The Annals of Statistics
2004, Vol. 32, No. 1, 135–166

OPTIMAL AGGREGATION OF CLASSIFIERS IN STATISTICAL LEARNING

BY ALEXANDRE B. TSYBAKOV

Université Paris 6

Classification can be considered as nonparametric estimation of sets, where the risk is defined by means of a specific distance between sets associated with misclassification error. It is shown that the rates of convergence of classifiers depend on two parameters: the complexity of the class of candidate sets and the margin parameter. The dependence is explicitly given, indicating that optimal *fast* rates approaching $O(n^{-1})$ can be attained, where n is the sample size, and that the proposed classifiers have the property of robustness to the margin. The main result of the paper concerns optimal aggregation of classifiers: we suggest a classifier that automatically adapts both to the complexity and to the margin, and attains the optimal fast rates, up to a logarithmic factor.

1. Introduction. Let (X_i, Y_i), $i = 1, \dots, n$, be i.i.d. random pairs of observations, where $X_i \in \mathbf{R}^d$ and $Y_i \in \{0, 1\}$. Denote by P_X the probability distribution of X_i and by $\pi = \pi_{X,Y}$ the joint distribution of (X_i, Y_i).

Let (X, Y) be a random pair distributed according to π and independent of $(X_1, \dots, X_n, Y_1, \dots, Y_n)$, and let the component X of the pair be observed. The problem of statistical learning in classification (pattern recognition) consists of predicting the corresponding value $Y \in \{0, 1\}$. A prediction rule decides that $Y = 1$ if $X \in G$ and $Y = 0$ if $X \notin G$, where G is a Borel subset of $\mathbf{R}^d$. The corresponding *classifier* is $\hat{Y} = I(X \in G)$, where $I(\cdot)$ denotes the indicator function. Since a classifier is uniquely determined by the set G, the name classifier will be attributed without loss of generality to G as well.

The misclassification error associated with G is

$$R(G) = P(Y \neq \hat{Y}) = P\big(Y \neq I(X \in G)\big) = E\big[\big(Y - I(X \in G)\big)^2\big].$$

It is well known [see, e.g., Devroye, Györfi and Lugosi (1996)] that

$$\min_G R(G) = R(G^*),$$

where

$$G^* = G^*_\pi = \{x : \eta(x) \geq 1/2\} \tag{1}$$

Received September 2001; revised February 2003.

AMS 2000 subject classifications. Primary 62G07; secondary 62G08, 62H30, 68T10.

Key words and phrases. Classification, statistical learning, aggregation of classifiers, optimal rates, empirical processes, margin, complexity of classes of sets.

and

$$\eta(x) = E(Y|X = x) = P(Y = 1|X = x).$$

In general, $R(G^*) \neq 0$, and the efficiency of a classifier G is measured by the difference

$$R(G) - R(G^*) = \int_{G \triangle G^*} |2\eta(x) - 1| P_X(dx) \stackrel{\text{def}}{=} d(G, G^*). \tag{2}$$

Note that $d(G, G^*)$ is a pseudodistance between the sets G and G^*; that is, it satisfies the axioms of the distance except $d(G, G') = 0 \Rightarrow G = G'$.

A classifier based on the data $(X_1, \dots, X_n, Y_1, \dots, Y_n)$ is denoted by

$$\hat{G}_n = \hat{G}_n(X_1, \dots, X_n, Y_1, \dots, Y_n).$$

A key characteristic of $\hat{G}_n$ is the value $R(\hat{G}_n)$ known in learning theory under the name of *generalization error*:

$$R(\hat{G}_n) = P\big(Y \neq I(X \in \hat{G}_n) \big| X_1, \dots, X_n, Y_1, \dots, Y_n\big).$$

The aim of statistical learning is to construct a classifier $\hat{G}_n$ such that $d(\hat{G}_n, G^*) = R(\hat{G}_n) - R(G^*)$ is as small as possible. Since $\hat{G}_n$ is random, the smallness of $d(\hat{G}_n, G^*)$ will be expressed in terms of the expected risk

$$E_{\pi,n}\big(d(\hat{G}_n, G^*)\big) = E_{\pi,n}\big(R(\hat{G}_n) - R(G^*)\big).$$

Here and later $E_{\pi,n}$ denotes the expectation w.r.t. the joint distribution $P_{\pi,n}$ of $(X_1, \dots, X_n, Y_1, \dots, Y_n)$.

Two basic families of classifiers are the plug-in rules and the empirical risk minimization (ERM) rules [see, e.g., Devroye, Györfi and Lugosi (1996) and Vapnik (1998)]. Plug-in classifiers have the form

$$\hat{G}_n = \{x : \hat{\eta}_n(x) \geq 1/2\},$$

where $\hat{\eta}_n(x)$ is a nonparametric estimator of the regression function $\eta(x)$ [i.e., $\hat{\eta}_n$ is plugged into (1) instead of η]. ERM classifiers are defined as

$$\hat{G}_n = \arg\min_{G \in \mathcal{C}} R_n(G), \tag{3}$$

where R_n is the empirical risk

$$R_n(G) = \frac{1}{n} \sum_{i=1}^{n} I\big(Y_i \neq I(X_i \in G)\big) = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - I(X_i \in G)\big)^2,$$

and $\mathcal{C}$ is a given collection of subsets of $\mathbf{R}^d$. Statistical properties of these two types of classifiers as well as of other related ones have been extensively studied [see Aizerman, Braverman and Rozonoer (1970), Vapnik and Chervonenkis (1974), Vapnik (1982, 1998), Breiman, Friedman, Olshen and Stone (1984), Devroye, Györfi and Lugosi (1996), Anthony and Bartlett (1999), Cristianini and

Shawe-Taylor (2000) and Schölkopf and Smola (2002) and the references therein]. Results concerning the convergence of the risk $E_{\pi,n}(d(\hat{G}_n, G^*))$ obtained in the literature are of the form

$$E_{\pi,n}\bigl(d(\hat{G}_n, G^*)\bigr) = O(n^{-\beta}), \qquad n \to \infty,$$

where $\beta > 0$ is some exponent, and typically $\beta \le 1/2$ if $R(G^*) \ne 0$. Mammen and Tsybakov (1999) first showed that one can attain fast rates, approaching n^{-1}, that is, with β arbitrarily close to 1 under mild assumptions on the behavior of $\eta(x)$ in a neighborhood of the boundary $\partial G^* = \{x : \eta(x) = 1/2\}$. For further results about the fast rates see Massart (2000) and Catoni (2001). The effect of behavior of the regression function η around ∂G^* on the classification error has been discussed earlier under different assumptions by Devroye, Györfi and Lugosi (1996) and Horváth and Lugosi (1998). Mammen and Tsybakov (1999) considered nonparametric discrimination which is slightly different from the pattern recognition problem studied here, but translation of their results in terms of pattern recognition is straightforward. They obtained lower bounds on the risks and optimal exponents β and showed that the optimal rates are determined by two parameters: the *complexity* of the class $\mathcal{G}^*$ of possible sets G^* and the *margin parameter* that characterizes the behavior of $\eta(x)$ in a neighborhood of the boundary ∂G^*. For "massive" sets $\mathcal{G}^*$, they showed that optimal rates are attained either by ERM over the class of candidates $\mathcal{C}$ that coincides with $\mathcal{G}^*$ or by ERM over a sieve $\mathcal{C}$ on $\mathcal{G}^*$ (in the latter case the margin parameter is supposed to be known and $\mathcal{C}$ depends on it). This allows one to establish optimal rates, but the classifiers are not always feasible: solution of (3) for "massive" sets $\mathcal{C} = \mathcal{G}^*$ is known only for a few examples (sets with convex or monotone boundaries); if $\mathcal{C}$ is a finite sieve, such a solution is available, but assuming the margin parameter to be known is not always realistic.

The aim of this paper is twofold. First, it will be shown that the results of Mammen and Tsybakov (1999) can be extended to feasible classifiers: for some finite sieves $\mathcal{C} = \mathcal{N}$ on $\mathcal{G}^*$ that *do not depend on the margin parameter* the ERM classifier (3) attains the optimal rate (fast rate, up to n^{-1}). This can be interpreted as *robustness to the margin* property of statistical learning in the case where $\mathcal{G}^*$, a class of sets containing G^*, is known.

Second, an adaptive setup with unknown class $\mathcal{G}^*$ will be studied. It will only be assumed that $\mathcal{G}^*$ is a member of a known collection $\{\mathcal{G}_1, \dots, \mathcal{G}_N\}$, where $\mathcal{G}_j$ are different classes of sets. Let $G_{n1}, \dots, G_{nN}$ denote the respective optimal classifiers. The aim is optimal aggregation of these classifiers, that is, selection of a data-dependent $\hat{j}$ such that:

1. The adaptive classifier $G_{n\hat{j}}$ attains the optimal (fast) rate, up to a logarithmic factor, simultaneously on all the classes $\mathcal{G}_1, \dots, \mathcal{G}_N$, or, equivalently, it mimics the rate of the best among the classifiers $G_{n1}, \dots, G_{nN}$, up to a logarithmic factor.

2. The definition of $\hat{j}$ is independent of the "true" set $\mathcal{G}^*$ and of the margin parameter.

Such an optimal aggregation procedure is suggested below. In particular, the adaptive classifier can attain the rates up to n^{-1} (depending on the configuration complexity/margin), and it has the *robustness to the margin* property, as in the case of known $\mathcal{G}^*$.

2. Results for known $\mathcal{G}^*$. In this section we suppose that a class $\mathcal{G}^*$ is known such that $G^* \in \mathcal{G}^*$. Introduce the following pseudodistance between the sets $G, G' \subseteq \mathbf{R}^d$:

$$d_\triangle(G, G') = P_X(G \triangle G').$$

Clearly,

$$d(G, G') \le d_\triangle(G, G') \le 1. \tag{4}$$

The relation between $d(G, G^*)$ and $d_\triangle(G, G^*)$ is crucial to characterize the margin, that is, the distribution of the random variable $\eta(X)$ when X is near the boundary $\partial G^* = \{x : \eta(x) = 1/2\}$. Assume the following:

(A1) Assumption on the margin. There exist $\kappa \ge 1$, $c_0 > 0$, $0 < \varepsilon_0 \le 1$ such that

$$d(G, G^*) \ge c_0 d_\triangle^\kappa(G, G^*)$$

for all G such that $d_\triangle(G, G^*) \le \varepsilon_0$.

The parameter κ appearing in assumption (A1) is called the *margin parameter*. The values $\kappa < 1$ are impossible in view of (4). As opposed to data-dependent notions of margin that have been introduced in the literature on classification recently [see, e.g., Schölkopf and Smola (2002)], assumption (A1) is a condition on the joint distribution of X and Y. The following proposition explains the origin of assumption (A1) and its relation to the behavior of $\eta(x)$ near the level $1/2$.

PROPOSITION 1. *Assume that*

$$P\big(|\eta(X) - 1/2| \le t\big) \le C_\eta t^\alpha \tag{5}$$

for some finite $C_\eta > 0$, $\alpha > 0$ and all $0 < t \le t_*$, where $t_* \le 1/2$. Then assumption (A1) *holds with $c_0 = 2C_\eta^{-1/\alpha}\alpha(\alpha+1)^{-1-1/\alpha}$, $\varepsilon_0 = C_\eta(\alpha+1)t_*^\alpha$ and*

$$\kappa = \frac{1+\alpha}{\alpha}. \tag{6}$$

PROOF. For any $t > 0$ and $\mathcal{A} = \{x : |\eta(x) - 1/2| > t\}$, in view of (2),

$$\begin{aligned} d(G, G^*) &\geq 2t\, P_X\big((G \triangle G^*) \cap \mathcal{A}\big) \\ &\geq 2t\big[P_X(G \triangle G^*) - P_X(\bar{\mathcal{A}})\big] \\ &\geq 2t\big[d_\triangle(G, G^*) - C_\eta t^\alpha\big]. \end{aligned}$$

Maximizing the last expression with respect to t we get the result. Note that the maximizer is $t_0 = [d_\triangle(G, G^*)/(\alpha + 1)C_\eta)]^{1/\alpha}$, and thus we have $t_0 \leq t_*$ for $d_\triangle(G, G^*) \leq \varepsilon_0$ if $\varepsilon_0 = C_\eta(\alpha + 1)t_*^\alpha$. $\square$

If $d = 1$ and X has a bounded density w.r.t. Lebesgue measure and the boundary ∂G^* reduces, for example, to one point 0, (5) may be interpreted as follows: $\eta(x) - 1/2 \sim x^{1/\alpha}$ for x close to 0. Then the best situation for learning is when $\alpha \to \infty$ ($\kappa \to 1$) and it corresponds to a jump of $\eta(x)$ at the boundary ∂G^*. The worst case corresponds to a plateau type behavior of $\eta(x)$ near $\partial G^* : \alpha \to 0$ (i.e., $\kappa \to \infty$). A typical intermediate case is $\alpha = 1$ (i.e., $\kappa = 2$).

The second basic assumption concerns complexity of the class of candidate sets $\mathcal{G}^*$. We will first need some definitions.

