad + \mathbb{P}\left\{R_n(\mathcal{F}_k) \leq \frac{2}{3}\mathbb{E}\left[R_n(\mathcal{F}_k)\right] - \frac{18\log k}{3n} - \frac{\delta}{3}\right\} \\
&\quad \text{(using the bounded differences inequality for the first term and Lemma 8.2 for the second term)} \\
&\leq \frac{1}{k^2}\exp(-2n\delta^2) + \frac{1}{k^2}\exp\left(-\frac{n\delta}{9}\right).
\end{aligned}$$

Integrating by parts and summing with respect to k leads to the oracle inequality of the theorem. □

Bibliographical remarks. Data-dependent penalties were suggested by Lugosi and Nobel [145], and in the closely related "luckiness" framework introduced by Shawe-Taylor, Bartlett, Williamson, and Anthony [197], see also Freund [92].

Penalization based on Rademacher averages was suggested by Bartlett, Boucheron, and Lugosi [24] and Koltchinskii [124]. For refinements and further development, see Koltchinskii and Panchenko [126], Lozano [142], [29], Bartlett, Bousquet and Mendelson [25], Bousquet, Koltchinskii and Panchenko [50]. Lugosi and

Wegkamp [147], Herbrich and Williamson [108], Mendelson and Philips [172]. The proof that Rademacher averages, empirical VC-entropy and empirical VC-dimension are sharply concentrated around their mean can be found in Boucheron, Lugosi, and Massart [45, 46].

Fromont [96] points out 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 [100] and Efron et al. [85–87] for general results on the bootstrap.

Empirical investigations on the performance of model selection based on Rademacher penalties can be found in Lozano [142] and Bartlett, Boucheron, and Lugosi [24]. Both papers build on a framework elaborated in Kearns, Mansour, Ng, and Ron [115]. Indeed, [115] is an early attempt to compare model selection criteria originating in structural risk minimization theory, MDL (Minimum Description Length principle), 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.

Lugosi and Wegkamp [147] 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 [126] when the optimal risk equals zero.

8.4. Ideal penalties

Naive penalties that tend to overestimate the excess risk in each model lead to conservative model selection strategies. For moderate sample sizes, they tend to favor small models. Encouraging results reported in simulation studies should not mislead the readers. Model selection based on naive Rademacher penalization manages to mimic the oracle when sample size is large enough to make the naive upper bound on the estimation bias small with respect to the approximation bias.

As model selection is ideally geared toward situations where sample size is not too large, one cannot feel satisfied by naive Rademacher penalties. We can guess quite easily what good penalties should be like. If we could build penalties in such a way that, with probability larger than $1 - \frac{1}{2n\,k^2}$,

$$L(g^*_{n,k}) - L^* \leq C' \left((L_n(g^*_{n,k}) - L_n(g^*) + \mathrm{pen}(n,k) \right) + C'' \frac{\log(2n\,k^2)}{n} ,$$

then by the definition of model selection by penalization, and a simple union bound, with probability larger than $1 - 1/n$, for any k, we would have

$$\begin{aligned} L(g^*_{n,\hat{k}}) - L^* &\leq L(g^*_{n,\hat{k}}) - L^* + C'(L_n(g^*_{n,k}) + \mathrm{pen}(n,k)) - C'(L_n(g^*_{n,\hat{k}}) + \mathrm{pen}(n,\hat{k})) \\ &\leq C' \left(L_n(g^*_{n,k}) - L_n(g^*) + \mathrm{pen}(n,k) \right) + C'' \frac{\log{(2n\,\hat{k}^2)}}{n} \\ &\leq C' \left(L_n(g^*_k) - L_n(g^*) + \mathrm{pen}(n,k) \right) + C'' \frac{\log(2n\,\hat{k}^2)}{n} . \end{aligned}$$

Assuming we only consider polynomially many models (as a function of n), this would lead to

$$\mathbb{E}L(g^*_{n,\hat{k}}) - L^* \leq \inf_k \left\{ C' \left[\mathbb{E}[L^*_k - L^* + \mathrm{pen}(n,k)] \right] + \frac{C'' \log(2en)}{n} \right\} + \frac{1}{n} .$$

Is this sufficient to meet the objectives set in Section 8.1?

This is where the robust analysis of empirical risk minimization (Section 5.3.5) comes into play. If we assume that, with high probability, the quantities ϵ^*_k defined at the end of Section 8.1 can be tightly estimated by data-dependent quantities and used as penalties, then we are almost done.

The following statement, that we abusively call a theorem, summarizes what could be achieved using such "ideal penalties." For the sake of brevity, we provide a theorem with $C' = 2$, but some more care allows one to develop oracle inequalities with arbitrary $C' > 1$.

Theorem 8.3. *If for every k*

$$\mathrm{pen}(n,k) \geq 32\varepsilon_k^* + \left(4\frac{(w(\varepsilon_k^*))^2}{\varepsilon_k^*} + \frac{32}{3}\right)\frac{\log(4nk^2)}{n}\,, \tag{35}$$

then

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

Most of the proof consists in checking that if $\mathrm{pen}(n,k)$ is chosen according to (35) then we can invoke the robust results on learning rates stated in Theorem 5.8 to conclude.

PROOF. Following the second bound in Theorem 5.8 (with $\theta = 1$), with probability at least $1 - \sum_k \frac{1}{2n\,k^2} \geq 1 - \frac{1}{n}$, we have, for every k,

$$L(g^*_{n,k}) - L(g^*) \leq 2\left(L_n(g^*_{n,k}) - L_n(g^*)\right) + \left(32\varepsilon_k^* + \left(4\frac{(w(\varepsilon_k^*))^2}{\varepsilon_k^*} + \frac{32}{3}\right)\frac{\log(4nk^2)}{n}\right)\,,$$

that is,

$$L(g^*_{n,k}) - L(g^*) \leq 2\left(L_n(g^*_{n,k}) - L_n(g^*) + \mathrm{pen}(n,k)\right)\,.$$

The Theorem follows by observing that $L(g^*_{n,\hat{k}}) - L^* \leq 1$. □

If $\mathrm{pen}(n,k)$ is about the same as the right-hand side of (35), then the oracle inequality of Theorem 8.3 has the same form as the ideal oracle inequality described at the end of Section 8.1. This should nevertheless not be considered as a definitive result but rather as an incentive to look for better penalties. It could also possibly point toward a dead end. Theorem 8.3 actually calls for building estimators of the sequence (ϵ_k^*), that is, of the sequence of fixed points of functions $\psi_k \circ w$. Recall that

$$\psi_k(w(r)) \approx \mathbb{E}\left[R_n\left\{f : f \in \mathcal{F}_k^*, Pf \leq r\right\}\right].$$

If the star-shaped loss class $\{f : f \in \mathcal{F}_k^*, Pf \leq r\}$ were known, given the fact that for a fixed class of functions $\mathcal{F}$, $R_n(\mathcal{F})$ is sharply concentrated around its expectation, estimating $\psi_k \circ w$ would be statistically feasible. But the loss class of interest depends not only on the k-th model $\mathcal{C}_k$, but also on the unknown Bayes classifier g^*. We will not pursue the search for ideal data-dependent penalties and look for roundabouts. In the next section, we will see that when $g^* \in \mathcal{C}_k$, even though g^* is unknown, sensible estimates of ε_k^* can be constructed. In Section 8.6, we will see how to use these estimates in model selection.

Bibliographical remarks. The results described in this section are inspired by Massart [160] where the concept of ideal penalty in classification is clarified. The notion that ideal penalties should be rooted in sharp risk estimates goes back to the pioneering works of Akaike [6] and Mallows [150]. As far as classification is concerned, a detailed account of these ideas can be found in the eighth chapter of Massart [161]. Various approaches to the excess risk estimation in classification can be found in Bartlett, Bousquet, and Mendelson [25] and Koltchinskii [125], where a discussion of the limits of penalization can also be found.

