The Annals of Statistics
1999, Vol. 27, No. 6, 1808–1829

SMOOTH DISCRIMINATION ANALYSIS[1]

By Enno Mammen and Alexandre B. Tsybakov

Ruprecht-Karls-Universität Heidelberg and Université Paris

Discriminant analysis for two data sets in $\mathbb{R}^d$ with probability densities f and g can be based on the estimation of the set $G = \{x: f(x) \geq g(x)\}$. We consider applications where it is appropriate to assume that the region G has a smooth boundary or belongs to another nonparametric class of sets. In particular, this assumption makes sense if discrimination is used as a data analytic tool. Decision rules based on minimization of empirical risk over the whole class of sets and over sieves are considered. Their rates of convergence are obtained. We show that these rules achieve optimal rates for estimation of G and optimal rates of convergence for Bayes risks. An interesting conclusion is that the optimal rates for Bayes risks can be very fast, in particular, faster than the "parametric" root-n rate. These fast rates cannot be guaranteed for plug-in rules.

1. Introduction. Assume that one observes two independent samples $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_m)$ of $\mathbb{R}^d$-valued i.i.d. observations with densities f or g [with respect to a σ-finite measure Q], respectively. The densities f and g are unknown and the measure Q need not be known. An additional random variable Z is observed that is assumed to have density f or g (and to be independent of X and Y). We consider the discrimination problem to classify if Z comes from f or g. Discrimination problems in this framework or in the framework of pattern recognition were studied by many authors [see, e.g., the recent books of Devroye, Györfi and Lugosi (1996) and Vapnik (1996) and the references cited therein].

A discrimination decision rule is defined by a set $G \subset \mathbb{R}^d$. We attribute Z to f if $Z \in G$ and to g otherwise. For a decision rule G the Bayes risk $R(G)$ (with prior probabilities $1/2$) is

$$R(G) = \tfrac{1}{2}\left\{\int_{G^c} f(x)Q(dx) + \int_G g(x)Q(dx)\right\},$$

where G^c is the complement of G. The Bayes risk is minimized by

$$G^* = \{x: f(x) \geq g(x)\}.$$

Denote $R^* = R(G^*) = \min_G R(G)$.

Received May 1997; revised July 1999.

[1] Supported in part by the Deutsche Forschungsgemeinschaft, Sonderforschungsbergich 373 "Quantifikation und Simulation ökonomischer Prozesse," Humboldt-Universität zu Berlin.

AMS 1991 *subject classifications*. Primary 62G05; secondary 62G20.

Key words and phrases. Discrimination analysis, optimal rates, empirical risk, Bayes risk, sieves.

Now that $R(G^*) = \frac{1}{2}\int \min\{f(x), g(x)\}Q(dx)$ and that

$$R(G) - R(G^*) = \tfrac{1}{2}d_{f,g}(G, G^*), \tag{1}$$

where

$$d_{f,g}(G_1, G_2) = \int_{G_1 \triangle G_2} |f - g|(x)Q(dx)$$

is a distance defined on measurable subsets of $\mathbb{R}^d$ and where $G_1 \triangle G_2 = [G_1^c \cap G_2] \cup [G_1 \cap G_2^c]$ is the symmetric difference of G_1 and G_2.

Since the densities f and g are assumed to be unknown, the Bayesian rule G^* is not available and one has to use empirical rules $\tilde{G}_{n,m}$, that is, set-valued functions based on observations $(X_1, \dots, X_n, Y_1, \dots, Y_m)$.

A standard way of assessing the quality of a decision rule $\tilde{G}_{n,m}$ is to estimate how fast $R(\tilde{G}_{n,m})$ converges to the minimal possible value R^*. The convergence $R(\tilde{G}_{n,m}) \to R^*$ (in probability, almost surely or in mean, respectively) was proved for various estimates $\tilde{G}_{n,m}$ of G^*. Moreover, certain bounds on the difference $\mathbf{E}(R(\tilde{G}_{n,m}) - R^*)$ are known for finite sample sizes [see Chapters 27, 28 of Devroye, Györfi and Lugosi (1996) and the papers by Barron (1991, 1994) and Barron, Birgé and Massart (1999)].

In this paper we study optimality of decision rule $\tilde{G}_{n,m}$:

1. How fast can $R(\tilde{G}_{n,m})$ converge to R^*, under [smoothness] assumptions on G^*?
2. Which decision rule $\tilde{G}_{n,m}$ attains the optimal rate of convergence?

To our knowledge, the only previous study of this problem was that of Marron (1983). Under smoothness assumptions on the densities f and g he proved that the optimal rates of convergence in discrimination are the same as those of the mean integrated squared error in density estimation. As an error criterion, Marron (1983) used the integrated (over all prior probabilities p from 0 to 1) difference $R_p(\tilde{G}_{n,m}) - R_p^*$, where $R_p(\cdot)$, R_p^* are the respective Bayes risks when the prior probabilities of f and g are p and $1 - p$. Our approach is quite different. We do not suppose that f and g are smooth. Instead, we put conditions on the possible sets G^*. In particular, we consider the case where the sets G^* are smooth enough (more precisely, that the boundary of G^* is smooth). This leads to different optimal decision rules and different optimal rates of convergence. In the setup of Marron (1983) plug-in rules $\{x: \hat{f}_n(x) > \hat{g}_m(x)\}$ show up as asymptotically optimal. Here $\hat{f}_n$ and $\hat{g}_m$ are properly chosen nonparametric estimators of f and g. Our decision rules are direct minimum contrast estimators of G^*. The intermediate density estimation step is avoided.

We consider nonparametric discrimination as a problem of set estimation. One of its specific features, as compared to other set estimation problems [see, e.g., Korostelev and Tsybakov (1993), Rudemo and Stryhn (1994), Mammen and Tsybakov (1995), Polonik (1995), Tsybakov (1997)], is that the nonstandard distance $d_{f,g}$ is inherent for the definition of the risk. We show that, under assumptions on the ε-entropy of the class of possible sets G^*, the empirical risk minimization rules of Vapnik and Chervonenkis (1974) type converge with optimal rate, in the distance $d_{f,g}$ and in the distance $d_{\triangle}$ [measure of symmetric difference; see (2) below]. The rate of convergence depends on the smoothness (ε-entropy, respectively) of the class of possible sets G^* and on the local slope of the difference $f - g$ around the boundary $\{x\colon f(x) = g(x)\}$. It is interesting that if the class of sets is not too large the convergence of Bayes risks turns out to be rather fast: the optimal rate is better than the "parametric" rate $(n \wedge m)^{-1/2}$.

This leads to the conclusion that direct estimation procedures (e.g.), empirical risk minimization) can achieve better performance in terms of Bayes risks than plug-in rules. In fact, plug-in rules are always affected by slow nonparametric rates, as in multivariate regression or density estimation. Their rates of convergence depend only on the smoothness of the underlying densities f and g (respectively, of the conditional probability function if the setup of pattern recognition is considered). We argue that smoothness of f and g is not a crucial point for discrimination analysis, and we impose smoothness assumptions on the sets G directly. This leads to faster convergence of Bayes risks. Our results about optimal convergence rates for Bayes risks confirm the remark by Devroye, Györfi and Lugosi [(1996), Section 6.7] that *classification is easier than regression function estimation.*

We prove upper and lower bounds on minimax risks of estimators for G^*. The proof of the upper bounds uses general results from empirical process theory [cf. Alexander (1984), Birgé and Massart (1993) and van de Geer (1998)]. The proof of lower bounds is based on Assouad's lemma and is inspired by the approaches in Korostelev and Tsybakov (1993) and Tsybakov (1997).

2. The results.

In this section we introduce empirical decision rules and state our results on optimal rates for discrimination. We start with some definitions.

Accuracy of set estimates will be measured by the distance $d_{f,g}(G, G')$ and

$$d_{\triangle}(G, G^*) = Q(G \triangle G^*), \tag{2}$$

where G and G' are (measurable) sets in $\mathbb{R}^d$. We will consider estimation of $G_K^* = G^* \cap K$, rather than G^*, where K is a subset of $\mathbb{R}^d$ with $0 < Q(K) < +\infty$. In particular, this includes compact sets K if Q is chosen as the Lebesgue measure and it allows the choice $K = \mathbb{R}^d$ if Q is a probability measure. In this context the following modified definition of Bayes risks is natural:

$$R_K(G) = \tfrac{1}{2}\left\{\int_{K\setminus G} f(x)Q(dx) + \int_G g(x)Q(dx)\right\}.$$

This definition of Bayes risks will be used everywhere below.

A basic element of the model is the class $\mathscr{G}$ of possible "candidate" sets G. This class is assumed to be given. It imposes, in turn, restrictions on the class $\mathscr{F}$ of possible pairs (f, g). Our results are given in a minimax framework, over the class $\mathscr{F}$. For a specified class $\mathscr{G}$ of subsets of K, for positive constants c_1,

c_2, η_0 and α, and for a σ-finite measure Q the class $\mathscr{F}$ is defined as

$$\begin{aligned}\mathscr{F} = \{(f, g)\colon & f \text{ and } g \text{ are probability densities on } \mathbb{R}^d \text{ w.r.t. } Q,\\ &\{x \in K\colon f(x) \geq g(x)\} \in \mathscr{G},\ f(x), g(x) \leq c_1 \text{ for } x \in K,\\ &Q\{x \in K\colon |f(x) - g(x)| \leq \eta\} \leq c_2\eta^\alpha \text{ for } 0 < \eta \leq \eta_0\}.\end{aligned} \tag{3}$$

This definition makes sense if the constants c_1, c_2, η_0 and α are chosen such that the class $\mathscr{F}$ is not empty. This is what we assume in the sequel (without explicitly stating the restrictions on these parameters.) Also, we assume for convenience that $0 < \eta_0 < 1/2$. The condition

$$Q\{x \in K\colon |f(x) - g(x)| \leq \eta\} \leq c_2\eta^\alpha \tag{4}$$

for $0 < \eta \leq \eta_0$ is related to the behavior of $g - f$ at the boundary of G. Under some additional regularity conditions the coefficient α can be easily calculated if $g - f$ has partial derivatives up to order $r + 1$ in a neighborhood of the boundary ∂G of G, if the first r derivatives vanish at ∂G and if not all partial derivatives of order $r + 1$ do vanish. Then the inequality (4) holds with $\alpha = (r + 1)^{-1}$ provided Q has a bounded density near ∂G (w.r.t. Lebesgue measure). The most interesting case may be here that $r = 0$, that is, $\alpha = 1$. Another interpretation of (4) could be the following: $(g - f)(x)$ is $O(|x|_G^{1/\alpha})$ near ∂G with $\alpha > 0$ where $|x|_G$ is the Euclidean distance of x from the boundary ∂G. This may be satisfied, in particular, for nonsmooth f and g.