DEFINITION 1. Let $\delta > 0$ be a given number, let $\bar{d}(\cdot, \cdot)$ be a pseudodistance between subsets of $\mathbf{R}^d$ and let $\mathcal{N}$ and $\mathcal{G}$ be classes of subsets of $\mathbf{R}^d$ such that for any $G \in \mathcal{G}$ there exists $G^{\mathcal{N}} \in \mathcal{N}$ satisfying $\bar{d}(G, G^{\mathcal{N}}) \leq \delta$. Then $\mathcal{N}$ is called *δ-net* on $\mathcal{G}$ for the pseudodistance $\bar{d}$.

DEFINITION 2. Let $\delta > 0$ be a given number, let $\mathcal{G}$ be a class of subsets of $\mathbf{R}^d$ and let $\bar{d}(\cdot, \cdot)$ be a pseudodistance between subsets of $\mathbf{R}^d$. Let $N_B(\delta, \mathcal{G}, \bar{d})$ be the smallest value m for which there exist pairs of sets (G_j^L, G_j^U), $j = 1, \ldots, m$, such that $\bar{d}(G_j^L, G_j^U) \leq \delta$ for all $j = 1, \ldots, m$, and for any $G \in \mathcal{G}$ there exists $j(G) \in \{1, \ldots, m\}$ for which $G_{j(G)}^L \subseteq G \subseteq G_{j(G)}^U$. Then the value $\mathcal{H}_B(\delta, \mathcal{G}, \bar{d}) = \log N_B(\delta, \mathcal{G}, \bar{d})$ is called the *δ-entropy with bracketing* of $\mathcal{G}$ for the pseudodistance $\bar{d}$.

DEFINITION 3. A class $\mathcal{G}$ of subsets of $\mathbf{R}^d$ is said to have *complexity bound* $\rho > 0$ if there exists a constant $A > 0$ such that

$$\mathcal{H}_B(\delta, \mathcal{G}, d_\triangle) \leq A\delta^{-\rho} \qquad \forall\, 0 < \delta \leq 1,$$

where $\mathcal{H}_B(\delta, \mathcal{G}, d_\triangle)$ is the δ-entropy with bracketing of $\mathcal{G}$ for the pseudodistance $d_\triangle$.

Note that the restriction $0 < \delta \leq 1$ in this definition is natural since $d_\triangle$ is always less than 1. Also, Definition 3 does not define ρ uniquely since it operates only

with an upper bound: if $\mathcal{G}$ has complexity bound ρ, then $\mathcal{G}$ also has any higher complexity bound $\rho' > \rho$.

Dudley (1974), Korostelev and Tsybakov (1993) and Mammen and Tsybakov (1995, 1999) give various examples of classes $\mathcal{G}$ satisfying Definition 3. These are typically classes of sets with smooth boundaries and their complexity bound is $\rho = (d-1)/\gamma$, where γ denotes the smoothness index of the boundary (roughly speaking, γ is the number of bounded derivatives of boundary surfaces).

(A2) Assumption on the complexity. The class of sets $\mathcal{G}^*$ has complexity bound $0 < \rho < 1$.

Now, putting together assumptions (A1) and (A2), we define the corresponding class of joint distributions of (X, Y).

DEFINITION 4. A class $\mathcal{P}$ of joint distributions π of (X, Y) is called a $(\mathcal{G}^*, \kappa, \rho)$-*class* if

$$G^* = G^*_\pi = \{x : \eta(x) \geq 1/2\} \in \mathcal{G}^*,$$

where $\mathcal{G}^*$ is such that assumptions (A1) and (A2) are satisfied and either $\varepsilon_0 = 1$ in assumption (A1) or

$$\lim_{t \to 0} \sup_{\pi \in \mathcal{P}} P\big(|\eta(X) - 1/2| \leq t\big) = 0. \tag{7}$$

Without loss of generality, we will assume that the constants ε_0 and c_0 appearing in assumption (A1) are the same for all $(\mathcal{G}^*, \kappa, \rho)$-classes that we will consider. For $G, G' \subseteq \mathbf{R}^d$ define the empirical analogue of $d_\triangle$:

$$d_{\triangle,e}(G, G') \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} I\,(X_i \in G \triangle G').$$

THEOREM 1. *Let $\kappa \geq 1$, $0 < \rho < 1$, $a > 0$, $\varepsilon = an^{-1/(1+\rho)}$. Let $\mathcal{P}$ be a $(\mathcal{G}^*, \kappa, \rho)$-class of joint distributions of (X, Y) and let $\mathcal{N}$ be an ε-net on $\mathcal{G}^*$ for the pseudometric $d_\triangle$ or $d_{\triangle,e}$, such that $\mathcal{N}$ has complexity bound ρ. Then the classifier*

$$\hat{G}_n = \arg\min_{G \in \mathcal{N}} R_n(G) \tag{8}$$

satisfies

$$\sup_{\pi \in \mathcal{P}} E_{\pi,n}\big(d(\hat{G}_n, G^*)\big) = O\big(n^{-\kappa/(2\kappa+\rho-1)}\big), \qquad n \to \infty.$$

If we suppress assumption (A1), that is, if there are no assumptions on the margin, we can only ensure that the classifier (8) attains the rate $n^{-1/2}$, independently of ρ:

PROPOSITION 2. *Let* $0 < \rho < 1$, $a > 0$, $\varepsilon = an^{-1/(1+\rho)}$. *Let* $\mathcal{P}^0$ *be a class of joint distributions* π *of* (X, Y) *for which the sets* $G^* = G^*_\pi = \{x : \eta(x) \geq 1/2\}$ *belong to a class* $\mathcal{G}^*$ *having complexity bound* ρ. *Let* $\mathcal{N}$ *be as in Theorem* 1. *Then the classifier* (8) *satisfies*

$$\sup_{\pi \in \mathcal{P}^0} E_{\pi,n}\big(d(\hat{G}_n, G^*)\big) = O(n^{-1/2}), \qquad n \to \infty.$$

Proofs of Theorem 1 and Proposition 2 are given in Section 5.

REMARKS. (i) The classifier (8) is feasible if, for example, the set $\mathcal{N}$ is finite. Construction of a finite ε-net $\mathcal{N}$ for the empirical distance $d_{\triangle,e}$ can be done in a distribution free way. This is also possible for the distance $d_\triangle$ if we suppose that P_X has a bounded density w.r.t. Lebesgue measure in $\mathbf{R}^d$. Then $\mathcal{N}$ can be taken as an ε-net on $\mathcal{G}^*$ for the distance λ defined as Lebesgue measure of symmetric difference of sets. For many examples of $\mathcal{G}^*$ an ε-net $\mathcal{N}$ w.r.t. λ can be chosen finite [see, e.g., Korostelev and Tsybakov (1993)]; hence a solution of (8) exists. Another possibility is to select $\mathcal{N}$ as a nonfinite Vapnik–Chervonenkis (VC) class of sets for which a solution of (8) exists, for instance, a class of sets with piecewise-polynomial boundaries, finite series boundary approximations and so on. Note that we do not require the inclusion $\mathcal{N} \subseteq \mathcal{G}^*$. Finally, $\mathcal{N} = \mathcal{G}^*$ obviously satisfies the assumptions of Theorem 1, that is, the result holds for empirical risk minimizers over the whole class $\mathcal{G}^*$, although they might be difficult to compute.

(ii) Robustness to the margin property holds: knowledge of the margin parameter κ is not needed to construct the classifier $\hat{G}_n$.

(iii) The rates in Theorem 1 are always faster than $n^{-1/2}$. They approach $n^{-1/2}$ as $\rho \to 1$, and they approach n^{-1} as $\rho \to 0, \kappa \to 1$.

(iv) If $\kappa \neq 1$ assumption (A1) and (7) follow from (5), where α is related to κ by means of (6); thus assumption (A1) and (7) in Definition 4 can be replaced directly by (5).

(v) If there is no assumption on the margin, the rate does not depend on the complexity ρ and equals $n^{-1/2}$ (cf. Proposition 2). This fact is naturally related to Theorem 1. Indeed, if there are no assumptions on the margin, arbitrarily large values of κ are admitted, which means that the rate $n^{-\kappa/(2\kappa+\rho-1)}$ given in Theorem 1 can become arbitrarily close from below to $n^{-1/2}$.

Inspection of the proofs shows that the result of Theorem 1 is true not only for the exact minimizer of the empirical risk, but also for any approximate minimizer $\hat{G}_{n,\mathrm{app}} \in \mathcal{N}$ satisfying

$$R_n(\hat{G}_{n,\mathrm{app}}) - \min_{G \in \mathcal{N}} R_n(G) \leq \frac{C}{\sqrt{n}} \sup_{\hat{G}_n} d^{(1-\rho)/2}_{\triangle,e}(\hat{G}_{n,\mathrm{app}}, \hat{G}_n)$$

for some $C > 0$, where the supremum in the right-hand side is taken over all

minimizers $\hat{G}_n$ of the empirical risk R_n if the minimizer is not unique. A particular example of such an approximate ERM classifier is

$$\hat{G}_{n,\mathrm{app}} = \arg \max_{G \in \mathcal{N}_{\mathrm{app}}} R_n(G),$$

where

$$\mathcal{N}_{\mathrm{app}} = \left\{ G \in \mathcal{N} : R_n(G) - \min_{G \in \mathcal{N}} R_n(G) \le \frac{C}{\sqrt{n}} \sup_{\hat{G}_n} d_{\triangle,e}^{(1-\rho)/2}(G, \hat{G}_n) \right\}.$$

The result of Theorem 1 cannot be improved in a minimax sense, as the next theorem shows. Let $\mathcal{G}^* = \mathcal{G}_{\mathrm{frag}}$, where $\mathcal{G}_{\mathrm{frag}}$ is the class of boundary fragments with boundaries of Hölder smoothness $\gamma > 0$ defined as follows. For given $\gamma > 0$ and $d \ge 2$ 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

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

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

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

Such sets are called boundary fragments. Define the class

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

The complexity bound of $\mathcal{G}_{\mathrm{frag}}$ is $\rho = (d-1)/\gamma$ [see, e.g., Mammen and Tsybakov (1999)]. Denote by $\mathcal{P} = \mathcal{P}_{\mathrm{frag}}$ the class of all joint distributions of (X, Y) such that $G^* \in \mathcal{G}_{\mathrm{frag}}$ and assumption (A1) and (7) hold.

THEOREM 2. *For $\rho = (d-1)/\gamma$, $\kappa \ge 1$ we have*

$$\liminf_{n \to \infty} \inf_{\tilde{G}_n} \sup_{\pi \in \mathcal{P}_{\mathrm{frag}}} E_{\pi,n}\big(d(\tilde{G}_n, G^*)\big) n^{\kappa/(2\kappa+\rho-1)} \ge c_{\min},$$

where $\inf_{\tilde{G}_n}$ denotes the infimum over all the classifiers and $c_{\min} > 0$ is a constant.

Proof of Theorem 2 follows Mammen and Tsybakov (1999) with minor modifications. In fact, it suffices to consider the joint distributions such that $P(Y = 0) = P(Y = 1) = 1/2$ and to assume that there exist densities $f(x)$ and $g(x)$ of the conditional distributions $P(X|Y = 1)$ and $P(X|Y = 0)$, respectively. Then G^* has the form $G^* = \{x : f(x) \ge g(x)\}$, and we are in the framework of Theorem 3 in Mammen and Tsybakov (1999) (if $\kappa \ne 1$). The rest of the proof for $\kappa \ne 1$ follows the same lines as the proof of Theorem 3 in Mammen and Tsybakov (1999). For

the case $\kappa = 1$ the lower bounds are obtained using constructions as in Korostelev and Tsybakov (1993) or Mammen and Tsybakov (1995) for estimators of supports of densities with jump discontinuities.

Note that Theorem 2 is true for all $\rho > 0$ (there is no restriction that $\rho < 1$, as in Theorem 1).

3. Aggregation. In this section we assume that $G^* \in \mathcal{G}^*$, where $\mathcal{G}^*$ is an unknown class of subsets of $\mathbf{R}^d$. Consider first the case where $\mathcal{G}^*$ belongs to a finite collection $\{\mathcal{G}_1, \dots, \mathcal{G}_N\}$, where $\mathcal{G}_j \neq \mathcal{G}_k$ are known classes of sets and $N = N_n$ is an integer that may depend on n.

The following assumption on $\mathcal{G}_j$ is used.

(A3) Assumption on the complexities. The classes $\mathcal{G}_j$ have complexity bounds ρ_j, respectively, such that

$$0 < \rho_1 \le \rho_2 \le \cdots \le \rho_N < 1.$$

To each $\mathcal{G}_j$ we associate an approximating set $\mathcal{N}_j$ and consider the ERM classifiers

$$G_{nk} = \arg\min_{G \in \mathcal{N}_k} R_n(G), \qquad k = 1, \dots, N.$$

We suppose that the sets $\mathcal{N}_j$ satisfy the following:

(A4) Assumption on the approximating sets. (i) $\mathcal{N}_j$ is an ε-net on $\mathcal{G}_j$ for the pseudodistance $d_\triangle$ or $d_{\triangle,e}$, where $\varepsilon = a_j n^{-1/(1+\rho_j)}$, $a_j > 0$, $j = 1, \dots, N$, and $\sup_j a_j \le a < \infty$.