8.5. **Localized Rademacher complexities**

The purpose of this section is to show how the distribution-dependent upper bounds on the excess risk of empirical risk minimization derived in Section 5.3 can be estimated from above and from below when the Bayes classifier belongs to the model. This is not enough to make penalization work but it will prove convenient when investigating a pre-testing method in the next section.

In this section, we are concerned with a single model $\mathcal{C}$ which contains the Bayes classifier g^*. The minimizer of the empirical risk will is denoted by g_n. The loss class is $\mathcal{F} = \{\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y} \ : \ g \in \mathcal{C}\}$. The functions $\psi(\cdot)$ and $w(\cdot)$ are defined as in Section 5.3. The quantity ε^* is defined as the solution of the fixed-point equation $r = \psi(w(r))$. As, thanks to Theorems 5.5 and 5.8, ε^* contains relevant information on the excess risk, it may be tempting to try to estimate ε^* from the data. However this is not the easiest way to proceed. As we will need bounds with prescribed accuracy and confidence for the excess risk, we will rather try to estimate from above and below the bound $r^*(\delta)$ defined in the statement of Theorem 5.5. Recall that $r^*(\delta)$ is defined as the solution of the equation

$$r = 4\psi(w(r)) + w(r)\sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n}\,.$$

In order to estimate $r^*(\delta)$, we estimate $\psi(\cdot)$ and $w(\cdot)$ by some functions $\hat{\psi}$ and $\hat{w}$ and solve the corresponding fixed-point equations. The rationale for this is contained in the following proposition.

Proposition 8.4. *Assume that the functions $\hat{\psi}$ and $\hat{w}$ satisfy the following conditions:*

1. *$\hat{\psi}$ and $\hat{w}$ are non-negative, non-decreasing on* $[0,1]$.
2. *The function $r \mapsto \hat{w}(r)/\sqrt{r}$ is non-increasing.*
3. *The function $r \mapsto \hat{\psi}(r)/r$ is non-increasing.*
4. $\hat{\psi} \circ w(r^*(\delta)) > \psi(w(r^*(\delta)))$.
5. *There exist constants $\kappa_1, \kappa_2 \geq 1$ such that $\hat{\psi}(w(r^*(\delta))) \leq \kappa_1 \psi(\kappa_2 w(r^*(\delta)))$.*
6. $\hat{w}(r^*_\delta) \geq w(r^*_\delta)$.
7. *There exist constants $\kappa_3, \kappa_4 \geq 1$ such that $\hat{w}(r^*_\delta) \leq \kappa_3 w(\kappa_4 r^*(\delta))$.*

Then the following holds:

1. *There exists $\hat{r}^*(\delta) > 0$, that solves $r = 4\hat{\psi}(\hat{w}(r)) + \hat{w}(r)\sqrt{2\frac{\log\frac{2}{\delta}}{n}} + \frac{8\log\frac{2}{\delta}}{3n}$.*
2. *If $\kappa = \kappa_1\kappa_2\kappa_3\sqrt{\kappa_4}$, then*
$$r^*(\delta) \leq \hat{r}^*(\delta) \leq \kappa r^*(\delta).$$

The proof of the proposition relies on elementary calculus and is left to the reader. A pleasant consequence of this lemma is that we may focus on the behavior of $\hat{\psi}$ and $\hat{w}$ at $w(r^*(\delta))$ and $r^*(\delta)$. In order to build estimates for $\hat{w}$, we will assume that w is defined by $w(r) \geq v(r) = \sup\{\sqrt{P\alpha|g-g^*|} \ : \ \alpha \in [0,1],\ \alpha(L(g)-L^*) \leq r\}$. This ensures that $w(r)/\sqrt{r}$ is non-increasing.

Before describing data-dependent functions $\hat{\psi}$ and $\hat{w}$ that satisfy the conditions of the lemma, we check that, within model $\mathcal{C}$, above a critical threshold related to $r^*(\delta)$, the empirical excess risk $L_n(g) - L_n(g_n)$ faithfully reflects the excess risk $L(g) - L(g^*)$. The following lemma and corollary could have been stated right after the proof of Theorem 5.5 in Section 5. It should be considered as a collection of ratio-type concentration inequalities.

Lemma 8.5. *With probability larger than $1-2\delta$ for all g in $\mathcal{C}$:*

$$L_n(g) - L_n(g_n) \leq L(g) - L(g^*) + \left(r^*(\delta) \vee \sqrt{(L(g)-L(g^*))\, r^*(\delta)}\right)$$

and

$$L(g) - L(g^*) - \left(r^*(\delta) \vee \sqrt{(L(g)-L(g^*))\, r^*(\delta)}\right) \leq L_n(g) - L_n(g_n)\,.$$

The proof consists in revisiting the proof of Theorem 5.5. An interesting consequence of this observation is the following corollary:

Corollary 8.6. *There exists $K \geq 1$ such that, with probability larger than $1-\delta$,*

$$\{g \in \mathcal{C} : \ L(g) - L(g^*) \leq K r^*(\delta)\} \subseteq \left\{g \in \mathcal{C} : \ L_n(g) - L_n(g_n) \leq K\left(2 + \frac{1}{\sqrt{K}}\right) r^*(\delta)\right\}\,.$$

and, with probability larger than $1-\delta$,

$$\{g \in \mathcal{C}:\ L(g) - L(g^*) \geq K r^*(\delta)\} \subseteq \left\{ g \in \mathcal{C}:\ L_n(g) - L_n(g_n) \geq K\left(1 - \frac{1}{\sqrt{K}}\right) r^*(\delta) \right\} .$$

In order to compute approximations for $\psi(\cdot)$ and $w(\cdot)$ it will also be useful to rely on the fact that the $L_2(P_n)$ metric structure of the loss class $\mathcal{F}$ faithfully reflect the $L_2(P)$ metric on $\mathcal{F}$. Note that, for any classifier $g \in \mathcal{C}$, $(g(x) - g^*(x))^2 = |\mathbb{1}_{g(x)\neq y} - \mathbb{1}_{g^*(x)\neq y}|$. As a matter of fact, this is even easier to establish than the preceding lemma. Squares of empirical L_2 distances to g^* are sums of i.i.d. random variables. So we are again in a position to invoke tools from empirical process theory. Moreover the connection between $P|g-g^*|$ and the variance of $|g-g^*|$ is obvious: $\sqrt{\mathrm{Var}[g-g^*]} \leq \sqrt{P|g-g^*|}$.

Lemma 8.7. *Let* $s^*(\delta)$ *denote the solution of the fixed-point equation*

$$s = 4\psi(\sqrt{s}) + \sqrt{s}\sqrt{\frac{2\log\frac{1}{\delta}}{n} + \frac{8\log\frac{1}{\delta}}{3n}} .$$

Then, with probability larger than $1-2\delta$, *for all* $g \in \mathcal{C}$,

$$\left(1 - \frac{\theta}{2}\right) P|g-g^*| - \frac{1}{2\theta} s^*(\delta) \leq P_n|g-g^*| \leq \left(1 + \frac{\theta}{2}\right) P|g-g^*| + \frac{1}{2\theta} s^*(\delta) .$$

The proof repeats again the proof of Theorem 5.5. This lemma will be used thanks to the following corollary:

Corollary 8.8. *For* $K \geq 1$, *with probability larger than* $1-\delta$,

$$\{g \in \mathcal{C}:\ P|g-g^*| \leq K s^*(\delta)\} \subseteq \left\{ g \in \mathcal{C}:\ P_n|g - g*| \leq K\left(1 + \frac{1}{\sqrt{K}}\right) s^*(\delta) \right\} ,$$

and, with probability larger than $1-\delta$,

$$\{g \in \mathcal{C}:\ P|g-g^*| \geq K s^*(\delta)\} \subseteq \left\{ g \in \mathcal{C}:\ P_n|g-g^*| \geq K\left(1 - \frac{1}{\sqrt{K}}\right) s^*(\delta) \right\} .$$

We are now equipped to build estimators of $w(\cdot)$ and $\psi(\cdot)$. When building an estimator of $w(\cdot)$, the guideline consists of two simple observations:

$$\begin{aligned} &\frac{1}{2}\sup\{P|g-g'|\ :\ L(g) \vee L(g') \leq L(g^*) + r\} \\ &\quad\leq\ \sup\{P|g-g^*|\ :\ L(g) \leq L(g^*) + r\} \\ &\quad\leq\ \sup\{P|g-g'|\ :\ L(g) \vee L(g') \leq L(g^*) + r\} . \end{aligned}$$

This prompts us to try to estimate $\sup\{P\alpha|g-g^*|\ :\ \alpha(L(g) - L(g^*)) \leq r\}$. This will prove to be feasible thanks to the results described above.