Consider now the following decision rule:

$$\hat{G}_{n,m} = \arg\min_{G \in \mathscr{G}} R_{n,m}(G), \tag{5}$$

where

$$R_{n,m}(G) = \frac{1}{2n}\sum_{i=1}^{n} \mathbf{I}(X_i \in K \setminus G) + \frac{1}{2m}\sum_{i=1}^{m} \mathbf{I}(Y_i \in G)$$

denotes the empirical risk. Here and below $\mathbf{I}$ is the indicator function. Clearly, $R_{n,m}(G)$ is an unbiased estimator of $R_K(G)$. We remark that the construction of $\hat{G}_{n,m}$ does not use knowledge of the dominating measure Q.

Although the definition of the empirical decision rule $\hat{G}_{n,m}$ is similar to that of Vapnik and Chervonenkis (1974) [see also Vapnik (1996) and the references therein], there is an important difference. Their minimization of empirical risks runs over a parametric class $\mathscr{C}$ of possible sets (rather than over a nonparametric class) and it is not supposed that $G_K^* \in \mathscr{C}$. Typical results of Vapnik–Chervonenkis theory focus on the convergence of $R(\tilde{G}_{n,m})$ to $\inf_{G \in \mathscr{C}} R(G)$ where $\tilde{G}_{n,m} = \arg\min_{G \in \mathscr{C}} R_{n,m}(G)$. If the true set $G_K^* \notin \mathscr{C}$, this corresponds only to evaluating the "variance term" in the total error $R(\tilde{G}_{n,m}) - R(G_K^*)$. However, the bias term is equally important, and the balance between bias and variance yields optimality. In contrast, our nonparametric approach allows efficient approximation of G_K^* by elements of a rich nonparametric class and so it accounts for both bias and variance terms. A similar remark applies to the discussion of sieve estimators considered below.

Our set estimation procedures (5) are closely related to maximum likelihood estimators of the support of a density [see Mammen and Tsybakov (1995)] and to excess mass estimators of density level sets [studied by Hartigan (1987), Müller and Sawitzki (1991), Müller (1993) and Polonik (1995)].

We study now the rate of convergence of $\hat{G}_{n,m}$ to G_K^*. This rate depends on the δ-entropy $H_B(\delta)$ (with bracketing) of the metric space $(\mathscr{G}, d_\triangle)$. For $\delta > 0$, the quantity $H_B(\delta) = H_B(\delta, \mathscr{G}, d_\triangle)$ is defined as the minimal number, such that $N_B(\delta) = \exp H_B(\delta)$ is an integer and such that there exist pairs (U_j, V_j), $j = 1, \dots, N_B(\delta)$, of subsets of $\mathscr{G}$ satisfying:

1. $U_j \subset V_j$, for $j = 1, \dots, N_B(\delta)$.
2. $d_\triangle(U_j, V_j) \le \delta$, for $j = 1, \dots, N_B(\delta)$.
3. For any $G \in \mathscr{G}$ there exists a $j \in \{1, \dots, N_B(\delta)\}$ such that $U_j \subset G \subset V_j$.

In the sequel we denote the probability measure and the expectation in case of underlying densities f and g by $\mathbf{P}_{f,g}$ or $\mathbf{E}_{f,g}$, respectively.

THEOREM 1. *For a class $\mathscr{G}$ of subsets of a set $K \subset \mathbb{R}^d$ and for positive constants c_1, c_2, η_0, α and σ-finite measure Q define the class $\mathscr{F}$ of pairs of densities (f, g) according to* (3). *Suppose that $0 < Q(K) < \infty$ and that there exist positive constants ρ and A such that*

$$H_B(\delta, \mathscr{G}, d_\triangle) \le A\delta^{-\rho}, \tag{6}$$

for $\delta > 0$ small enough. Then, for all $p \ge 1$,

$$\lim \sup_{n \wedge m \to \infty} \sup_{(f,g) \in \mathscr{F}} \tau(n,m)^p \mathbf{E}_{f,g} d_\triangle^p\left(\hat{G}_{n,m}, G_K^*\right) < \infty, \tag{7}$$

$$\lim \sup_{n \wedge m \to \infty} \sup_{(f,g) \in \mathscr{F}} \tau(n,m)^{p[1+\alpha]/\alpha} \mathbf{E}_{f,g} d_{f,g}^p\left(\hat{G}_{n,m}, G_K^*\right) < \infty, \tag{8}$$

where $n \wedge m$ denotes the minimum of n and m and where

$$\tau(n,m) = \begin{cases} (n \wedge m)^{\alpha/[2+\alpha+\rho\alpha]}, & \text{if } \rho < 1, \\ (n \wedge m)^{\alpha/[2(\alpha+1)]}[\log(n \wedge m)]^{-\alpha/(\alpha+1)}, & \text{if } \rho = 1, \\ (n \wedge m)^{\alpha/[(\alpha+1)(\rho+1)]}, & \text{if } \rho > 1. \end{cases}$$

Theorem 1 allows treating a number of interesting special cases. First, a rather general example where (6) holds is that of Dudley's classes $\mathscr{G}$ [Dudley (1974); see also Mammen and Tsybakov (1995)]. These classes contain sets (possibly disconnected) with piecewise smooth boundaries.

Another example is given by the class $\mathscr{G}$ of convex subsets G of $K = [0,1]^2$. The bound (6) for this class holds with $\rho = 1/2$ [Dudley (1974)]. A computationally efficient algorithm for constructing $\hat{G}_{n,m}$ (which is in this case a convex set with piecewise linear boundary) is proposed by Müller (1995).

Bloch and Silverman (1997) discuss classes of two-dimensional sets with monotone boundaries. They propose an algorithm for the calculation of $\hat{G}_{n,m}$ based on dynamic programming. For this class (6) holds with $\rho = 1$ [Dudley (1974)].

Finally, Theorem 1 covers the case where $\mathscr{G}$ is a class of boundary fragments with smooth boundaries [cf. Korostelev and Tsybakov (1993)]. For these models, that we will discuss in detail, we derive lower bounds on the minimax risks and show that the rates of Theorem 1 cannot be improved for $\rho < 1$. We define now boundary fragments with Hölder continuous boundaries. For given $\gamma > 0$ and $d \geq 2$, consider the functions $b(x_1, \dots, x_{d-1})$, b: $[0,1]^{d-1} \to [0,1]$ having continuous partial derivatives up to order l, where l is the maximal integer that is strictly less than γ. For such functions b, we denote the Taylor polynomial of order l at a point $x \in [0,1]^{d-1}$ by $p_{b,x}(\cdot)$. For a given $L > 0$, let $\Sigma(\gamma, L)$ be the class of functions b such that

$$|b(y) - p_{b,x}(y)| \leq L|y - x|^{\gamma} \text{ for all } x, y \in [0,1]^{d-1},$$

where $|y|$ stands for the Euclidean norm of $y \in [0,1]^{d-1}$. A function b in $\Sigma(\gamma,L)$ defines a set

$$G_b = \{(x_1, \dots, x_d) \in [0,1]^d : 0 \leq x_d \leq b(x_1, \dots, x_{d-1})\}. \tag{9}$$

Such sets are called boundary fragments. Define the class

$$\mathscr{G}_{\text{frag}} = \{G_b : b \in \Sigma(\gamma, L)\}. \tag{10}$$

For given positive constants γ, L, c_1, c_2, η_0 and α and a σ-finite measure Q we define the class $\mathscr{F} = \mathscr{F}_{\text{frag}}$ of pairs (f, g) of probability densities satisfying (3) with $\mathscr{G} = \mathscr{G}_{\text{frag}}$.

It is well known [see, e.g., Dudley (1974)] that the δ-entropy with bracketing of $\mathscr{G}_{\text{frag}}$ satisfies

$$H_B(\delta, \mathscr{G}_{\text{frag}}, d_{\triangle}) \leq A\delta^{-[d-1]/\gamma}, \tag{11}$$

for some $A > 0$ and all $\delta > 0$ small enough. Thus, (6) is satisfied with $\rho = (d-1)/\gamma$.