(ii) The approximating sets are nested:

$$\mathcal{N}_1 \subseteq \mathcal{N}_2 \subseteq \cdots \subseteq \mathcal{N}_N.$$

(iii) $\mathcal{N}_j$ has complexity bound ρ_j, $j = 1, \dots, N$.

Introduce the thresholds

$$T_{nk}(G, G') = (\log^2 n) \max\left\{ n^{-1/(1+\rho_k)}, \frac{1}{\sqrt{n}} d_{\triangle,e}^{(1-\rho_k)/2}(G, G') \right\}, \qquad k = 1, \dots, N.$$

Define, for any $j \in \{1, \dots, N\}$ and $k \ge j$,

$$\mathcal{N}_{kj} = \left\{ G \in \mathcal{N}_j : |R_n(G) - R_n(G_{nk})| \le T_{nk}(G, G_{nk}) \right\}.$$

We call the index j *admissible* if $\mathcal{N}_{kj} \neq \varnothing$ for all $k \ge j$. Note that the set of all admissible j is nonempty since it contains at least $j = N$. In fact, $\mathcal{N}_{NN} \neq \varnothing$ ($\mathcal{N}_{NN}$ contains G_{nN}). Set

$$\hat{j} = \min\{\text{admissible } j\},$$

and define the adaptive classifier

$$G_n^* \stackrel{\text{def}}{=} G_{n\hat{j}}.$$

THEOREM 3. *Let assumptions* (A3) *and* (A4) *hold and let, for some* $\kappa \geq 1$, $\mathcal{P}_j$ *be* $(\mathcal{G}_j, \kappa, \rho_j)$*-classes of joint distributions of* (X, Y), $j = 1, \ldots, N$. *Then, if* $N = O(n^\beta)$, *as* $n \to \infty$, *for some finite* $\beta > 0$ *and if* $\rho_1 \stackrel{def}{=} \rho_{\min}$, $\rho_N \stackrel{def}{=} \rho_{\max}$ *do not depend on* n, *the adaptive classifier* G_n^* *satisfies*

$$\sup_{\pi \in \mathcal{P}_j} E_{\pi,n}\big(d(G_n^*, G^*)\big) \leq C\left(\frac{\log^4 n}{n}\right)^{\kappa/(2\kappa+\rho_j-1)}$$

for any $j = 1, \ldots, N$, *where* $n > 1$ *and* C *is a finite constant that does not depend on* n.

The message of Theorem 3 is that, up to a logarithmic factor, the adaptive classifier attains the optimal rates simultaneously on all the classes $\mathcal{G}_1, \ldots, \mathcal{G}_N$. If the "true" class where G^* lies is $\mathcal{G}_{j^*}$, Theorem 3 asserts that G_n^* has at most a logarithmically worse rate than the optimal one $n^{-\kappa/(2\kappa+\rho_{j^*}-1)}$. In other words, G_n^* mimics (up to a logarithmic factor) the rate of the best classifier among $G_{n1}, \ldots, G_{nN}$.

Theorem 3 can be extended to a continuous scale of values ρ, under an additional nestedness assumption. In fact, instead of a finite collection of classes $\{\mathcal{G}_1, \ldots, \mathcal{G}_N\}$, consider a collection $\{\mathcal{G}_\rho\}_{\rho \in \mathcal{I}}$, where $\mathcal{I} = [\rho_{\min}, \rho_{\max}]$, with known constants $\rho_{\min}$ and $\rho_{\max}$ such that $0 < \rho_{\min} < \rho_{\max} < 1$. The statistician only knows that $G^* \in \mathcal{G}_\rho$ for some unknown $\rho \in \mathcal{I}$.

COROLLARY 1. *Assume that* $0 < \rho_{\min} < \rho_{\max} < 1$, $\kappa \geq 1$, *and that the* $\mathcal{G}_\rho$*'s are classes of subsets of* $\mathbf{R}^d$ *such that* $\mathcal{G}_\rho \subset \mathcal{G}_{\rho'}$ *for* $\rho < \rho'$, *where* ρ, $\rho' \in [\rho_{\min}, \rho_{\max}]$ *and the class* $\mathcal{G}_\rho$ *has complexity bound* ρ. *Let assumption* (A4) *hold with* $\mathcal{G}_j = \mathcal{G}_{\rho_j}$, $\rho_j = \rho_{\min} + (j/N)(\rho_{\max} - \rho_{\min})$, *and let the* $\mathcal{P}_\rho$*'s be* $(\mathcal{G}_\rho, \kappa, \rho)$*-classes of joint distributions of* (X, Y), $\rho \in [\rho_{\min}, \rho_{\max}]$. *Then, if* $N = O(n^\beta)$, *as* $n \to \infty$, *and* $N \geq n^{\beta'}$ *for some finite* $\beta \geq \beta' > 0$, *the adaptive classifier* G_n^* *satisfies*

$$\sup_{\pi \in \mathcal{P}_\rho} E_{\pi,n}\big(d(G_n^*, G^*)\big) \leq C\left(\frac{\log^4 n}{n}\right)^{\kappa/(2\kappa+\rho-1)},$$

for any $\rho \in [\rho_{\min}, \rho_{\max}]$, *where* $n > 1$ *and* C *is a finite constant that does not depend on* n.

To prove this corollary, it is enough to note that since $\mathcal{P}_\rho \subset \mathcal{P}_{\rho'}$ for $\rho < \rho'$ and in view of Theorem 3 we have, for any $\rho \in [\rho_{\min}, \rho_{\max}]$ and $\rho' = \min\{\rho_j : \rho_j > \rho\}$,

$$\sup_{\pi \in \mathcal{P}_\rho} E_{\pi,n}\big(d(G_n^*, G^*)\big) \le \sup_{\pi \in \mathcal{P}'_\rho} E_{\pi,n}\big(d(G_n^*, G^*)\big) \le C\left(\frac{\log^4 n}{n}\right)^{\kappa/(2\kappa+\rho'-1)}$$

$$\le C\left(\frac{\log^4 n}{n}\right)^{\kappa/(2\kappa+\rho+n^{-\beta'}-1)} \le C'\left(\frac{\log^4 n}{n}\right)^{\kappa/(2\kappa+\rho-1)},$$

where C' is a finite constant that does not depend on n and we used that $\rho' \le \rho + N^{-1}$.

Finally, if we suppress assumption (A1), that is, if there are no assumptions on the margin, we can nevertheless ensure that the same adaptive classifier G_n^* as in Theorem 3 or in Corollary 1 attains the rate $n^{-1/2}\log^2 n$:

PROPOSITION 3. *Let assumptions* (A3) *and* (A4) *be satisfied, with* $\rho_1 \stackrel{def}{=} \rho_{\min}$, $\rho_N \stackrel{def}{=} \rho_{\max}$ *that do not depend on* n. *Let* $\mathcal{P}_j^0$ *be a class of joint distributions* π *of* (X, Y) *for which the sets* $G^* = G_\pi^* = \{x : \eta(x) \ge 1/2\}$ *belong to* $\mathcal{G}_j$. *Then, if* $N = O(n^\beta)$, *as* $n \to \infty$, *for some finite* $\beta > 0$, *the adaptive classifier* G_n^* *satisfies*

$$\sup_{\pi \in \mathcal{P}_j^0} E_{\pi,n}\big(d(G_n^*, G^*)\big) \le C n^{-1/2}\log^2 n, \tag{11}$$

for any $j = 1, \dots, N$, *where* $n > 1$ *and* C *is a finite constant that does not depend on* n.

Proof of Proposition 3 is given in Section 5. One can also get a result similar to Proposition 3 with a continuous scale of ρ, under the conditions described in Corollary 1. One of these conditions, namely the nestedness of sets $\mathcal{G}_\rho$, implies that in this case it is sufficient to state (11) for the largest set $\mathcal{P}_j^0 = \mathcal{P}_{\rho_{\max}}^0$.

REMARKS. (i) The same comments as after Theorem 1 can be made about the feasibility of the method; in particular, the classifiers G_{nk} and G_n^* can be readily constructed. However, it is preferable here that the sets $\mathcal{N}_j$ be finite, since otherwise it is difficult to check the conditions $\mathcal{N}_{kj} \ne \varnothing$. For this reason, it might be helpful to consider ε-nets $\mathcal{N}_j$ for the empirical distance $d_{\triangle,e}$ rather than for $d_\triangle$.

(ii) The aggregation procedure proposed in this section realizes adaptation to the unknown complexity ρ (similar to adaptation to unknown smoothness in the usual nonparametric problems, such as regression and density estimation). However, unlike those problems, we have another parameter, namely the characteristic

of the margin κ, which is also unknown. It is important that this parameter appear nowhere in the above aggregation procedure; in other words, the property of robustness to the margin holds. Moreover, the optimality properties of the aggregation procedure stated in Theorem 3 are preserved if instead of assumption (A1) [or instead of (5)] we assume a more general behavior of the margin distribution near the boundary:

$$P\big(|\eta(X) - 1/2| \le t\big) \le \varphi(t), \tag{12}$$

where $\varphi(t)$ is a monotone function and $\varphi(t) \downarrow 0$ as $t \to 0$. In this case the rates in Theorems 1–3 will be, in general, different from the actually stated ones, and they will depend on φ. The only point to emphasize is that the aggregation procedure remains the same as above and that the robustness to the margin property holds under these much more general conditions.

(iii) The aggregation method proposed here is a member of the general family of procedures that can be called pretesting aggregation schemes [cf. Tsybakov (2002)]. Another well-known member of this family is the method of Lepski (1990), which was originally defined for the Gaussian white noise model but can be extended to the classification framework as well. An important difference from the above aggregation scheme is that the Lepski method would use pairwise comparisons between classifiers G_{nk} to define its own set of admissible j's. This would induce too high a bias and would not allow us to get the result of Theorem 3 unless $\kappa = 1$.

(iv) The sets $\mathcal{G}_j$ in Theorem 3 are not supposed to be nested. They only need to have the ordered *upper bounds* ρ_j on complexities [assumption (A3)]. The nestedness assumption applies to the approximating sets $\mathcal{N}_j$ only and causes no problem because these sets can be chosen by the statistician.

(v) The factor $\log^2 n$ in the definition of the threshold T_{nk} can be replaced by $C^* \log n$ for some constant $C^* > 0$ large enough. This, in turn, improves the logarithmic factor in Theorem 3, which becomes $(\log^2 n)^{\kappa/(2\kappa+\rho_j-1)}$. However, the precise value of C^* is not known since it relies on unknown accurate constants in the empirical process inequalities used in proofs. Also, inspection of the proofs shows that $\log^2 n$ can be readily replaced by $\ell_n \log n$, where ℓ_n is some sequence that tends to ∞ slower than $\log n$ (with the respective minor improvement for the rate in Theorem 3).

4. Discussion. In this section we discuss some questions related to optimality of classification methods.

The first question is: how fast can the convergence of classifiers be and how does one construct the classifiers that have optimal convergence rates? Several papers discussing this question arrive at conclusions that are, at first glance, contradictory. Yang (1999) claims that the optimal rates are quite slow (substantially slower than $n^{-1/2}$), and they are attained with plug-in rules; Mammen and Tsybakov (1999)

claim that the rates are fast (between $n^{-1/2}$ and n^{-1}) and they are attained by ERM and related classifiers (this is also the message of Theorem 1 above); in the papers deriving oracle inequalities [Barron (1991), Lugosi and Nobel (1999), Koltchinskii and Panchenko (2002), Koltchinskii (2001) and Bartlett, Boucheron and Lugosi (2002)] ERM-based and other related classifiers are shown to converge with the rate at best $n^{-1/2}$ (up to a log-factor), if $R(G^*) \neq 0$. In fact, there is no contradiction since different classes of joint distributions π of (X, Y) are considered. Yang (1999) and the papers on oracle inequalities cited above do not impose assumption (A1) (or any other assumption) on the margin. Therefore, it is not surprising that they get rates slower than $n^{-1/2}$: one cannot obtain a rate faster than $n^{-1/2}$ with no assumptions on the margin [cf. the lower bound given by Devroye, Györfi and Lugosi (1996), page 240]. The same effect is observed in Proposition 2: with no assumptions on the margin we get the slow rate $n^{-1/2}$. However, note that Proposition 2 applies only if $\rho < 1$. If no restrictions on the joint distribution π are imposed, one cannot exclude the case where $\rho \geq 1$, that is, the sets G which are extremely complex (the boundaries ∂G are very nonregular), and the rate of convergence can be arbitrarily slow. Yang (1999) works with π such that the function η is smooth, and expresses the rates in terms of the smoothness of η. This assumption does not imply that the boundary of G is regular. For example, using linear combinations of infinitely smooth functions on small hypercubes, it is easy to construct a function η that is infinitely smooth in every coordinate, but the boundary of G is not regular. This suggests that Yang's assumptions do not exclude the values $\rho \geq 1$ for the complexity of possible sets G, explaining why the optimal rates in the situation considered by Yang (1999) are slower than $n^{-1/2}$. On the contrary, Mammen and Tsybakov (1999) and Theorem 1 above show what can be achieved when the situation is "nice," that is, assumption (A1) on the margin holds. In this case the fast rates (up to n^{-1}) are realizable. Similar results for penalized ERM classifiers are given by Massart (2000).