Lemma 8.9. *Let* $K > 2$ *and let* $\hat{v}$ *be defined as*

$$\hat{w}^2(r) = \frac{\sqrt{K-1}}{\sqrt{K-1}-1}\sup\left\{ P_n\alpha|g-g'|\ :\ \alpha \in [0,1],\ g, g' \in \mathcal{C},\ L_n(g) \vee L_n(g') \leq L_n(g_n) + \frac{1}{\alpha}K\left(2 + \frac{1}{\sqrt{K}}\right) r \right\} .$$

Let $\kappa_3 = \sqrt{\frac{2(1+1/\sqrt{K})}{1-1/\sqrt{K-1}}}$ *and* $\kappa_4 = K\frac{2\sqrt{K}+1}{\sqrt{K}-1}$. *Then, with probability larger than* $1-4\delta$,

$$w(r^*(\delta)) \leq \hat{w}(r^*(\delta)) \leq \kappa_3 w(\kappa_4 r^*(\delta)) .$$

PROOF. Let r be such that $r \geq r^*(\delta)$. Thanks to Lemma 8.5, with probability larger than $1-\delta$, $L_n(g^*)-L_n(g_n) \leq r^*(\delta) \leq r$ and $L(g)-L(g^*) \leq \frac{K}{\alpha} r$ implies $L_n(g)-L_n(g_n) \leq \frac{K}{\alpha}\left(2+\frac{1}{\sqrt{K}}\right) r$, so

$$\hat{w}^2(r) \quad \geq \quad \frac{1}{1-1/\sqrt{K-1}} \sup\left\{P_n \alpha|g-g^*| \ : \ \alpha \in [0,1], g \in \mathcal{C}, \ \alpha(L(g)-L(g^*)) \leq Kr\right\} .$$

Furthermore, by Lemma 8.7, with probability larger than $1-2\delta$,

$$\begin{aligned} \hat{w}^2(r) \quad & \geq \quad \frac{1}{1-1/\sqrt{K-1}} \sup\left\{\alpha(1-\frac{1}{\sqrt{K-1}}) P|g-g^*| \ : \ \alpha \in [0,1], \ g \in \mathcal{C}, \ \frac{K-1}{r} r \leq L(g)-L(g^*) \leq \frac{K}{\alpha} r\right\} \\ & \geq \quad \frac{1}{1-1/\sqrt{K-1}}\left((1-\frac{1}{\sqrt{K-1}}\right) w^2(Kr) \\ & \geq \quad w^2(r) . \end{aligned}$$

On the other hand, applying the elementary observation above, Lemma 8.5, and then Lemma 8.7, with probability larger than $1-2\delta$,

$$\begin{aligned} \hat{w}^2(r) \quad & \leq \quad \frac{2}{1-1/\sqrt{K-1}} \sup\left\{P_n \alpha|g-g^*| \ : \ \alpha \in [0,1], \ g \in \mathcal{C}, \ L_n(g) \leq L_n(g_n)+\frac{K}{\alpha}\left(2+\frac{1}{\sqrt{K}}\right) r\right\} \\ & \leq \quad \frac{2}{1-1/\sqrt{K-1}} \sup\left\{P_n \alpha|g-g^*| \ : \ \alpha \in [0,1], \ g \in \mathcal{C}, \ L(g) \leq L(g^*)+\frac{K}{\alpha} \frac{2\sqrt{K}+1}{\sqrt{K}-1} r\right\} \\ & \leq \quad \frac{2(1+1/\sqrt{K})}{1-1/\sqrt{K-1}} \sup\left\{P \alpha|g-g^*| \ : \alpha \in [0,1], \ g \in \mathcal{C}, \ L(g) \leq L(g^*)+\frac{K}{\alpha} \frac{2\sqrt{K}+1}{\sqrt{K}-1} r\right\} \\ & \leq \quad \frac{2(1+1/\sqrt{K})}{1-1/\sqrt{K-1}} w^2\left(K \frac{2\sqrt{K}+1}{\sqrt{K}-1} r\right) . \end{aligned}$$

□

Lemma 8.10. *Assume that* $\psi(w(r^*(\delta)) \geq \frac{8 \log \frac{1}{\delta}}{n}$. *Let* $K \geq 4$. *Let*

$$\hat{\psi}(r)=2 R_n\left\{\alpha(\mathbb{1}_{g(X) \neq Y}-\mathbb{1}_{g'(X) \neq Y}) \ : \ \alpha \in [0,1], \ \alpha^2 P_n|g-g_n| \leq r^2 \ , \alpha^2 P_n|g'-g_n| \leq K r^2 \ \right\} .$$

Then, with probability larger than $1-6\delta$,

$$\psi(w(r^*(\delta))) \leq \hat{\psi}(w(r^*(\delta))) \leq 8 \psi(\sqrt{2(K+2)} w(r^*(\delta))) .$$

PROOF. Note that $\hat{\psi}$ is positive, non-decreasing, and because it is defined with respect to star-hulls, $\hat{\psi}(r)/r$ is non-decreasing.

First recall that, by Theorem 5.5 and Lemma 8.7, taking $\theta=1$ there, with probability larger than $1-2\delta$,

$$\begin{aligned} P|g^*-g_n| \quad & \leq \quad w^2(r^*(\delta)) \\ P_n|g^*-g_n| \quad & \leq \quad \frac{3}{2} w^2(r^*(\delta))+\frac{s^*(\delta)}{2} \\ P_n|g-g_n| \quad & \leq \quad \frac{3}{2}\left(P|g-g^*|+w^2(r^*(\delta))\right)+s^*(\delta) . \end{aligned}$$

Let us first establish that, with probability larger than $1-2\delta$, $\hat{\psi}(w(r^*(\delta)))$ is larger than the empirical Rademacher complexity of the star-hull of a fixed class of loss-functions.

For $K \geq 4$ we have $Kw^2(r^*(\delta)) \geq \frac{6}{2}w^2(r^*(\delta)) + s^*(\delta)$. Invoking the observations above, with probability larger than $1-2\delta$,

$$
\begin{aligned}
&\hat{\psi}(w(r^*(\delta))) \\
&\quad\geq\ 2R_n\left\{\alpha(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y})\ :\ \alpha\in[0,1],\ \alpha^2 P_n|g-g_n| \leq Kw^2(r^*(\delta))\ \right\} \\
&\quad\geq\ 2R_n\left\{\alpha(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y})\ :\ \alpha\in[0,1],\ \alpha^2\frac{3}{2}\left(P|g-g^*| + w^2(r^*(\delta))\right) + \alpha^2 s^*(\delta) \leq Kw^2(r^*(\delta))\ \right\} \\
&\quad\geq\ 2R_n\left\{\alpha(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y})\ :\ \alpha\in[0,1],\ \alpha^2 P|g-g^*| \leq w^2(r^*(\delta))\ \right\}
\end{aligned}
$$

By Lemma 8.2, with probability larger than $1-\delta$, the empirical Rademacher complexity is larger than half of its expected value.