The rate $\tau(n, m)$ defined in Theorem 1 is not optimal for $\rho \geq 1$, that is, for the case of a very huge class $\mathscr{G}$. This follows from the next theorem where estimates are defined that achieve faster (and, in fact, optimal) rates. Examples of models with huge parameter sets are known where optimal rates cannot be achieved by minimum contrast estimates (e.g., least squares estimates, maximum likelihood estimates); see Birgé and Massart (1993). We conjecture that the same phenomenon appears here. Alternative estimates are sieve estimates. Sieve estimates achieve optimal rates in several models; see, for example, Birgé and Massart (1993), Wong and Shen (1995), van de Geer (1995), Mammen and Tsybakov (1995), Birgé and Massart (1998), Barron, Birgé and Massart (1999). In our setup, a sieve estimate is defined by

$$\overline{G}_{n,m} = \arg\min_{G \in \mathscr{N}_{n,m}} R_{n,m}(G), \tag{12}$$

where $\mathscr{N}_{n,m}$ is a finite class of subsets of K. The next theorem states upper bounds on the rates of sieve estimates for both cases, $\rho \geq 1$ and $\rho < 1$. This is done for a subclass $\mathscr{F}'$ of $\mathscr{F}$. In the theorem we do not exclude the case $\rho < 1$ (where the rates of $\hat{G}_{n,m}$ are the same as those for $\overline{G}_{n,m}$) because sieve estimates $\overline{G}_{n,m}$ are easier to compute than $\hat{G}_{n,m}$. Note that the computation of $\hat{G}_{n,m}$ needs a minimization over a nonparametric class of sets, and this requires the development of special numerical procedures [cf. Müller (1995), Bloch and Silverman (1997)]. So also for $\rho < 1$ (in particular when no procedure for the calculation of $\hat{G}_{n,m}$ is available) it makes sense to apply sieve estimates with finite $\mathscr{N}_{n,m}$.

THEOREM 2. *Let $\mathscr{G}$ be a class of subsets of a set $K \subset \mathbb{R}^d$, let c_1, c_2, η_0 and α be positive constants and let Q be a σ-finite measure with $0 < Q(K) < \infty$. Let, as in Theorem* 1, *the class $\mathscr{F}$ of pairs of densities (f, g) be defined according to* (3).

Suppose that $\mathscr{F}'$ is a subset of $\mathscr{F}$ and that $\mathscr{N}_{n,m}$ is a family of subsets of K, such that for every $(f, g) \in \mathscr{F}'$ one can find a set $G' \in \mathscr{N}_{n,m}$ with

$$d_{f,g}(G', \{x \in K\colon f(x) \geq g(x)\}) \leq C\tau_0(n,m)^{-(1+\alpha)/\alpha}, \tag{13}$$

where C is a positive constant and

$$\tau_0(n,m) = (n \wedge m)^{\alpha/[2+\alpha+\rho\alpha]}. \tag{14}$$

Finally, suppose that one of the following two conditions holds:

(i) *There exist constants $A > 0$ and $0 < \rho < 1$ such that* (6) *holds for $\delta > 0$ small enough.*

(ii) *There exist constants $B > 0$ and $\rho > 0$ such that the family of sets $\mathscr{N}_{n,m}$ has a finite number $N_{n,m}$ of elements with*

$$\log N_{n,m} \leq B\tau_0(n,m)^{\rho}. \tag{15}$$

Define the sieve estimate $\overline{G}_{n,m}$ according to (12). *Then, for all $p \geq 1$,*

$$\lim \sup_{n \wedge m \to \infty} \sup_{(f,g) \in \mathscr{F}'} \tau_0(n,m)^p \mathbf{E}_{f,g} d_{\triangle}^p \left(\overline{G}_{n,m}, G_K^*\right) < \infty, \tag{16}$$

$$\lim \sup_{n \wedge m \to \infty} \sup_{(f,g) \in \mathscr{F}'} \tau_0(n,m)^{[1+\alpha]p/\alpha} \mathbf{E}_{f,g} d_{f,g}^p \left(\overline{G}_{n,m}, G_K^*\right) < \infty. \tag{17}$$

Let us first remark that in Theorem 2 for $\rho < 1$ [if the entropy bound (6) applies] we do not require that (15) holds. So in this case, arbitrarily huge sets $\mathscr{N}_{n,m}$ can be chosen. Under condition (ii) because of (13), $\mathscr{N}_{n,m}$ is an ε-net of $\mathscr{G}$ with respect to the metric $d_{f,g}$ and with $\varepsilon = C\tau_0(n,m)^{-(1+\alpha)/\alpha}$. This fact and Lemma 2 in the Appendix imply that $\mathscr{N}_{n,m}$ is also an ε-net with respect to the metric $d_{\triangle}$ where now $\varepsilon = C'\tau_0(n,m)^{-1}$ with a constant C'. So

(15) implies that

$$H(\delta, \mathscr{G}, d_{\triangle}) \leq A\delta^{-\rho}, \tag{18}$$

for $\delta = C'\tau_0(n, m)^{-1}$ with the same ρ and with some constant A. Here $H(\delta, \mathscr{G}, d_{\triangle})$ is the δ-entropy (without bracketing), that is, the minimal number N such that there exist sets $U_1, \dots, U_{[\exp(N)]}$ with $\min_{1 \leq j \leq [\exp(N)]} d_{\triangle}(U_j, G) \leq \delta$ for all $G \in \mathscr{G}$. Note that (18) differs from (6) because now entropy without bracketing instead of entropy with bracketing is used. On the other hand (18) does not imply that there exists an ε-net $\mathscr{N}_{n,m}$ with (13) and (15) because not all ε-nets of $\mathscr{G}$ (with respect to the metric $d_{\triangle}$) satisfy (13).

We discuss now conditions that imply (13) and (15). Suppose that K is a compact set and that for pairs $(f, g) \in \mathscr{F}'$ the first partial derivatives of $f - g$ are uniformly bounded in a neighborhood of the boundary of $G = \{x \in K\colon f(x) \geq g(x)\}$ and that $\mathscr{F}$ is defined with $\alpha = 1$. Assume furthermore that the entropy $H(\delta, \mathscr{G}, d_H)$ (without bracketing) fulfills

$$H(\delta, \mathscr{G}, d_H) \leq A\delta^{-\rho}, \tag{19}$$

where d_H is the Haussdorff distance, that is, $d_H(G_1, G_2) = \max\{\sup_{x \in G_1} \inf_{y \in G_2}|x - y|, \sup_{y \in G_2} \inf_{x \in G_1}|x - y|\}$ and $|\cdot|$ denotes the Euclidean norm. Then if one chooses $\mathscr{N}_{n,m}$ as an ε-net of $\mathscr{G}$ [with respect to the metric d_H and with $\varepsilon = c\tau_0(n, m)^{-1}$ for some $c > 0$] it is easy to see that this choice of $\mathscr{N}_{n,m}$ fulfills (13). Furthermore, because of (19) $\mathscr{N}_{n,m}$ can be chosen such that (15) holds. This result could be generalized to other choices of α under appropriate uniform smoothness conditions on $f - g$ for $(f, g) \in \mathscr{F}'$.

We now come back to the discussion of the classes $\mathscr{G}_{\text{frag}}$ of boundary fragments with Hölder continuous boundary; see (10). Let Q be the Lebesgue measure. For the class $\mathscr{G}_{\text{frag}}$ the assumptions of Theorem 1 hold with $\rho = (d - 1)/\gamma$; see (11). For two sets G_{b_1} and G_{b_2} [see (9)] we have that $d_{\triangle}(G_{b_1}, G_{b_2}) = \int|b_1(u) - b_2(u)|\,du$. Therefore the δ-entropy of $\mathscr{G}$ with respect to $d_{\triangle}$ coincides with the δ-entropy of $\Sigma(\gamma, L)$ with respect to the L_1-norm. This is the reason why (11) holds. We discuss now the assumptions of Theorem 2 for these classes for $\alpha = 1$. We consider the subclass $\mathscr{F}'_{\text{frag}}$ of pairs $(f, g) \in \mathscr{F}_{\text{frag}}$ with the property that $f - g$ is uniformly Lipschitz continuous with respect to its last argument,

$$\begin{aligned}\mathscr{F}'_{\text{frag}} = \Big\{(f, g) \in \mathscr{F}_{\text{frag}}\colon |(f - g)(x_1, \dots, x_{d-1}, x_d) \\ -(f - g)(x_1, \dots, x_{d-1}, x'_d)| \\ \leq C|x_d - x'_d| \text{ for all } 0 \leq x_1, \dots, x_{d-1}, x_d, x'_d \leq 1\Big\}\end{aligned} \tag{20}$$

for some constant C (which is assumed to be large enough as compared to c_2; otherwise $\mathscr{F}'_{\text{frag}}$ may be empty). For two pairs (f_1, g_1) and $(f_2, g_2) \in \mathscr{F}'_{\text{frag}}$, choose b_1 and $b_2 \in \Sigma(\gamma, L)$ such that $G_{b_j} = \{x \in K\colon f_j(x) \geq g_j(x)\} \in \mathscr{G}_{\text{frag}}$ for $j = 1, 2$. Then we have

$$\begin{aligned}d_{f,g}(G_{b_1}, G_{b_2}) &\leq C\int|b_1(u) - b_2(u)|^2\,du \\ &= C\|b_1 - b_2\|_2^2.\end{aligned} \tag{21}$$

Choose now an ε-net in $(\Sigma(\gamma, L), \|\cdot\|_2)$ with $\varepsilon = c\tau_0(n,m)^{-1}$ for a constant $c > 0$ and define $\mathscr{N}_{n,m}$ as the net of the corresponding sets in $\mathscr{G}_{\text{frag}}$. There exists such an ε-net with the bound (15) for the number of its elements. This follows from the entropy bound $H(\delta, \Sigma(\gamma, L), \|\cdot\|_2) \le A\delta^{-[d-1]/\gamma}$ for δ small enough. Furthermore (13) holds because of (21). Therefore, for $\alpha = 1$ the assumptions of Theorem 2 are fulfilled for classes of boundary fragments if one chooses $\mathscr{F}'_{\text{frag}}$ according to (20). A similar discussion applies for other choices of α.

The next theorem states that for classes of boundary fragments no better rates can be achieved than the rates given in the upper bounds of Theorem 1 (for $0 < \rho < 1$) and of Theorem 2 (for all $\rho > 0$), where $\rho = (d-1)/\gamma$.