The second question is whether the aggregation of classifiers is possible with such fast rates. Theorem 3 answers this question affirmatively. It shows that fast rates are achievable without any knowledge of the complexity and of the margin of the underlying joint distribution π. Note that the usual aggregation procedures based on penalization with penalty terms (random or nonrandom) of order greater than or equal to $n^{-1/2}$ cannot achieve such fast rates, irrespective of whether the underlying joint distribution is nice or not. If the penalty is of the order $n^{-1} \log n$ [as in Bartlett, Boucheron and Lugosi (2002), Massart (2000) or Catoni (2001)], the question remains open. Another approach to aggregation is to mimic the best member of a convex combination of classifiers and related boosting and bagging techniques [see Breiman (1996), Schapire, Freund, Bartlett and Lee (1998) and Bühlmann and Yu (2002)]. The results of Koltchinskii and Panchenko (2002) and Koltchinskii (2001) suggest that, in the general case, the best one can guarantee for a convex combination of classifiers is the rate $n^{-1/2}$ (up to log-factors).

In the present paper classification is considered as a special case of nonparametric estimation of sets. A specific point is that here one uses a particular distance d given in (2) to define the risk, unlike in the usual set estimation problems where one works with Lebesgue measure of symmetric difference distance or with the Hausdorff distance [see Korostelev and Tsybakov (1993)]. The principal difficulty in treating the classification problem, as compared to other well-known nonparametric problems, such as regression or density estimation, is the lack of precise bias–variance decomposition of the risk. The form of this decomposition cannot be claimed to be additive, and the bias does not appear in a closed form. To illustrate this, the bias–variance trade-off for ERM classifiers (cf. the proof of Theorem 1) can be expressed by the inequality

$$d(\hat{G}_n, G^*) \le \frac{C}{\sqrt{n}} d_{\triangle}^{(1-\rho)/2}(\hat{G}_n, G^*) \tag{13}$$

(with some $C > 0$ and only asymptotically, with a probability close to 1). The right-hand side of this inequality plays the role of the "variance term." Note also that the "empirical variance" $d_{\triangle,e}^{(1-\rho)/2}(\cdot,\cdot)/\sqrt{n}$ logically appears in the threshold of the aggregation procedure of Section 3. The left-hand side of (13) represents the total loss, while the bias term in the proper sense is not visible. Thus, the usual nonparametric argument suggesting that one balance the approximation error and the stochastic error, as well as its adaptive extensions, is not directly applicable.

The aggregation procedure proposed in Section 3 relies strongly upon the approximation properties of the chosen sets $\mathcal{N}_j$. To make it work efficiently, approximation properties should be available for typical systems of sets (usually VC-classes). The question, how well concrete systems of VC-classes approximate the classes of sets with smooth boundaries, is crucial in this context. To answer this question, it would be helpful to have an approximation theory for sets, analogous to that for functions.

5. Proofs of the theorems. Without loss of generality assume in the proofs that $a = 1$, and write for brevity $P = P_{\pi,n}$, $E = E_{\pi,n}$.

PROOF OF THEOREM 1. Write $\mathcal{G}' = \mathcal{G}^* \cup \mathcal{N}$. Clearly, $\mathcal{G}'$ has the same complexity bound ρ as $\mathcal{G}^*$ and $\mathcal{N}$. Fix an element $G^{\mathcal{N}} \in \mathcal{N}$ such that $d_{\triangle}(G^{\mathcal{N}}, G^*) \le n^{-1/(1+\rho)}$ or $d_{\triangle,e}(G^{\mathcal{N}}, G^*) \le n^{-1/(1+\rho)}$ (such a $G^{\mathcal{N}}$ exists since $\mathcal{N}$ is an $n^{-1/(1+\rho)}$-net on $\mathcal{G}^*$ for the pseudodistance $d_{\triangle}$ or $d_{\triangle,e}$). Note that, if $d_{\triangle,e}$ is used, $G^{\mathcal{N}}$ is random. Consider the random event

$$\Omega = \{d_{\triangle}(G^{\mathcal{N}}, G^*) \le bn^{-1/(1+\rho)}\},$$

where $b = \max(2, c_1)$ and c_1 is the constant in Lemma 7 from the Appendix. If $\mathcal{N}$ is an $n^{-1/(1+\rho)}$-net on $\mathcal{G}^*$ for $d_{\triangle}$ we have $P(\overline{\Omega}) = 0$. If $\mathcal{N}$ is an $n^{-1(1+\rho)}$-net

on $\mathcal{G}^*$ for $d_{\triangle,e}$, using (63) we get

$$
\begin{aligned}
P(\overline{\Omega}) &\le P\Big(d_\triangle(G^{\mathcal{N}}, G^*) > bn^{-1/(1+\rho)}, d_{\triangle,e}(G^{\mathcal{N}}, G^*) > d_\triangle(G^{\mathcal{N}}, G^*)/2\Big) \\
&\quad + P\Big(d_\triangle(G^{\mathcal{N}}, G^*) > c_1 n^{-1/(1+\rho)}, d_{\triangle,e}(G^{\mathcal{N}}, G^*) \le d_\triangle(G^{\mathcal{N}}, G^*)/2\Big) \\
&= P\Big(d_\triangle(G^{\mathcal{N}}, G^*) > c_1 n^{-1/(1+\rho)}, d_{\triangle,e}(G^{\mathcal{N}}, G^*) \le d_\triangle(G^{\mathcal{N}}, G^*)/2\Big) \\
&\le A_1 \exp\bigl(-A_2 n^{\rho/(1+\rho)}\bigr).
\end{aligned} \tag{14}
$$

We conclude that (14) holds in both cases: when $\mathcal{N}$ is an $n^{-1(1+\rho)}$-net on $\mathcal{G}^*$ either for $d_\triangle$ or for $d_{\triangle,e}$. Now, $R_n(G^{\mathcal{N}}) \ge R_n(\hat{G}_n)$, and therefore on Ω,

$$
\begin{aligned}
d(\hat{G}_n, G^*) &\le \bigl[R_n(G^{\mathcal{N}}) - R_n(G^*) - d(G^{\mathcal{N}}, G^*)\bigr] + d(G^{\mathcal{N}}, G^*) \\
&\quad + \bigl[R_n(G^*) - R_n(\hat{G}_n) + d(\hat{G}_n, G^*)\bigr] \\
&\le bn^{-1/(1+\rho)} + \bigl[R_n(G^{\mathcal{N}}) - R_n(G^*) - d(G^{\mathcal{N}}, G^*)\bigr] \\
&\quad + \bigl[R_n(G^*) - R_n(\hat{G}_n) + d(\hat{G}_n, G^*)\bigr],
\end{aligned} \tag{15}
$$

where we have used (4).

Our local aim now is to show that the probability $P(d_\triangle(\hat{G}_n, G^*) > \varepsilon_0)$ is negligible as $n \to \infty$, so that we can apply assumption (A1). If $\varepsilon_0 = 1$, this probability is 0. If $\varepsilon_0 < 1$, we will bound this probability using (7). From (15) and the fact that $G^{\mathcal{N}}, \hat{G}_n \in \mathcal{G}'$, we get on Ω,

$$
d(\hat{G}_n, G^*) \le bn^{-1/(1+\rho)} + 2 \sup_{G \in \mathcal{G}'} |R_n(G) - R_n(G^*) - d(G, G^*)|,
$$

and for all $t \ge 2bn^{-1/(1+\rho)}$,

$$
\begin{aligned}
&P\bigl(d(\hat{G}_n, G^*) > t\bigr) \\
&\quad \le P\Bigl(\sup_{G \in \mathcal{G}'} |R_n(G) - R_n(G^*) - d(G, G^*)| > t/4\Bigr) + P(\overline{\Omega}).
\end{aligned} \tag{16}
$$

Next, for any $t > 0$,

$$
\begin{aligned}
d(\hat{G}_n, G^*) &\ge 2 \int_{\hat{G}_n \triangle G^*} |\eta(x) - 1/2| I\bigl(|\eta(x) - 1/2| \ge t\bigr) P_X(dx) \\
&\ge 2t\bigl[d_\triangle(\hat{G}_n, G^*) - P\bigl(|\eta(X) - 1/2| < t\bigr)\bigr].
\end{aligned}
$$

In view of (7) there exists $0 < t_0 < 1/2$ such that $\sup_{\pi \in \mathcal{P}} P(|\eta(X) - 1/2| < t_0) < \varepsilon_0/2$. Hence

$$
d_\triangle(\hat{G}_n, G^*) \le \frac{1}{2t_0} d(\hat{G}_n, G^*) + \frac{\varepsilon_0}{2}, \tag{17}
$$

and therefore, by (16), for all n large enough,

$$
\begin{aligned}
&P\big(d_\triangle(\hat G_n, G^*) > \varepsilon_0\big)\\
&\quad\le P\big(d(\hat G_n, G^*) > \varepsilon_0 t_0\big)\\
&\quad\le P\Big(\sup_{G\in\mathcal{G}'} |R_n(G) - R_n(G^*) - d(G, G^*)| > \varepsilon_0 t_0/4\Big) + P(\overline{\Omega}).
\end{aligned}
$$

Since $\varepsilon_0 t_0/4 < 1$, we can apply Lemma 11, which yields, together with (14), for n large enough,

$$
\begin{aligned}
&P\big(d_\triangle(\hat G_n, G^*) > \varepsilon_0\big)\\
&\quad\le C_2 \exp\big(-n(\varepsilon_0 t_0)^2/(16C_2)\big) + A_1 \exp\big(-A_2 n^{\rho/(1+\rho)}\big).
\end{aligned}
\tag{18}
$$

Now we proceed to the main evaluation of the risk of the classifier $\hat G_n$. Define the random event $\Omega_0 = \{d_\triangle(\hat G_n, G^*) \le \varepsilon_0\}$, where $c_1 > 0$ is a constant such that $c_1 n^{-1/(1+\rho)} < \varepsilon_0$. Using (14) and (18) we get

$$
\begin{aligned}
E\big(d(\hat G_n, G^*)\big) &\le E\big[d(\hat G_n, G^*) I(\Omega_0\Omega)\big]\\
&\quad + E\big[d(\hat G_n, G^*) I\big(d_\triangle(\hat G_n, G^*) > \varepsilon_0\big)\big] + P(\overline{\Omega})\\
&\le E\big[d(\hat G_n, G^*) I(\Omega_0\Omega)\big] + C_2 \exp\big(-n(\varepsilon_0 t_0)^2/(16C_2)\big)\\
&\quad + 2A_1 \exp\big(-A_2 n^{\rho/(1+\rho)}\big).
\end{aligned}
\tag{19}
$$

On the event $\Omega_0\Omega$ we have $d_\triangle(G^{\mathcal{N}}, G^*) \le b n^{-1/(1+\rho)}$. Thus, using (15) and the fact that $G^{\mathcal{N}}, \hat G_n \in \mathcal{G}'$, we obtain that, on $\Omega_0\Omega$,

$$
\begin{aligned}
&d(\hat G_n, G^*)\\
&\quad\le b n^{-1/(1+\rho)} + 2 \sup_{G\in\mathcal{G}' : d_\triangle(G, G^*) \le b n^{-1/(1+\rho)}} |R_n(G^*) - R_n(G) + d(G, G^*)|\\
&\qquad + \frac{d_\triangle^{(1-\rho)/2}(\hat G_n, G^*)}{\sqrt{n}} \sup_{G\in\tilde{\mathcal{G}}'} \frac{\sqrt{n}|R_n(G^*) - R_n(G) + d(G, G^*)|}{d_\triangle^{(1-\rho)/2}(G, G^*)}\\
&\quad\le b n^{-1/(1+\rho)} + 2V_{0n}(\mathcal{G}') + \frac{V_n(\mathcal{G}')}{\sqrt{n}} d_\triangle^{(1-\rho)/2}(\hat G_n, G^*),
\end{aligned}
\tag{20}
$$

where $\tilde{\mathcal{G}}' = \{G \in \mathcal{G}' : d_\triangle(G, G^*) \ge b n^{-1/(1+\rho)}\}$, V_{0n} and V_n are defined in the Appendix and we set $c_2 = b$ in the definition of V_{0n}. This and assumption (A1)

imply that on $\Omega_0\Omega$ we have

$$d(\hat{G}_n, G^*) \le bn^{-1/(1+\rho)} + 2V_{0n}(\mathcal{G}') + \frac{c_0^{-1/\kappa} V_n(\mathcal{G}')}{\sqrt{n}} d^{(1-\rho)/2\kappa}(\hat{G}_n, G^*). \tag{21}$$

Introduce the random events

$$\Omega_1 = \{V_{0n}(\mathcal{G}') \le \max(B_3, b) n^{-1/(1+\rho)}\},$$

$$\Omega_2 = \{d(\hat{G}_n, G^*) \ge 4\max(B_3, b) n^{-1/(1+\rho)}\},$$

where $B_3 > 0$ is the constant from Lemma 9 in the Appendix. Then

$$\begin{aligned} &E\big(d(\hat{G}_n, G^*) I(\Omega_0\Omega)\big) \\ &\quad \le E\big(d(\hat{G}_n, G^*) I(\Omega\Omega_0\Omega_1\Omega_2)\big) + E\big(d(\hat{G}_n, G^*) I(\overline{\Omega}_1 \cup \overline{\Omega}_2)\big) \\ &\quad \le E\big(d(\hat{G}_n, G^*) I(\Omega\Omega_0\Omega_1\Omega_2)\big) + P(\overline{\Omega}_1) \\ &\qquad + 4\max(B_3, b) n^{-1/(1+\rho)}. \end{aligned} \tag{22}$$