Let us now check that $\hat{\psi}(w(r^*(\delta))$ can be upper bounded by a multiple of $\psi(w(r^*(\delta)))$. Invoking again the observations above, with probability larger than $1-2\delta$,

$$
\begin{aligned}
&\hat{\psi}(w(r^*(\delta))) \\
&\quad\leq\ 4R_n\left\{\alpha(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y})\ :\ \alpha\in[0,1],\ \alpha^2 P_n|g-g_n| \leq Kw^2(r^*(\delta))\ \right\} \\
&\quad\leq\ 4R_n\left\{\alpha(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y})\ :\ \alpha\in[0,1],\ \alpha^2\left(\frac{1}{2}P|g-g^*| - s^*(\delta) - w^2(r^*(\delta))\right) \leq Kw^2(r^*(\delta))\ \right\} \\
&\quad\leq\ 4R_n\left\{\alpha(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y})\ :\ \alpha\in[0,1],\ \alpha^2 P|g-g^*| \leq 2\left(s^*(\delta) + (K+1)w^2(r^*(\delta))\right)\ \right\} \\
&\quad\leq\ 4R_n\left\{\alpha(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y})\ :\ \alpha\in[0,1],\ \alpha^2 P|g-g^*| \leq 2(K+2)w^2(r^*(\delta))\ \right\}.
\end{aligned}
$$

Now the last quantity is again the conditional Rademacher average with respect to a fixed class of functions. By Lemma 8.2, with probability larger than $1-\delta^3 \geq 1-\delta$,

$$
\begin{aligned}
&R_n\left\{\alpha(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y})\ :\ \alpha\in[0,1],\ \alpha^2 P|g-g^*| \leq 2(K+2)w^2(r^*(\delta))\ \right\} \\
&\quad\leq\ 2\mathbb{E}\left[R_n\left\{\alpha(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g^*(X)\neq Y})\ :\ \alpha\in[0,1],\ \alpha^2 P|g-g^*| \leq 2(K+2)w^2(r^*(\delta))\ \right\}\right].
\end{aligned}
$$

Hence, with probability larger than $1-3\delta$,

$$
\hat{\psi}(w(r^*(\delta))) \leq 8\psi(\sqrt{2(K+2)}w(r^*(\delta))).
$$

□

We may now conclude this section by the following result. We combine Proposition 8.4, Lemma 8.9 and Lemma 8.10 and choose $K=4$ in both Lemmas.

Proposition 8.11. *. Let ψ and w be defined as $\psi(r) = \mathbb{E}\left[R_n\left\{f\ :\ f\in\mathcal{F}^*, Pf^2 \leq r^2\right\}\right]$, and $w = (r) = \sup\left\{\sqrt{P|f|}\ :\ f\in\mathcal{F}^*, Pf \leq r\right\}$. Let $\hat{w}$ be defined by*

$$
\hat{w}^2(r) = \frac{\sqrt{3}}{\sqrt{3}-1}\sup\left\{P_n\alpha|g-g'|\ :\ \alpha\in[0,1],\ g,g'\in\mathcal{C},\ L_n(g)\vee L_n(g') \leq L_n(g_n) + \frac{1}{\alpha}4\left(2+\frac{1}{\sqrt{4}}\right)r\right\},
$$

and $\hat{\psi}$ be defined by

$$
\hat{\psi}(r) = 2R_n\left\{\alpha(\mathbb{1}_{g(X)\neq Y} - \mathbb{1}_{g'(X)\neq Y})\ :\ \alpha\in[0,1],\ \alpha^2 P_n|g-g_n| \leq 4r^2\ , \alpha^2 P_n|g'-g_n| \leq 4r^2\ \right\}.
$$

Let $\hat{r}^*(\delta)$ *be defined as the solution of equation* $r = 4\hat{\psi}(\hat{w}(r)) + \hat{w}(r)\sqrt{2\frac{2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n}$. *Then, with probability at least* $1 - 10\delta$,

$$r^*(\delta) \leq \hat{r}^*(\delta) \leq 480 r^*(\delta)\,.$$

Note although we give explicit constants, no attempt has been made to optimize the value of these constants. It is believed that the last constant 480 can be dramatically improved, at least by being more careful.

Bibliographical remarks. Analogs of Theorem 8.5 can be found in Koltchinskii [125] and Bartlett, Mendelson, and Philips [30]. The presentation given here is inspired by [125]. The idea of estimating the δ-reliable excess risk bounds $r^*(\delta)$ is put forward in [125] where several variants are exposed.

8.6. **Pre-testing**

In classification, the difficulties encountered by model selection through penalization partly stem from the fact that penalty calibration compels us to compare each $L(g^*_{n,k})$ with the unaccessible golden standard $L(g^*)$, although we actually only need to compare $L(g^*_{n,k})$ with $L(g^*_{n,k'})$, and to calibrate a threshold $\tau(k, k', D_n)$ so that $L_n(g^*_{n,k}) - L_n(g^*_{n,k'}) \leq \tau(k, k', D_n)$ when $L(g^*_{n,k})$ is not significantly larger than $L(g^*_{n,k'})$. As estimating the excess risk looks easier when the Bayes classifier belongs to the model, we will present in this section a setting where the performance of the model selection method essentially relies on the ability to estimate the excess risk when the Bayes classifier belongs to the model.

Throughout this section, we will rely on a few non-trivial assumptions.

Assumption 8.1.

(1) *The sequence of models* $(\mathcal{C}_k)_k$ *is nested:* $\mathcal{C}_k \subseteq \mathcal{C}_{k+1}$.

(2) *There exists some index* k^* *such that for all* $k \geq k^*$, *the Bayes classifier* g^* *belongs to* $\mathcal{C}_k$. *That is, we assume that the approximation bias vanishes for sufficiently large models. Conforming to a somewhat misguiding tradition, we call model* $\mathcal{C}_{k^*}$ *the true model.*

(3) *There exists a constant* Γ *such that, for each* $k \geq k^*$, *with probability larger than* $1 - \frac{\delta}{12k^2}$, *for all* $j \leq k$

$$r^*_k\left(\frac{\delta}{12k^2}\right) \leq \tau(j, k, D_n) \leq \Gamma r^*_k\left(\frac{\delta}{12k^2}\right)$$

where $r^*_k(\cdot)$ *is a distribution-dependent upper bound on the excess risk in model* $\mathcal{C}_k$ *with tunable reliability, defined as in Section 5.3.3.*

For each pair of indices $j \leq k$, let the threshold $\tau(j, k, D_n)$ be defined by

$$\tau(j, k, D_n) = \hat{r}^*_k\left(\frac{12}{k^2}\right)\,,$$

where $\hat{r}^*_k(\cdot)$ is defined as in Proposition 8.11 from Section 8.5. Hence we may take $\Gamma = 480$. Note that for $k \geq k^*$, the threshold looks like the ideal penalty described by (35).

The pre-testing method consists in first determining which models are admissible. Model $\mathcal{C}_j$ is said to be admissible if for all k larger than j, there exists some $g \in \mathcal{C}_j \subseteq \mathcal{C}_k$, such that $L_n(g) - L_n(g^*_{n,k}) \leq \tau(j, k, D_n)$. The aggregation procedure then selects the smallest admissible index

$$\hat{k} = \min\left\{j \;:\; \forall k > j,\ \exists\, g \in \mathcal{C}_j,\ L_n(g) - L_n(g^*_{n,k}) \leq \tau(j, k, D_n)\right\}$$

and outputs the minimizer $g^*_{n,\hat{k}}$ of the empirical risk in $\mathcal{C}_{\hat{k}}$.

Note that the pre-testing procedure does not fit exactly in the framework of the comparison method mentioned in Section 8.2. There, model selection was supposed to be based on comparisons between empirical risk minimizers. Here, model selection is based on the (estimated) ability to approximate $g^*_{n,j}$ by classifiers from $\mathcal{C}_k$.

Theorem 8.12. *Let $\delta > 0$. Let $(\mathcal{C}_k)_k$ denote a collection of nested models that satisfies Assumption 8.1. Let the index $\hat{k}$ and the classifier $g^*_{n,\hat{k}}$ be chosen according to the pre-testing procedure. Then, with probability larger than $1-\delta$,*

$$L(g^*_{n,\hat{k}}) - L^* \leq (\Gamma + 1 + \sqrt{1+4\Gamma})\, r^*_{k^*}\left(\frac{\delta}{12(k^*)^2}\right).$$

The theorem implies that, with probability larger than $1-\delta$, the excess risk of the selected classifier is of the same order of magnitude as the available upper bound on the excess risk of the "true" model. Note that this statement does not exactly match the goal we assigned ourselves in Section 8.1. The excess risk of the aggregated classifier is not compared with the excess risk of the oracle. Although the true model may coincide with the oracle for large sample sizes, this may not be the case for small and moderate sample sizes.