THEOREM 3. *Let $K = [0,1]^d$ and let $\mathscr{F} = \mathscr{F}_{\text{frag}}$ be as in* (3) *with $\mathscr{G} = \mathscr{G}_{\text{frag}}$. Suppose that Q is the Lebesgue measure on K. Then*

$$(22)\qquad \lim_{n\wedge m\to\infty}\inf\ \inf_{\tilde{G}_{n,m}}\ \sup_{(f,g)\in\mathscr{F}_{\text{frag}}} \tau_0(n,m)^p\mathbf{E}_{f,g}d^p_{\triangle}\big(\tilde{G}_{n,m}, G^*_K\big) > 0,$$

$$(23)\qquad \lim_{n\wedge m\to\infty}\inf\ \inf_{\tilde{G}_{n,m}}\ \sup_{(f,g)\in\mathscr{F}_{\text{frag}}} \tau_0(n,m)^{[1+\alpha]p/\alpha}\mathbf{E}_{f,g}d^p_{f,g}\big(\tilde{G}_{n,m}, G^*_K\big) > 0,$$

for every $p \ge 1$. In (22) *and* (23) *the infimum runs over all estimates $\tilde{G}_{n,m}$ of G. The rate $\tau_0(n,m)$ is defined as in* (14) *where now $\rho = (d-1)/\gamma$. For $\alpha = 1$* (22) *and* (23) *hold with $\mathscr{F}_{\text{frag}}$ replaced by $\mathscr{F} = \mathscr{F}'_{\text{frag}}$; see* (20).

Theorems 1–3, together with (11), show that the rates of convergence $\tau_0(n,m)^{-1}$ and $\tau_0(n,m)^{-[1+\alpha]/\alpha}$ are optimal in the minimax sense for the distances $d_\triangle$ and $d_{f,g}$ respectively. This holds for the class $\mathscr{F}_{\text{frag}}$ when $\gamma > d-1$ and for the class $\mathscr{F}'_{\text{frag}}$ when $\alpha = 1$ and $\gamma > 0$. Furthermore, this rate is achieved by $\hat{G}_{m,n}$ whenever $\gamma > d-1$ or by $\overline{G}_{n,m}$ for $\gamma > d-1$ or for $\alpha = 1$ and any $\gamma > 0$. Note that for the distance $d_\triangle$ the rate is exactly the same as the optimal rate in the problem of level sets estimation [cf. Tsybakov (1997)].

The value of $\mathbf{E}_{f,g}d_{f,g}(\tilde{G}_{n,m}, G^*_K)$ is of particular interest because it corresponds to the Bayes risk for the discrimination rule $\tilde{G}_{n,m}$; see (1). The following corollary shows that the Bayes risk of the decision rule $\hat{G}_{n,m}$ or $\overline{G}_{n,m}$ converges with optimal rate to the minimal Bayes risk $R_K(G^*_K)$ (under appropriate conditions).

COROLLARY 1. *Suppose that $K = [0,1]^d$ and Q is the Lebesgue measure on K. Assume that $\gamma > d-1$ (Case 1) or that $\alpha = 1$ and $\gamma > 0$ (Case 2). For Case 1 choose $\mathscr{F} = \mathscr{F}_{frag}$ and $\hat{G}^*_{n,m} = \overline{G}_{n,m}$ or $\hat{G}^*_{n,m} = \hat{G}_{n,m}$. For Case 2 choose $\mathscr{F} = \mathscr{F}'_{frag}$ and $\hat{G}^*_{n,m} = \overline{G}_{n,m}$. Then*

$$\frac{\sup_{(f,g)\in\mathscr{F}}\mathbf{E}_{f,g}\big[R_K\big(\hat{G}^*_{n,m}\big) - R_K(G^*_K)\big]}{\inf_{\tilde{G}_{n,m}}\sup_{(f,g)\in\mathscr{F}}\mathbf{E}_{f,g}\big[R_K\big(\tilde{G}_{n,m}\big) - R_K(G^*_K)\big]} = \rho_{n,m},$$

where $\rho_{n,m}$ *is a sequence with* $\rho_{n,m} = O(1)$ *and* $\rho_{n,m}^{-1} = O(1)$, *as* $n \wedge m \to \infty$. *Moreover*,

$$\inf_{\tilde{G}_{n,m}} \sup_{(f,g)\in\mathscr{F}} \mathbf{E}_{f,g}\big[R_K(\tilde{G}_{n,m}) - R_K(G_K^*)\big] = \tilde{\rho}_{n,m}(n \wedge m)^{-(1+\alpha)\gamma/[(2+\alpha)\gamma+\alpha(d-1)]},$$

where $\tilde{\rho}_{n,m}$ *is a sequence with* $\tilde{\rho}_{n,m} = O(1)$ *and* $\tilde{\rho}_{n,m}^{-1} = O(1)$, *as* $n \wedge m \to \infty$.

It is interesting that the optimal rate of convergence of Bayes risks is rather fast. It is always faster than the "parametric" rate $(n \wedge m)^{-1/2}$, whenever $\gamma > d - 1$, since $(1 + \alpha)\gamma/[(2 + \alpha)\gamma + \alpha(d - 1)] > 1/2$. This phenomenon is explained by the structure of the distances $d_{f,g}$ and $d_\triangle$ For G in a small vicinity of the true set G_K^* we have $d_{f,g}(G, G_K^*) \sim d_\triangle^{(1+\alpha)/\alpha}(G, G_K^*)$ (cf. Lemma 2 below). So also when the rates for $d_\triangle$ are slow, the power $(1 + \alpha)/\alpha$ accelerates the convergence for $d_{f,g}$. Note that $d_\triangle$ is the typical distance in set estimation problems; see, for example, Korostelev and Tsybakov (1993). We may conclude therefore that *classification is easier than set estimation* [see a related discussion for pattern recognition problems in Section 6.7 of Devroye, Györfi and Lugosi (1996)]. Furthermore, note that for two sets, G and G', it holds that $d_\triangle(G, G') = \|\mathbf{I}_G - \mathbf{I}_{G'}\|_Q^2$ where $\|\cdot\|_Q$ is the norm in $L_2(Q)$. This means that $d_\triangle$ is equal to the *squared* $L_2(Q)$-norm. The $L_2(Q)$ norm appears naturally in the proofs of Theorems 1 and 2. Clearly, in the $L_2(Q)$ norm, without squares, the convergence of decision rules is slower and, in particular, always slower than the "parametric" rate $(n \wedge m)^{-1/2}$.

We end this section by some remarks on generalizations and extensions of the results.

REMARK 1. The lower bounds of Theorem 3 hold trivially for classes $\mathscr{F}$ that contain a class $\mathscr{F}_{\text{frag}}$ of boundary fragments. For example, Theorem 3 and Corollary 1 remain valid if one replaces $\mathscr{F}_{\text{frag}}$ by $\mathscr{F}_{\text{Dudley}}$ where $\mathscr{F}_{\text{Dudley}}$ is the class $\mathscr{F}$ defined as in (3), with $\mathscr{G}$ being a Dudley class. This follows immediately from the fact that the Dudley class of sets with smoothness γ contains the class of boundary fragments with the same degree of smoothness. Thus, for the empirical rules (5), Bayes risks attain the optimal rates of convergence on Dudley classes, too.

REMARK 2. Theorem 3 can be easily extended to boundary fragments where the boundaries belong to other function classes. We now briefly discuss convex or monotone boundaries $b(\cdot)$, when $d = 2$. For the case of convex $b(\cdot)$, one should set $\gamma = 2$ and choose g_0 in the proof of Theorem 3 with a parabolic level profile instead of a constant profile [cf. the proof of Theorem 5.2 in Mammen and Tsybakov (1995)]. Together with Theorem 1, this shows rate optimality of the rule (5) for the case where

$$\mathscr{G} = \mathscr{G}_{\text{conv}} = \big\{\text{all closed convex subsets of } [0,1]^2\big\}.$$

The corresponding optimal rates are $(n \wedge m)^{-2\alpha/(4+3\alpha)}$ and $(n \wedge m)^{-2(1+\alpha)/(4+3\alpha)}$ for the distances $d_\triangle$ and $d_{f,g}$, respectively. For $\alpha = 1$, $m = n$ the optimal rate of convergence of the Bayes risks is $n^{-4/7}$, faster than the parametric rate $n^{-1/2}$.

If $\mathscr{G} = \mathscr{G}_{\mathrm{mon}} = \{G_b \subset [0,1]^2\colon\ b(\cdot)$ is monotone nondecreasing$\}$, then Theorem 3 remains valid with $\gamma = 1$. This easily follows if one performs the proof of Theorem 3 with a density g_0 having a linear level profile (instead of a constant profile). Furthermore, using entropy bounds for monotone functions with respect to the L_2 norm [see Mammen (1991) and van de Geer (1991)] we get that the assumptions of Theorem 2 are fulfilled with $\rho = 1$. This follows by the same argument as for boundary fragments with Hölder continuous boundary; see the discussion before the statement of Theorem 3. So, together with Theorem 2, this shows rate optimality of sieve estimates. As an interesting conclusion we get that for $\mathscr{G}_{\mathrm{mon}}$ the optimal rate for the Bayes risk is $n^{-1/2}$, independently of α.

We mention briefly some other straightforward generalizations. Analogous results hold for the choice of Bayes prior probabilities p and $1 - p$, with $p \neq \frac{1}{2}$; then the set G should be defined as $\{x\colon\ pf(x) \geq (1-p)g(x)\}$. Furthermore, pattern recognition problems, as studied in Vapnik (1996), Devroye, Györfi and Lugosi (1996), can easily be covered. Another generalization concerns models with more than two populations.

3. Proofs.

PROOF OF THEOREM 1. We give first the proof for $\rho < 1$. For this case we use a result of van de Geer (1998) that we state, for convenience, as a lemma. For related results, see also Birgé and Massart (1993, 1998), Barron, Birgé and Massart (1999).