By Lemma 9,

$$P(\overline{\Omega}_1) \le B_1 \exp\big(-B_2 n^{\rho/(1+\rho)}\big). \tag{23}$$

Now, (21) implies that on the event $\Omega\Omega_0\Omega_1$ we have

$$d(\hat{G}_n, G^*) \le 3\max(B_3, b) n^{-1/(1+\rho)} + \frac{c_0^{-1/\kappa} V_n(\mathcal{G}')}{\sqrt{n}} d^{(1-\rho)/2\kappa}(\hat{G}_n, G^*).$$

From this inequality and the definition of Ω_2 we get that, on $\Omega\Omega_0\Omega_1\Omega_2$,

$$d(\hat{G}_n, G^*) \le \frac{4c_0^{-1/\kappa} V_n(\mathcal{G}')}{\sqrt{n}} d^{(1-\rho)/2\kappa}(\hat{G}_n, G^*).$$

Thus there exists a constant $C > 0$ such that on $\Omega\Omega_0\Omega_1\Omega_2$ one has

$$d(\hat{G}_n, G^*) \le C V_n(\mathcal{G}')^{2\kappa/(2\kappa+\rho-1)} n^{-\kappa/(2\kappa+\rho-1)}.$$

This and (19), (22), (23) imply

$$\begin{aligned} &E\big(d(\hat{G}_n, G^*)\big) \\ &\quad \le C E\big(V_n(\mathcal{G}')^{2\kappa/(2\kappa+\rho-1)}\big) n^{-\kappa/(2\kappa+\rho-1)} + 4\max(B_3, b) n^{-1/(1+\rho)} \\ &\qquad + C_2 \exp\big(-n(\varepsilon_0 t_0)^2/(16C_2)\big) + B_1 \exp\big(-B_2 n^{\rho/(1+\rho)}\big) \\ &\qquad + 2A_1 \exp\big(-A_2 n^{\rho/(1+\rho)}\big). \end{aligned}$$

To finish the proof it remains to note that $n^{-1/(1+\rho)} = O(n^{-\kappa/(2\kappa+\rho-1)})$, $n \to \infty$, if $0 < \rho < 1$, $\kappa \ge 1$, and $\sup_n E(V_n(\mathcal{G}')^q) < \infty$ for any $q > 0$, in view of Lemma 8

in the Appendix. □

PROOF OF PROPOSITION 2. We follow the same lines as in the proof of Theorem 1 up to (15), and we skip the part of the proof from (15) to (18) needed to deal with assumption (A1), which is not present in this proposition. Next, similarly to (19) and (20) we obtain

$$\begin{aligned} E\big(d(\hat{G}_n, G^*)\big) &\le E[d(\hat{G}_n, G^*)I(\Omega)] + P(\overline{\Omega}) \\ &\le E[d(\hat{G}_n, G^*)I(\Omega)] + A_1 \exp\big(-A_2 n^{\rho/(1+\rho)}\big) \end{aligned} \tag{24}$$

and, on the event Ω,

$$d(\hat{G}_n, G^*) \le b n^{-1/(1+\rho)} + 2V_{0n}(\mathcal{G}') + V_n(\mathcal{G}')/\sqrt{n}, \tag{25}$$

where V_n and V_{0n} are the same as in (20) and we have used the rough bound $d_\triangle(\hat{G}_n, G^*) \le 1$. Now, the result follows from (24) and (25) if we observe that $\sup_n E(V_n(\mathcal{G}')) < \infty$ and $\sup_n n^{1/(1+\rho)} E(V_{0n}(\mathcal{G}')) < \infty$, in view of Lemmas 8 and 9 in the Appendix. □

PROOF OF THEOREM 3. Fix j^* such that $G^* \in \mathcal{G}_{j^*}$, and write $\rho^* = \rho_{j^*}$. The proof of Theorem 3 will consist in showing that

$$\sup_{\pi \in \mathcal{P}_{j^*}} E_{\pi,n}\big(d(G_n^*, G^*)\big) = O\big((\log n)^{4\kappa/(2\kappa+\rho^*-1)} n^{-\kappa/(2\kappa+\rho^*-1)}\big), \quad n \to \infty. \tag{26}$$

Introduce some notation. Write $\mathcal{G}'_j = \mathcal{G}_j \cup \mathcal{N}_j$. Clearly, $\mathcal{G}'_j$ has complexity bound ρ_j. For brevity, write

$$\begin{aligned} V_n(j) &= V_n(\mathcal{G}'_j), \qquad & V_{0n}(j) &= V_{0n}(\mathcal{G}'_j), \\ W_n(j) &= W_n(\mathcal{G}'_j), \qquad & W_{0n}(j) &= W_{0n}(\mathcal{G}'_j), \end{aligned}$$

where V_{0n}, V_n, W_{0n}, W_n are defined in the Appendix, with $\rho = \rho_j$, with c_1 being the constant from Lemma 7 and $c_2 = c_1$.

For any $j \in \{1, \dots, N\}$ and $k \ge j$ define the pseudoclassifier

$$G_{nkj} = \begin{cases} \arg\min_{G \in \mathcal{N}_{kj}} R_n(G), & \text{if } \mathcal{N}_{kj} \ne \varnothing, \\ G_{nk}, & \text{if } \mathcal{N}_{kj} = \varnothing. \end{cases}$$

We will need some lemmas. The first lemma states that, with probability close to 1, the value j^* is admissible.

LEMMA 1.

$$\sup_{\pi \in \mathcal{P}_{j^*}} P_{\pi,n}(\hat{j} > j^*) = o(1/n), \qquad n \to \infty.$$

The second lemma contains the basic observation needed for the proof.

LEMMA 2. *If j is admissible then, for all $k \geq j$,*

$$0 \leq R_n(G_{nkj}) - R_n(G_{nj}) \leq T_{nk}(G_{nkj}, G_{nk}), \tag{27}$$

$$|R_n(G_{nkj}) - R_n(G_{nk})| \leq T_{nk}(G_{nkj}, G_{nk}). \tag{28}$$

The next two lemmas are technical.

LEMMA 3.

$$\sup_{\pi \in \mathcal{P}_{j^*}} \frac{1}{\sqrt{n}} E_{\pi,n}\big(d_{\triangle}^{(1-\rho^*)/2}(G_{nj^*}, G^*)\big) = O\big(n^{-\kappa/(2\kappa+\rho^*-1)}\big), \qquad n \to \infty.$$

LEMMA 4.

$$\sup_{\pi \in \mathcal{P}_{j^*}} E_{\pi,n}\big(T_{nj^*}\big(G_{nj^*\hat{j}}, G_{nj^*}\big)\big) = O\big((\log n)^{4\kappa/(2\kappa+\rho^*-1)} n^{-\kappa/(2\kappa+\rho^*-1)}\big), \qquad n \to \infty.$$

Proofs of the lemmas are given in Section 6.

Now we proceed to the proof of (26). First, note that, in view of Lemma 1,

$$\begin{aligned} E\big(d(G_n^*, G^*)\big) &\leq E\big(d\big(G_{n\hat{j}}, G^*\big) I(\hat{j} \leq j^*)\big) + P(\hat{j} > j^*) \\ &= E\big(d\big(G_{n\hat{j}}, G^*\big) I(\hat{j} \leq j^*)\big) + o(1/n). \end{aligned} \tag{29}$$

Thus, it is sufficient to work on the random event $\{\hat{j} \leq j^*\}$. If $\hat{j} \leq j^*$, using the definition of $\hat{j}$ and Lemma 2 we find

$$0 \leq R_n\big(G_{nj^*\hat{j}}\big) - R_n\big(G_{n\hat{j}}\big) \leq T_{nj^*}\big(G_{nj^*\hat{j}}, G_{nj^*}\big), \tag{30}$$

$$\big|R_n\big(G_{nj^*\hat{j}}\big) - R_n\big(G_{nj^*}\big)\big| \leq T_{nj^*}\big(G_{nj^*\hat{j}}, G_{nj^*}\big). \tag{31}$$

Note also that if $\hat{j} \leq j^*$, we have $G_{n\hat{j}} \in \mathcal{N}_{\hat{j}} \subseteq \mathcal{N}_{j^*} \subseteq \mathcal{G}'_{j^*}$. Thus, if $\hat{j} \leq j^*$, acting similarly to (20), we get

$$\begin{aligned} &d\big(G_{n\hat{j}}, G^*\big) \\ &\quad \leq \big|R_n\big(G_{n\hat{j}}\big) - R_n(G^*)\big| + \big|R_n(G^*) - R_n\big(G_{n\hat{j}}\big) + d\big(G_{n\hat{j}}, G^*\big)\big| \\ &\quad \leq \big|R_n\big(G_{n\hat{j}}\big) - R_n(G^*)\big| + V_{0n}(j^*) + \frac{V_n(j^*)}{\sqrt{n}} d_{\triangle}^{(1-\rho^*)/2}\big(G_{n\hat{j}}, G^*\big). \end{aligned} \tag{32}$$

Now we come to the main step in the upper bounds. Using (30) and (31) we obtain

$$
\begin{aligned}
&\bigl|R_n\bigl(G_{n\hat{j}}\bigr) - R_n(G^*)\bigr| \\
&\quad\le \bigl|R_n(G^*) - R_n\bigl(G_{nj^*}\bigr)\bigr| + \bigl|R_n\bigl(G_{nj^*}\bigr) - R_n\bigl(G_{nj^*\hat{j}}\bigr)\bigr| \\
&\qquad + \bigl|R_n\bigl(G_{nj^*\hat{j}}\bigr) - R_n\bigl(G_{n\hat{j}}\bigr)\bigr| \\
&\quad\le d\bigl(G_{nj^*}, G^*\bigr) + \bigl|R_n(G^*) - R_n\bigl(G_{nj^*}\bigr) + d\bigl(G_{nj^*}, G^*\bigr)\bigr| \\
&\qquad + 2T_{nj^*}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr) \\
&\quad\le d\bigl(G_{nj^*}, G^*\bigr) + V_{0n}(j^*) + \frac{V_n(j^*)}{\sqrt{n}} d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*}, G^*\bigr) \\
&\qquad + 2T_{nj^*}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr).
\end{aligned} \tag{33}
$$

Substitution of (33) into (32) yields (for $\hat{j} \le j^*$)

$$
\begin{aligned}
d\bigl(G_{n\hat{j}}, G^*\bigr) \le d\bigl(G_{nj^*}, G^*\bigr) + 2\Biggl[& V_{0n}(j^*) + \frac{V_n(j^*)}{\sqrt{n}} d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*}, G^*\bigr) \\
& + T_{nj^*}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr)\Biggr].
\end{aligned} \tag{34}
$$

Now, in view of Lemmas 8 and 9 from the Appendix and Lemma 3,

$$
E\bigl(V_{0n}(j^*)\bigr) = O\bigl(n^{-1/(1+\rho^*)}\bigr) = O\bigl(n^{-\kappa/(2\kappa+\rho^*-1)}\bigr) \tag{35}
$$

and

$$
\begin{aligned}
&E\Biggl(\frac{V_n(j^*)}{\sqrt{n}} d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*}, G^*\bigr)\Biggr) \\
&\quad\le \frac{1}{\sqrt{n}}\bigl[P\bigl(V_n(j^*) \ge \log^2 n\bigr) + (\log^2 n) E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*}, G^*\bigr)\bigr)\bigr] \\
&\quad\le \frac{\log^2 n}{\sqrt{n}} E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*}, G^*\bigr)\bigr) + o\Biggl(\frac{1}{n}\Biggr) \\
&\quad= O\bigl(n^{-\kappa/(2\kappa+\rho^*-1)} \log^2 n\bigr).
\end{aligned} \tag{36}
$$

From (34)–(36) we get

$$
\begin{aligned}
&E\bigl(d\bigl(G_{n\hat{j}}, G^*\bigr) I(\hat{j} \le j^*)\bigr) \\
&\qquad\le E\bigl(d\bigl(G_{nj^*}, G^*\bigr)\bigr) + 2E\bigl(T_{nj^*}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr)\bigr) \\
&\qquad\quad + O\bigl(n^{-\kappa/(2\kappa+\rho^*-1)} \log^2 n\bigr).
\end{aligned} \tag{37}
$$

Theorem 3 follows now from (29), (37), Theorem 1 and Lemma 4. $\square$

PROOF OF PROPOSITION 3. Note first that Lemmas 1 and 2 remain valid since their proofs do not use assumption (A1); see Section 6. Therefore, we get (29) and (34) exactly in the same way as in the proof of Theorem 3. Using (34) and the rough bounds $d_\triangle \le 1$ and

$$T_{nj^*}\big(G_{nj^*\hat{j}}, G_{nj^*}\big) \le \max\big(n^{-1/(1+\rho^*)}, n^{-1/2}\big)\log^2 n = n^{-1/2}\log^2 n, \tag{38}$$

we obtain that, on the event $\{\hat{j} \le j^*\}$,

$$d\big(G_{n\hat{j}}, G^*\big) \le d\big(G_{nj^*}, G^*\big) + 2\big[V_{0n}(j^*) + V_n(j^*)n^{-1/2} + n^{-1/2}\log^2 n\big]. \tag{39}$$