The proof is organized into three lemmas.

Lemma 8.13. *With probability larger than $1-\frac{\delta}{3}$, model $\mathcal{C}_{k^*}$ is admissible.*

PROOF. From Theorem 5.5, for each $k \geq k^*$, with probability larger than $1-\frac{\delta}{12k^2}$:

$$L_n(g^*_{n,k^*}) - L_n(g^*_{n,k}) \leq r^*_k\left(\frac{\delta}{12k^2}\right).$$

The proof of Lemma 8.13 is then completed by using the assumption that for each index k larger than k^*, with probability larger than $1-\frac{\delta}{12k^2}$, $r^*_k\left(\frac{\delta}{12k^2}\right) \leq \tau(k^*, k, D_n)$ holds, and resorting to the union bound. □

The next lemma deals with models which suffer an excessive approximation bias. The proof of this lemma will again rely on Theorem 5.5. But, this time, the model under investigation is $\mathcal{C}_{k^*}$.

Lemma 8.14. *Under Assumption 8.1, let κ be such that $\kappa - \sqrt{\kappa} \geq \Gamma$, then with probability larger than $1-\frac{\delta}{3}$, no index $k < k^*$ such that*

$$\inf_{g\in\mathcal{C}_k} L(g) \geq L^* + \kappa r^*_{k^*}\left(\frac{\delta}{(k^*)^2}\right)$$

is admissible.

PROOF. As all models $\mathcal{C}_k$ satisfying the condition in the lemma are included in

$$\left\{ g \in \mathcal{C}^*_k :\ L(g) \geq L^* + \kappa r^*_{k^*}\left(\frac{\delta}{(k^*)^2}\right)\right\},$$

it is enough to focus on the empirical process indexed by $\mathcal{C}_{k^*}$, and to apply Lemma 8.5 to $\mathcal{C}_{k^*}$. Choosing $\theta = 1/\sqrt{\kappa}$, for all k of interest, with probability larger than $1-\frac{\delta}{12(k^*)^2}$, we have

$$\begin{aligned} L_n(g^*_{n,k}) - L_n(g^*_{n,k^*}) &\geq L_n(g^*_{n,k}) - L_n(g^*) \\ &\geq (\kappa_3 - \sqrt{\kappa_3})\, r^*_{k^*}\left(\frac{\delta}{(k^*)^2}\right) \\ &\geq \Gamma\, r^*_{k^*}\left(\frac{\delta}{(k^*)^2}\right). \end{aligned}$$

Now, with probability larger than $1-\frac{\delta}{12(k^*)^2}$, the right hand side is larger than $\tau(k, k^*, D_n)$. □

The third lemma is a direct consequence of Theorem 5.8. It ensures that, with high probability, the pre-testing procedure provides a trade-off between estimation bias and approximation bias which is not much worse than the one provided by model $\mathcal{C}_{k^*}$.

Lemma 8.15. *Let κ be such that $\kappa - \sqrt{\kappa} \le \Gamma$. Under Assumption 8.1, for any $k \le k^*$ such that*

$$\inf_{g \in \mathcal{C}_k} L(g) \le L^* + \kappa\, r^*_{k^*}\left(\frac{\delta^2}{12k^2}\right),$$

with probability larger than $1 - \frac{\delta}{k^2}$,

$$L(g^*_{n,k}) - L(g^*) \le (\kappa + \sqrt{\kappa}) r^*_{k^*}\left(\frac{\delta^2}{12k^2}\right).$$

Bibliographical remarks. Pre-testing procedures were proposed by Lepskii [136], [137], [135] for performing model selection in a regression context. They are also discussed by Birgé [35]. Their use in model selection for classification was pioneered by Tsybakov [221] which is the main source of inspiration for this section. Koltchinskii [125] also revisits comparison-based methods using concentration inequalities and provides a unified account of penalty-based and comparison-based model selection techniques in classification.

In this section we presented model selection from a hybrid perspective, mixing the efficiency viewpoint advocated at the beginning of Section 8 (trying to minimize the classification risk without assuming anything about the optimal classifier g^*) and the consistency viewpoint. In the latter perspective, it is assumed that there exists a true model, that is, a minimal model without approximation bias and the goal is to first identify this true model (see Csiszár and Shields [67], Csiszár [66] for examples of recent results in the consistency approach for different problems), and then perform estimation in this hopefully true model.

The main tools in the construction of data-dependent thresholds for determining admissibility are ratio-type uniform deviation inequalities. The introduction of Talagrand's inequality for suprema of empirical processes greatly simplified the derivation of such ratio-type inequalities. An early account of ratio-type inequalities, predating [216], can be found in Chapter V of van de Geer [226]. Bartlett, Mendelson, and Philips [30] provide a concise and comprehensive comparison between the random empirical structure and the original structure of the loss class. This analysis is geared toward the analysis of empirical risk minimization.

The use and analysis local Rademacher complexities was promoted by Koltchinskii and Panchenko [126] (in the special case where $L(g^*) = 0$) and reached a certain level of maturity in Bartlett, Bousquet, and Mendelson [25], where Rademacher complexities of L_2 balls around g^* are considered. Koltchinskii [125] went one step further and pointed out that there is no need to estimate separately complexity and noise conditions: what matters is $\phi(w(\cdot))$. Koltchinskii [125] (as well as Bartlett, Mendelson, and Philips [30]) proposed to compute localized Rademacher complexities on the level sets of the empirical risk. Lugosi and Wegkamp [147] propose penalties based on empirical Rademacher complexities of the class of classifiers reduced to those with small empirical risk and obtain oracle inequalities that do not need the assumption that the optimal classifier is in one of the models.

Van de Geer and Tsybakov [223] recently pointed out that in some special cases, penalty-based model selection can achieve adaptivity to the noise conditions.

8.7. Revisiting hold-out estimates

Designing and assessing model selection policies based on either penalization or pre-testing 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 risk estimates that are at the core of the model selection techniques 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 much harder than empirical risk minimization. Obtaining non-asymptotic oracle inequalities for such cross-validation methods remains a challenge.

The simplest cross validation method is hold-out. It consists in splitting the sample of size $n+m$ in two parts: a training set of length n and a test set of length m. Let us denote by $L'_m(g)$ the average loss of g on the test set. Note that once the training set has been used to derive a collection of candidate classifiers $(g^*_{n,k})_k$, the model selection problem turns out to look like the problem we considered at the beginning of Section 5.2: picking a classifier from a finite collection. Here the collection is data-dependent but we may analyze the problem by reasoning conditionally on the training set. A second difficulty is raised by the fact we may not assume anymore that the Bayes classifier belongs to the collection of candidate classifiers. We need to robustify the argument of Section 5.2. Henceforth, let $g^*_{n,\tilde{k}}$ denote the minimizer of the probability of error in the collection $(g^*_{n,k})_k$.

The following theorem is a strong incentive to theoretically investigate and practically use resampling methods. Moreover its proof is surprisingly simple.