LEMMA 1 [Van de Geer (1998), Lemma 5.13]. *For a probability measure P, let $\mathscr{H}$ be a class of uniformly bounded functions h in $L_2(P)$. Suppose that the δ-entropy with bracketing $H_B(\delta, \mathscr{H}, L_2(P))$ satisfies, for some $0 < \nu < 2$ and $A > 0$, the inequality*

$$H_B(\delta, \mathscr{H}, L_2(P)) \leq A\delta^{-\nu} \tag{24}$$

for all $\delta > 0$ small enough. Let h_0 be a fixed element in $\mathscr{H}$. Then there exist constants $D_1 > 0$, $D_2 > 0$ such that for a sequence of i.i.d. random variables $Z_1, \dots, Z_n$ with distribution P it holds that

$$P\left(\sup_{h \in \mathscr{H}} \frac{\left| n^{-1/2} \sum_{i=1}^n \{(h - h_0)(Z_i) - \mathbf{E}(h - h_0)(Z_i)\}\right|}{\left\{\|h - h_0\| \vee n^{-1/(2+\nu)}\right\}^{1-\nu/2}} > D_1 x \right) \leq D_2 e^{-x} \tag{25}$$

for $x \geq 1$. Here $\|\cdot\|$ denotes the $L_2(P)$-norm, and $x \vee y = \max(x, y)$.

W.l.o.g. assume in the sequel that $n \leq m$. For a given set G denote

$$h_G(x) = \mathbf{I}(x \in G), \qquad h_0(x) = \mathbf{I}(x \in G_K^*).$$

Note that

$$R_{n,m}(G) - R_{n,m}(G_K^*) = \frac{1}{2n}\sum_{i=1}^{n}(h_0 - h_G)(X_i) + \frac{1}{2m}\sum_{i=1}^{m}(h_G - h_0)(Y_i). \tag{26}$$

Clearly,

$$\mathbf{E}(R_{n,m}(G) - R_{n,m}(G_K^*)) = \tfrac{1}{2}d_{f,g}(G_K^*, G). \tag{27}$$

Here and later $\mathbf{E} = \mathbf{E}_{f,g}$, $\mathbf{P} = \mathbf{P}_{f,g}$ for brevity.

Observe also that

$$\|h_G - h_0\|_f^2 = \int_{G_K^* \triangle G} f(x)Q(dx) \le c_1 d_\triangle(G_K^*, G) \tag{28}$$

and

$$\|h_G - h_0\|_g^2 = \int_{G_K^* \triangle G} g(x)Q(dx) \le c_1 d_\triangle(G_K^*, G), \tag{29}$$

where $\|h\|_f^2 = \int h^2(x)f(x)Q(dx)$.

Consider the random variable

$$V_{n,m} = \sqrt{n}\,\frac{R_{n,m}(G_K^*) - R_{n,m}\big(\hat{G}_{n,m}\big) + d_{f,g}\big(\hat{G}_{n,m}, G_K^*\big)/2}{d_\triangle^{(1-\rho)/2}\big(\hat{G}_{n,m}, G_K^*\big)}.$$

By definition of $\hat{G}_{n,m}$ we have $R_{n,m}(\hat{G}_{n,m}) \le R_{n,m}(G_K^*)$. This implies

$$\sqrt{n}\,\frac{d_{f,g}\big(\hat{G}_{n,m}, G_K^*\big)}{d_\triangle^{(1-\rho)/2}\big(\hat{G}_{n,m}, G_K^*\big)} \le 2V_{n,m}. \tag{30}$$

We consider now the event

$$E = \Big\{d_\triangle\big(\hat{G}_{n,m}, G^*{}_K\big) > c_1^{-1}n^{-2/(2+\nu)}\Big\}$$

where $\nu = 2\rho$. Taking into account (26) and (27), we obtain that, if E holds,

$$\begin{aligned} V_{n,m} \le & \sup_{\substack{G\in\mathscr{G}:\, d_\triangle(G,G_K^*)\\ > c_1^{-1}n^{-2/(2+\nu)}}} \frac{\sqrt{n}\,\big|(1/2n)\Sigma_{i=1}^n\{(h_G - h_0)(X_i) - \mathbf{E}(h_G - h_0)(X_i)\}\big|}{d_\triangle^{(1-\rho)/2}(G, G_K^*)} \\ & + \sup_{\substack{G\in\mathscr{G}:\, d_\triangle(G,G_K^*)\\ > c_1^{-1}n^{-2/(2+\nu)}}} \frac{\sqrt{n}\,\big|(1/2m)\Sigma_{i=1}^m\{(h_G - h_0)(Y_i) - \mathbf{E}(h_G - h_0)(Y_i)\}\big|}{d_\triangle^{(1-\rho)/2}(G, G_K^*)} \\ \le & \sup_{h\in\mathscr{H}} \frac{\sqrt{n}\,\big|(1/2n)\Sigma_{i=1}^n\{(h - h_0)(X_i) - \mathbf{E}(h - h_0)(X_i)\}\big|}{c_1^{(\rho-1)/2}\big\{\|h - h_0\|_f \vee n^{-1/(2+\nu)}\big\}^{1-\nu/2}} \\ & + \sup_{h\in\mathscr{H}} \frac{\sqrt{m}\,\big|(1/2m)\Sigma_{i=1}^m\{(h - h_0)(Y_i) - \mathbf{E}(h - h_0)(Y_i)\}\big|}{c_1^{(\rho-1)/2}\big\{\|h - h_0\|_g \vee m^{-1/(2+\nu)}\big\}^{1-\nu/2}}, \end{aligned} \tag{31}$$

where $\mathscr{H} = \{h(x) = \mathbf{I}(x \in G): G \in \mathscr{G}\}$ and we used (28) and (29) to get the last inequality. Furthermore, (6), (28) and (29) entail that (24) holds with $\nu = 2\rho$. Thus, Lemma 1 can be applied, and consequently

$$\lim \sup_{n \wedge m \to \infty} \mathbf{E}\big[V_{n,m}^k \mathbf{I}(E)\big] \le C(k) \tag{32}$$

for all $k > 0$ and finite constants $C(k)$ depending on k.

We use now the following lemma.

LEMMA 2. *There exists a constant* $c(\alpha)$ *depending on* α *such that for Lebesgue measurable subsets* G_1 *and* G_2 *of* K *and for* $(f, g) \in \mathscr{F}$,

$$c(\alpha) d_{\triangle}^{(1+\alpha)/\alpha}(G_1, G_2) \le d_{f,g}(G_1, G_2) \le 2c_1 d_{\triangle}(G_1, G_2).$$

PROOF. The second inequality of the lemma is trivial. To prove the first inequality, note that the condition $Q(|f-g| < \eta) \le c_2\eta^{\alpha}$, $0 < \eta \le \eta_0$, and the boundedness of $Q(K)$ implies $Q(|f-g| < \eta) \le \tilde{c}_2\eta^{\alpha}$, $\forall\ \eta > 0$, where $\tilde{c}_2 > 0$ depends only on $c_2, \eta_0, Q(K)$ and α. Hence, choosing $\eta = [d_{\triangle}(G_1, G_2)/(2\tilde{c}_2)]^{1/\alpha}$, we get

$$\begin{aligned} d_{f,g}(G_1, G_2) &\ge \int_{G_1 \triangle G_2} |f-g| \mathbf{I}(|f-g| \ge \eta)\, dQ \\ &\ge \eta\big[Q(G_1 \triangle G_2) - Q(|f-g| < \eta)\big] \\ &\ge \eta d_{\triangle}(G_1, G_2) - \tilde{c}_2\eta^{1+\alpha} \\ &\ge c(\alpha) d_{\triangle}^{(1+\alpha)/\alpha}(G_1, G_2), \end{aligned}$$

where $c(\alpha) = 2^{-1-1/\alpha}(\tilde{c}_2)^{-1/\alpha}$. □

On E^c we have $d_{\triangle}(G_K, \hat{G}_{n,m}) \le c_1 n^{-1/(1+\rho)}$ and, because of the second inequality of Lemma 2, $d_{f,g}(G_K, \hat{G}_{n,m}) \le 2c_1^2 n^{-1/(1+\rho)}$. Since

$$n^{-1/(1+\rho)} = o(n^{-[1+\alpha]/[2+\alpha+\rho\alpha]})$$

and

$$n^{-1/(1+\rho)} = o(n^{-\alpha/[2+\alpha+\rho\alpha]}),$$

it suffices to consider the event E.

The first inequality of Lemma 2 and (30) imply

$$d_{\triangle}\big(G_K^*, \hat{G}_{n,m}\big) \le \big(2V_{n,m}/c(\alpha)\big)^{2\alpha/[2+\alpha+\rho\alpha]} n^{-\alpha/[2+\alpha+\rho\alpha]}. \tag{33}$$

Together with (32) this shows (7). Inequality (8) can be proved by plugging (33) into (30). Thus we have shown Theorem 1 for $0 < \rho < 1$.

We now come to the proof of Theorem 1 for $\rho \ge 1$. For this part of the proof we use the following result of Alexander (1984).

LEMMA 3 [(Alexander (1984)]. *Let a class* $\mathscr{H}$ *consist of functions with values in* $\{0, 1\}$, *let* P *be a probability measure and let* (24) *with* $\nu \ge 2$ *be satisfied. Then there exist constants* D_3, D_4, D_5, $D_6 > 0$ (*depending only on* A

and ν) such that for a sequence of i.i.d. random variables $Z_1, \dots, Z_n$ with distribution P it holds that

$$P\left(\sup_{h \in \mathscr{H}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{h(Z_i) - \mathbf{E}h(Z_i)\} \right| > xq_n \right) \le D_3 \exp\left(-D_4 x^2 q_n^2\right)$$

for $D_5 \le x \le D_6 n^{1/2}/q_n$ where

$$q_n = \begin{cases} n^{(\nu-2)/2(\nu+2)}, & \text{if } \nu > 2, \\ \log n, & \text{if } \nu = 2. \end{cases}$$

Here again $\|\cdot\|$ denotes the $L_2(P)$-norm.