To finish the proof, it remains to substitute (39) in (29), to apply Proposition 2 with $\rho = \rho^*$ and to note that $\sup_n E(V_n(j^*)) < \infty$ and $\sup_n n^{1/(1+\rho^*)} E(V_{0n}(j^*)) < \infty$, in view of Lemmas 8 and 9 in the Appendix. □

6. Proofs of the lemmas.

PROOF OF LEMMA 1. We have

$$P(\hat{j} > j^*) \le \sum_{k \ge j^*} P\big(\mathcal{N}_{kj^*} = \varnothing\big) \le N \max_{k \ge j^*} P\big(\mathcal{N}_{kj^*} = \varnothing\big).$$

Since $N = O(n^\beta)$, it is sufficient to show that

$$\sup_{k \ge j^*} P\big(\mathcal{N}_{kj^*} = \varnothing\big) = o(1/n^{1+\beta}), \tag{40}$$

as $n \to \infty$. Without loss of generality we will assume that all $a_j = 1$ in assumption (A4)(i). Fix an element $G^{\mathcal{N}}$ of $\mathcal{N}_{j^*}$ such that $d_\triangle(G^{\mathcal{N}}, G^*) \le n^{-1/(1+\rho^*)}$ if $\mathcal{N}_{j^*}$ is an ε-net on $\mathcal{G}_{j^*}$ for the distance $d_\triangle$ or $d_{\triangle,e}(G^{\mathcal{N}}, G^*) \le n^{-1/(1+\rho^*)}$ if $\mathcal{N}_{j^*}$ is an ε-net on $\mathcal{G}_{j^*}$ for the distance $d_{\triangle,e}$ [such an element $G^{\mathcal{N}}$ exists in view of assumption (A4)(i) and it is random if the distance $d_{\triangle,e}$ is used]. We have

$$P\big(\mathcal{N}_{kj^*} = \varnothing\big) \le P\big(|R_n(G^{\mathcal{N}}) - R_n(G_{nk})| > T_{nk}(G^{\mathcal{N}}, G_{nk})\big). \tag{41}$$

Since we consider $k \ge j^*$, then also $\mathcal{N}_k \supseteq \mathcal{N}_{j^*}$, and $G^{\mathcal{N}} \in \mathcal{N}_k$. Therefore, $R_n(G^{\mathcal{N}}) \ge R_n(G_{nk})$, and

$$\begin{aligned} 0 &\le R_n(G^{\mathcal{N}}) - R_n(G_{nk}) \\ &\le \big[R_n(G^{\mathcal{N}}) - R_n(G^*) - d(G^{\mathcal{N}}, G^*)\big] + d(G^{\mathcal{N}}, G^*) \\ &\quad + \big[R_n(G^*) - R_n(G_{nk}) + d(G_{nk}, G^*)\big]. \end{aligned} \tag{42}$$

Consider the random event $\Omega_* = \{d_\triangle(G^{\mathcal{N}}, G^*) \le bn^{-1/(1+\rho^*)}\}$. Repeating the argument of (14) with $\rho = \rho^*$ we get

$$P(\overline{\Omega}_*) \le A_1 \exp\big(-A_2 n^{\rho^*/(1+\rho^*)}\big). \tag{43}$$

Since $d(G^{\mathcal{N}}, G^*) \le d_\triangle(G^{\mathcal{N}}, G^*)$ [cf. (4)] we deduce from (42) that, on the event Ω_*,

$$
\begin{aligned}
|R_n(G^{\mathcal{N}}) - R_n(G_{nk})| \le{}& bn^{-1/(1+\rho^*)} + |R_n(G^{\mathcal{N}}) - R_n(G^*) - d(G^{\mathcal{N}}, G^*)| \\
&+ |R_n(G^*) - R_n(G_{nk}) + d(G_{nk}, G^*)|.
\end{aligned} \tag{44}
$$

The right-hand side of (44) is analogous to that of (15). Applying the argument as in (20) and observing that $G_{nk}, G^* \in \mathcal{G}_k'' \stackrel{\text{def}}{=} \mathcal{G}_{j^*} \cup \mathcal{N}_k$ (clearly, $\mathcal{G}_k''$ has complexity bound ρ_k, since $\rho_k \ge \rho^*$), $G^{\mathcal{N}}, G^* \in \mathcal{G}_{j^*}' \subseteq \mathcal{G}_k''$ and $n^{-1/(1+\rho^*)} \le n^{-1/(1+\rho_k)}$ we get that, on Ω_*,

$$
\begin{aligned}
&|R_n(G^{\mathcal{N}}) - R_n(G_{nk})| \\
&\quad \le bn^{-1/(1+\rho^*)} + 2V_{0n}(\mathcal{G}_k'') + \frac{V_n(\mathcal{G}_k'')}{\sqrt{n}} d_\triangle^{(1-\rho_k)/2}(G_{nk}, G^*),
\end{aligned} \tag{45}
$$

where V_{0n} and V_n are defined in the Appendix, with $\rho = \rho_k$ and with the constant $c_2 = b$. Next, the triangle inequality and the fact that $(1 - \rho_k)/2 < 1$, $\rho_k \ge \rho^*$ imply that, on Ω_*,

$$
\begin{aligned}
d_\triangle^{(1-\rho_k)/2}(G_{nk}, G^*) &\le d_\triangle^{(1-\rho_k)/2}(G^{\mathcal{N}}, G_{nk}) + d_\triangle^{(1-\rho_k)/2}(G^{\mathcal{N}}, G^*) \\
&\le d_\triangle^{(1-\rho_k)/2}(G^{\mathcal{N}}, G_{nk}) + \bigl(bn^{-1/(1+\rho^*)}\bigr)^{(1-\rho_k)/2} \\
&\le d_\triangle^{(1-\rho_k)/2}(G^{\mathcal{N}}, G_{nk}) + b' n^{-(1-\rho_k)/(2(1+\rho_k))}
\end{aligned} \tag{46}
$$

where $b' = \max(b^{1/2}, 1)$. From (45) and (46) we obtain that, on Ω_*,

$$
\begin{aligned}
|R_n(G^{\mathcal{N}}) - R_n(G_{nk})| \le{}& \bigl(b + b' V_n(\mathcal{G}_k'')\bigr) n^{-1/(1+\rho_k)} + 2V_{0n}(\mathcal{G}_k'') \\
&+ \frac{V_n(\mathcal{G}_k'')}{\sqrt{n}} d_\triangle^{(1-\rho_k)/2}(G^{\mathcal{N}}, G_{nk}).
\end{aligned} \tag{47}
$$

Now,

$$
P\bigl(|R_n(G^{\mathcal{N}}) - R_n(G_{nk})| > T_{nk}(G^{\mathcal{N}}, G_{nk})\bigr) \le p_1 + p_2 + P(\overline{\Omega}_*), \tag{48}
$$

where

$$
\begin{aligned}
p_1 = P\Bigl(\bigl\{&|R_n(G^{\mathcal{N}}) - R_n(G_{nk})| \\
&> T_{nk}(G^{\mathcal{N}}, G_{nk}), d_\triangle(G^{\mathcal{N}}, G_{nk}) \le c_1 n^{-1/(1+\rho_k)}\bigr\} \cap \Omega_*\Bigr),
\end{aligned}
$$

$$
\begin{aligned}
p_2 = P\Bigl(\bigl\{&|R_n(G^{\mathcal{N}}) - R_n(G_{nk})| \\
&> T_{nk}(G^{\mathcal{N}}, G_{nk}), d_\triangle(G^{\mathcal{N}}, G_{nk}) > c_1 n^{-1/(1+\rho_k)}\bigr\} \cap \Omega_*\Bigr).
\end{aligned}
$$

Using (47) we obtain

$$
\begin{aligned}
p_1 &\le P\Big(\big[b + \big(b' + c_1^{(1-\rho_k)/2}\big)V_n(\mathcal{G}_k'')\big]n^{-1/(1+\rho_k)} \\
&\qquad + 2V_{0n}(\mathcal{G}_k'') > T_{nk}(G^{\mathcal{N}}, G_{nk})\Big) \\
&\le P\Big(\big[b + \big(b' + c_1^{(1-\rho_k)/2}\big)V_n(\mathcal{G}_k'')\big]n^{-1/(1+\rho_k)} \\
&\qquad + 2V_{0n}(\mathcal{G}_k'') > n^{-1/(1+\rho_k)}\log^2 n\Big) \qquad (49) \\
&\le P\Big(\big(b' + c_1^{(1-\rho_k)/2}\big)V_n(\mathcal{G}_k'') > \frac{1}{2}(\log^2 n - b)\Big) \\
&\quad + P\Big(V_{0n}(\mathcal{G}_k'') > \frac{1}{4}n^{-1/(1+\rho_k)}\log^2 n\Big) \\
&= o\Big(\frac{1}{n^{1+\beta}}\Big),
\end{aligned}
$$

as $n \to \infty$ uniformly in k, in view of Lemmas 8 and 9. Using (47), the probability p_2 is bounded as follows:

$$
\begin{aligned}
p_2 &\le P\Big(\big(b + b'V_n(\mathcal{G}_k'')\big)n^{-1/(1+\rho_k)} + 2V_{0n}(\mathcal{G}_k'') + \frac{V_n(\mathcal{G}_k'')}{\sqrt{n}}d_{\triangle}^{(1-\rho_k)/2}(G^{\mathcal{N}}, G_{nk}) \\
&\qquad > \frac{\log^2 n}{2}n^{-1/(1+\rho_k)} \\
&\qquad\quad + \frac{\log^2 n}{2\sqrt{n}}d_{\triangle,e}^{(1-\rho_k)/2}(G^{\mathcal{N}}, G_{nk}), d_{\triangle}(G^{\mathcal{N}}, G_{nk}) > c_1 n^{-1/(1+\rho_k)}\Big) \\
&\le p_3 + p_4,
\end{aligned}
$$

where

$$
\begin{aligned}
p_3 &= P\Big(\big(b + b'V_n(\mathcal{G}_k'')\big)n^{-1/(1+\rho_k)} + 2V_{0n}(\mathcal{G}_k'') > \frac{\log^2 n}{2}n^{-1/(1+\rho_k)}\Big), \\
p_4 &= P\Big(V_n(\mathcal{G}_k'')d_{\triangle}^{(1-\rho_k)/2}(G^{\mathcal{N}}, G_{nk}) > \frac{\log^2 n}{2}d_{\triangle,e}^{(1-\rho_k)/2}(G^{\mathcal{N}}, G_{nk}), \\
&\qquad\qquad d_{\triangle}(G^{\mathcal{N}}, G_{nk}) > c_1 n^{-1/(1+\rho_k)}\Big).
\end{aligned}
$$

Note that $\sup_k p_3 = o(1/n^{1+\beta})$, similarly to (49). Also, for n large enough,

p_4 satisfies (since $G^{\mathcal{N}}, G_{nk} \in \mathcal{G}_k''$)

$$p_4 \le P\biggl(V_n(\mathcal{G}_k'') > \inf_{(G,G')\in\mathcal{S}_k}\biggl(\frac{d_{\triangle,e}(G,G')}{d_\triangle(G,G')}\biggr)^{(1-\rho_k)/2}\frac{\log^2 n}{2}\biggr)$$

$$\le P\biggl(V_n(\mathcal{G}_k'') > \frac{\log^2 n}{4}\biggr) + P\biggl(\inf_{(G,G')\in\mathcal{S}_k}\biggl(\frac{d_{\triangle,e}(G,G')}{d_\triangle(G,G')}\biggr)^{(1-\rho_k)/2} < \frac{1}{2}\biggr)$$

$$\le D_1\exp(-D_2\log^2 n) + P\biggl(W_n(\mathcal{G}_k'') \ge \frac{1}{2}\biggr) = o\biggl(\frac{1}{n^{1+\beta}}\biggr), \qquad n\to\infty,$$

where $\mathcal{S}_k = \{(G,G') : G, G' \in \mathcal{G}_k'', d_\triangle(G,G') \ge c_1 n^{-1/(1+\rho_k)}\}$ and we used Lemmas 7 and 8. Here again $o(1/n^{1+\beta})$ converges to 0 uniformly in k. Thus,

$$\sup_k p_2 = o(1/n^{1+\beta}), \qquad n\to\infty,$$

which, together with (41), (43), (48) and (49), yields (40). □

PROOF OF LEMMA 2. If j is admissible, then

$$G_{nkj} = \arg\min_{G\in\mathcal{N}_{kj}} R_n(G)$$

for all $k \ge j$. Since $\mathcal{N}_{kj} \subset \mathcal{N}_j$ we have $R_n(G_{nkj}) \ge R_n(G_{nj})$, which is the left inequality in (27). On the other hand, $G_{nkj} \in \mathcal{N}_{kj}$, and (28) follows from the definition of $\mathcal{N}_{kj}$. Finally, $\mathcal{N}_j \subset \mathcal{N}_k$ for $k \ge j$, and thus $R_n(G_{nkj}) - R_n(G_{nj}) \le R_n(G_{nkj}) - R_n(G_{nk}) \le T_{nk}(G_{nkj}, G_{nk})$, where the last inequality is again due to $G_{nkj} \in \mathcal{N}_{kj}$. This proves the right-hand side inequality in (27). □

PROOF OF LEMMA 3. In view of assumption (A1) we have

$$\begin{aligned} E\bigl(d_\triangle^{(1-\rho^*)/2}(G_{nj^*}, G^*)\bigr) &= E\bigl(d_\triangle^{(1-\rho^*)/2}(G_{nj^*}, G^*) I\bigl(d_\triangle(G_{nj^*}, G^*) \ge \varepsilon_0\bigr)\bigr) \\ &\quad + c_0^{-1/\kappa} E\bigl(d^{(1-\rho^*)/2\kappa}(G_{nj^*}, G^*)\bigr) \\ &\le P\bigl(d_\triangle(G_{nj^*}, G^*) \ge \varepsilon_0\bigr) \\ &\quad + c_0^{-1/\kappa} E\bigl(d^{(1-\rho^*)/2\kappa}(G_{nj^*}, G^*)\bigr). \end{aligned}$$

To finish the proof it remains to apply Theorem 1 and (18) (with $\rho = \rho^*$, $\hat{G}_n = G_{nj^*}$). □

To prove Lemma 4 we need the following auxiliary result.