Theorem 8.16. *Let $(g^*_{n,k})_{k\leq N}$ denote a collection of classifiers obtained by processing a random training sample of length n. Let $\tilde{k}$ denote the index k that minimizes*

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

Let $\hat{k}$ denote the index k that minimizes $L'_m(g^*_{n,k})$ where the empirical risk L'_m is evaluated on an independent test sample of length m. Let $w(\cdot)$ be such that, for any classifier g,

$$\sqrt{\mathrm{Var}[\mathbf{1}_{g\neq g^*}]} \leq w\left(L(g) - L^*\right)$$

and such that $w(x)/\sqrt{x}$ is non-increasing. Let τ^* denote the smallest positive solution of $w(\epsilon) = \sqrt{m}\,\epsilon$. If $\theta \in (0,1)$, then

$$\mathbb{E}\left[L(g^*_{n,\hat{k}}) - L(g^*)\right] \leq (1+\theta)\inf_k \left[\mathbb{E}\left[L(g^*_{n,k}) - L(g^*)\right] + \left(\frac{8}{3m} + \frac{\tau^*}{\theta}\right)\log N\right] .$$

Remark 8.17. Assume the Mammen-Tsybakov noise conditions with exponent α hold, that is, we can choose $w(r) = \left(\frac{r}{h}\right)^{\alpha/2}$ for some positive h. Then, as $\tau^* = \left(\frac{1}{m\,h^\alpha}\right)^{-1/(2-\alpha)}$, the theorem translates into

$$\mathbb{E}\left[L(g^*_{n,\hat{k}}) - L(g^*)\right] \leq (1+\theta)\inf_k \left[\mathbb{E}\left[L(g^*_{n,k}) - L(g^*)\right] + \left(\frac{4}{3m} + \frac{1}{\theta\,(m\,h^\alpha)^{1/(2-\alpha)}}\right)\log N\right] .$$

Note that the hold-out based model selection method does not need to estimate the function $w(\cdot)$.

Using the notation of (31), the oracle inequality of Theorem 8.16 is almost optimal as far as the additive terms are concerned. Note however that the multiplicative factor on the right-hand side depends on the ratio between the minimal excess risk for samples of length n and samples of length $n+m$. This ratio depends on the setting of the learning problem, that is, on the approximation capabilities of the model collection and the noise conditions. As a matter of fact, the choice of a good trade-off between training and test sample sizes is still a matter of debate.

PROOF. By Bernstein's inequality and a union bound over the elements of $\mathcal{C}$, with probability at least $1-\delta$, for all $g^*_{n,k}$,

$$L(g^*_{n,k}) - L(g^*) \leq L'_m(g^*_{n,k}) - L'_m(g^*) + \sqrt{\frac{2\log\frac{N}{\delta}}{m}} \times w(L(g^*_{n,k}) - L(g^*)) + \frac{4\log\frac{N}{\delta}}{3m} ,$$

and

$$L(g^*) - L(g^*_{n,\tilde{k}}) \leq L'_m(g^*) - L'_m(g^*_{n,\tilde{k}}) + \sqrt{\frac{2\log\frac{N}{\delta}}{m}} \times w(L(g^*_{n,\tilde{k}}) - L(g^*)) + \frac{4\log\frac{N}{\delta}}{3m} .$$

Summing the two inequalities, we obtain

$$L(g^*_{n,k}) - L(g^*_{n,\tilde{k}}) \leq L'_m(g^*_{n,k}) - L'_m(g^*_{n,\tilde{k}}) + 2\sqrt{\frac{2\log\frac{N}{\delta}}{m}} \times w(L(g^*_{n,k}) - L(g^*)) + \frac{8\log\frac{N}{\delta}}{3m}\,, \tag{36}$$

As $L'_m(g^*_{n,\hat{k}}) - L'_m(g^*_{n,\tilde{k}}) \leq 0$, with probability larger than $1-\delta$,

$$L(g^*_{n,\hat{k}}) - L(g^*_{n,\tilde{k}}) \leq 2\sqrt{\frac{2\log\frac{N}{\delta}}{m}} \times w(L(g^*_{n,\hat{k}}) - L(g^*)) + \frac{8\log\frac{N}{\delta}}{3m}\,. \tag{37}$$

Let τ^* be defined as in the statement of the theorem. If $L(g^*_{n,\hat{k}}) - L(g^*) \geq \tau^*$, then $w(L(g^*_{n,\hat{k}}) - L(g^*))/\sqrt{m} \leq \sqrt{(L(g^*_{n,\hat{k}}) - L(g^*))\tau^*}$, and we have

$$\begin{aligned} L(g^*_{n,\hat{k}}) - L(g^*_{n,\tilde{k}}) &\leq 2\sqrt{2\log\frac{N}{\delta}} \times \sqrt{\tau^*} \times \sqrt{L(g^*_{n,\hat{k}}) - L(g^*)} + \frac{8\log\frac{N}{\delta}}{3m} \\ &\leq \frac{\theta}{2}(L(g^*_{n,\hat{k}}) - L(g^*)) + \frac{8}{2\theta}\tau^* \log\frac{N}{\delta} + \frac{8\log\frac{N}{\delta}}{3m}\,. \end{aligned}$$

Hence, with probability larger than $1-\delta$ (with respect to the test set),

$$L(g^*_{n,\hat{k}}) - L(g^*) \leq \frac{1}{1-\theta/2}(L(g^*_{n,\tilde{k}}) - L(g^*)) + \frac{1}{(1-\theta/2)} \times 4\log\frac{N}{\delta} \times \left(\frac{\tau^*}{\theta} + \frac{2}{3m}\right)\,.$$

Finally, taking expectation with respect to the training set and the test set, we get the oracle inequality stated in the theorem.

Bibliographical remarks. Hastie, Tibshirani and Friedman [104] provide an application-oriented discussion of model selection strategies. They provide an argument in defense of the hold-out methodology. An early account of using hold-out estimates in model selection can be found in Lugosi and Nobel [145] and in Bartlett, Boucheron, and Lugosi [24]. A sharp use of hold-out estimates in an adaptive regression framework is described by Wegkamp in [237]. This section essentially comes from the course notes by P. Massart [161] where better constants and exponential inequalities for excess risk can be found.

Acknowledgments. We thank Anestis Antoniadis for encouraging us to write this survey. We are indebted to the associate editor and the referees for the excellent suggestions that significantly improved the paper.

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] M.A. Aizerman, E.M. Braverman, and L.I. Rozonoer. *Method of potential functions in the theory of learning machines.* Nauka, Moscow, 1970.

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

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

[8] 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.

[9] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

[10] M. Anthony and N. Biggs. *Computational Learning Theory*. Cambridge Tracts in Theoretical Computer Science (30). Cambridge University Press, Cambridge, 1992.

[11] M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1993.

[12] A Antos, L. Devroye, and L. Györfi. Lower bounds for Bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:643–645, 1999.

[13] 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.

[14] A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Machine learning*, 30:31–56, 1998.

[15] P. Assouad. Densité et dimension. *Annales de l'Institut Fourier*, 33:233–282, 1983.

[16] J.-Y. Audibert and O. Bousquet. Pac-Bayesian generic chaining. In L. Saul, S. Thrun, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, 16, Cambridge, Mass., MIT Press. 2004.

[17] J.-Y. Audibert. PAC-Bayesian Statistical Learning Theory. Ph.D. Thesis, Université Paris 6, Pierre et Marie Curie, 2004.

[18] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.

[19] Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117(4):467–493, 2000.

[20] A.R. Barron, L. Birgé, and P. Massart. Risks bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–415, 1999.

[21] A.R. Barron. Logically smooth density estimation. Technical Report TR 56, Department of Statistics, Stanford University, 1985.

[22] 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.

[23] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.

[24] P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2001.

[25] P. Bartlett, O. Bousquet, and S. Mendelson. Localized Rademacher complexities. *The Annals of Statistics*, 33:1497–1537, 2005.

[26] P. L. Bartlett and S. Ben-David. Hardness results for neural network approximation problems. *Theoretical Computer Science*, 284:53–66, 2002.

[27] P.L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, to appear, 2005.

[28] 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.

[29] P.L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[30] P. L. Bartlett, S. Mendelson, and P. Philips. Local Complexities for Empirical Risk Minimization. In *Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, Springer, 2004.

[31] 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.

[32] S. Ben-David, N. Eiron, and H.-U. Simon. Limitations of learning via embeddings in Euclidean half spaces. *Journal of Machine Learning Research*, 3:441–461, 2002.

[33] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

[34] S.N. Bernstein. *The Theory of Probabilities*. Gostehizdat Publishing House, Moscow, 1946.

[35] 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.

[36] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

[37] 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.

[38] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4:329–375, 1998.

[39] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *The Annals of Statistics*, to appear, 2006.

[40] 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.

[41] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.