Lemma 3 follows immediately from Corollary 2.4 in Alexander (1987). To see this, one chooses $\psi = \psi_1$ in (2.10) of Alexander (1984) and applies (1.6) of Alexander (1984).

Consider now the random variable

$$W_{n,m} = \sqrt{n}\, \frac{R_{n,m}(G_K^*) - R_{n,m}\left(\hat{G}_{n,m}\right) + d_{f,g}\left(\hat{G}_{n,m}, G_K^*\right)/2}{q_n}$$

where q_n is defined as in Lemma 3, with $\nu = 2\rho$. By definition of $\hat{G}_{n,m}$ we have $R_{n,m}(\hat{G}_{n,m}) \le R_{n,m}(G_K^*)$. This implies

$$\sqrt{n}\, q_n^{-1} d_{f,g}\left(\hat{G}_{n,m}, G_K^*\right) \le 2W_{n,m}. \tag{34}$$

We argue that

$$\mathbf{E}W_{n,m}^k \le C'(k) \tag{35}$$

for all $k > 0$ with some finite constants $C'(k)$ depending on k. For the proof of (35) note first that

$$\mathbf{E}W_{n,m}^k \le \mathbf{E}T_{n,m}^k$$

with

$$\begin{aligned} T_{n,m} &= \sup_{G \in \mathscr{G}} \sqrt{n}\, q_n^{-1} \left| R_{n,m}(G_K^*) - R_{n,m}(G) + d_{f,g}(G, G_K^*)/2 \right| \\ &\le \sup_{h \in \mathscr{H}} \sqrt{n} \left| \frac{1}{2nq_n} \sum_{i=1}^{n} \{(h - h_0)(X_i) - \mathbf{E}(h - h_0)(X_i)\} \right| \\ &\quad + \sup_{h \in \mathscr{H}} \sqrt{m} \left| \frac{1}{2mq_m} \sum_{i=1}^{m} \{(h - h_0)(Y_i) - \mathbf{E}(h - h_0)(Y_i)\} \right|, \end{aligned}$$

where we acted similarly to (31), and $\mathscr{H}$ is the same as in (31). Clearly, $T_{n,m} \le 3n^{1/2}/q_n$. This gives, for any $D > 0$,

$$\mathbf{E}W_{n,m}^k \le D^k + \left[3n^{1/2}/q_n\right]^k \mathbf{P}\{T_{n,m} > D\}.$$

Applying Lemma 3 we find that, if $D > 2D_5$, the last probability is bounded by $D_7 \exp(-D_8 D^2 q_n^2)$, for some positive constants D_7 and D_8. This entails (35).

Statement (8) of Theorem 1 (for $\rho \geq 1$) follows now from (34) and (35). For the proof of claim (7) one applies the first inequality of Lemma 2. □

PROOF OF THEOREM 2. Under the assumption (i) the result is shown by an easy modification of the proof of Theorem 1. We will consider only case (ii).

As in (26), (27) we get, for any subset G of K,

$$R_{n,m}(G) - R_{n,m}(G_K^*) - \tfrac{1}{2} d_{f,g}(G_K^*, G) = Z_{n,m}(G), \tag{36}$$

where

$$\begin{aligned} Z_{n,m}(G) &= \frac{1}{2n}\sum_{i=1}^{n} U_i(G) + \frac{1}{2m}\sum_{i=1}^{m} V_j(G),\\ U_i(G) &= (h_0 - h_G)(X_i) - \mathbf{E}(h_0 - h_G)(X_i),\\ V_j(G) &= (h_G - h_0)(Y_j) - \mathbf{E}(h_G - h_0)(Y_j). \end{aligned}$$

By Bernstein's inequality we have for all $a > 0$ and all $G \subseteq K$ that

$$\mathbf{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} U_i(G) \right| \geq a \right) \leq 2\exp\left(-\frac{k_1 n a^2}{a + d_{f,g}(G_K^*, G)^{\alpha/(1+\alpha)}} \right), \tag{37}$$

where $k_1 > 0$ is a constant. This holds because $|U_i(G)| \leq 2$ and $\mathbf{E}(U_i(G)^2) = \int_{G \triangle G_K^*} f\,dQ \leq c_1 d_\triangle(G_K^*, G) \leq c_1 c(\alpha)^{-\alpha/(1+\alpha)} d_{f,g}(G_K^*, G)^{\alpha/(1+\alpha)}$ where we used the assumption that f is bounded by c_1 (see the definition of $\mathscr{F}$) and Lemma 2.

We now apply (37) with $a = \frac{1}{8} d_{f,g}(G_K^*, G)$, and note that $d_{f,g}(G_K^*, G) \leq 2$. This gives, with a constant $k_2 > 0$,

$$\begin{aligned} &\mathbf{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} U_i(G) \right| \geq \frac{1}{8} d_{f,g}(G_K^*, G) \right)\\ &\qquad \leq 2\exp\left(-k_2 n d_{f,g}(G_K^*, G)^{(2+\alpha)/(1+\alpha)} \right). \end{aligned} \tag{38}$$

Using (38) and the analogous inequality with $V_j(G)$ in place of $U_i(G)$, we get

$$\mathbf{P}\left(|Z_{n,m}(G)| \geq \tfrac{1}{4} d_{f,g}(G_K^*, G) \right) \leq 2\exp\left(-k_3 (n \wedge m) d_{f,g}(G_K^*, G)^{(2+\alpha)/(1+\alpha)} \right)$$

for some $k_3 > 0$ and any $G \subseteq K$. Denote, for $t > 0$,

$$\mathscr{S} = \left\{ G \in \mathscr{N}_{n,m} : d_{f,g}(G_K^*, G) \geq t \tau_0(n,m)^{-(1+\alpha)/\alpha} \right\}.$$

We argue now that

$$\begin{aligned} &\mathbf{P}\left(\exists\, G \in \mathscr{S} : |Z_{n,m}(G)| \geq \tfrac{1}{4} d_{f,g}(G_K^*, G) \right)\\ &\quad \leq 2\exp\left(B\tau_0(n,m)^\rho \right) \exp\left(-k_3 (n \wedge m) t^{(2+\alpha)/(1+\alpha)} \tau_0(n,m)^{-(2+\alpha)/\alpha} \right)\\ &\quad \leq 2\exp\left(\left[B - k_3 t^{(2+\alpha)/(1+\alpha)} \right] \tau_0(n,m)^\rho \right)\\ &\quad \leq 2\exp\left(-(k_3/2) t \tau_0(n,m)^\rho \right)\ \forall\ t \geq \max(1, 2B/k_3), \end{aligned} \tag{39}$$

where the equality $(n \wedge m)\tau_0(n,m)^{-(2+\alpha)/\alpha} = \tau_0(n,m)^\rho$ has been used.

Choose now $G_{n,m} \in \mathscr{N}_{n,m}$ with

$$d_{f,g}(G_K^*, G_{n,m}) \le C\tau_0(n,m)^{-(1+\alpha)/\alpha}. \tag{40}$$

Note that for $t > 4C$ we have

$$\begin{aligned} &\frac{1}{4} d_{f,g}(G_K^*, G) - \frac{1}{2} d_{f,g}(G_K^*, G_{n,m}) \\ &\qquad \ge \frac{C}{2}\tau_0(n,m)^{-(1+\alpha)/\alpha} \qquad \forall\, G \in \mathscr{S}. \end{aligned} \tag{41}$$

In view of (36) and (41) we find, for any $t > 4C$,

$$\begin{aligned} &\mathbf{P}\Big(d_{f,g}\big(G_K^*, \overline{G}_{n,m}\big) \ge t\tau_0(n,m)^{-(1+\alpha)/\alpha}\Big) \\ &\quad \le \mathbf{P}(\exists\, G \in \mathscr{S}\colon R_{n,m}(G) - R_{n,m}(G_{n,m}) \le 0) \\ &\quad = \mathbf{P}\bigg(\exists\, G \in \mathscr{S}\colon \frac{1}{2} d_{f,g}(G_K^*, G) + Z_{n,m}(G) \\ &\qquad\qquad - \frac{1}{2} d_{f,g}(G_K^*, G_{n,m}) - Z_{n,m}(G_{n,m}) \le 0\bigg) \\ &\quad = \mathbf{P}\bigg(\exists\, G \in \mathscr{S}\colon Z_{n,m}(G) \le -\frac{1}{4} d_{f,g}(G_K^*, G)\bigg) \\ &\qquad + \mathbf{P}\bigg(Z_{n,m}(G_{n,m}) \ge \frac{C}{2}\tau_0(n,m)^{-(1+\alpha)/\alpha}\bigg). \end{aligned} \tag{42}$$