LEMMA 5. *For any* $\pi \in \mathcal{P}_{j^*}$,

$$\frac{1}{\sqrt{n}} E_{\pi,n}\bigl(d_{\triangle,e}^{(1-\rho^*)/2}(G_{nj^*\hat{j}}, G_{nj^*})\bigr)$$

$$\le \frac{3}{2\sqrt{n}} E_{\pi,n}\bigl(d_\triangle^{(1-\rho^*)/2}(G_{nj^*\hat{j}}, G_{nj^*})\bigr) + Cn^{-\kappa/(2\kappa+\rho^*-1)},$$

where $C > 0$ does not depend on $\pi \in \mathcal{P}_{j^*}$.

PROOF. By Lemma 1 and since $d_{\triangle,e} \le 1$, we have

$$
\begin{aligned}
&E\bigl(d_{\triangle,e}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr)\bigr)\\
&\qquad \le E\bigl(d_{\triangle,e}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr) I\bigl(\hat{j} \le j^*\bigr)\bigr) + o(1/n).
\end{aligned} \tag{50}
$$

If $\hat{j} \le j^*$, we have $G_{nj^*\hat{j}} \in \mathcal{N}_{j^*} \subseteq \mathcal{G}'_{j^*}$, $G_{nj^*} \in \mathcal{G}'_{j^*}$, and thus

$$
\begin{aligned}
&d_{\triangle,e}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr)\\
&\qquad \le W_{0n}(j^*)^{(1-\rho^*)/2} + \bigl(W_n(j^*)+1\bigr) d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr),
\end{aligned} \tag{51}
$$

where $W_{0n}(j^*)$ and $W_n(j^*)$ are defined at the beginning of the proof of Theorem 3. Using Lemma 10, we find

$$
\begin{aligned}
E\bigl(W_{0n}(j^*)^{(1-\rho^*)/2}\bigr) &\le \bigl[E\bigl(W_{0n}(j^*)\bigr)\bigr]^{(1-\rho^*)/2}\\
&= O\bigl(n^{(1-\rho^*)/(2(1+\rho^*))}\bigr), \qquad n \to \infty.
\end{aligned} \tag{52}
$$

Also, by Lemma 7,

$$
\begin{aligned}
&E\bigl(\bigl(W_n(j^*)+1\bigr) d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr)\bigr)\\
&\quad \le \tfrac{3}{2} E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr)\bigr)\\
&\qquad + E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr) I\bigl(W_n(j^*) \ge \tfrac{1}{2}\bigr)\bigr)\\
&\quad \le \tfrac{3}{2} E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr)\bigr) + A_1 \exp\bigl(-A_2 n^{\rho^*/(1+\rho^*)}\bigr).
\end{aligned} \tag{53}
$$

Combining (50)–(53) and observing that

$$
n^{-1/(1+\rho^*)} = O\bigl(n^{-\kappa/(2\kappa+\rho^*-1)}\bigr) \qquad \forall \kappa \ge 1,\ 0 < \rho^* < 1, \tag{54}
$$

we obtain the lemma. The constant C in the formulation of Lemma 5 does not depend on π since the remainder terms in (50)–(53) are uniform in $\pi \in \mathcal{P}_{j^*}$. □

PROOF OF LEMMA 4. Using Lemmas 3 and 5 and (54) we get

$$
\begin{aligned}
&E\bigl(T_{nj^*}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr)\bigr)\\
&\quad \le \log^2 n \left[n^{-1/(1+\rho^*)} + \frac{1}{\sqrt{n}} E\bigl(d_{\triangle,e}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr)\bigr)\right]\\
&\quad \le \log^2 n \left[n^{-1/(1+\rho^*)} + \frac{3}{2\sqrt{n}} E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G_{nj^*}\bigr)\bigr) + C n^{-\kappa/(2\kappa+\rho^*-1)}\right]
\end{aligned} \tag{55}
$$

$$
\begin{aligned}
&\le O\bigl(n^{-\kappa/(2\kappa+\rho^*-1)}\log^2 n\bigr)\\
&\quad+\frac{3\log^2 n}{2\sqrt{n}}\Bigl[E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}},G^*\bigr)\bigr)+E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*},G^*\bigr)\bigr)\Bigr]\\
&=O\bigl(n^{-\kappa/(2\kappa+\rho^*-1)}\log^2 n\bigr)+\frac{3\log^2 n}{2\sqrt{n}}E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}},G^*\bigr)\bigr).
\end{aligned}
$$

Next, note that if $\hat{j}\le j^*$, we have $G_{nj^*\hat{j}}\in\mathcal{N}_{j^*}\subseteq\mathcal{G}'_{j^*}$, and thus

$$
\begin{aligned}
&d\bigl(G_{nj^*\hat{j}},G^*\bigr)\\
&\quad\le\bigl|R_n\bigl(G_{nj^*\hat{j}}\bigr)-R_n(G^*)\bigr|\\
&\qquad+\bigl|R_n(G^*)-R_n\bigl(G_{nj^*\hat{j}}\bigr)+d\bigl(G_{nj^*\hat{j}},G^*\bigr)\bigr|\\
&\quad\le\bigl|R_n\bigl(G_{nj^*\hat{j}}\bigr)-R_n(G^*)\bigr|+V_{0n}(j^*)+\frac{V_n(j^*)}{\sqrt{n}}d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}},G^*\bigr),
\end{aligned}
\tag{56}
$$

where $V_{0n}(j^*)$ and $V_n(j^*)$ are defined at the beginning of the proof of Theorem 3. Applying (31) and acting similarly to (56) we find that, for $\hat{j}\le j^*$,

$$
\begin{aligned}
&\bigl|R_n\bigl(G_{nj^*\hat{j}}\bigr)-R_n(G^*)\bigr|\\
&\quad\le\bigl|R_n\bigl(G_{nj^*}\bigr)-R_n(G^*)\bigr|+\bigl|R_n\bigl(G_{nj^*\hat{j}}\bigr)-R_n\bigl(G_{nj^*}\bigr)\bigr|\\
&\quad\le\bigl|R_n\bigl(G_{nj^*}\bigr)-R_n(G^*)-d\bigl(G_{nj^*},G^*\bigr)\bigr|+d\bigl(G_{nj^*},G^*\bigr)\\
&\qquad+T_{nj^*}\bigl(G_{nj^*\hat{j}},G_{nj^*}\bigr)\\
&\quad\le d\bigl(G_{nj^*},G^*\bigr)+V_{0n}(j^*)+\frac{V_n(j^*)}{\sqrt{n}}d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*},G^*\bigr)\\
&\qquad+T_{nj^*}\bigl(G_{nj^*\hat{j}},G_{nj^*}\bigr).
\end{aligned}
$$

This and (56) imply that, on the event $\{\hat{j}\le j^*\}$,

$$
\begin{aligned}
d\bigl(G_{nj^*\hat{j}},G^*\bigr)\le{}&d\bigl(G_{nj^*},G^*\bigr)+2V_{0n}(j^*)+\frac{V_n(j^*)}{\sqrt{n}}d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*},G^*\bigr)\\
&+\frac{V_n(j^*)}{\sqrt{n}}d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}},G^*\bigr)+T_{nj^*}\bigl(G_{nj^*\hat{j}},G_{nj^*}\bigr).
\end{aligned}
\tag{57}
$$

Taking expectations, treating the event $\{\hat{j}\le j^*\}$ via Lemma 1 and taking into

account (35), (36), (54), (55) and Theorem 1 (with $\rho = \rho^*$, $\hat{G}_n = G_{nj^*}$) we get

$$
\begin{aligned}
&E\bigl(d(G_{nj^*\hat{j}}, G^*)\bigr)\\
&\quad\le o\biggl(\frac{1}{n}\biggr) + E\bigl(d(G_{nj^*}, G^*)\bigr) + 2E\bigl(V_{0n}(j^*)\bigr)\\
&\qquad + E\biggl(\frac{V_n(j^*)}{\sqrt{n}} d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G^*\bigr)\biggr)\\
&\qquad + E\biggl(\frac{V_n(j^*)}{\sqrt{n}} d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*}, G^*\bigr)\biggr) \qquad (58)\\
&\qquad + \frac{3\log^2 n}{2\sqrt{n}}\bigl[E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G^*\bigr)\bigr)\bigr] + O\bigl(n^{-\kappa/(2\kappa+\rho^*-1)}\log^2 n\bigr)\\
&\quad= \frac{1}{\sqrt{n}} E\biggl(\biggl(\frac{3}{2}\log^2 n + V_n(j^*)\biggr) d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G^*\bigr)\biggr)\\
&\qquad + O\bigl(n^{-\kappa/(2\kappa+\rho^*-1)}\log^2 n\bigr).
\end{aligned}
$$

Acting as in (36), we get

$$
\frac{1}{\sqrt{n}} E\bigl(V_n(j^*) d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G^*\bigr)\bigr) \le \frac{\log^2 n}{\sqrt{n}} E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G^*\bigr)\bigr) + o\biggl(\frac{1}{n}\biggr).
$$

This and (58) yield

$$
\begin{aligned}
&E\bigl(d(G_{nj^*\hat{j}}, G^*)\bigr)\\
&\quad\le \frac{5\log^2 n}{2\sqrt{n}} E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G^*\bigr)\bigr) + O\bigl(n^{-\kappa/(2\kappa+\rho^*-1)}\log^2 n\bigr). \qquad (59)
\end{aligned}
$$

The following intermediate lemma holds.

LEMMA 6. *Let ε_0 be the constant from assumption* (A1). *Then*

$$
\sup_{\pi\in\mathcal{P}_{j^*}} P_{\pi,n}\bigl(d_{\triangle}\bigl(G_{nj^*\hat{j}}, G^*\bigr) \ge \varepsilon_0\bigr) = O\bigl(n^{-\kappa/(2\kappa+\rho^*-1)}\bigr), \qquad n\to\infty.
$$

Proof of Lemma 6 will be given at the end of this section.

Using Lemma 6 and assumption (A1) we find

$$
\begin{aligned}
&E\bigl(d_{\triangle}^{(1-\rho^*)/2}\bigl(G_{nj^*\hat{j}}, G^*\bigr)\bigr)\\
&\quad\le c_0^{-(1-\rho^*)/2\kappa} E\bigl(d^{(1-\rho^*)/2\kappa}\bigl(G_{nj^*\hat{j}}, G^*\bigr) I\bigl(d_{\triangle}\bigl(G_{nj^*\hat{j}}, G^*\bigr) < \varepsilon_0\bigr)\bigr)\\
&\qquad + P\bigl(d_{\triangle}\bigl(G_{nj^*\hat{j}}, G^*\bigr) \ge \varepsilon_0\bigr) \qquad (60)\\
&\quad\le c_0^{-(1-\rho^*)/2\kappa} S^{(1-\rho^*)/2\kappa} + O\bigl(n^{-\kappa/(2\kappa+\rho^*-1)}\bigr),
\end{aligned}
$$

where $S = E(d(G_{nj^*\hat{j}}, G^*))$. This and (59) imply

$$S \le C \log^2 n \big[S^{(1-\rho^*)2\kappa}/\sqrt{n} + n^{-\kappa/(2\kappa+\rho^*-1)} \big],$$

where $C > 0$ is a constant independent of π. Any solution S of this inequality satisfies

$$S = O\big((\log n)^{4\kappa/(2\kappa+\rho^*-1)} n^{-\kappa/(2\kappa+\rho^*-1)}\big).$$

Substitution of this result into (60) yields

$$\frac{1}{\sqrt{n}} E\big(d_\triangle^{(1-\rho^*)/2}(G_{nj^*\hat{j}}, G^*)\big) = O\big((\log n)^{2(1-\rho^*)/(2\kappa+\rho^*-1)} n^{-\kappa/(2\kappa+\rho^*-1)}\big).$$

Combining this with (55) completes the proof of Lemma 4. □

PROOF OF LEMMA 6. Using (57) and the rough bounds $d_\triangle \le 1$ and (38) we obtain that, on the event $\{\hat{j} \le j^*\}$,

$$d\big(G_{nj^*\hat{j}}, G^*\big) \le d\big(G_{nj^*}, G^*\big) + 2\big[V_{0n}(j^*) + V_n(j^*) n^{-1/2}\big] + n^{-1/2} \log^2 n. \tag{61}$$