[42] 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.
[43] 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 (COLT)*, pages 144–152. Association for Computing Machinery, New York, NY, 1992.
[44] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *The Annals of Probability*, 33:514–560, 2005.
[45] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000.
[46] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31:1583–1614, 2003.
[47] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris*, 334:495–500, 2002.
[48] 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.
[49] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.
[50] 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 (COLT)*, pages 59–73. Springer, 2002.
[51] L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26:801–849, 1998.
[52] L. Breiman. Some infinite theory for predictor ensembles. *The Annals of Statistics*, 2004.
[53] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth International, Belmont, CA, 1984.
[54] P. Bühlmann and B. Yu. Boosting with the l_2-loss: Regression and classification. *Journal of the American Statistical Association*, 98:324–339, 2004.
[55] A. Cannon, J.M. Ettinger, D. Hush, and C. Scovel. Machine learning with data dependent hypothesis classes. *Journal of Machine Learning Research*, 2:335–358, 2002.
[56] G. Castellan. Density estimation via exponential model selection. *IEEE Transactions on Information Theory*, 49(8):2052–2060, 2003.
[57] O. Catoni. Randomized estimators and empirical complexity for pattern recognition and least square regression. preprint PMA-677.
[58] O. Catoni. *Statistical learning theory and stochastic optimization*. Ecole d'été de Probabilités de Saint-Flour XXXI. Springer-Verlag, Lecture Notes in Mathematics, Vol. 1851, 2004.
[59] O. Catoni. Localized empirical complexity bounds and randomized estimators, 2003. Preprint.
[60] N. Cesa-Bianchi and D. Haussler. A graph-theoretic generalization of the Sauer-Shelah lemma. *Discrete Applied Mathematics*, 86:27–35, 1998.
[61] M. Collins, R.E. Schapire, and Y. Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48:253-285, 2002.
[62] C. Cortes and V.N. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.
[63] 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.
[64] 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.
[65] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, UK, 2000.
[66] I. Csiszár. Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Transactions on Information Theory*, 48:1616–1628, 2002.
[67] I. Csiszár and P. Shields. The consistency of the BIC Markov order estimator. *The Annals of Statistics*, 28:1601–1619, 2000.
[68] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, pages 1–50, January 2002.
[69] A. Dembo. Information inequalities and concentration of measure. *The Annals of Probability*, 25:927–939, 1997.
[70] P.A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice-Hall, Englewood Cliffs, NJ, 1982.
[71] L. Devroye. Automatic pattern recognition: A study of the probability of error. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:530–543, 1988.
[72] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.
[73] L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28:1011–1018, 1995.
[74] 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.

[75] L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.

[76] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[77] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley, New York, 1973.

[78] R.O. Duda, P.E. Hart, and D.G.Stork. *Pattern Classification*. John Wiley and Sons, 2000.

[79] R.M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6:899–929, 1978.

[80] R.M. Dudley. Balls in R^k do not cut all subsets of $k+2$ points. *Advances in Mathematics*, 31 (3):306–308, 1979.

[81] R.M. Dudley. Empirical processes. In *Ecole de Probabilité de St. Flour 1982*. Lecture Notes in Mathematics #1097, Springer-Verlag, New York, 1984.

[82] R.M. Dudley. Universal Donsker classes and metric entropy. *The Annals of Probability*, 15:1306–1326, 1987.

[83] R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge, 1999.

[84] R.M. Dudley, E. Giné, and J. Zinn. Uniform and universal Glivenko-Cantelli classes. *Journal of Theoretical Probability*, 4:485–510, 1991.

[85] B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979.

[86] B. Efron. *The jackknife, the bootstrap, and other resampling plans*. SIAM, Philadelphia, 1982.

[87] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1994.

[88] 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.

[89] 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.

[90] P. Frankl. On the trace of finite sets. *Journal of Combinatorial Theory, Series A*, 34:41–45, 1983.

[91] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285, 1995.

[92] Y. Freund. Self bounding learning algorithms. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 127–135, 1998.

[93] 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.

[94] 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.

[95] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28:337–374, 2000.

[96] M. Fromont. *Some problems related to model selection: adaptive tests and bootstrap calibration of penalties*. Thèse de doctorat, Université Paris-Sud, december 2003.

[97] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 1972.

[98] E. Giné. Empirical processes and applications: an overview. *Bernoulli*, 2:1–28, 1996.

[99] E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, 12:929–989, 1984.

[100] E. Giné. 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.

[101] P. Goldberg and M. Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parametrized by real numbers. *Machine Learning*, 18:131–148, 1995.

[102] U. Grenander. *Abstract inference*. John Wiley & Sons Inc., New York, 1981.

[103] P. Hall. Large sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*, 11(4):1156–1174, 1983.

[104] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer Verlag, New York, 2001.

[105] D. Haussler. Decision theoretic generalizations of the PAC model for neural nets and other learning applications. *Information and Computation*, 100:78–150, 1992.

[106] 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.

[107] 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.

[108] R. Herbrich and R.C. Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3:175–212, 2003.

[109] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[110] P. Huber. The behavior of the maximum likelihood estimates under non-standard conditions. In *Proc. Fifth Berkeley Symposium on Probability and Mathematical Statistics*, pages 221–233. Univ. California Press, 1967.

[111] W. Jiang. Process consistency for adaboost. *The Annals of Statistics*, 32:13–29, 2004.

[112] D.S. Johnson and F.P. Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6:93–107, 1978.

[113] I. Johnstone. Function estimation and gaussian sequence models. Technical Report. Department of Statistics, Stanford University, 2002.
[114] 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.
[115] 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.
[116] 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.
[117] M.J. Kearns and U.V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, Massachusetts, 1994.
[118] A. G. Khovanskii. *Fewnomials*. Translations of Mathematical Monographs, vol. 88, American Mathematical Society, 1991.
[119] J.C. Kieffer. Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Transactions on Information Theory*, 39:893–902, 1993.
[120] G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41:495–502, 1970.
[121] P. Koiran and E.D. Sontag. Neural networks with quadratic VC dimension. *Journal of Computer and System Science*, 54, 1997.
[122] 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.
[123] 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.
[124] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47:1902–1914, 2001.
[125] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. Manuscript, september 2003.
[126] 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.
[127] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30, 2002.
[128] 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.
[129] S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. In *UAI-2002: Uncertainty in Artificial Intelligence*, 2002.
[130] J. Langford and M. Seeger. Bounds for averaging classifiers. CMU-CS 01-102, Carnegie Mellon University, 2001.
[131] 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.
[132] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1997. `http://www.emath.fr/ps/`.
[133] M. Ledoux and M. Talagrand. *Probability in Banach Space*. Springer-Verlag, New York, 1991.
[134] 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.
[135] O. V. Lepskiĭ, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Annals of Statistics*, 25(3):929–947, 1997.
[136] O.V. Lepskiĭ. A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470, 1990.
[137] O.V. Lepskiĭ. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659, 1991.
[138] 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.
[139] Y. Lin. A note on margin-based loss functions in classification. Technical Report 1029r, Department of Statistics, University of Wisconsin, Madison, 1999.
[140] Y. Lin. Some asymptotic properties of the support vector machine. Technical Report 1044r, Department of Statistics, University of Wisconsin, Madison, 1999.
[141] Y. Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.
[142] F. Lozano. Model selection using Rademacher penalization. In *Proceedings of the Second ICSC Symposia on Neural Computation (NC2000)*. ICSC Adademic Press, 2000.
[143] M.J. Luczak and C. McDiarmid. Concentration for locally acting permutations. *Discrete Mathematics*, 265:159–171, 2003.
[144] G. Lugosi. Pattern classification and learning theory. In L. Györfi, editor, *Principles of Nonparametric Learning*, pages 5–62. Springer, Wien, 2002.

[145] G. Lugosi and A. Nobel. Adaptive model selection using empirical complexities. *The Annals of Statistics*, 27:1830–1864, 1999.

[146] G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, 32:30–55, 2004.

[147] G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *The Annals of Statistics*, 2:1679–1697 , 2004.