Now, by virtue of inequality (37) and its analogue with $V_j(G)$, and using (40) we obtain

$$\begin{aligned} &\mathbf{P}\bigg(Z_{n,m}(G_{n,m}) \ge \frac{C}{2}\tau_0(n,m)^{-(1+\alpha)/\alpha}\bigg) \\ &\quad \le 2\exp\left(-\frac{k_1 n (C/2)^2 \tau_0(n,m)^{-2(1+\alpha)/\alpha}}{(C/2)\tau_0(n,m)^{-(1+\alpha)/\alpha} + d_{f,g}(G_K^*, G_{n,m})^{\alpha/(1+\alpha)}}\right) \\ &\quad \le 2\exp\big(-k_4 \tau_0(n,m)^{\rho}\big) \end{aligned} \tag{43}$$

with some constant $k_4 > 0$. Combining (39), (42) and (43) we get

$$\begin{aligned} &\mathbf{P}\Big(d_{f,g}\big(G_K^*, \overline{G}_{n,m}\big) \ge t\tau_0(n,m)^{-(1+\alpha)/\alpha}\Big) \\ &\qquad \le 2\exp\big(-(k_3/2)t\tau_0(n,m)^{\rho}\big) + \exp\big(-k_4\tau_0(n,m)^{\rho}\big) \end{aligned} \tag{44}$$

for $t > \max(1, 2B/k_3, 4C)$. The result (17) of Theorem 2 for (ii) now follows from (44) and from the fact that $d_{f,g}(G_K^*, \overline{G}_{n,m}) \le 2$. Inequality (16) is a consequence of (17) and of Lemma 2. □

PROOF OF THEOREM 3. We first give the proof for $\mathscr{F} = \mathscr{F}_{\text{frag}}$. Suppose w.l.o.g. that $n \le m$. We consider the subset of $\mathscr{F}_{\text{frag}}$ that contains all pairs

(f, g_0), where g_0 is a fixed density on K and f belongs to a finite class of densities $\mathscr{F}_1$ that will be defined below. Then

$$\begin{aligned}
&\sup_{(f,g)\in\mathscr{F}_{\mathrm{frag}}} \mathbf{E}_{f,g} d_\triangle^p\big(\tilde{G}_{n,m}, G\big)\\
&\qquad \geq \sup_{(f,g_0):\, f\in\mathscr{F}_1} \mathbf{E}_{f,g_0} d_\triangle^p\big(\tilde{G}_{n,m}, G\big) \qquad (45)\\
&\qquad \geq \mathbf{E}_{g_0}\left[\frac{1}{\#\mathscr{F}_1}\sum_{f\in\mathscr{F}_1} \mathbf{E}_f\Big\{d_\triangle^p\big(\tilde{G}_{n,m}, G\big)|Y_1,\dots,Y_m\Big\}\right],
\end{aligned}$$

where $\mathbf{E}_f$ and $\mathbf{E}_{g_0}$ denote the expectations w.r.t the distributions of $(X_1,\dots,X_n)$ and $(Y_1,\dots,Y_m)$ when the underlying densities are f and g_0. Here and later, $\#\mathscr{F}_1$ denotes the number of elements of $\mathscr{F}_1$.

For simplicity, we give the proof only for the case $d = 2$ (an extension to higher dimensions is straightforward). For this case, it suffices to bound the term in squared brackets in (45) from below by $cn^{-\alpha\gamma p/[(2+\alpha)\gamma+\alpha]}$, where $c > 0$ is a constant that does not depend on the sample $Y_1,\dots,Y_m$. This would prove (22). The lower bound (23) follows from (22) and Lemma 2. Furthermore, it suffices to consider the case $p = 1$, since it implies the result for $p \geq 1$ by application of the Hölder inequality. Hence for the proof of the theorem it suffices to show that for any estimator $\tilde{G}_{n,m}$ and any n, m large enough,

$$n^{\alpha\gamma/[(2+\alpha)\gamma+\alpha]}\frac{1}{\#\mathscr{F}_1}\sum_{f\in\mathscr{F}_1} \mathbf{E}_f\Big\{d_\triangle\big(\tilde{G}_{n,m}, G\big)|Y_1,\dots,Y_m\Big\} \geq c \qquad \text{a.s.,} \tag{46}$$

where $c > 0$ does not depend on n, m (and $Y_1,\dots,Y_m$).

Before we come to the proof of (46) we define g_0 and the class $\mathscr{F}_1$. For this purpose, let φ be an infinitely many times differentiable function on $\mathbb{R}^1$ with the following properties: $\varphi(t) = 0$ for $|t| \geq 1$, $\varphi(t) \geq 0$ for all t, $\max \varphi(t) = 1$ and $\varphi(0) = 1$. For a fixed integer $M \geq 2$ and a constant τ with $0 < \tau < 1$ we define $b_1 = [\tau/c_2]^{1/\alpha}M^{-\gamma/\alpha}$. For $x = (x_1, x_2)$ in $K = [0,1]^2$ we now put

$$\begin{aligned}
g_0(x) = (1 + \eta_0 + b_1)\mathbf{I}\big\{0 < x_2 < \tfrac{1}{2}\big\} + \mathbf{I}\big\{\tfrac{1}{2} \leq x_2 < \tfrac{1}{2} + \tau M^{-\gamma}\big\}\\
+(1 - \eta_0 - b_2)\mathbf{I}\big\{\tfrac{1}{2} + \tau M^{-\gamma} \leq x_2 \leq 1\big\},
\end{aligned}$$

where $b_2 > 0$ is chosen such that $\int g_0(x)\,dx = 1$ and $0 < \eta_0 < 1/2$. W.l.o.g. we assume that c_1 [see (3)] is large enough, so that $g_0(x) < c_1$ for all x in K. For $j = 1,\dots,M$ we put

$$\varphi_j(t) = \tau M^{-\gamma}\varphi\left(M\left[t - \frac{2j-1}{M}\right]\right).$$

For vectors $\omega = (\omega_1,\dots,\omega_M)$ of elements $\omega_j \in \{0,1\}$ and for $t \in [0,1]$, we define

$$b(t,\omega) = \tfrac{1}{2} + \sum_{j=1}^{M} \omega_j\varphi_j(t).$$

Put $\Omega = \{0,1\}^M$. With this notation, define for $\omega \in \Omega$ and $x \in [0,1]^2$,

$$f_\omega(x) = g_0(x) + \left[\frac{b(x_1,\omega) - x_2}{c_2}\right]^{1/\alpha} \mathbf{I}\left\{\frac{1}{2} \le x_2 \le b(x_1,\omega)\right\}$$
$$- b_3(\omega)\mathbf{I}\left\{\frac{1}{2} + \tau M^{-\gamma} < x_2 \le 1\right\},$$

where $b_3(\omega) > 0$ is chosen such that $\int f_\omega(x)\,dx = 1$. Set now

$$\mathscr{F}_1 = \{f_\omega : \omega \in \Omega\}.$$

Let us first show that

$$(f_\omega, g_0) \in \mathscr{F}_{\text{frag}} \tag{47}$$

for all $\omega \in \Omega$.

Proof of (47). The equality $\int [g_0(x) - f_\omega(x)]\,dx = 0$ entails

$$\int_0^1 \int_{1/2}^{b(x_1,\omega)} \left[\frac{b(x_1,\omega) - x_2}{c_2}\right]^{1/\alpha} dx_2\,dx_1 = b_3(\omega)\left[\frac{1}{2} - \tau M^{-\gamma}\right].$$

This gives

$$\begin{aligned} b_3(\omega) &= \frac{1}{\frac{1}{2} - \tau M^{-\gamma}} \sum_{j=1}^M \omega_j \int_0^1 \int_{1/2}^{1/2+\varphi_j(t)} \left[\frac{\frac{1}{2} + \varphi_j(t) - u}{c_2}\right]^{1/\alpha} du\,dt \\ &= \frac{c_2^{-1/\alpha}}{\frac{1}{2} - \tau M^{-\gamma}} \sum_{j=1}^M \omega_j \int_0^1 \int_0^{\varphi_j(t)} v^{1/\alpha}\,dv\,dt \\ &= \frac{c_2^{-1/\alpha}}{\frac{1}{2} - \tau M^{-\gamma}} \frac{\alpha}{\alpha+1} \sum_{j=1}^M \omega_j \int_0^1 \varphi_j(t)^{1+\alpha^{-1}}\,dt \qquad (48) \\ &\le \frac{c_2^{-1/\alpha}}{\frac{1}{2} - \tau M^{-\gamma}} \frac{\alpha}{\alpha+1} M[\tau M^{-\gamma}]^{1+\alpha^{-1}} \int \varphi(Mt)^{1+\alpha^{-1}}\,dt \\ &= O(M^{-\gamma(1+\alpha^{-1})}). \end{aligned}$$

Hence $f_\omega \le c_1$ for c_1 and M large enough. Next, the set

$$\{x : f_\omega(x) \ge g_0(x)\} = \{x : 0 \le x_2 \le b(x_1,\omega)\}$$

belongs to $\mathscr{G}_{\text{frag}}$ since $b(\cdot,\omega) \in \Sigma(\gamma, L)$ for $\tau > 0$ small enough. To guarantee (47), it remains to show

$$\lambda\{x \in K : |f_\omega(x) - g_0(x)| \le \eta\} \le c_2\eta^\alpha,$$

for $0 < \eta \le \eta_0$. But this follows from the fact that, for $0 < \eta \le \eta_0$,

$$\begin{aligned}
&\{x \in K: |f_\omega(x) - g_0(x)| \le \eta\} \\
&\quad = \left\{ x \in K: 1/2 \le x_2 \le b(x_1, \omega), \left[\frac{b(x_1, \omega) - x_2}{c_2}\right]^{1/\alpha} \le \eta \right\} \\
&\quad = \{x \in K: b(x_1, \omega) - c_2 \eta^\alpha \le x_2 \le b(x_1, \omega)\}.
\end{aligned}$$

Proof of (46). We use Assouad's lemma [see Bretagnolle and Huber (1979) and Assouad (1983)]. For our purposes it will be more convenient to apply the version of this lemma stated in Korostelev and Tsybakov (1993), which is adapted to the problem of estimation of sets.