On the other hand, note that the argument leading to (17) holds for any set in place of $\hat{G}_n$, in particular,

$$d_\triangle\big(G_{nj^*\hat{j}}, G^*\big) \le \frac{1}{2t_0} d\big(G_{nj^*\hat{j}}, G^*\big) + \frac{\varepsilon_0}{2}. \tag{62}$$

Combining (61) and (62) yields that on $\{\hat{j} \le j^*\}$ we have

$$\begin{aligned} &d_\triangle\big(G_{nj^*\hat{j}}, G^*\big) \\ &\quad \le \frac{1}{2t_0}\big[d\big(G_{nj^*}, G^*\big) + 2\big[V_{0n}(j^*) + V_n(j^*) n^{-1/2}\big] + n^{-1/2} \log^2 n\big] + \frac{\varepsilon_0}{2}. \end{aligned}$$

Thus, for n large enough,

$$\begin{aligned} &P\big(d_\triangle\big(G_{nj^*\hat{j}}, G^*\big) \ge \varepsilon_0\big) \\ &\quad \le P(\hat{j} > j^*) \\ &\qquad + P\Big(d\big(G_{nj^*}, G^*\big) + 2\big[V_{0n}(j^*) + V_n(j^*) n^{-1/2}\big] \ge \varepsilon_0 t_0 - n^{-1/2} \log^2 n\Big) \\ &\quad \le o(1/n) + P\Big(d\big(G_{nj^*}, G^*\big) \ge (\varepsilon_0 t_0 - n^{-1/2} \log^2 n)/3\Big) \\ &\qquad + P\big(V_{0n}(j^*) \ge (\varepsilon_0 t_0 - n^{-1/2} \log^2 n)/6\big) \\ &\qquad + P\big(V_n(j^*) \ge (n^{1/2} \varepsilon_0 t_0 - \log^2 n)/6\big) \\ &\quad \le o(1/n) + C\big[E\big(d\big(G_{nj^*}, G^*\big)\big) + E\big(V_{0n}(j^*)\big)\big] + D_1 \exp\big(-\sqrt{n}/C\big) \\ &\quad = O\big(n^{-\kappa/(2\kappa+\rho^*-1)}\big), \end{aligned}$$

for some $C > 0$, where we used Lemmas 1 and 8, the Markov inequality, (35) and Theorem 1 (with $\rho = \rho^*$, $\hat{G}_n = G_{nj^*}$). □

APPENDIX

Assume everywhere in this appendix that $\mathcal{G}$ is a class of sets in $\mathbf{R}^d$ having complexity bound $\rho \in [\rho_{\min}, \rho_{\max}]$, where $0 < \rho_{\min} < \rho_{\max} < 1$, and that $G^* \in \mathcal{G}$. Write

$$W_n(\mathcal{G}) = \sup_{G,G' \in \mathcal{S}} \left| \left(\frac{d_{\triangle,e}(G, G')}{d_{\triangle}(G, G')} \right)^{(1-\rho)/2} - 1 \right|,$$

where $c_1 > 0$ and $\mathcal{S} = \{(G, G') : G, G' \in \mathcal{G}, d_{\triangle}(G, G') \geq c_1 n^{-1/(1+\rho)}\}$.

Note that $R_n(G)$ and $d_{\triangle,e}(G, G')$ are sums of i.i.d. random variables bounded by 1 and

$$d(G, G^*) = E\big(R_n(G) - R_n(G^*)\big), \qquad d_{\triangle}(G, G') = E\big(d_{\triangle,e}(G, G')\big)$$

for every G, G'.

The results given below are related to the theory of empirical processes and based on the book of van de Geer (2000). Some of them can be also deduced from van der Vaart and Wellner (1996), Alexander (1984, 1987) and Birgé and Massart (1993).

LEMMA 7. *There exist constants $c_1 > 0$, $A_1 > 0$, $A_2 > 0$, depending only on A, $\rho_{\min}$, $\rho_{\max}$ such that*

$$P\big(W_n(\mathcal{G}) \geq \tfrac{1}{2}\big) \leq A_1 \exp\big(-A_2 n^{\rho/(1+\rho)}\big).$$

PROOF. The proof follows the lines of Lemma 5.16 [van de Geer (2000), pages 82 and 83], where one sets $g(x) = I(x \in G \triangle G')$. Then $\|g\|^2 = d_{\triangle}(G, G')$, $\|g\|_n^2 = d_{\triangle,e}(G, G')$ and there exists $\delta_0 > 0$ such that condition $n\delta_n^2 \geq 2\mathcal{H}_B(\delta_n, \mathcal{G}, P_X)$ required in Lemma 5.16 of van de Geer (2000) is satisfied with $\delta_n = \delta_0 n^{-1/(2(1+\rho))}$ [in fact, $\mathcal{H}_B(\delta, \mathcal{G}, P_X) = \mathcal{H}_B(\delta^2, \mathcal{G}, d_{\triangle}) \leq A\delta^{-2\rho}$]. Therefore, one can apply the argument on page 83 of van de Geer (2000), which yields, for any $\eta > 0$,

$$\begin{aligned} &P\left(\sup_{G,G' \in \mathcal{S}} \left| \left(\frac{d_{\triangle,e}(G, G')}{d_{\triangle}(G, G')} \right)^{1/2} - 1 \right| \geq \eta \right) \\ &\quad \leq 4 \sum_{s \geq 2^5/\eta} \exp\left(-\frac{n^{\rho/(1+\rho)} s^2 \eta^2}{2^7} \right) \qquad (63) \\ &\quad \leq A_1 \exp\big(-A_2 n^{\rho/(1+\rho)}\big), \end{aligned}$$

where $A_1 > 0$, $A_2 > 0$ and $c_1 = 2^{10}\delta_0^2/\eta$. Here put $\eta = \min(1 - (1/2)^{1/(1-\rho)}, (3/2)^{1/(1-\rho)} - 1)$. Then $|x - 1| < \eta$ implies $|x^{1-\rho} - 1| < 1/2$ for $x > 0$, and thus (63) implies the lemma. The constants $c_1 > 0$, $A_1 > 0$, $A_2 > 0$ can be chosen

depending only on $A, \rho_{\min}, \rho_{\max}$ since they are continuous positive functions of A, ρ and $\rho \in [\rho_{\min}, \rho_{\max}]$. $\square$

In what follows c_2 is an arbitrary positive constant. Define

$$W_{0n}(\mathcal{G}) = \sup_{G,G' \in \mathcal{G}\,:\, d_\triangle(G,G') \le c_1 n^{-1/(1+\rho)}} d_{\triangle,e}(G, G'),$$

$$V_n(\mathcal{G}) = \sup_{G \in \mathcal{G}\,:\, d_\triangle(G,G^*) \ge c_1 n^{-1/(1+\rho)}} \frac{\sqrt{n}|R_n(G^*) - R_n(G) + d(G, G^*)|}{d_\triangle^{(1-\rho)/2}(G, G^*)},$$

$$V_{0n}(\mathcal{G}) = \sup_{G \in \mathcal{G}\,:\, d_\triangle(G,G^*) \le c_2 n^{-1/(1+\rho)}} |R_n(G^*) - R_n(G) + d(G, G^*)|.$$

LEMMA 8. *There exist $D_1 > 0$, $D_2 > 0$, $D_3 > 0$ depending only on $A, \rho_{\min}, \rho_{\max}$ such that*

$$P\big(V_n(\mathcal{G}) \ge x\big) \le D_1 \exp(-D_2 x) \qquad \forall\, x \ge D_3.$$

LEMMA 9. *There exist $B_1 > 0$, $B_2 > 0$, $B_3 > 0$ depending only on $A, \rho_{\min}, \rho_{\max}$ such that*

$$P\big(V_{0n}(\mathcal{G}) \ge x n^{-1/(1+\rho)}\big) \le B_1 \exp\big(-B_2 n^{\rho/(1+\rho)}\big) \qquad \forall\, x \ge B_3.$$

PROOF OF LEMMAS 8 AND 9. Apply, respectively, the inequalities (5.43) and (5.42) of van de Geer (2000) with $g(X, Y) = (Y - I(X \in G))^2 - (Y - I(X \in G^*))^2$, $g_0(X, Y) \equiv 0$. Then $\nu_n(g) = \nu_n(g) - \nu_n(g_0) = \sqrt{n}(R_n(G^*) - R_n(G) + d(G, G^*))$, $\|g - g_0\|^2 = d_\triangle(G, G^*)$ [$\|\cdot\|$ being the $L_2(\pi)$-norm], $\beta = 0$, $\alpha = 2\rho$ (cf. the proof of Lemma 7). $\square$

LEMMA 10. *There exist $Q_1 > 0$, $Q_2 > 0$, $Q_3 > 0$ depending only on $A, \rho_{\min}, \rho_{\max}$ such that*

$$P\big(W_{0n}(\mathcal{G}) \ge x n^{-1/(1+\rho)}\big) \le Q_1 \exp\big(-Q_2 n^{\rho/(1+\rho)}\big) \qquad \forall\, x \ge Q_3.$$

PROOF. Act as in the previous proof using (5.42) of van de Geer (2000), but with g as in Lemma 7. $\square$

LEMMA 11. *There exists $C_2 > 0$ depending only on $A, \rho_{\min}, \rho_{\max}$ such that*

$$P\left(\sup_{G \in \mathcal{G}} |R_n(G^*) - R_n(G) + d(G, G^*)| \ge t\right) \le C_2 \exp(-nt^2/C_2) \qquad \forall\, 0 < t \le 1.$$

PROOF. Use Theorem 5.11 in van de Geer (2000) with g, $\nu_n(g)$ as in the proof of Lemmas 8 and 9. It is easy to see that in this case the assumptions of Theorem 5.11 are satisfied with $K = 1$, some $R \ge 1$ and the constant a such that

$a_0 \le a \le \sqrt{n}$, provided both $a_0 > 0$ and the constant C_1 in Theorem 5.11 are chosen large enough. Thus, denoting $t = a/\sqrt{n}$ and using (5.35) of van de Geer (2000), we obtain Lemma 11 for $a_0/\sqrt{n} \le t \le 1$. For $t < a_0/\sqrt{n}$ the inequality is trivial, by the choice of C_2 large enough. □

REFERENCES

AIZERMAN, M. A., BRAVERMAN, E. M. and ROZONOER, L. I. (1970). *Method of Potential Functions in the Theory of Learning Machines.* Nauka, Moscow (in Russian).

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.]

ALEXANDER, K. S. (1987). Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probab. Theory Related Fields* **75** 379–423.

ANTHONY, M. and BARTLETT, P. L. (1999). *Neural Network Learning*: *Theoretical Foundations*. Cambridge Univ. Press.

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.

BARTLETT, P. L., BOUCHERON, S. and LUGOSI, G. (2002). Model selection and error estimation. *Machine Learning* **48** 85–113.

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

BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. and STONE, C. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

BÜHLMANN, P. and YU, B. (2002). Analyzing bagging. *Ann. Statist.* **30** 927–961.

CATONI, O. (2001) Randomized estimators and empirical complexity for pattern recognition and least square regression. Prépublication 677, Laboratoire de Probabilités et Modèles Aléatoires, Univ. Paris 6/7. Available at www.proba.jussieu.fr.

CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000). *An Introduction to Support Vector Machines.* Cambridge Univ. Press.

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. Approximation Theory* **10** 227–236.

HORVÁTH, M. and LUGOSI, G. (1998). Scale-sensitive dimensions and skeleton estimates for classification. *Discrete Appl. Math.* **86** 37–61.

KOLTCHINSKII, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory* **47** 1902–1914.

KOLTCHINSKII, V. and PANCHENKO, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.* **30** 1–50.

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

LEPSKI, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

LUGOSI, G. and NOBEL, A. (1999). Adaptive model selection using empirical complexities. *Ann. Statist.* **27** 1830–1864.

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

MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829.

MASSART, P. (2000). Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math*. **9** 245–303.

SCHAPIRE, R. E., FREUND, Y., BARTLETT, P. L. and LEE, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** 1651–1686.

SCHÖLKOPF, B. and SMOLA, A. (2002). *Learning with Kernels*. MIT Press.

TSYBAKOV, A. B. (2002). Discussion of "Random rates in anisotropic regression," by M. Hoffmann and O. Lepskii. *Ann. Statist.* **30** 379–385.

VAN DE GEER, S. (2000). *Applications of Empirical Process Theory*. Cambridge Univ. Press.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

VAPNIK, V. N. (1982). *Estimation of Dependencies Based on Empirical Data.* Springer, New York.

VAPNIK, V. N. (1998). *Statistical Learning Theory.* Wiley, New York.

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

YANG, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. II. Model selection for adaptation. *IEEE Trans. Inform. Theory* **45** 2271–2292.

LABORATOIRE DE PROBABILITÉS
ET MODÈLES ALÉATOIRES
UNIVERSITÉ PARIS 6
4 PLACE JUSSIEU
BOÎTE COURRIER 188
75252 PARIS CEDEX 05
FRANCE
E-MAIL: tsybakov@ccr.jussieu.fr