[148] G. Lugosi and K. Zeger. Concept learning using complexity regularization. *IEEE Transactions on Information Theory*, 42:48–54, 1996.

[149] 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.

[150] C.L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1997.

[151] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[152] 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.

[153] 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.

[154] K. Marton. A simple proof of the blowing-up lemma. *IEEE Transactions on Information Theory*, 32:445–446, 1986.

[155] K. Marton. Bounding $\bar{d}$-distance by informational divergence: a way to prove measure concentration. *The Annals of Probability*, 24:857–866, 1996.

[156] K. Marton. A measure concentration inequality for contracting Markov chains. *Geometric and Functional Analysis*, 6:556–571, 1996. Erratum: 7:609–613, 1997.

[157] 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.

[158] P. Massart. Optimal constants for Hoeffding type inequalities. Technical report, Mathematiques, Université de Paris-Sud, Report 98.86, 1998.

[159] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28:863–884, 2000.

[160] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciencies de Toulouse*, IX:245–303, 2000.

[161] P. Massart. *Ecole d'Eté de Probabilité de Saint-Flour XXXIII*, chapter Concentration inequalities and model selection. LNM. Springer-Verlag, 2003.

[162] P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, to appear.

[163] D. A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 230–234. ACM Press, 1998.

[164] D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*. ACM Press, 1999.

[165] D. A. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

[166] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge, 1989.

[167] 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.

[168] C. McDiarmid. Concentration for independent permutations. *Combinatorics, Probability, and Computing*, 2:163–178, 2002.

[169] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley, New York, 1992.

[170] S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48:1977–1991, 2002.

[171] 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.

[172] S. Mendelson and P. Philips. On the importance of "small" coordinate projections. *Journal of Machine Learning Research*, 5:219–238, 2004.

[173] S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Inventiones Mathematicae*, 152:37–55, 2003.

[174] V. Milman and G. Schechman. *Asymptotic theory of finite-dimensional normed spaces*. Springer-Verlag, New York, 1986.

[175] B.K. Natarajan. *Machine Learning: A Theoretical Approach*. Morgan Kaufmann, San Mateo, CA, 1991.

[176] D. Panchenko. A note on Talagrand's concentration inequality. *Electronic Communications in Probability*, 6, 2001.

[177] D. Panchenko. Some extensions of an inequality of Vapnik and Chervonenkis. *Electronic Communications in Probability*, 7, 2002.

[178] D. Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *The Annals of Probability*, 31:2068–2081, 2003.

[179] T. Poggio, S. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428:419–422, 2004.

[180] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.
[181] D. Pollard. Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics*, 22:271–278, 1995.
[182] W. Polonik. Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *The Annals of Statistics*, 23(3):855–881, 1995.
[183] E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probability Theory and Related Fields*, 119:163–175, 2001.
[184] 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, Annales de l'Institut Henri Poincaré*, 2001.
[185] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
[186] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.
[187] W.H. Rogers and T.J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *Annals of Statistics*, 6:506–514, 1978.
[188] M. Rudelson, R. Vershynin. Combinatorics of random processes and sections of convex bodies. *The Annals of Math*, to appear, 2004.
[189] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory Series A*, 13:145–147, 1972.
[190] R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.
[191] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26:1651–1686, 1998.
[192] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
[193] D. Schuurmans. Characterizing rational versus exponential learning curves. In *Computational Learning Theory: Second European Conference. EuroCOLT'95*, pages 272–286. Springer Verlag, 1995.
[194] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.
[195] C. Scovel and I. Steinwart. Fast rates for support vector machines. Los Alamos National Laboratory Technical Report LA-UR 03-9117, 2003.
[196] M. Seeger. PAC-Bayesian generalisation error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.
[197] 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.
[198] S. Shelah. A combinatorial problem: Stability and order for models and theories in infinity languages. *Pacific Journal of Mathematics*, 41:247–261, 1972.
[199] G.R. Shorack and J. Wellner. *Empirical Processes with Applications in Statistics*. Wiley, New York, 1986.
[200] 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.
[201] A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors. *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 2000.
[202] A. J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.
[203] D.F. Specht. Probabilistic neural networks and the polynomial Adaline as complementary techniques for classification. *IEEE Transactions on Neural Networks*, 1:111–121, 1990.
[204] J.M. Steele. Existence of submatrices with all possible columns. *Journal of Combinatorial Theory, Series A*, 28:84–88, 1978.
[205] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, pages 67–93, 2001.
[206] I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Transactions on Information Theory*, 51:128–142, 2005.
[207] I. Steinwart. Support vector machines are universally consistent. *Journal of Complexity*, 18:768–791, 2002.
[208] I. Steinwart. On the optimal parameter choice in ν-support vector machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1274–1284, 2003.
[209] I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.
[210] S.J. Szarek and M. Talagrand. On the convexified Sauer-Shelah theorem. *Journal of Combinatorial Theory, Series B*, 69:183–192, 1997.
[211] M. Talagrand. The Glivenko-Cantelli problem. *The Annals of Probability*, 15:837–870, 1987.
[212] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22:28–76, 1994.
[213] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.*, 81:73–205, 1995.
[214] M. Talagrand. The Glivenko-Cantelli problem, ten years later. *Journal of Theoretical Probability*, 9:371–384, 1996.
[215] M. Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, 24:1049–1103, 1996. (Special Invited Paper).

[216] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996.

[217] M. Talagrand. A new look at independence. *The Annals of Probability*, 24:1–34, 1996. (Special Invited Paper).

[218] M. Talagrand. Vapnik-Chervonenkis type conditions and uniform Donsker classes of functions. *The Annals of Probability*, 31:1565–1582, 2003.

[219] M. Talagrand. *The generic chaining: upper and lower bounds for stochastic processes.* Springer-Verlag, New York, 2005.

[220] A. Tsybakov. On nonparametric estimation of density level sets. *Ann. Stat.*, 25(3):948–969, 1997.

[221] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32:135–166, 2004.

[222] A. B. Tsybakov. *Introduction l'estimation non-paramétrique*. Springer, 2004.

[223] A. Tsybakov and S. van de Geer Square root penalty: adaptation to the margin in classification and in edge estimation. *The Annals of Statistics*, to appear, 2005.

[224] S. Van de Geer. A new approach to least-squares estimation, with applications. *The Annals of Statistics*, 15:587–602, 1987.

[225] S. Van de Geer. Estimating a regression function. *The Annals of Statistics*, 18:907–924, 1990.

[226] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, UK, 2000.

[227] A.W. van der Waart and J.A. Wellner. *Weak convergence and empirical processes.* Springer-Verlag, New York, 1996.

[228] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780, 1963.

[229] V.N. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

[230] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[231] V.N. Vapnik. *Statistical Learning Theory*. John Wiley, New York, 1998.

[232] 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.

[233] 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.

[234] 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.

[235] M. Vidyasagar. *A Theory of Learning and Generalization*. Springer, New York, 1997.

[236] V. Vu. On the infeasibility of training neural networks with small mean squared error. *IEEE Transactions on Information Theory*, 44:2892–2900, 1998.

[237] Marten Wegkamp. Model selection in nonparametric regression. *Annals of Statistics*, 31(1):252–273, 2003.

[238] R.S. Wenocur and R.M. Dudley. Some special Vapnik-Chervonenkis classes. *Discrete Mathematics*, 33:313–318, 1981.

[239] Y. Yang. Minimax nonparametric classification. I. Rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999.

[240] Y. Yang. Minimax nonparametric classification. II. Model selection for adaptation. *IEEE Transactions on Information Theory*, 45(7):2285–2292, 1999.

[241] Y. Yang. Adaptive estimation in pattern recognition by combining different procedures. *Statistica Sinica*, 10:1069–1089, 2000.

[242] V.V. Yurinksii. Exponential bounds for large deviations. *Theory of Probability and its Applications*, 19:154–155, 1974.

[243] V.V. Yurinksii. Exponential inequalities for sums of random vectors. *Journal of Multivariate Analysis*, 6:473–499, 1976.

[244] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–85, 2004.

[245] D.-X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49:1743–1752, 2003.