For $j = 1, \dots, M$ and for a vector $\omega = (\omega_1, \dots, \omega_M)$, we write

$$\begin{aligned}
\omega_{j0} &= (\omega_1, \dots, \omega_{j-1}, 0, \omega_{j+1}, \dots, \omega_M), \\
\omega_{j1} &= (\omega_1, \dots, \omega_{j-1}, 1, \omega_{j+1}, \dots, \omega_M).
\end{aligned}$$

For $i = 0$ and $i = 1$, let P_{ji} be the probability measure corresponding to the distribution of $X_1, \dots, X_n$ when the underlying density is $f_{\omega_{ji}}$. The expectation w.r.t. P_{ji} is denoted by $\mathbf{E}_{ji}$. Arguing as in (5.3)–(5.6) in Korostelev and Tsybakov (1993), we find that the sum

$$S = \frac{1}{\#\mathscr{F}_1} \sum_{f \in \mathscr{F}_1} \mathbf{E}_f \left\{ d_\triangle \left(\tilde{G}_{n,m}, G \right) | Y_1, \dots, Y_m \right\}$$

is bounded as follows:

$$\begin{aligned}
S &\ge \tfrac{1}{2} \sum_{j=1}^{M} \lambda\left\{x: \tfrac12 \le x_2 \le \tfrac12 + \varphi_j(x_1)\right\} \int \min\{dP_{j1}, dP_{j0}\} \\
&= \tfrac12 \sum_{j=1}^{M} \tau M^{-\gamma} \int \varphi(Mt)\, dt \int \min\{dP_{j1}, dP_{j0}\}.
\end{aligned} \tag{49}$$

Now,

$$\int \min\{dP_{j1}, dP_{j0}\} \ge \tfrac12 \left[1 - H^2(P^0, P^1)/2\right]^n,$$

where $H(\cdot, \cdot)$ denotes the Hellinger distance and P^0, P^1 denote the probability distributions of X_1 under the densities $f_{\omega_{10}}$ or $f_{\omega_{11}}$, respectively. We have

$$\begin{aligned}
&H^2(P^0, P^1) \\
&\quad = \int \left[\sqrt{f_{\omega_{10}}(x)} - \sqrt{f_{\omega_{11}}(x)} \right]^2 dx \\
&\quad = \int_0^1 \left[\int_{1/2}^{(1/2)+\varphi_1(x_1)} \left\{ 1 - \sqrt{1 + \left(\frac{\frac12 + \varphi_1(x_1) - x_2}{c_2} \right)^{1/\alpha}} \right\}^2 dx_2 \right.
\end{aligned}$$

(50)

$$\left. + \int_{(1/2)+\tau M^{-\gamma}}^{1} \left\{ \sqrt{1-\eta_0-b_2-b_3(\omega_{10})} - \sqrt{1-\eta_0-b_2-b_3(\omega_{11})} \right\}^2 dx_2 \right] dx_1$$

$$\le \int_0^1 \left[\int_0^{\varphi_1(x_1)} \left\{ 1 - \sqrt{1 + \left(\frac{v}{c_2}\right)^{1/\alpha}} \right\}^2 dv \right] dx_1 + \frac{1}{2} \left| b_3(\omega_{10}) - b_3(\omega_{11}) \right|^2.$$

Here

$$\begin{aligned} &\int_0^1 \int_0^{\varphi_1(x_1)} \left\{ 1 - \sqrt{1 + \left(\frac{v}{c_2}\right)^{1/\alpha}} \right\}^2 dv\, dx_1 \\ &\qquad \le \frac{1}{2} \int_0^1 \int_0^{\varphi_1(x_1)} \left(\frac{v}{c_2}\right)^{2/\alpha} dv\, dx_1 \\ &\qquad = \frac{\alpha}{2(\alpha+2)} c_2^{-2/\alpha} \int_0^1 [\varphi_1(x_1)]^{1+2\alpha^{-1}}\, dx_1 \qquad (51) \\ &\qquad = \frac{\alpha}{2(\alpha+2)} c_2^{-2/\alpha} [\tau M^{-\gamma}]^{1+2\alpha^{-1}} \int [\varphi(Mt)]^{1+2\alpha^{-1}}\, dt \\ &\qquad \le C^* M^{-\gamma(1+2\alpha^{-1})-1}, \end{aligned}$$

where C^* depends only on α, c_2, τ and φ.

On the other hand, similarly to (48), one gets

$$\begin{aligned} |b_3(\omega_{10}) - b_3(\omega_{11})| &\le \frac{c_2^{-1/\alpha}}{\frac{1}{2} - \tau M^{-\gamma}} \frac{\alpha}{\alpha+1} \int_0^1 [\varphi_j(t)]^{1+\alpha^{-1}}\, dt \qquad (52) \\ &= O(M^{-\gamma(1+\alpha^{-1})-1}). \end{aligned}$$

Combining (50)–(52), one gets

$$H^2(P^0, P^1) \le C^* M^{-\gamma(1+2\alpha^{-1})-1} [1 + o(1)].$$

Choose now M as the smallest integer that is larger or equal to $n^{\alpha/[(2+\alpha)\gamma+\alpha]}$. Then

$$H^2(P^0, P^1) \le C^* n^{-1} [1 + o(1)].$$

This gives, with a constant $C_1^* > 0$,

$$\int \min\{dP_{j1}, dP_{j0}\} \ge \frac{1}{2} \left[1 - \frac{C^*}{2} n^{-1} \{1 + o(1)\} \right]^n \ge C_1^*$$

for all n large enough. This inequality and (49) yield

$$S \ge \tfrac{1}{2} C_1^* \tau M^{-\gamma} \int \varphi(t)\, dt \ge C_2^* n^{-\alpha\gamma/[(2+\alpha)\gamma+\alpha]},$$

for all n large enough. The constant $C_2^* > 0$ depends only on α, c_2, τ and φ.

This completes the proof of (46). Thus, the theorem is proved for $\mathscr{F} = \mathscr{F}_{\mathrm{frag}}$.

The proof for $\mathscr{F} = \mathscr{F}'_{\mathrm{frag}}$ follows the same lines. We only have to modify the definition of f_ω: in fact, now $f_\omega - g_0$ must satisfy the Lipschitz condition in the definition of $\mathscr{F}'_{\mathrm{frag}}$. We therefore set (recall that now $\alpha = 1$):

$$f_\omega(x) = g_0(x) + \min\left\{C\left(x_2 - \frac{1}{2}\right), \frac{b(x_1, \omega) - x_2}{c_2}\right\} I\left\{\frac{1}{2} \le x_2 \le b(x_1, \omega)\right\}$$
$$- b'_3(\omega)\min\left\{C\left(x_2 - \frac{1}{2} - \tau M^{-\gamma}\right), 1\right\}\mathbf{I}\left\{\frac{1}{2} + \tau M^{-\gamma} < x_2 \le 1\right\},$$

where C is the constant from the definition of $\mathscr{F}'_{\mathrm{frag}}$, and $b'_3(\omega)$ is an appropriate constant chosen so that f_ω is a probability density. With this new definition of f_ω the above calculations carry through, up to simple modifications, and we omit them. □

REFERENCES

ALEXANDER, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12** 1041–1067. [Correction: (1987) **15** 428–430.]

ASSOUAD, P. (1983). Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris* **296** 1021–1024.

BARRON, A. (1991). Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed.) 561–576. Kluwer, Dordrecht.

BARRON, A. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning* **14** 115–133.

BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413.

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150.

BIRGÉ, L. and MASSART, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4** 329–375.

BLOCH, D. A. and SILVERMAN, B. W. (1997). Monotone discriminant functions and their applications in rheumathology. *J. Amer. Statist. Assoc.* **92** 144–153.

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densités: risque minimax. *Z. Warsch. Verw. Gebiete* **47** 119–137.

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

DUDLEY, R. M. (1974). Metric entropy of some classes of sets with differentiable boundaries. *J. Approx. Theory* **10** 227–236.

HARTIGAN, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267–270.

KOROSTELEV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction. Lecture Notes in Statist.* **82**. Springer, New York.

MAMMEN, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759.

MAMMEN, E. and TSYBAKOV, A. B. (1995). Asymptotic minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23** 502–524.

MARRON, J. S. (1983). Optimal rates of convergence to Bayes risk in nonparametric discrimination. *Ann. Statist.* **11** 1142–1155.

MÜLLER, D. W. (1993). The excess mass approach in statistics. *Beiträge zur Statistik* **3**. Inst. Math. für Angewandte, Univ. Heidelberg.

MÜLLER, D. W. (1995). A backward-induction algorithm for computing the best convex contrast of two bivariate samples. *Beiträge zur Satistik* **29**. Inst. für Angewandte, Univ. Heidelberg.

MÜLLER, D. W. and SAWITZKI, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86** 738–746.

POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters: an excess mass approach. *Ann. Statist.* **23** 855–881.

RUDEMO, M. and STRYHN, H. (1994). Approximating the distributions of maximum likelihood contour estimates in two-region images. *Scand. J. Statist.* **21** 41–56.

TSYBAKOV, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969.

VAN DE GEER, S. (1991). The entropy bound for monotone functions. Technical Report 91–100. Univ. Leiden.

VAN DE GEER, S. (1995). The method of sieves and minimum contrast estimates. *Math. Methods Statist.* **4** 20–38.

VAN DE GEER, S. (1998). *Applications of Empirical Process Theory to M-estimation*. Unpublished manuscript.

VAPNIK, V. N. (1996). *The Nature of Statistical Learning Theory*. Springer, New York.

VAPNIK, V. N. and CHERVONENKIS, A. JA. (1974). *Theory of Pattern Recognition*. Nauka, Moscow (in Russian).

WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362.

INSTITUT FÜR ANGEWANDTE MATHEMATIK
UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 294
69120 HEIDELBERG
GERMANY
E-MAIL: enno@statlab.uni-heidelberg.de

LABORATOIRE DE PROBABILITÉS
ET MODÈLES ALÉATOIRES
UMR CNRS 7599
UNIVERSITÉ PARIS VI
BP 188, 4 PLACE DE JUSSUEU
F-75252 PARIS
FRANCE